Isotonic Distributional Regression

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Alexander Henzi

von Günsberg

Leiterin der Arbeit:

Prof. Dr. Johanna Ziegel

Institut für mathematische Statistik und Versicherungslehre
der Universität Bern

Urheberrechtlicher Hinweis

Isotonic Distributional Regression

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Alexander Henzi

von Günsberg

Leiterin der Arbeit:

Prof. Dr. Johanna Ziegel

Institut für mathematische Statistik und Versicherungslehre
der Universität Bern

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, 03. Juni 2022

Der Dekan:
Prof. Dr. Zoltan Balogh

Acknowledgments

First of all, I would like to express my deep gratitude to my supervisor, Johanna Ziegel. Over the past years, during which I have completed both my Bachelor's and Master's thesis and now my PhD under her supervision, I have benefited from her continuous support and guidance. I thank her for providing uncountably many opportunities for me to develop my skills as a statistician and researcher. It has been a pleasure to collaborate with her.

I sincerely appreciate Tilmann Gneiting for his enthusiasm and positivity when working together, and for inviting me to visit the Heidelberg Institute of Theoretical Studies. I also want to thank Lutz Dümbgen, Alexandre Mösching, Timo Dimitriadis and Marius Puke for stimulating and fruitful collaboration. Many thanks to Alexandre Mösching for careful proofreading of the introduction of my thesis. This thesis includes research with medical data, which would not have been possible without Gian-Reto Kleger, Matthias Hilty and Pedro Wendel Garcia, to whom I express my gratefulness. My gratitude goes to Peter Grünwald, Wouter Koolen, Tyron Lardy and Muriel Pérez for their hospitality during my stay at the CWI Amsterdam. I thank Bodhisattva Sen for accepting to be referee of this thesis. My work has been supported financially by the Swiss National Science Foundation, which I gratefully acknowledge.

I would like to thank all my colleagues at the Institute of Mathematical Statistics and Actuarial Science of the University of Bern. It was a pleasure to work at the institute in its friendly and familiar atmosphere. Special thanks to the participants of the institute running group for pleasurable runs in the Bremgarten forest.

I feel very happy in the light of the constant support of my family. I am thankful to my sister Anna for exchange on the joys and sorrows of life as a PhD student. My wife Qingqing has helped me over numerous periods of doubt and frustration, and words cannot express my gratitude for her patience, support and love.

Bern, 14.03.2022
Alexander Henzi

Abstract

Distributional regression estimates the probability distribution of a response variable conditional on covariates. The estimated conditional distribution comprehensively summarizes the available information on the response variable, and allows to derive all statistical quantities of interest, such as the conditional mean, threshold exceedance probabilities, or quantiles.

This thesis develops isotonic distributional regression, a method for estimating conditional distributions under the assumption of a monotone relationship between covariates and a response variable. The response variable is univariate and real-valued, and the covariates lie in a partially ordered set. The monotone relationship is formulated in terms of stochastic order constraints, that is, the response variable increases in a stochastic sense as the covariates increase in the partial order. This assumption alone yields a shape-constrained non-parametric estimator, which does not involve any tuning parameters.

The estimation of distributions under stochastic order restrictions has already been studied for various stochastic orders, but so far only with totally ordered covariates. Apart from considering more general partially ordered covariates, the first main contribution of this thesis lies in a shift of focus from estimation to prediction. Distributional regression is the backbone of probabilistic forecasting, which aims at quantifying the uncertainty about a future quantity of interest comprehensively in the form of probability distributions. When analyzed with respect to predominant criteria for probabilistic forecast quality, isotonic distributional regression is shown to have desirable properties. In addition, this thesis develops an efficient algorithm for the computation of isotonic distributional regression, and proposes an estimator under a weaker, previously not thoroughly studied stochastic order constraint.

A main application of isotonic distributional regression is the uncertainty quantification for point forecasts. Such point forecasts sometimes stem from external sources, like physical models or expert surveys, but often they are generated with statistical models. The second contribution of this thesis is the extension of isotonic distributional regression to allow covariates that are point predictions from a regression model, which may be trained on the same data to which isotonic distributional regression is to be applied. This combination yields a so-called distributional index model. Asymptotic consistency is proved under suitable assumptions, and real data applications demonstrate the usefulness of the method.

Isotonic distributional regression provides a benchmark in forecasting problems, as it allows to quantify the merits of a specific, tailored model for the application at hand over a generic method which only relies on monotonicity. In such comparisons it is vital to assess the significance of forecast superiority or of forecast misspecification. The third contribution of this thesis is the development of new, safe methods for forecast evaluation, which require no or minimal assumptions on the data generating processes.

Contents

Chapter 1

Introduction

One of the main goals of statistical modeling is to provide forecasts for the future. Because of its intrinsic uncertainty, the most natural and consistent way to predict the future is to quantify its uncertainty in the form of probability distributions (Dawid, 1984), rather than issuing deterministic point forecasts.

The tool that allows statisticians to predict the future is regression analysis, the study of the relationship between a response variable and covariates with statistical models. Once a regression model has been formulated and its unknowns have been estimated with data, it yields predictions for future realizations of the response variable conditional on the information provided by newly observed covariates. In linear regression analysis the model output is a single number, the predicted mean of the response variable, and the prediction is deterministic and does not directly imply any quantification of uncertainty. This distinguishes classical regression for the mean from distributional regression, which aims at estimating the conditional probability distribution of the response variable given the covariates, thereby providing a full quantification of the probabilities of all possible outcomes.

Probabilistic forecasting is one of the major areas of application of distributional regression methods, and many methods have been developed specifically for the purpose of forecasting. Therefore, criteria and measures for probabilistic forecast quality are essential for the understanding and evaluation of distributional regression techniques. Of central importance is the paradigm of maximizing sharpness subject to calibration (Gneiting et al., 2007), which requires that probabilistic forecasts should be as informative and concentrated as possible, while still maintaining consistency between predicted probabilities and observed event frequencies.

This thesis develops distributional regression methods and new tools to evaluate and compare probabilistic forecasts. In Chapter 2, isotonic distributional regression (IDR) is introduced as a generic method for estimating probability distributions when there is a monotone relationship between a response variable and covariates. Chapter 3 extends isotonic distributional regression to provide uncertainty quantification for point forecasts from statistical models. The topic of Chapter 4 is the evaluation of probabilistic forecasts, with a special focus on significance testing in sequential settings. The following introduction motivates these topics, elaborates on their background and connections, and gives a glance at some of the main results.

Isotonic distributional regression

Motivation and main results. The area with the most mature practical implementation of distributional regression and probabilistic forecasting is arguably weather forecasting (Gneiting and Katzfuss, 2014). Nowadays, weather forecasts are produced with advanced physical models and powerful computing systems, and ensemble forecasts are the current state-of-the-art (Bauer et al., 2015). An ensemble forecast is a collection of point forecasts, generated by running a numerical weather prediction model several times with slightly different initial conditions, each time producing a different forecast. These resulting forecasts, typically 20 up to 50, provide both estimates for the future value of a weather variable, and at the same time quantify the forecast uncertainty. Let $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ denote ensemble forecasts for a variable $Y \in \mathbb{R}$, such as accumulated precipitation. One could try to estimate event probabilities for Y by counting ensemble members, that is, for a set $B \subseteq \mathbb{R}$, define

$$\hat{P}(B) = \frac{1}{d} \sum_{i=1}^{d} \mathbb{1}\{X_i \in B\}, \tag{1}$$

where the indicator function $\mathbb{1}\{\cdot\}$ equals 1 if the statement in brackets is true and 0 otherwise. Unfortunately this approach often yields unsatisfactory results, because the predicted probabilities $\hat{P}(B)$ may strongly deviate from observed event frequencies. In spite of tremendous progress, numerical weather predictions remain subject to biases and errors, which require statistical correction. The goal of statistical post-processing is to estimate the distribution of the observation conditional on the weather predictions, $\mathcal{L}(Y \mid \mathbf{X} = \mathbf{x})$, which correctly specifies all event probabilities given the forecasts.

The post-processing of ensemble forecasts is an active field of research (Vannitsem et al., 2018). Post-processing is often done with specific models for the variable(s) at hand. For example, a model for accumulated precipitation amounts proposed in the literature relies on censored generalized extreme value (cGEV) distributions,

$$\mathcal{L}(Y \mid \mathbf{X} = \mathbf{x}) = \mathrm{cGEV}\Big(m = \alpha_0 + \alpha_1 \bar{\mathbf{x}} + \alpha_2 \sum_{i=1}^{d} \mathbb{1}\{x_i = 0\},\ \sigma = \beta_0 + \beta_1 \mathrm{MD}(\mathbf{x}),\ \xi\Big), \tag{2}$$

where the location parameter m is affine in the ensemble mean $\bar{\mathbf{x}} = (x_1 + \cdots + x_d)/d$ and the number of zero precipitation forecasts, the scale parameter σ is an affine function of the mean difference $\mathrm{MD}(\mathbf{x}) = \sum_{k,l=1,\ldots,d} |x_k - x_l|/(d(d-1))$, and the shape parameter ξ does not depend on the forecasts (Scheuerer, 2014). Censoring of the distribution at zero ensures that the model only predicts non-negative precipitation amounts. All model parameters are estimated on a training data set.

Parametric distributional regression methods like the cGEV model above have been applied with success, but this example also indicates that their development and implementation require a lot of expertise and fine-tuning. This motivates the question whether there might be a universal post-processing approach for all types of weather variables, which is free from tuning parameters and still yields reasonably precise predictions. This goal can indeed be achieved by borrowing ideas from shape-constrained regression.

Shape-constrained regression refers to non-parametric estimators under qualitative constraints such as monotonicity or convexity, see for instance the survey by Guntuboyina and Sen (2018). Such methods do not involve tuning parameters, as desired above. For the post-processing of ensemble forecasts, or more generally any point forecasts, a natural and safe qualitative assumption is that if the forecasts increase in a certain sense, then the actual observation should also tend to attain higher values. This requirement can be formalized as

$$\mathbb{P}(Y > y \mid \mathbf{X} = \mathbf{x}) \ \leq \ \mathbb{P}(Y > y \mid \mathbf{X}' = \mathbf{x}'), \ y \in \mathbb{R}, \quad \text{if } \ x_i \leq x_i', \ i = 1, \dots, d.$$

In words, when all d forecasts $\mathbf{x}'$ are greater than $\mathbf{x}$, then the conditional probability that the observation exceeds any threshold y should be higher when the forecast is $\mathbf{x}'$ than when it is $\mathbf{x}$.[1] Denoting the conditional cumulative distribution functions (CDFs) $\mathbb{P}(Y \leq y \mid \mathbf{X} = \mathbf{x})$ by $F_{\mathbf{x}}(y)$ and the componentwise ordering $x_i \leq x_i', \ i = 1, \dots, d$, by $\mathbf{x} \preceq \mathbf{x}'$, the above condition can be compactly written as

$$F_{\mathbf{x}}(y) \ \geq \ F_{\mathbf{x}'}(y), \ y \in \mathbb{R}, \quad \text{if } \ \mathbf{x} \preceq \mathbf{x}'. \tag{3}$$

Figure 1 illustrates with a real data example that (3) is indeed a plausible assumption. Condition (3) is known as stochastic dominance (Lehmann, 1955), and the estimation of CDFs under this constraint is not a new problem in statistics. It has already been analyzed in the setting of univariate covariates, that is, when $d = 1$. Brunk et al. (1966) and El Barmi and Mukerjee (2005) consider the case when the covariate takes at most two or $K < \infty$ different values, respectively, and Mösching and Dümbgen (2020) the more general case of a continuously distributed covariate. When $d = 1$, all pairs $\mathbf{x}$, $\mathbf{x}'$ of realizations of the covariate can be ordered, and thus one speaks of a total order. On the other hand, when $d > 1$, it can occur that neither $\mathbf{x} \preceq \mathbf{x}'$ nor $\mathbf{x}' \preceq \mathbf{x}$, which makes the order "$\preceq$" on $\mathbb{R}^d$ an instance of a partial order relation.

In the first part of Chapter 2 of this thesis, the problem of estimating conditional distributions under restriction (3) is analyzed in detail, and it is shown that the proposed estimation method is consistent and has desirable properties when applied to probabilistic forecasting. To give an impression of the challenges and results, let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be a training data set from which the conditional CDFs $F_{\mathbf{x}}$ are to be estimated. For a fixed threshold y, one approach to this problem is to define the least squares estimator

$$(\hat{F}_{\mathbf{x}_1}(y), \dots, \hat{F}_{\mathbf{x}_n}(y)) \ = \ \underset{p_i \geq p_j \text{ if } \mathbf{x}_i \preceq \mathbf{x}_j}{\arg\min} \sum_{i=1}^{n} (p_i - \mathbb{1}\{y_i \leq y\})^2 \ \in [0,1]^n. \tag{4}$$

For covariates $\mathbf{x} \notin \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, any interpolation method satisfying the monotonicity constraints can be applied. The estimator (4) is referred to as isotonic distributional regression, from now on abbreviated IDR. IDR satisfies the constraints in (3) by definition. However, it is not self-evident that the functions $y \mapsto \hat{F}_{\mathbf{x}_i}(y)$ define CDFs, but

[1] For ensemble forecasts this is a slight simplification for the ease of exposition. Ensemble forecasts are regarded as exchangeable, that is, the ordering of the components in $\mathbf{x}$ is arbitrary. In this case the proposed requirement is natural, but it can be weakened. See the detailed analysis in Section 2.1.

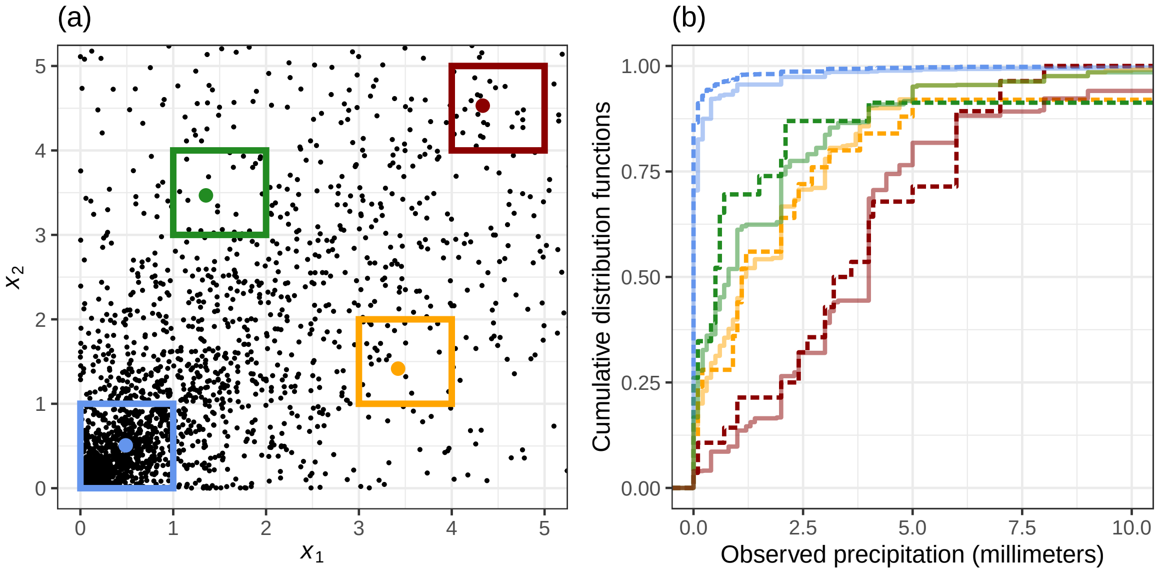


Figure 1: (a) Two ensemble forecasts for daily accumulated precipitation at Frankfurt airport, Germany. Each dot corresponds to forecasts for one calendar day from the years 2007 to 2016. Data available in the R package `isodistrreg` (Henzi. et al., 2021). (b) Empirical distribution functions of the observed precipitation (saturated colors, dashed), conditional on the precipitation forecasts taking values in the boxes with the same colors as in (a), and IDR CDFs (shaded colors) when the forecasts (x_1, x_2) are the colored larger dots in (a).

this follows from the celebrated min-max formula for monotone regression,

$$\hat{F}_{\mathbf{x}_i}(y) \;=\; \min_{U \in \mathcal{U}:\, \mathbf{x}_i \in U} \; \max_{V \in \mathcal{U}:\, \mathbf{x}_i \notin V} \; \frac{1}{\#(U \cap V^c)} \sum_{j:\, x_j \in U \cap V^c} \mathbb{1}\{y_j \leq y\}, \tag{5}$$

where $\mathcal{U}$ are the upper sets in $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and V^c is the complement $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \setminus V$. A set U is an upper set if $u \in U$ implies that $v \in U$ for all v with $u \preceq v$. Formula (5) can be found, for instance, in the monographs by Barlow et al. (1972) or Robertson et al. (1988), and it sheds more light on IDR. Panel (b) of Figure 1 depicts IDR CDFs $\hat{F}_{\mathbf{x}}$ for a real data example on the post-processing of precipitation forecasts. The CDFs are piecewise constant, an immediate consequence of (5), and on each constant piece the value of the IDR CDF is the empirical frequency of observations y_j satisfying $y_j \leq y$, where the corresponding $\mathbf{x}_j$ lie in a certain set $U \cap V^c$ depending both on the covariate $\mathbf{x}$ and on the threshold y. The IDR CDFs $\hat{F}_{\mathbf{x}}$ look similar to the empirical CDFs of the observations when the covariates take values in the boxes around $\mathbf{x}$ in panel (a). However, IDR does not require the manual, arbitrary specification of such a neighborhood, and avoids unintuitive crossings of the conditional CDFs when $\mathbf{x} \preceq \mathbf{x}'$.

With the application of post-processing weather forecasts in mind, is natural to further investigate IDR with respect to criteria for probabilistic forecast quality. Gneiting et al. (2007) propose the paradigm that probabilistic forecasts should maximize sharpness subject to calibration. Calibration means that observed event frequencies should conform with the probabilities derived from a probabilistic forecast. The sharpness principle states that probabilistic forecasts ought to be informative, ideally with predicted probabilities close to zero or one for most events of interest.

It will be shown that IDR has the remarkable property that conditional probabilities of threshold (non-)exceedance in the training data set are always equal to the predicted probabilities. Namely, for all thresholds y, it holds

$$P_n(Y \leq y \mid \hat{F}_{\mathbf{X}}(y)) = \hat{F}_{\mathbf{X}}(y).$$

Here P_n denotes the empirical distribution of the training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, and $(\mathbf{X}, Y) \sim P_n$. While this is no guarantee for correct calibration out-of-sample, it indicates that IDR CDFs should have good calibration properties provided that the training data set is large enough. The unconditional version of the above equation is

$$P_n(Y \leq y) = \frac{1}{n} \sum_{i=1}^{n} \hat{F}_{\mathbf{x}_i}(y),$$

known as marginal calibration (Gneiting et al., 2007). It shows that the IDR CDFs decompose the unconditional empirical distribution function $P_n(Y \leq y)$ into sharper conditional distributions.

A standard tool for evaluating and comparing probabilistic forecasts are proper scoring rules (Gneiting and Raftery, 2007), which assess calibration and sharpness simultaneously. A proper scoring rule is a loss function $S = S(P, Y)$ mapping a probabilistic forecast P and an observation Y to a numerical score, such that

$$\mathbb{E}_P[S(P, Y)] \; \leq \; \mathbb{E}_Q[S(P, Y)]$$

for all P, Q in a certain family of probability measures such that the expectations $\mathbb{E}_P[\cdot]$, $\mathbb{E}_Q[\cdot]$ with respect to P, Q are well-defined. Under proper scoring rules, the correct forecast for Y attains a minimal expected score, which makes them suitable loss functions for estimation and forecast comparison. A popular choice is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976), defined for real-valued outcomes and predictive CDFs,

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 \, dz.$$

It can be seen that IDR, which was defined rather ad-hoc in (4) as least squares estimator, also minimizes the CRPS over the training data, $\sum_{i=1}^{n} \mathrm{CRPS}(G_{\mathbf{x}_i}, y_i)$, among all distribution functions $G_{\mathbf{x}_1}, \dots, G_{\mathbf{x}_n}$ satisfying the stochastic order constraints (3). Even more, since IDR minimizes the integrand of the CRPS pointwise, it simultaneously minimizes all weighted versions of the CRPS (Gneiting and Ranjan, 2011),

$$\mathrm{CRPS}_w(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 \, w(z) dz,$$

which are also proper scoring rules with non-negative weight functions w.

It is well known that monotone regression for the mean has similar optimality properties, see Barlow et al. (1972). Precisely, in the same way as IDR simultaneously minimizes all weighted versions of the CRPS, monotone regression for the mean is simultaneously optimal with respect to so-called consistent scoring functions for the mean (Gneiting, 2011). Jordan et al. (2021+) have shown that such results hold for monotone regression for more functionals than only the mean. Their results, in conjunction with the elementary score decompositions for quantiles and expectiles by Ehm et al. (2016), yield a more comprehensive characterization of the proper scoring rules which IDR simultaneously minimizes. Invariance under the choice of loss function is a special property of monotone regression estimators. Most regression techniques, such as the parametric cGEV model introduced earlier, yield different estimates on training data when the loss function is changed. This may be problematic since there is often no stringent criterion for preferring one loss function over another one (Patton, 2020).

Guaranteed in-sample calibration and invariance with respect to the choice of the loss function make IDR an ideal benchmark for probabilistic forecasting problems, in particular in the post-processing of weather forecasts but more generally in any situation with a monotone relationship between covariates and observations. Due to its generality, it cannot be expected that IDR outperforms models tailored to a specific problem, but it gives a benchmark relative to which the merits of such tailored methods can be assessed. To illustrate this and complement the simulations and data application from Section 2.1, consider the results of a study on the post-processing of ensemble forecasts for wind speed in Germany by Schulz and Lerch (2021+, Table 3). Averaged over different stations and forecast lead times, the CRPS of the raw ensemble, converted to a probabilistic forecast as in (1), equals 1.33. A parametric post-processing model reduces this error to 0.95, and the best-performing method, a distributional neural network, achieves a CRPS of 0.84. The CRPS of IDR is 0.98 and thereby higher than that of the other post-processing methods. But the relative reduction in CRPS compared to the raw ensemble differs only by 2.3 percentage points between the parametric model and IDR, which shows that the average gains of the parametric model over a generic benchmark are not large in this application.

Computational aspects. A challenge in the computation of IDR is that it requires to solve the minimization problem (4) for all thresholds y. If $\tilde{y}_1 < \cdots < \tilde{y}_m$ denote the distinct values of $\{y_1, \ldots, y_n\}$, then the squared error $\sum_{i=1}^{n}(p_i - \mathbb{1}\{y_i \leq y\})^2$ as a function of y is constant in between $\tilde{y}_j$ and $\tilde{y}_{j+1}$, so it is sufficient to compute IDR only for $y = \tilde{y}_1, \ldots, \tilde{y}_m$. But even with this simplification one still has to solve up to n constrained minimization problems with n variables.

In Section 2.2 the relationship between the solutions of the minimization problem

$$A(\mathbf{z}) = \operatorname*{arg\,min}_{\theta_i \geq \theta_j \text{ if } x_i \preceq x_j} \sum_{i=1}^{n} (\theta_i - z_i)^2$$

as a function of the vector $\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{R}^n$ is investigated, where $x_1, \ldots, x_n$ are covariates in a general space $\mathcal{X}$ equipped with some binary relation "$\preceq$". It is shown that when vectors $\mathbf{z}$ and $\tilde{\mathbf{z}}$ only differ in one or few components, then also $A(\mathbf{z})$ and $A(\tilde{\mathbf{z}})$ often are equal in most components, and if $A(\mathbf{z})$ is already available, then $A(\tilde{\mathbf{z}})$ can be computed from it and $\tilde{\mathbf{z}}$ with few operations. This directly applies to IDR, where the vectors $(\mathbb{1}\{y_1 \leq y\}, \ldots, \mathbb{1}\{y_n \leq y\}) \in \{0, 1\}^n$ only change in one component when $y_1, \ldots, y_n$ are all distinct and y is increased from $\tilde{y}_j$ to $\tilde{y}_{j+1}$. In the case of a total order, the results give rise to an abridged version of the Pool-Adjacent-Violators Algorithm (PAVA; de Leeuw et al., 2009) that reduces the computation time of IDR by a factor of up to 100 for relevant sample sizes, when compared to a naive implementation.

Weaker stochastic orders. Isotonic distributional regression estimates conditional CDFs under the constraint that $F_{\mathbf{x}}(y) \geq F_{\mathbf{x}'}(y)$ for all $y \in \mathbb{R}$ if $\mathbf{x} \preceq \mathbf{x}'$. It was claimed that this approach is natural for the post-processing of point forecasts, because higher forecasts should imply that the observation also more likely attains higher values. However, there are special situations where this is not true, because the variability of an outcome variable strongly increases or decreases with the forecast. The example given in Section 2.3 are income expectations. In economic surveys, respondents who expect a very low income in future are sometimes overly pessimistic and, when questioned in the second round of the survey, ultimately have an income above a substantial percentage of respondents who expressed higher expectations. Formally, if x denotes the expected future income and F_x the conditional CDF of the realized future income, then the CDFs F_x and $F_{x'}$ may cross in the upper tail, for certain values of x and x'.

The stochastic dominance relation considered until now in this introduction is only one instance of a stochastic order, and many more are studied in the monograph by Shaked and Shanthikumar (2007). A solution to the problem of crossing CDFs is to perform distributional regression under second order stochastic dominance (SSD), a weaker constraint than stochastic dominance. If F and G are CDFs, then F is smaller than G in SSD if

$$\int_{-\infty}^{y} F(t)\,dt \ \geq \ \int_{-\infty}^{y} G(t)\,dt, \ y \in \mathbb{R}.$$

This condition allows that F and G cross in the upper tail. The estimation of conditional distributions under SSD constraints is more involved than under stochastic dominance. It has been considered previously by Rojo and El Barmi (2003) and El Barmi and Marchev (2009), but only for two samples, i.e. a binary covariate. In Section 2.3, consistent estimators for conditional CDFs under SSD constraints for general real-valued covariates are developed.

Distributional index models

Motivation and main results. IDR assumes that the covariate $\mathbf{X}$ exhibits a monotone relationship with the response variable Y. Point forecasts from an external source, such as numerical weather predictions, are the leading example for such a situation. But what if the point forecasts are generated by the statistician, with a regression model that is based on parameters estimates from the same data which is intended for the estimation of the conditional distributions with IDR?

More formally, assume that a data set $(z_1, y_1), \ldots, (z_n, y_n) \in \mathcal{Z} \times \mathbb{R}$ with a general covariate space $\mathcal{Z}$ is used to estimate a regression function $\hat{\theta}_n : \mathcal{Z} \mapsto \mathbb{R}$ generating point forecasts. One could now again argue that as $\hat{\theta}_n(z)$ increases, the outcome variable should increase too, and apply IDR to the data $(\hat{\theta}_n(z_1), y_1), \ldots, (\hat{\theta}_n(z_n), y_n) \in \mathbb{R} \times \mathbb{R}$. However, $\hat{\theta}_n$ is a function of $(z_1, y_1), \ldots, (z_n, y_n)$, and it is not sensible to impose assumptions on the relationship between the outcome variable and an estimator $\hat{\theta}_n$

which depends on the data itself. A better model is to assume that the dependency between Y and Z is fully described by an unknown underlying function $\theta : \mathcal{Z} \mapsto \mathbb{R}$, that is, $\mathbb{P}(Y \leq y \mid Z = z) = F_{\theta(z)}(y)$ for some family of CDFs $F_{\theta(z)}$, $z \in \mathcal{Z}$, and that the conditional distributions are increasing in stochastic dominance as $\theta(z)$ increases,

$$F_{\theta(z)}(y) \ \geq \ F_{\theta(z')}(y), \ y \in \mathbb{R}, \quad \text{ if } \theta(z) \leq \theta(z'). \tag{6}$$

Many classical models, such as the linear model with homoscedastic Gaussian errors and certain generalized linear models (McCullagh and Nelder, 1989), impose assumptions which are stronger than (6). The difference to these parametric models is that the distribution functions F_u for $u \in \theta(\mathcal{Z})$ are not specified in (6), resulting in a semi-parametric model. This is similar to single index models for the mean (Härdle et al., 1993), which postulate that $Y = g(\alpha^\top Z) + \varepsilon$ for $\alpha, Z \in \mathbb{R}^p$, an unspecified (smooth) function g and a zero-mean error term ε. The model with assumption (6) is therefore called distributional (single) index model (DIM).

Having formulated model (6), the question of main interest is whether the application of IDR to the transformed data $(\hat{\theta}_n(z_1), y_1), \ldots, (\hat{\theta}_n(z_n), y_n)$ can possibly yield a consistent estimator for the conditional distribution functions $F_{\theta(z)}$. As shown in Section 3.1, this is indeed the case when $\hat{\theta}_n$ is consistent for θ at a sufficiently fast rate and certain regularity conditions hold. This justifies the use of IDR for the post-processing of statistical point forecasts, which may even be generated with a regression model that is estimated on the same data to which IDR is applied in a second step.

The application of the DIM in Section 3.1 is the prediction of the length of stay (LoS) of patients in intensive care units (ICUs), based on patient data available at the latest 24 hours after admission. Models generating point forecasts for the LoS have already been proposed in the literature (Verburg et al., 2017; Kramer, 2017), but they are of limited usefulness for predicting individual patients' LoS because, even conditional on many patient specific covariates, the uncertainty in the LoS is high. However, the point forecasts from these models can be combined with IDR to a distributional index model. In Section 3.1, it is shown that this combination produces calibrated probabilistic forecasts, which outperform other distributional regression methods in an application on predicting the LoS of patients in Swiss ICUs.

Application to COVID-19 patients' intensive care unit length of stay. In 2019, when the article on the distributional index model was written, no one anticipated what relevance the problem of predicting ICU patients' length of stay would gain only a few months later. The COVID-19 pandemic, which began in late 2019 and started spreading around the world in early 2020, induced a considerable strain on intensive care unit resources. Apart from the high number of patients requiring treatment, a key problem is the frequent need of prolonged ICU treatment of severely ill COVID-19 patients. Bed planning in ICUs and estimating ICU capacity therefore depend on knowledge about how long patients are expected to be in an ICU. The article in Section 3.2, written during and after the first COVID-19 wave in Switzerland, applies the DIM to provide probabilistic forecasts for COVID-19 patients' ICU length of stay.

New methods for forecast evaluation

Motivation. Probabilistic forecasts should be calibrated and sharp, and these properties are usually evaluated on a test data set where the forecasts are compared with actual observations. But how can one verify if forecast miscalibration is statistically significant, or if a forecast is significantly superior to a benchmark like IDR? Regarding calibration testing, Gneiting et al. (2007) state that

> *the use of formal tests is often hindered by complex dependence structures,*

and this statement also applies to forecast comparison. In many real situations, such as the evaluation of weather forecasts, observations are not independent and identically distributed, but rather characterized by (mostly) unknown dependence over space and time. This complex dependence makes the application of many statistical tests impossible, or at least questionable.

To illustrate these difficulties, consider one of the most influential contributions to testing forecast superiority, the Diebold-Mariano test (Diebold and Mariano, 1995). For error series $e_{1,t}, e_{2,t}$, $t = 1, \dots, T$, of two competing forecasts, such as the CRPS of of probabilistic weather forecasts, the Diebold-Mariano test is based on the statistic

$$\frac{1}{(T\hat{\sigma}_T^2)^{1/2}} \sum_{t=1}^{T} (e_{1,t} - e_{2,t}),$$

where $\hat{\sigma}_T^2$ is a variance estimator for the error differences $d_t = e_{1,t} - e_{2,t}$, $t = 1, \dots, T$. Negative values of the test statistic imply that forecast 1 is superior to forecast 2, because it achieves a smaller error on average. Under the null hypothesis of equal expected forecast errors and under certain regularity conditions, the asymptotic distribution of this test statistic is standard Gaussian, which allows to test the significance of forecast superiority.

Unfortunately, it is often not clear in practice if the asymptotic theory behind the Diebold-Mariano test applies. For example, Diebold and Mariano (1995) state that asymptotic normality holds if the score differentials d_t, $t = 1, \dots, T$ are a stationary process and if the variance estimator is consistent. But stationarity is often unplausible, for example in weather forecasting, where forecasts exhibit different errors depending on the season and on weather regimes. The selection of an unsuitable variance estimator may further impair the test validity. Giacomini and White (2006) show that similar tests of forecast superiority are valid under much weaker assumptions than stationarity, but still, their asymptotic theory rules out many practically relevant situations. For instance, it is not allowed that parameters of regression models generating the forecasts are estimated in an expanding window fashion, that is, use all past observations from times $1, \dots, t-1$ to produce the forecast for time t. However, this procedure is frequently applied in forecasting, also in the case study of Section 2.1, and it is unclear whether the p-values derived from these tests are valid in such situations.

Hypothesis testing with e-values. Most forecasting situations are sequential. Predictions are issued at discrete time points $t = 1, 2, 3, \dots$, and the prediction at time t

refers to an observation which is revealed at $t+h$ for some lag $h \in \mathbb{N}$. For example, this setting encloses daily weather forecasts issued for the next day ($h = 1$), or quarterly inflation forecasts two quarters ahead ($h = 2$). The observations are not independent, neither are the forecast errors, and often little is known about the dependency structure. Moreover, when producing the forecast at time t the forecaster knows all past observations and may use this information in the prediction.

While this setting makes the application of classical statistical tests difficult, it is suitable for sequential testing procedures. Starting from 2019, there has been a surge in new contributions to the field of sequential testing by various authors (Grünwald et al., 2019; Ramdas et al., 2020; Shafer, 2021; Vovk and Wang, 2021). Their methods rely on the concepts of e-values and test martingales, and in Sections 4.1, 4.2, and 4.3 of this thesis, these tools are used to develop new tests for forecast evaluation.

An e-value is a non-negative random variable E such that for all probability distributions $\mathbb{P}$ in a set $\mathcal{H}$ representing the null hypothesis,

$$\mathbb{E}_{\mathbb{P}}[E] \ \leq \ 1.$$

In words, E has expected value less or equal to 1 under the null hypothesis, and high values of E provide evidence against the null hypothesis, since by Markov's inequality,

$$\sup_{\mathbb{P} \in \mathcal{H}} \mathbb{P}(E \geq 1/\alpha) \leq \alpha, \quad \alpha \in (0, 1).$$

In particular, $1/E$ is a conservative p-value. E-values also have a financial interpretation in terms of betting (Shafer, 2021). If one unit of money is invested into an e-value to bet against the null hypothesis, then E gives the factor by which the investment is multiplied after the observations are revealed. If the null hypothesis is true, then one cannot expect to gain money with the bet E, on average.

One main motivation for using e-values is their simple behavior under combinations. The arithmetic mean of e-values is again an e-value, and so is the product of independent e-values. Moreover, and most importantly for forecast evaluation, if $(E_t)_{t \in \mathbb{N}}$ is sequence of conditional e-values adapted to a filtration $(\mathcal{F})_{t \in \mathbb{N}}$, that is, $E_t \geq 0$ almost surely and $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \leq 1$ for all t, then the cumulative product $e_t = \prod_{i=1}^{t} E_i$ is an e-value, and the process $(e_t)_{t \in \mathbb{N}}$ is a non-negative supermartingale, satisfying

$$\mathbb{P}\Big(\sup_{t \in \mathbb{N}} e_t \geq 1/\alpha \Big) \leq \alpha, \tag{7}$$

by optional stopping theorems for martingales (see for example Ramdas et al., 2020). Non-negative supermartingales with initial value one are also called test martingales. The above inequality implies that if $(e_t)_{t \in \mathbb{N}}$ exceeds the level $1/\alpha$ at least once, then the null hypothesis $\mathcal{H}$ can be rejected at the level α. More generally, when deciding to stop or continue the hypothesis testing at a time point τ, one may take into account the values e_t, $t = 1, \dots, \tau$, observed so far, which does not impair the validity of the test. This is because the process $(e_t)_{t \in \mathbb{N}}$ satisfies $\mathbb{E}[e_\tau] \leq 1$ for any stopping time τ. This property is in strong contrast to non-sequential testing procedures, where data-dependent optional stopping or continuation may dramatically inflate the rate of false rejections of the null hypothesis.

Probability forecast comparison. Section 4.1 addresses the problem of testing probability forecast superiority. Two probability forecasts $p_t, q_t \in [0,1]$ for a binary event $Y_{t+h} \in \{0,1\}$ are compared with a proper scoring rule S. For the ease of exposition, consider the case of forecast lag $h = 1$. The null hypothesis is that p_t achieves a smaller error than q_t at all times t, given the information at the time of forecasting, which is represented by

$$\mathcal{H} = \{\mathbb{P}\colon \mathbb{E}_{\mathbb{P}}[S(p_t, Y_{t+1}) - S(q_t, Y_{t+1}) \mid \mathcal{F}_t] \leq 0 \text{ almost surely, } t \in \mathbb{N}\}. \tag{8}$$

Here the distributions $\mathbb{P}$ are distributions generating the process $(p_t, q_t, Y_t)_{t\in\mathbb{N}}$ of forecasts and observations, which is adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. The null hypothesis (8) is of interest when the p_t are a benchmark, like predictions with IDR, since then the new or more specific prediction method q_t should significantly outperform p_t.

An e-value with the null hypothesis (8) should be smaller than 1 if the observed score differences $S(p_t, Y_{t+1}) - S(q_t, Y_{t+1})$ are negative, and greater than 1 if they are positive. This suggests to define the e-value at time point t by

$$E_{p_t,q_t;\lambda_t}(Y_{t+1}) = 1 + \lambda_t \frac{S(p_t, Y_{t+1}) - S(q_t, Y_{t+1})}{|S(p_t, \mathbb{1}\{p_t > q_t\}) - S(q_t, \mathbb{1}\{p_t > q_t\})|}, \tag{9}$$

for some $\lambda_t \in (0,1]$, where the normalization factor in the denominator ensures non-negativity. Interestingly, as shown in Section 4.1, this is essentially the only way for constructing e-values at a single time point t. The parameter λ_t has to be specified by the test user. Its role is fundamentally different from parameters in classical statistical tests, such as the variance estimator $\hat{\sigma}_T^2$ in the Diebold-Mariano test, which may crucially influence the validity of the test. Under the null hypothesis, $E_{p_t,q_t;\lambda_t}(Y_{t+1})$ is an e-value for any choice of λ_t, but λ_t needs to be tuned well in order to maximize power when the null hypothesis is violated. Furthermore, λ_t can be chosen sequentially based on past data and on (p_t, q_t), since for the validity of the e-value, it is is only necessary that λ_t is $\mathcal{F}_t$-measurable.

E-values of the form (9) can be combined into the following product,

$$e_T = \prod_{i=1}^{T-1} E_{p_i,q_i;\lambda_i}(Y_{i+1}),$$

and this cumulative product should grow fast with time T if the null hypothesis is not true. A good strategy is to choose $\log(e_T) = \sum_{i=1}^{T-1} \log(E_{p_i,q_i;\lambda_i}(Y_{i+1}))$ as a target criterion, penalizing small or even zero e-values which may have a devastating impact on the power of the test. The (expected) logarithm of an e-value is also referred to as the growth rate, and maximizing the growth rate under an alternative hypothesis is suggested by Grünwald et al. (2019) as a method for constructing powerful e-values.

The e-values introduced above yield a valid test for forecast superiority without imposing any assumptions on the data generating process, which is in strong contrast to other methods such as the Diebold-Mariano test. Also, thanks to property (7), the null hypothesis can be rejected as soon as the process $(e_t)_{t\in\mathbb{N}}$ exceeds the level $1/\alpha$ for the first time, without having to fix a sample size in advance. These advantages come at the price of reduced power. But as simulations and a case study in Section 4.1 demonstrate, it is usually possible to draw the same conclusions with e-values as with extant tests for forecast superiority.

Testing probabilistic calibration. Calibration requires that probabilistic forecasts are consistent with observed event frequencies. This definition is unambiguous if the event of interest is binary, but for $Y \in \mathbb{R}$ many different probabilities can be derived from a predictive CDF F, and so there exist many different notions of calibration. A popular notion is probabilistic calibration (see for example Gneiting et al., 2007), which requires that

$$Z_F(Y) = F(Y-) + V(F(Y) - F(Y-)) \sim \mathrm{UNIF}(0,1),$$

where $F(Y-) = \lim_{z \uparrow Y} F(z)$, $\mathrm{UNIF}(0,1)$ denotes the uniform distribution on $(0,1)$, and $V \sim \mathrm{UNIF}(0,1)$ is independent of (F, Y). The quantity $Z_F(Y)$ is called probability integral transform (PIT). In words, this definition states that the probability that Y is less or equal to the predicted α-quantile should equal α, for all $\alpha \in (0,1)$, with suitable randomization in case the predictive CDF F has discontinuities. An ubiquitous tool in forecast evaluation are histograms of PIT samples, where forecast biases become visible as skewed PIT distributions, and errors in the variability, so-called dispersion errors, in the form of U- or inverse U-shaped histograms.

The definition of probabilistic calibration above is non-sequential, because it is only based on a single random forecast-observation pair (F, Y). In practice, one observes a sequence $(F_t, Y_t)_{t \in \mathbb{N}}$ of forecasts and observations, and the definition of calibration must specify properties of the whole sequence. Considering again a forecast with lag 1 for simplicity, a reasonable extension of the definition is to require

$$\mathcal{L}(Z_{F_t}(Y_{t+1}) \mid Z_{F_j}(Y_{j+1}),\, j < t) = \mathrm{UNIF}(0,1),\ \ t \in \mathbb{N}.$$

This means that, for a forecast to be calibrated, the distribution of the PIT at forecast time t should be uniform, conditional on all values for the PIT that have been observed in the past, which implies independence of $(Z_{F_t}(Y_{t+1}))_{t \in \mathbb{N}}$.

To test this hypothesis with e-values one can follow a similar strategy as for the comparison of probability forecasts, namely, first consider the task of testing uniformity of $Z_{F_t}(Y_{t+1})$ for a single t. An e-value $E_t = E_t(z)$ must satisfy $E_t \geq 0$ and

$$\mathbb{E}[E_t(Z_{F_t}(Y_{t+1})) \mid Z_{F_j}(Y_{j+1}),\, j < t] = \int_0^1 E_t(z)\, dz = 1,$$

since $Z_{F_t}(Y_{t+1})$ is uniformly distributed conditional on $Z_{F_j}(Y_{j+1})$, $j < t$, under the null hypothesis. That means, E_t can be constructed by applying a density estimator $\hat{f}$ to the available PIT observations and setting $E_t(z) = \hat{f}(z;\, Z_{F_j}(Y_{t+j}), j < t)$. In Section 4.2, it is suggested to compute $\hat{f}$ with beta distributions with parameters estimated from $Z_{F_j}(Y_{t+j})$, $j < t$, or with suitable kernel density estimators,

$$\hat{f}(z;\, Z_{F_j}(Y_{t+j}), j < t) \ = \ \frac{1}{b(t-1)} \sum_{j=1}^{t-1} \kappa\left(\frac{z - Z_{F_j}(Y_{t+j})}{b}\right), \tag{10}$$

for some kernel density κ and bandwidth $b > 0$ that may depend on t. If the forecast is not probabilistically calibrated and the miscalibration persists over time, then $Z_{F_j}(Y_{t+1})$

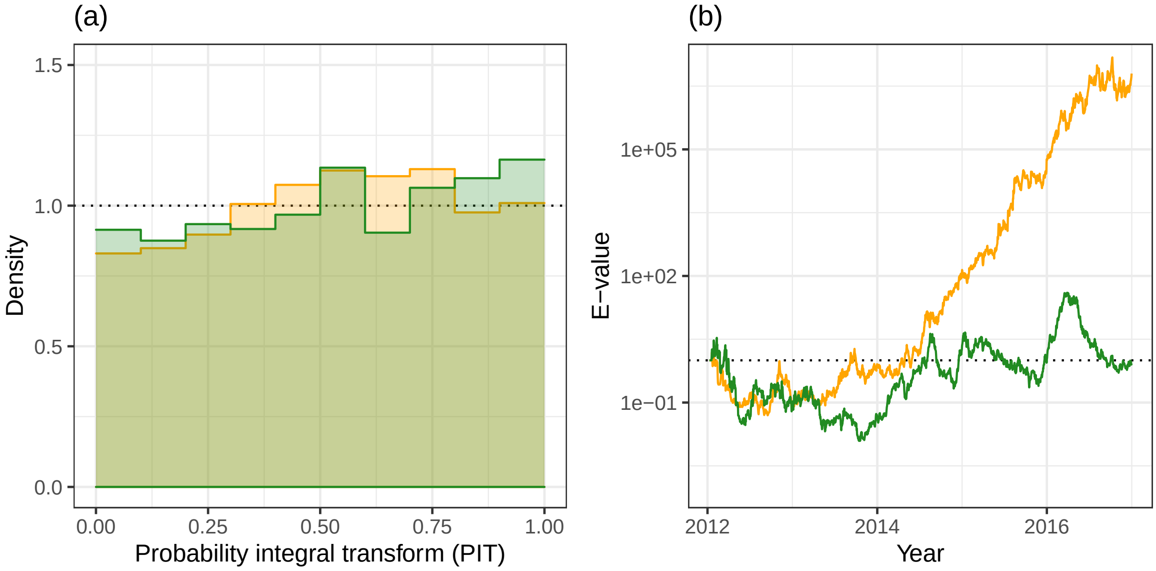


Figure 2: (a) PIT histograms for IDR (green) and cGEV (orange) post-processed accumulated precipitation forecasts for Frankfurt airport, Germany. (b) Cumulative products of e-values for testing probabilistic calibration of the forecasts (IDR: green, cGEV: orange). The dotted black line shows the level 1.

should be more likely to attain values in regions where $\hat{f}$ is greater than 1, accumulating evidence against the null hypothesis. Furthermore, for forecast lag 1, the e-values can again be combined into the cumulative product $e_T = \prod_{t=1}^{T-1} E_t(Z_{F_t}(Y_{t+1}))$.

Figure 2 shows PIT histograms and e-values for testing probabilistic calibration of post-processed daily accumulated precipitation forecasts. The data are weather station observations for Frankfurt airport, Germany, from the years 2007 to 2016, and ensemble forecasts by the European Centre for Medium-Range Weather forecasts (Molteni et al., 1996). The ensemble consists of 50 forecasts. Post-processing models are estimated on data from the years 2007 to 2011 and validated on 2012 to 2016. A cGEV model (2) is estimated by CRPS minimization, and compared with a simple univariate IDR taking only the ensemble mean as covariate. The e-values are constructed sequentially with the kernel density estimator in (10) and described in detail in Section 4.2. One advantage of e-values compared to classical tests is that the evidence against the null hypothesis can be monitored over time. Both post-processing methods have small e-values in the years 2012 to 2014, and start to increase more steeply after 2014. This is an indication that the dependence between the ensemble forecasts and observations changes over time, causing miscalibration, and that model parameters should be re-estimated with more recent data. Such information is valuable for practitioners, who need to understand when and why miscalibration occurs in order to improve the forecast quality. The e-values for the cGEV post-processed forecasts reach a maximum of $1.5 \cdot 10^7$, corresponding to a p-value of about $6.7 \cdot 10^{-8}$. For IDR the maximum is 40, which by (7) allows rejection of the null hypothesis at the level 0.025, even if the e-value at the end of the validation period is close to 1 and does not give any evidence against the null hypothesis.

Testing probabilistic calibration amounts to testing if the PIT follows the continuous uniform distribution on $[0, 1]$. There are other, closely related notions of calibration which require testing the null hypothesis of a discrete uniform distribution, and of stochastic order relations compared to the uniform distribution. These problems are also analyzed in Section 4.2.

Assessing calibration of probability forecasts. The last two chapters of this thesis treat calibration assessment for probability forecasts. A probability forecast $P \in [0, 1]$ for $Y \in \{0, 1\}$ is calibrated if

$$\mathbb{P}(Y = 1 \mid P = x) = x,$$

for all x in the support of P. To test calibration, one compares the predicted and observed event frequencies for all values of the predictions P. This is a standard problem if P only attains few different values, but it becomes delicate if P is continuously distributed, in which case a data set $(x_i, Y_i) \in [0, 1] \times \{0, 1\}$, $i = 1, \dots, n$, only contains one observation for each value of $x_1, \dots, x_n$.

Probably the most popular test of calibration is due to Hosmer and Lemeshow (1980). To compute it, one starts by partitioning the interval $[0, 1]$ into g bins I_k, $k = 1, \dots, g$, typical choices being $[0, 0.1], (0.1, 0.2], \dots, (0.9, 1]$ or intervals delimited by quantiles of $x_1, \dots, x_n$. Then, one compares the observed and expected event frequencies in each bin,

$$o_{jk} = \sum_{i\colon x_i \in I_k} \mathbb{1}\{Y_i = j\},\ j = 0, 1, \quad \hat{e}_{1k} = \sum_{i\colon x_i \in I_k} x_i, \quad \hat{e}_{0k} = \sum_{i\colon x_i \in I_k} (1 - x_i).$$

Under independence and calibration, the statistic

$$\sum_{j=1}^{k} \left(\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right)$$

asymptotically follows a χ^2-distribution with $g - 2$ degrees of freedom. The graphical counterpart of the Hosmer-Lemeshow test are reliability diagrams (Murphy and Winkler, 1977), where observed event frequencies for each bin are plotted against the bin midpoint. Both the Hosmer-Lemeshow test and reliability diagrams have the drawback that they may be strongly influenced by the choice of the bins, which is usually arbitrary and may cause untenable instabilities; see Bertolini et al. (2000) and Section S2 in the supplementary material of Dimitriadis et al. (2021).

The instability problem of the Hosmer-Lemeshow test and reliability diagrams lead us back to isotonic distributional regression. One of the main motivations for isotonic distributional regression is that it provides a tuning-parameter free benchmark method for generating probabilistic forecasts. To construct stable reliability diagrams without manual binning, Dimitriadis et al. (2021) propose to plot $x_1, \dots, x_n$ against their isotonic re-calibration $\hat{p}(x_1), \dots, \hat{p}(x_n)$,

$$(\hat{p}(x_1), \dots, \hat{p}(x_n)) \;=\; \underset{p_1 \le \cdots \le p_n}{\arg\min} \sum_{i=1}^{n} (p_i - Y_i)^2, \tag{11}$$

assuming $x_1 < \cdots < x_n$ for simplicity. This is equivalent to isotonic distributional regression, since for a binary outcome variable the expected value equals the probability of outcome 1. The isotonicity assumption is not a strong restriction, because most probability predictions exhibit a monotone relationship with the event rate, and otherwise the predictions can often be discarded without any deeper analysis.

The proposal by Dimitriadis et al. (2021) solves the instability problem of reliability diagrams, and the last two sections of this thesis complement their work by developing stable alternatives to the Hosmer-Lemeshow test. In Section 4.3, the e-Hosmer-Lemeshow (eHL) test is proposed, a test based on e-values which is extensible to sequential settings. If $Y_1, \ldots, Y_n$ are independent, then for any validation subset $\mathcal{V} \subseteq \{1, \ldots, n\}$ and given $q_i \in [0, 1]$, $i \in \mathcal{V}$, the likelihood ratio

$$E_{\mathcal{V}} = \prod_{i \in \mathcal{V}} \frac{q_i^{Y_i}(1 - q_i^{1-Y_i})}{x_i^{Y_i}(1 - x_i^{1-Y_i})}$$

is an e-value under the null hypothesis of calibration. To specify q_i, $i \in \mathcal{V}$, one may estimate the true conditional event probabilities $p(x) = \mathbb{P}(Y = 1 \mid P = x)$ on the remaining part (x_i, Y_i), $i \notin \mathcal{V}$, of the data. Any estimation method is admissible for this purpose, but to complement the reliability diagrams by Dimitriadis et al. (2021) and avoid tuning parameters or dependency on the loss function, the proposal in Section 4.3 is to use isotonic regression.[2] This is also a sensible choice from the perspective of e-values. For binary outcomes, isotonic regression yields the maximum likelihood estimator of the event probabilities under the constraint of isotonicity (Barlow et al., 1972), and hence it maximizes the growth rate

$$\log(E_{\{1,\ldots,n\}\setminus\mathcal{V}}) = \sum_{i \in \{1,\ldots,n\}\setminus\mathcal{V}} \left(Y_i \log \frac{q_i}{x_i} + (1 - Y_i) \log \frac{1 - q_i}{1 - x_i} \right),$$

over all q_i, $i \notin \mathcal{V}$, such that $q_i \leq q_j$ if $x_i \leq x_j$.

Usually, there is no generic choice for a validation subset $\mathcal{V}$. Instead, one often splits the data randomly into two subsets $\mathcal{V}$ and $\{1, \ldots, n\} \setminus \mathcal{V}$. To make the e-value independent of the data split, one can repeat this procedure B times with different splits $\mathcal{V}_1, \ldots, \mathcal{V}_B$, and average the resulting e-values, $E = \sum_{b=1}^{B} E_{\mathcal{V}_b}/B$. For example, $\mathcal{V}_1, \ldots, \mathcal{V}_b$ could be all (or sufficiently many) subsets of $\mathcal{V}$ of size $\lceil n/2 \rceil$ drawn without replacement. Such a data splitting and de-randomization is not necessary if (x_i, Y_i), $i = 1, \ldots, n$, are observed at sequential time points or have another natural ordering. In this case, one can apply analogous strategies as in Sections 4.1 and 4.2, that is, estimate the conditional event probabilities with the pairs $(x_1, Y_1), \ldots, (x_{i-1}, Y_{i-1})$ to generate q_i, and combine the e-values with the cumulative product. In such a sequential setting, independence is not required.

The Hosmer-Lemeshow and eHL test only allow to reject the null hypothesis of calibration, but they do not give an indication of how serious the miscalibration is. With large sample size n, the tests tend to reject the null hypothesis even for acceptably

[2] To be fair, also isotonic regression is not completely free of implementation decisions. It requires the specification of an interpolation method for out-of-sample predictions, and for e-values one should avoid predicted probabilities of exactly 0 or 1, which may cause the e-value to attain zero.

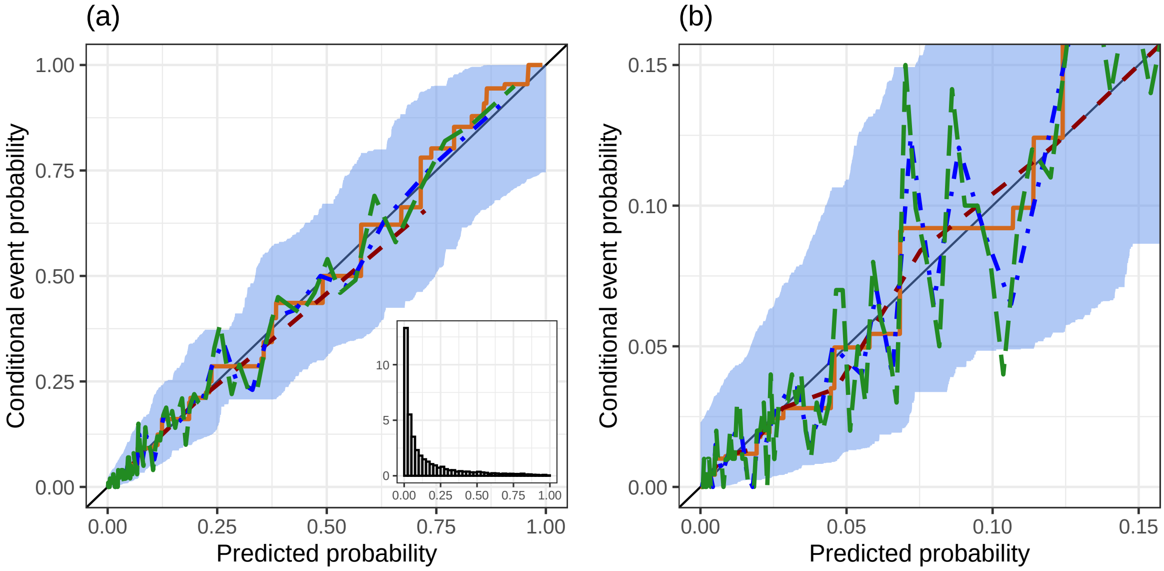


Figure 3: (a) Predicted and conditional event probabilities for the simulation example on calibration testing. Observed frequencies are computed by binning, with right-closed bins delimited by quantiles of the predicted probabilities (red, dashes: 10 bins; blue, dot-dashes: 50 bins; green, long dashes: 100 bins), and by isotonic regression (brown, solid line). The shaded blue region is a simultaneous 90% confidence band for the true conditional event probabilities, and the fine black line is the bisection line for perfect calibration. The inset plot in panel (a) shows a histogram of the distribution of the predicted probabilities. Panel (b) enlarges the region for probabilities up to 0.15 from panel (a).

well calibrated predictions and provide no practically useful information. Instead of rejecting calibration, one would often be more interested in showing that $|p(x) - x|$ is small for all values of x. This goal can be achieved with a simultaneous confidence band for the function p. A confidence band consists of data-dependent functions $U^\alpha(x) = U^\alpha(x;\, (x_i, Y_i), i = 1, \ldots, n)$, $L^\alpha(x) = L^\alpha(x;\, (x_i, Y_i), i = 1, \ldots, n)$ such that

$$\mathbb{P}(p(x) \in [L^\alpha(x), U^\alpha(x)] \text{ for all } x \in [0,1]) \;\geq\; 1 - \alpha,$$

for any small $\alpha \in (0,1)$. If the band $[L^\alpha(x), U^\alpha(x)]$, $x \in [0,1]$, contains the identity function, then the null hypothesis of calibration cannot be rejected. More generally the band allows to show that x is calibrated up to an error less than ε if $\max(U^\alpha(x) - x, x - L^\alpha(x)) \leq \varepsilon$ for all $x \in [0,1]$ simultaneously. In Section 4.4, a confidence band is developed solely under the assumption that $Y_1, \ldots, Y_n$ are independent and that the function p is increasing. The method requires large sample sizes to achieve sufficiently narrow band, but large sample sizes are exactly the situation where only rejecting the hypothesis of calibration becomes uninformative.

Figure 3 illustrates the different methods for assessing binary outcome predictions with a simulation example by Kramer and Zimmerman (2007). The simulation starts by defining the logit of binary event probabilities as a linear combination of 20 binary and 3 numerical covariates with certain coefficients. To introduce model misspecification, the true event probabilities are a slight modification of these base probabilities, with a relative distortion of at most 0.6%. The model coefficients are estimated with logistic regression on a training data set of size $n = 10'000$, and the predictions are evaluated on an independent validation data set of the same size.[3] The model is misspecified due to the distortion of the original probabilities, but the misspecification could be regarded as negligible in practice. Figure 3 assesses the calibration of the model predictions on the validation data for one simulation. Conditional event probabilities are computed by grouping the observations into bins delimited by 10, 50, and 100 quantiles of the predicted probabilities with equally spaced levels, so that each bin contains 1000, 200, or 100 observations, respectively. For only 10 bins the estimated conditional event probabilities are close to the bin midpoint, and the corresponding Hosmer-Lemeshow test has a p-value of 0.54. With 50 or 100 bins the estimates move erratically around the bisection line, and the p-value of the Hosmer-Lemeshow test drops to 0.02 or 0.05, respectively. The isotonic regression estimate of the conditional event probabilities avoids these instabilities. The eHL test, performed with $B = 50'000$ random splits of the validation data set into $n/2 = 5000$ observations for estimating the conditional event probabilities and computing the e-values, yields an e-value of 0.2 and hence no evidence against miscalibration. A simultaneous confidence band is constructed with the raw method from Section 4.4. It contains the diagonal and therefore also does not allow to reject calibration, but it provides much more information than this sole test. For example, if a threshold of 0.05 for the conditional event probability is used for decision making, say, deciding whether a patient requires a certain treatment, then the confidence band suggests that patients with predicted probability above 0.114 (where the lower bound crosses 0.05) require treatment, while patients with predicted probability below 0.017 (where the upper bound crosses 0.05) do not. For the remaining patients, the predictions alone do not give a sufficient basis for decision making.

Structure of the thesis

The remainder of this thesis consists of three chapters, as introduced before: Isotonic distributional regression (Chapter 2), Distributional index models (Chapter 3), and New methods for forecast evaluation (Chapter 4). The sections of each chapter contain published research papers or arXiv preprints (available on `https://arxiv.org`) in their original format, with the exact reference given at the beginning of each section.

[3] In the original setting by Kramer and Zimmerman (2007) the predictions are validated in-sample. An out-of-sample validation is performed here because none of the methods applied has guaranteed validity when the coefficients of the logistic regression model are estimated in-sample.

Bibliography for the introduction

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, London-New York-Sydney.

Bauer, P., Thorpe, A. and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** 47–55.

Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A. and Apolone, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics* **5** 251–253.

Brunk, H. D., Franck, W. E., Hanson, D. L. and Hogg, R. V. (1966). Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *Journal of the American Statistical Association* **61** 1067–1080.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)* **147** 278–290.

de Leeuw, J., Hornik, K. and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software* **32** 1–24.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13** 253–263.

Dimitriadis, T., Gneiting, T. and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences* **118** e2016191118.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 505–562.

El Barmi, H. and Marchev, D. (2009). New and improved estimators of distribution functions under second-order stochastic dominance. *Journal of Nonparametric Statistics* **21** 143–153.

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint. *Journal of the American Statistical Association* **100** 252–261.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1** 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29** 411–422.

Grünwald, P., de Heide, R. and Koolen, W. (2019). Safe Testing. *arXiv e-prints* arXiv:1906.07801.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science* **33** 568–594.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *The Annals of Statistics* **21** 157–178.

Henzi., A., Ziegel, J. F. and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 963–993.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods* **9** 1043–1069.

Jordan, A. I., Mühlemann, A. and Ziegel, J. F. (2021+). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Annals of the Institute of Statistical Mathematics* to appear.

Kramer, A. A. (2017). Are ICU length of stay predictions worthwhile? *Critical care medicine* **45** 379–380.

Kramer, A. A. and Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited. *Critical care medicine* **35** 2052–2056.

Lehmann, E. L. (1955). Ordered Families of Distributions. *The Annals of Mathematical Statistics* **26** 399–419.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* **22** 1087–1096.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996). The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society* **122** 73–119.

Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **26** 41–47.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics* **14** 24–49.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics* **38** 796–809.

Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv e-prints* arXiv:2009.03167.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order restricted statistical inference.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester.

Rojo, J. and El Barmi, H. (2003). Estimation of distribution functions under second order stochastic dominance. *Statistica Sinica* **13** 903–926.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society* **140** 1086–1096.

Schulz, B. and Lerch, S. (2021+). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review* to appear.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184** 407–431.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders.* Springer Series in Statistics, Springer, New York.

Vannitsem, S., Wilks, D. S. and Messner, J. W. (2018). *Statistical Postprocessing of Ensemble Forecasts.* Elsevier, Amsterdam.

Verburg, I. W., Atashi, A., Eslami, S., Holman, R., Abu-Hanna, A., de Jonge, E., Peek, N. and de Keizer, N. F. (2017). Which models can I use to predict adult ICU length of stay? A systematic review. *Critical care medicine* **45** e222–e231.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics* **49** 1736–1754.

Chapter 2

Isotonic distributional regression

2.1 Isotonic distributional regression

The content of this section is published as

Henzi, A., Ziegel, J. F. and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 963–993.

The version included in this thesis, which combines the paper and its supplementary material, is published as arXiv preprint with identifier *arXiv:1909.03725*. The original article is published under license Creative Commons CC BY-NC-ND 4.0).

Isotonic Distributional Regression

Alexander Henzi and Johanna F. Ziegel

University of Bern, Switzerland

E-mail: alexander.henzi@stat.unibe.ch johanna.ziegel@stat.unibe.ch

Tilmann Gneiting

Heidelberg Institute for Theoretical Studies (HITS) and Karlsruhe Institute of Technology (KIT), Germany

E-mail: tilmann.gneiting@h-its.org

Summary. Isotonic distributional regression (IDR) is a powerful nonparametric technique for the estimation of conditional distributions under order restrictions. In a nutshell, IDR learns conditional distributions that are calibrated, and simultaneously optimal relative to comprehensive classes of relevant loss functions, subject to isotonicity constraints in terms of a partial order on the covariate space. Nonparametric isotonic quantile regression and nonparametric isotonic binary regression emerge as special cases. For prediction, we propose an interpolation method that generalizes extant specifications under the pool adjacent violators algorithm. We recommend the use of IDR as a generic benchmark technique in probabilistic forecast problems, as it does not involve any parameter tuning nor implementation choices, except for the selection of a partial order on the covariate space. The method can be combined with subsample aggregation, with the benefits of smoother regression functions and gains in computational efficiency. In a simulation study, we compare methods for distributional regression in terms of the continuous ranked probability score (CRPS) and L_2 estimation error, which are closely linked. In a case study on raw and postprocessed quantitative precipitation forecasts from a leading numerical weather prediction system, IDR is competitive with state of the art techniques.

Keywords: conditional distribution estimation; monotonicity; probabilistic forecast; proper scoring rule; stochastic order; subagging; weather prediction

1. Introduction

There is an emerging consensus in the transdisciplinary literature that regression analysis should be distributional, with Hothorn et al. (2014) arguing forcefully that

> [t]he ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables.

This article is published as: Henzi, A., Ziegel, J.F. and Gneiting, T. (2021), Isotonic distributional regression. *Journal of the Royal Statistical Society Series B*, https://doi.org/10.1111/rssb.12450.

Distributional regression marks a clear break from the classical view of regression, which has focused on estimating the conditional mean of the response variable in terms of one or more explanatory variable(s) or covariate(s). Later extensions have considered other functionals of the conditional distributions, such as quantiles or expectiles (Koenker, 2005; Newey and Powell, 1987; Schulze Waltrup et al., 2015). However, the reduction of a conditional distribution to a single-valued functional results in tremendous loss of information. Therefore, from the perspectives of both estimation and prediction, regression analysis ought to be distributional.

In the extant literature, both parametric and nonparametric approaches to distributional regression are available. Parametric approaches assume that the conditional distribution of the response is of a specific type (e.g., Gaussian) with an analytic relationship between the covariates and the distributional parameters. Key examples include statistically postprocessed meteorological and hydrologic forecasts, as exemplified by Gneiting et al. (2005), Schefzik et al. (2013) and Vannitsem et al. (2018). In powerful semi-parametric variants, the conditional distributions remain parametric, but the influence of the covariates on the parameter values is modeled nonparametrically, e.g., by using generalized additive models (Rigby and Stasinopoulos, 2005; Klein et al., 2015; Umlauf and Kneib, 2018) or modern neural networks (Rasp and Lerch, 2018; Gasthaus et al., 2019). In related developments, semiparametric versions of quantile regression (Koenker, 2005) and transformation methods (Hothorn et al., 2014) can be leveraged for distributional regression.

Nonparametric approaches to distributional regression include kernel or nearest neighbor methods that depend on a suitable notion of distance on the covariate space. Then, the empirical distribution of the response for neighboring covariates in the training set is used for distributional regression, with possible weighting in dependence on the distance to the covariate value of interest. Kernel smoothing methods and mixture approaches allow for absolutely continuous conditional distributions (Hall et al., 1999; Dunson et al., 2007; Li and Racine, 2008). Classification and regression trees partition the covariate space into leaves, and assign constant regression functions on each leaf (Breiman et al., 1984). Linear aggregation via bootstrap aggregation (bagging) or subsample aggregation (subagging) yields random forests (Breiman, 2001), which are increasingly being used to generate conditional predictive distributions, as proposed by Hothorn et al. (2004) and Meinshausen (2006).

Isotonicity is a natural constraint in estimation and prediction problems. Consider, e.g., postprocessing techniques in weather forecasting, where the covariates stem from the output of numerical weather prediction (NWP) models, and the response variable is the respective future weather quantity. Intuitively, if the NWP model output indicates a larger precipitation accumulation, the associated regression functions ought to be larger as well. Isotonic relationships of this type hold in a plethora of applied settings. In fact, standard linear regression analysis rests on the assumption of isotonicity, in the form of monotonicity in the values of the covariate(s), save for changes in sign.

Concerning nonparametric regression for a conditional functional, such as the mean or a quantile, there is a sizable literature on estimation under the constraint of isotonicity. The classical work of Brunk (1955), Ayer et al. (1955), van Eeden (1958), Bartholomew (1959a,b) and Miles (1959) is summarized in Barlow et al. (1972), Robertson et al. (1988) and de Leeuw et al. (2009). Subsequent approaches include Bayesian and non-Bayesian smoothing techniques (e.g., Mammen, 1991; Neelon and Dunson, 2004; Dette et al., 2006; Shively et al., 2009), and reviews are available in Groeneboom and Jongbloed (2014) and Guntuboyina and Sen (2018).

In distributional regression it may not be immediately clear what is meant by isotonicity, and the literature typically considers one ordinal covariate only (e.g., Hogg, 1965; Rojo and El Barmi, 2003; El Barmi and Mukerjee, 2005; Davidov and Iliopoulos, 2012), with a notable exception being the work of Mösching and Dümbgen (2020b), whose considerations allow for a real-valued covariate. In the general case of a partially ordered covariate space, which we consider here, it is unclear whether semi- or nonparametric techniques might be capable of handling monotonicity contraints, and suitable notions of isotonicity remain to be developed.

To this end, we assume that the response Y is real-valued, and equip the covariate space $\mathcal{X}$ with a partial order $\preceq$. Our aim is to estimate the conditional distribution of Y given the covariate X, for short $\mathcal{L}(Y|X)$, on training data, in a way that respects the partial order, and we desire to use this estimate for prediction. Formally, a distributional regression technique generates a mapping from $x \in \mathcal{X}$ to a probability measure F_x, which serves to model the conditional distribution $\mathcal{L}(Y|X = x)$. This mapping is isotonic if $x \preceq x'$ implies $F_x \leq_{\text{st}} F_{x'}$, where $\leq_{\text{st}}$ denotes the usual stochastic order, i.e., $G \leq_{\text{st}} H$ if $G(y) \geq H(y)$ for $y \in \mathbb{R}$, where we use the same symbols for the probability measures G, H and their associated conditional cumulative distribution functions (CDFs). Equivalently, $G \leq_{\text{st}} H$ holds if $G^{-1}(\alpha) \leq H^{-1}(\alpha)$ for $\alpha \in (0, 1)$, where $G^{-1}(\alpha) = \inf\{y \in \mathbb{R} : G(y) \geq \alpha\}$ is the standard quantile function (Shaked and Shanthikumar, 2007).

Useful comparisons of predictive distributions are in terms of proper scoring rules, of which the most prominent and most relevant instance is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976; Gneiting and Raftery, 2007). We show that there is a unique isotonic distributional regression that is optimal with respect to the CPRS (Theorem 2.1), and refer to it as the *isotonic distributional regression* (IDR). As it turns out, IDR is a universal solution, in that the estimate is optimal with respect to a broad class of proper scoring rules (Theorem 2.2). Classical special cases such as nonparametric isotonic quantile regression and probabilistic classifiers for threshold-defined binary events are nested by IDR. Simultaneously, IDR avoids pitfalls commonly associated with nonparametric distributional regression, such as suboptimal partitions of the covariate space and level crossing (Athey et al., 2019, p. 1167).

For illustration, consider the joint distribution of (X, Y), where X is uniform on $(0, 10)$ and

$$Y \mid X \sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{scale} = \min\{\max\{X, 1\}, 6\}), \tag{1}$$

so that $\mathcal{L}(Y|X = x) \leq_{\text{st}} \mathcal{L}(Y|X = x')$ if $x \leq x'$. Figure 1 shows IDR conditional CDFs and quantiles as estimated on a training set of size $n = 600$. IDR is capable of estimating both the strongly right-skewed conditional distributions for lower values

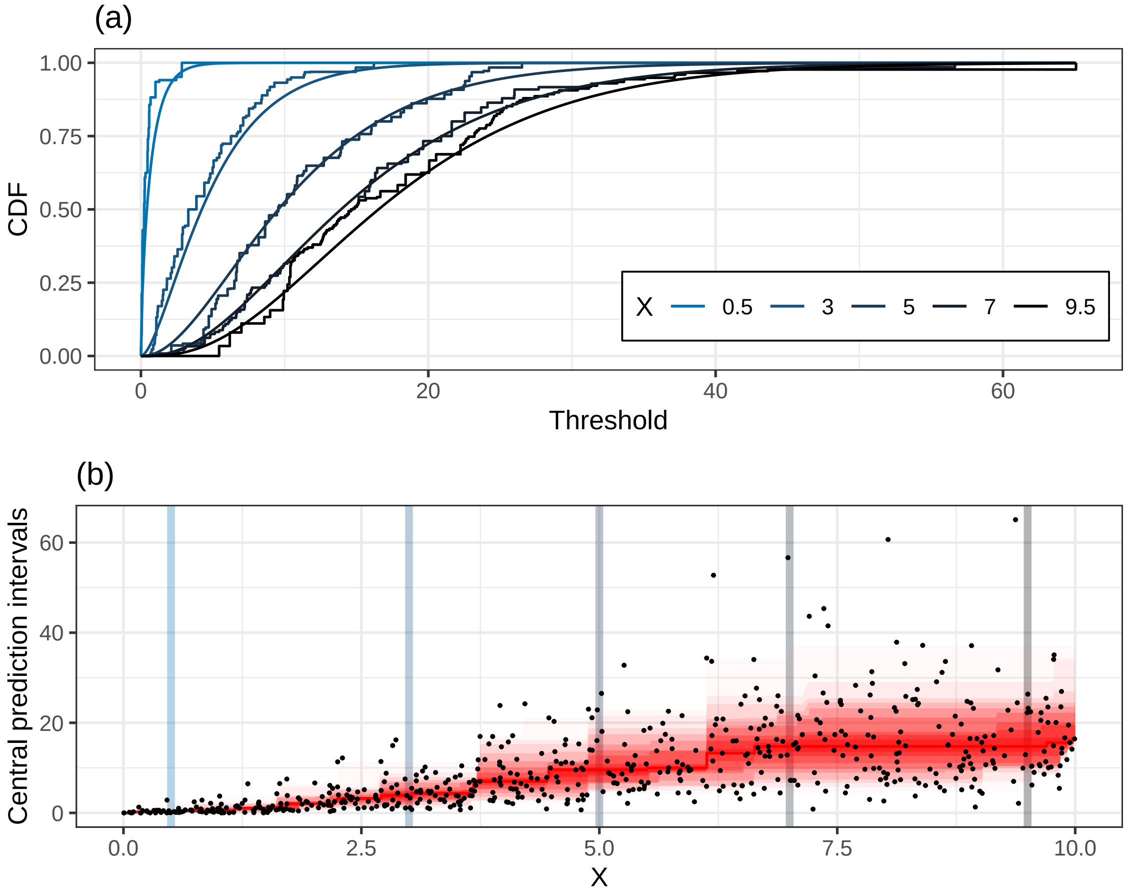


Fig. 1. Simulation example for a sample of size $n = 600$ from the distribution in (1): (a) True conditional CDFs (smooth) and IDR estimates (step functions) for selected values of the covariate. (b) IDR estimated conditional distributions. The shaded bands correspond to probability mass 0.10 each, with the darkest shade marking the central interval. Vertical strips indicate the cross-sections corresponding to the values of the covariate in panel (a).

of X and the more symmetric distributions as X increases. The CDFs are piecewise constant, and they never cross each other. The computational cost of IDR is of order at least $\mathcal{O}(n \log n)$ and may become prohibitive as n grows. However, IDR can usefully be combined with subsample aggregation (subagging), much in the spirit of random forests (Breiman, 2001), with the benefits of reduced computational cost under large training samples, smoother regression functions, and (frequently) improved predictive performance.

The remainder of the paper is organized as follows. The methodological core of the paper is in Section 2, where we prove existence, uniqueness and universality of the IDR solution, discuss computational issues and asymptotic consistency, and propose strategies for prediction. In Section 3 we turn to the critical issue of the choice of a partial order on the covariate space. Section 4 reports on a comparative simulation study that addresses both prediction and estimation, and Section 5 is devoted to a case study on probabilistic quantitative precipitation forecasts, with covariates provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble system. Precipitation accumulations feature unfavorable properties that challenge parametric approaches to distributional regression: The conditional distributions have a point mass at zero, and they are continuous and right skewed on the positive half-axis. In a comparison to state-of-the-art methods that have been developed specifically for the purpose, namely Bayesian Model Averaging (BMA; Sloughter et al., 2007), Ensemble Model Output Statistics (EMOS; Scheuerer, 2014), and Heteroscedastic Censored Logistic Regression (HCLR; Messner et al., 2014), the (out-of-sample) predictive performance of IDR is competitive, despite the method being generic, and being fully automatic once a partial order on the covariate space has been chosen.

We close the paper with a discussion in Section 6, where we argue that IDR provides a very widely applicable, competitive benchmark in probabilistic forecasting problems. The use of benchmark techniques has been called for across application domains (e.g., Rossi, 2013; Pappenberger et al., 2015; Basel Committee on Banking Supervision, 2016; Vogel et al., 2018), and suitable methods should be competitive in terms of predictive performance, while avoiding implementation decisions that may vary from user to user. IDR is well suited to this purpose, as it is entirely generic, does not involve any implementation decisions, other than the choice of the partial order, applies to all types of real-valued outcomes with discrete, continuous or mixed discrete-continuous distributions, and accommodates general types of covariate spaces.

2. Isotonic distributional regression

We proceed to introduce the isotonic distributional regression (IDR) technique. To this end, we first review basic facts on proper scoring rules and notions of calibration. Then we define the IDR solution, prove existence, uniqueness and universality, and discuss its computation and asymptotic consistency. Thereafter, we turn from estimation to prediction and describe how IDR can be used in out-of-sample forecasting. Throughout, we identify a Borel probability measure on the real line $\mathbb{R}$ with its cumulative distribution function (CDF), and we denote the extended real line by $\bar{\mathbb{R}} = [-\infty, \infty]$.

2.1. Preliminaries

Following Gneiting and Raftery (2007), we argue that distributional regression techniques should be compared and evaluated using proper scoring rules. A *proper scoring rule* is a function $\mathrm{S} : \mathcal{P} \times \mathbb{R} \to \bar{\mathbb{R}}$, where $\mathcal{P}$ is a suitable class of probability measures on $\mathbb{R}$, such that $\mathrm{S}(F, \cdot)$ is measurable for any $F \in \mathcal{P}$, the integral $\int \mathrm{S}(G, y) \,\mathrm{d}F(y)$ exists, and

$$\int \mathrm{S}(F, y) \,\mathrm{d}F(y) \leq \int \mathrm{S}(G, y) \,\mathrm{d}F(y)$$

for all $F, G \in \mathcal{P}$. A key example is the *continuous ranked probability score* (CRPS), which is defined for all Borel probability measures, and given as

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \le z\})^2 \, \mathrm{d}z.$$

Introduced by Matheson and Winkler (1976), the CRPS has become popular across application areas and methodological communities, both for the purposes of evaluating predictive performance and as a loss function in estimation; see, e,g., Hersbach (2000), Gneiting et al. (2005), Hothorn et al. (2014), Pappenberger et al. (2015), Rasp and Lerch (2018) and Gasthaus et al. (2019). The CRPS is reported in the same unit as the response variable, and it reduces to the absolute error, $|x - y|$, if F is the point or Dirac measure in $x \in \mathbb{R}$.

Results in Laio and Tamea (2007), Ehm et al. (2016) and Ben Bouallègue et al. (2018) imply that the CRPS can be represented equivalently as

$$\begin{aligned} \mathrm{CRPS}(F, y) &= 2 \int_{(0,1)} \mathrm{QS}_\alpha(F, y) \, \mathrm{d}\alpha && (2) \\ &= 2 \int_{(0,1)} \int_{\mathbb{R}} \mathrm{S}^Q_{\alpha,\theta}(F, y) \, \mathrm{d}\theta \, \mathrm{d}\alpha && (3) \\ &= \int_{\mathbb{R}} \int_{(0,1)} \mathrm{S}^P_{z,c}(F, y) \, \mathrm{d}c \, \mathrm{d}z, && (4) \end{aligned}$$

where the mixture representation (2) is in terms of the asymmetric piecewise linear or pinball loss,

$$\mathrm{QS}_\alpha(F, y) = \begin{cases} (1-\alpha)\,(F^{-1}(\alpha) - y), & y \le F^{-1}(\alpha), \\ \alpha\,(y - F^{-1}(\alpha)), & y \ge F^{-1}(\alpha), \end{cases} \tag{5}$$

which is customarily thought of as a quantile loss function, but can be identified with a proper scoring rule (Gneiting, 2011, Theorem 3). The representations (3) and (4) express the CRPS in terms of the *elementary* or *extremal scoring functions* for the α-quantile functional, namely,

$$\mathrm{S}^Q_{\alpha,\theta}(F, y) = \begin{cases} 1-\alpha, & y \le \theta < F^{-1}(\alpha), \\ \alpha, & F^{-1}(\alpha) \le \theta < y, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $\theta \in \mathbb{R}$; and for probability assessments of the binary outcome $\mathbb{1}\{y \le z\}$ at the threshold value $z \in \mathbb{R}$, namely

$$\mathrm{S}^P_{z,c}(F, y) = \begin{cases} 1-c, & F(z) < c, \; y \le z, \\ c, & F(z) \ge c, \; y > z, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $c \in (0, 1)$. For background information on elementary or extremal scoring functions and related concepts see Ehm et al. (2016).

Predictive distributions ought to be calibrated (Dawid, 1984; Diebold et al., 1998; Gneiting et al., 2007), in the broad sense that they should be statistically compatible with the responses, and various notions of calibration have been proposed and studied. In the spirit of Gneiting and Ranjan (2013), we consider the joint distribution $\mathbb{P}$ of the response Y and the distributional regression F_X. The most widely used criterion is *probabilistic calibration*, which requires that the *probability integral transform* (PIT), namely, the random variable

$$Z = F_X(Y-) + V\,(F_X(Y) - F_X(Y-)), \tag{8}$$

be standard uniform, where $F_X(Y-) = \lim_{y \uparrow Y} F_X(y)$ and V is a standard uniform variable that is independent of F_X and Y. If F_X is continuous the PIT is simply $Z = F_X(Y)$. Here we introduce the novel notion of *threshold calibration*, requiring that

$$\mathbb{P}(Y \le y \,|\, F_X(y)) = F_X(y) \tag{9}$$

almost surely for $y \in \mathbb{R}$, which implies *marginal calibration*, defined as $\mathbb{P}(Y \le y) = \mathbb{E}(F_X(y))$ for $y \in \mathbb{R}$. If $F_X = \mathcal{L}(Y|X)$ then it is calibrated in any of the above senses (Gneiting and Ranjan, 2013, Theorem 2.8).

2.2. Existence, uniqueness and universality

A partial order relation $\preceq$ on a set $\mathcal{X}$ has the same properties as a total order, namely reflexivity, antisymmetry and transitivity, except that the elements need not be comparable, i.e., there might be elements $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ such that neither $x \preceq x'$ nor $x' \preceq x$ holds. A key example is the componentwise order on $\mathbb{R}^n$.

For a positive integer n and a partially ordered set $\mathcal{X}$, we define the classes

$$\mathcal{X}^n_{\uparrow} = \{\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n : x_1 \preceq \cdots \preceq x_n\},$$
$$\mathcal{X}^n_{\downarrow} = \{\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n : x_1 \succeq \cdots \succeq x_n\}$$

of the increasingly and decreasingly (totally) ordered tuples in $\mathcal{X}$, respectively. Similarly, given a further partially ordered set $\mathcal{Q}$ and a vector $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$, the classes

$$\mathcal{Q}^n_{\uparrow,\boldsymbol{x}} = \{\boldsymbol{q} = (q_1, \dots, q_n) \in \mathcal{Q}^n : q_i \preceq q_j \text{ if } x_i \preceq x_j\},$$
$$\mathcal{Q}^n_{\downarrow,\boldsymbol{x}} = \{\boldsymbol{q} = (q_1, \dots, q_n) \in \mathcal{Q}^n : q_i \succeq q_j \text{ if } x_i \preceq x_j\}$$

comprise the increasingly and decreasingly (partially) ordered tuples in $\mathcal{Q}$, with the order induced by the tuple $\boldsymbol{x}$ and the partial order $\preceq$ on $\mathcal{X}$.

Let $I \subseteq \mathbb{R}$ be an interval, and let S be a proper scoring rule with respect to a class $\mathcal{P}$ of probability distributions on I that contains all distributions with finite support. Given training data in the form of a covariate vector $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ and response vector $\boldsymbol{y} = (y_1, \dots, y_n) \in I^n$, we may interpret any mapping from $\boldsymbol{x} \in \mathcal{X}^n$ to $\mathcal{P}^n$ as a distributional regression function. Throughout, we equip $\mathcal{P}$ with the usual stochastic order.

Definition 2.1 (S-based regression)**.** An element $\hat{\boldsymbol{F}} = (\hat{F}_1, \dots, \hat{F}_n) \in \mathcal{P}^n$ is an *S-based isotonic regression* of $\boldsymbol{y} \in I^n$ on $\boldsymbol{x} \in \mathcal{X}^n$, if it is a minimizer of the empirical loss

$$\ell_{\mathrm{S}}(\boldsymbol{F}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(F_i, y_i)$$

over all $\boldsymbol{F} = (F_1, \dots, F_n)$ in $\mathcal{P}^n_{\uparrow, \boldsymbol{x}}$.

In plain words, an S-based isotonic regression achieves the best fit in terms of the scoring rule S, subject to the conditional CDFs $\hat{F}_1, \dots, \hat{F}_n$ satisfying partial order constraints induced by the covariate values $x_1, \dots, x_n$. The definition and the subsequent results can be extended to losses of the form $\ell_{\mathrm{S}}(\boldsymbol{F}) = \sum_{i=1}^{n} w_i \mathrm{S}(F_i, y_i)$ with rational, strictly positive weights $w_1, \dots, w_n$. The adaptations are straightforward and left to the reader.

Furthermore, the definition of an S-based isotonic regression as a minimizer of ℓ_{S} continues to apply when $\mathcal{X}$ is equipped with a pre- or quasiorder $\preceq$ instead of a partial order. Preorders are not necessarily antisymmetric, and so there might be elements x, x' such that $x \preceq x'$ and $x' \preceq x$ but $x' \neq x$. In this setting, we define x and x' to be equivalent if $x \preceq x'$ and $x' \preceq x$, and set $[x] \preceq_p [x']$ if representatives u, u' of the equivalence classes $[x], [x']$ satisfy $u \preceq u'$. Then the binary relation $\preceq_p$ defines a partial order on the set of equivalence classes, and the S-based isotonic regression with the new covariates and the partial order $\preceq_p$ coincides with the original S-based isotonic regression.

In Appendix A we prove the following result.

Theorem 2.1 (existence and uniqueness)**.** *There exists a unique* CRPS*-based isotonic regression* $\hat{\boldsymbol{F}} \in \mathcal{P}^n$ *of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$*.*

We refer to this unique $\hat{\boldsymbol{F}}$ as the *isotonic distributional regression* (IDR) of $\boldsymbol{y}$ on $\boldsymbol{x}$. In the particular case of a total order on the covariate space, and assuming that $x_1 < \dots < x_n$, for each $z \in I$ the solution $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z), \dots, \hat{F}_n(z))$ is given by

$$\hat{F}_i(z) = \min_{k=1,\dots,i} \max_{j=k,\dots,n} \frac{1}{j-k+1} \sum_{l=k}^{j} \mathbb{1}\{y_l \leq z\} \tag{10}$$

for $i = 1, \dots, n$; see eqs. (1.9)–(1.13) of Barlow et al. (1972). A similar max–min formula applies under partial orders (Robertson and Wright, 1980; Jordan et al., 2021), and it follows that $\hat{F}_i$ is piecewise constant with any points of discontinuity at $y_1, \dots, y_n$.

At first sight, the specific choice of the CRPS as a loss function may seem arbitrary. However, the subsequent result, which we prove in Appendix A, reveals that IDR is simultaneously optimal with respect to broad classes of proper scoring rules that include all relevant choices in the extant literature. The popular logarithmic score allows for the comparison of absolutely continuous distributions with respect to a fixed dominating measure only and thus is not applicable here. Statements concerning calibration are with respect to the empirical distribution of the training data $(x_1, y_1), \dots, (x_n, y_n)$.

Theorem 2.2 (universality)**.** *The* IDR *solution* $\hat{\boldsymbol{F}}$ *of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$ *is threshold calibrated and has the following properties.*

i) The IDR *solution* $\hat{\boldsymbol{F}}$ *is an* S*-based isotonic regression of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$ *under any scoring rule of the form*

$$\mathrm{S}(F,y) = \int_{(0,1)\times\mathbb{R}} \mathrm{S}^{Q}_{\alpha,\theta}(F,y)\,\mathrm{d}H(\alpha,\theta) \tag{11}$$

or

$$\mathrm{S}(F,y) = \int_{\mathbb{R}\times(0,1)} \mathrm{S}^{P}_{z,c}(F,y)\,\mathrm{d}M(z,c), \tag{12}$$

where $\mathrm{S}^{Q}_{\alpha,\theta}$ *is the elementary quantile scoring function* (6)*,* $\mathrm{S}^{P}_{z,c}$ *is the elementary probability scoring rule* (7)*, and* H *and* M *are locally finite Borel measures on* $(0,1)\times\mathbb{R}$ *and* $\mathbb{R}\times(0,1)$*, respectively.*

ii) For every $\alpha \in (0,1)$ *it holds that* $\hat{\boldsymbol{F}}^{-1}(\alpha) = (\hat{F}_1^{-1}(\alpha),\dots,\hat{F}_n^{-1}(\alpha))$ *is a minimizer of*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{s}_\alpha(\theta_i, y_i) \tag{13}$$

over all $\boldsymbol{\theta} = (\theta_1,\dots,\theta_n) \in I^n_{\uparrow,\boldsymbol{x}}$*, under any function* $\mathrm{s}_\alpha : I\times I \to \bar{\mathbb{R}}$ *which is left-continuous in both arguments and such that* $\mathrm{S}(F,y) = \mathrm{s}_\alpha(F^{-1}(\alpha),y)$ *is a proper scoring rule on* $\mathcal{P}$*.*

iii) For every threshold value $z \in I$*, it is true that* $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z),\dots,\hat{F}_n(z))$ *is a minimizer of*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{s}(\eta_i, \mathbb{1}\{y_i \le z\}) \tag{14}$$

over all ordered tuples $\boldsymbol{\eta} = (\eta_1,\dots,\eta_n) \in [0,1]^n_{\downarrow,\boldsymbol{x}}$*, under any function* $\mathrm{s} : [0,1]\times\{0,1\} \to \bar{\mathbb{R}}$ *that is a proper scoring rule for binary events, which is left-continuous in its first argument, satisfies* $\mathrm{s}(0,y) = \lim_{p\to 0}\mathrm{s}(p,y)$*, and is real-valued, except possibly* $\mathrm{s}(0,1) = -\infty$ *or* $\mathrm{s}(1,0) = -\infty$*.*

The quantile weighted and threshold weighted versions of the CRPS studied by Gneiting and Ranjan (2011) arise from (11) and (12) with $H = G_0 \otimes \lambda$ and $M = \lambda \otimes G_1$, where λ denotes the Lebesgue measure, and G_0 and G_1 are σ-finite Borel measures on $(0,1)$ and $\mathbb{R}$, respectively. If G_0 and G_1 are Lebesgue measures, we recover the mixture representations (3) and (4) of the CRPS. By results of Ehm et al. (2016), if H is concentrated on $\{\alpha\}\times\mathbb{R}$ and M is concentrated on $\{z\}\times(0,1)$, these representations cover essentially all proper scoring rules that depend on the predictive distribution F via $F^{-1}(\alpha)$ or $F(z)$ only, yielding universal optimality in statements in parts ii) and iii) of Theorem 2.2.

In particular, as a special case of (13), the IDR solution is a minimizer of the quantile loss under the asymmetric piecewise linear or pinball function (5) that lies at the heart of quantile regression (Koenker, 2005). Consequently, as the mixture representation (2) of the CRPS may suggest, IDR nests classical nonparametric isotonic quantile regression as introduced and studied by Robertson and Wright (1975) and Casady and Cryer (1976). In other words, part ii) of Theorem 2.2 demonstrates that, if we (hypothetically) perform nonparametric isotonic quantile regression at every level $\alpha \in (0, 1)$ and piece the conditional quantile functions together, we recover the IDR solution. However, the IDR solution is readily computable (Section 2.3), without invoking approximations or truncations, unlike brute force approaches to simultaneous quantile regressions. Loss functions of the form (13) also include the interval score (Winkler, 1972; Gneiting and Raftery, 2007, eq. (43)), which constitutes the most used proper performance measure for interval forecasts.

In the special case of a binary response variable, we see from iii) and (14) that the IDR solution is an S-based isotonic regression under just any applicable proper scoring rule S. Furthermore, threshold calibration is the strongest possible notion of calibration in this setting (Gneiting and Ranjan, 2013, Theorem 2.11), so the IDR solution is universal in every regard. In the further special case of a total order on the covariate space, the IDR and pool adjacent violators (PAV) algorithm solutions coincide, and the statement in iii) is essentially equivalent to Theorem 1.12 of Barlow et al. (1972). In particular, the IDR or PAV solution yields both the nonparametric maximum likelihood estimate and the nonparametric least squares estimate under the constraint of isotonicity. The latter suggests a computational implementation via quadratic programming, to which we tend now.

2.3. Computational aspects

The key observation towards a computational implementation is the aforementioned special case of (14), according to which the IDR solution $\hat{\boldsymbol{F}} \in \mathcal{P}^n$ of $\boldsymbol{y} \in \mathbb{R}^n$ on $\boldsymbol{x} \in \mathcal{X}^n$ satisfies

$$\hat{\boldsymbol{F}}(z) = \arg \min_{\eta \in [0,1]^n_{\downarrow, \boldsymbol{x}}} \sum_{i=1}^{n} (\eta_i - \mathbb{1}\{y_i \le z\})^2 \tag{15}$$

at every threshold value $z \in \mathbb{R}$. In this light, the computation of the IDR CDF at any fixed threshold reduces to a quadratic programming problem. The above target function is constant in between the unique values of $y_1, \ldots, y_n$, say $\tilde{y}_1 < \cdots < \tilde{y}_m$, and so it suffices to estimate the CDFs at these points only. In contrast, exact implementations based on quantiles would need to consider all levels of the form i/j with integers $1 \le i < j \le n$, which is computationally prohibitive. In the threshold-based approach, the overall cost depends on the quadratic programming solver applied, and the computation becomes much faster if recursive relations between consecutive conditional CDFs $\hat{\boldsymbol{F}}(\tilde{y}_k)$ and $\hat{\boldsymbol{F}}(\tilde{y}_{k-1})$ are taken advantage of. In the case of a total order, Henzi et al. (2020) describe a recursive adaptation of the PAV algorithm to IDR that considerably reduces the computation time as compared to a naive implementation which does not take into account recursive relations. Under general partial orders, active set methods for solutions to the quadratic programming problem (15) have been discussed by de Leeuw et al. (2009). In our implementation, we use the powerful quadratic programming solver OSQP (Stellato et al., 2020) as supplied by the package `osqp` in the statistical programming environment R (Stellato

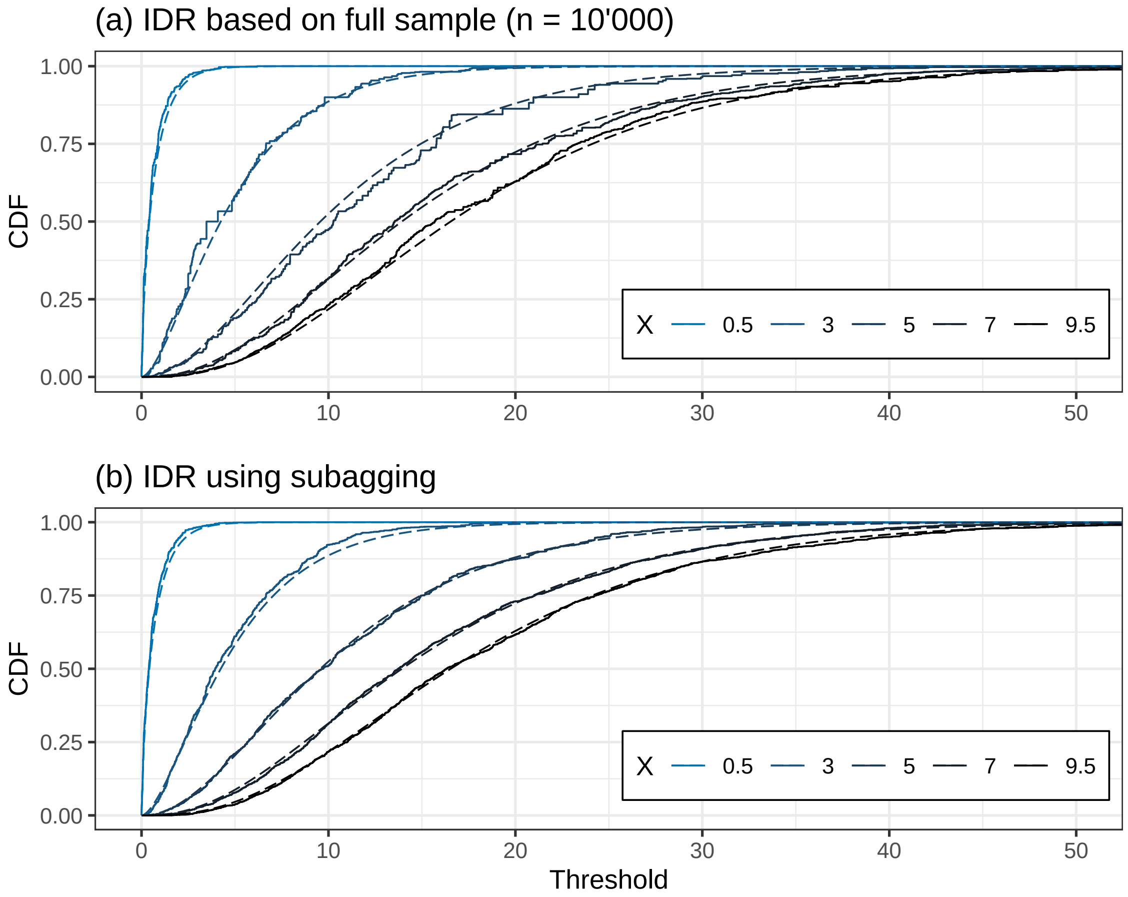


Fig. 2. Simulation example for a sample of size $n = 10\,000$ from the distribution in (1). The true conditional CDFs (smooth dashed graphs) are compared to IDR estimates (step functions) based on (a) the full training sample of size $n = 10\,000$ and (b) linear aggregation of IDR estimates on 100 subsamples of size $1\,000$ each.

et al., 2019; R Core Team, 2020), which can be warm-started efficiently by taking $\hat{\boldsymbol{F}}(\tilde{y}_{k-1})$ as a starting point for the computation of $\hat{\boldsymbol{F}}(\tilde{y}_k)$.

Clearly, a challenge in the computational implementation of IDR with general partial orders is that the number of variables in the quadratic programming problem (15) grows at a rate of $\mathcal{O}(n)$. As a remedy, we propose subsample aggregation, much in the spirit of random forests that rely on bootstrap aggregated (bagged) classification and regression trees (Breiman, 1996, 2001). It was observed early on that random forests generate conditional predictive distributions (Hothorn et al., 2004; Meinshausen, 2006), and recent applications include the statistical postprocessing of ensemble weather forecasts (Taillardat et al., 2016; Schlosser et al., 2019; Taillardat et al., 2019). Bühlmann and Yu (2002) and Buja and Stützle (2006) argue forcefully that subsample aggregation (subagging) tends to be equally effective as bagging, but at considerably lower computational cost.

In view of the superlinear computational costs of IDR, smart uses of subsample aggregation yield major efficiency gains, taking into account that the estimation on different subsamples can be performed in parallel. Isotonicity is preserved under linear aggregation, and the aggregated conditional CDFs can be inverted to generate isotonic conditional quantile functions, with the further benefit of smoother estimates in continuous settings. A detailed investigation of optimal subsample aggregation for IDR is a topic for future research. For illustration, Figure 2 returns to the simulation example in Figure 1, but now with a much larger training sample of size $n = 10\,000$ from the distribution in (1). Linear aggregation based on 100 subsamples (drawn without replacement) of size $n = 1\,000$ each is superior to the brute force approach on the full training set in terms of estimation accuracy. The computation on the full dataset for this simulation example takes 11.7 seconds for the naive implementation, but only 1.1 seconds for the sequential algorithm of Henzi et al. (2020). Subagging gives computation times of 9.9 and 2.5 seconds, respectively, or 1.8 and 0.5 seconds when parallelized over eight cores.†

2.4. Consistency

We proceed to prove uniform consistency of the IDR estimator. While strong consistency of nonparametric isotonic quantile regression for single quantiles was proved decades ago (Robertson and Wright, 1975; Casady and Cryer, 1976), uniform consistency and rates of convergence for the IDR estimator have been established only recently, and exclusively in the case of a total order, see El Barmi and Mukerjee (2005, Theorem 1) and Mösching and Dümbgen (2020b, Theorem 3.3).

For $x \in \mathcal{X}$ and $y \in \mathbb{R}$, let $\hat{F}_x(y)$ denote the IDR estimate based on fixed or random pairs $(X_1, Y_1), \dots, (X_n, Y_n)$. As introduced thus far, the IDR solution $\hat{\boldsymbol{F}} = (\hat{F}_1, \dots, \hat{F}_n)$ is defined at the covariate values $X_1, \dots, X_n \in \mathcal{X}$ only. For general $x \in \mathcal{X}$, we merely assume that $\hat{F}_x(y)$ is some value in between the bounds given by

$$\max_{i \in s(x)} \hat{F}_i(y) \le \hat{F}_x(y) \le \min_{i \in p(x)} \hat{F}_i(y). \tag{16}$$

Here, we define the sets of the indices of *direct predecessors* and *direct successors* of $x \in \mathcal{X}$ among the covariate values as

$$p(x) = \{i \in \{1, \dots, n\} : X_i \preceq X_j \preceq x \implies X_j = X_i,\, j = 1, \dots, n\} \tag{17}$$

and

$$s(x) = \{i \in \{1, \dots, n\} : x \preceq X_j \preceq X_i \implies X_j = X_i,\, j = 1, \dots, n\}, \tag{18}$$

respectively.

In Appendix B we establish the following consistency theorem, which covers key examples of partial orders and is based on strictly weaker assumptions than the results of Mösching and Dümbgen (2020b). However, in contrast to their work, we do not provide rates of convergence. The choice $\mathcal{X} = [0, 1]^d$ for the covariate space merely serves to simplify the presentation: As IDR is invariant under

†With Intel(R) Xeon(R) E5-2630 v4 2.20GHz CPUs, in R (R Core Team, 2020), using the `doParallel` package for parallelization. Times reported are averages over 100 replicates.

strictly isotonic transformations, any covariate vector $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ can be transformed to have support in $[0,1]^d$, and the componentwise partial order can be replaced by any weaker preorder. A key assumption uses the concept of an *antichain* in a partially ordered set $(\mathcal{S}, \preceq)$, which is a subset $A \subseteq \mathcal{S}$ that does not admit comparisons, in the sense that $u \preceq v$ for $u, v \in A$ implies $u = v$. As we discuss subsequently, results of Brightwell (1992) imply that the respective distributional condition is mild.

Theorem 2.3 (uniform consistency)**.** *Let $\mathcal{X} = [0,1]^d$ be endowed with the componentwise partial order and the norm $\|u\| = \max_{i=1,\dots,d} |u_i|$. Let further $(X_{ni}, Y_{ni}) \in [0,1]^d \times \mathbb{R}$, $n \in \mathbb{N}$, $i = 1, \dots, n$, be a triangular array such that $(X_{n1}, Y_{n1}), \dots, (X_{nn}, Y_{nn})$ are independent and identically distributed random vectors for each $n \in \mathbb{N}$, and let $S_n = \{X_{n1}, \dots, X_{nn}\}$. Assume that*

(i) for all non-degenerate rectangles $J \subseteq \mathcal{X}$, there exists a constant $c_J > 0$ such that

$$\#(S_n \cap J) \geq n c_J$$

with asymptotic probability one, i.e., if A_n denotes the event that $\#(S_n \cap J) \geq n c_J$, then $\mathbb{P}(A_n) \to 1$ as $n \to \infty$;

(ii) for some $\gamma \in (0,1)$,

$$\max\{\#A : \, A \subset S_n \text{ is antichain}\} \leq n^\gamma$$

with asymptotic probability one.

Assume further that the true conditional CDFs $F_x(y) = \mathbb{P}(Y_{ni} \leq y \mid X_{ni} = x)$ satisfy

(iii) $F_x(y)$ is decreasing in x for all $y \in \mathbb{R}$;

(iv) for every $\eta > 0$, there exists $r > 0$ such that

$$\sup\{|F_x(y) - F_{x'}(y)| : \, x, x' \in [0,1]^d, \|x - x'\| \leq r, \, y \in \mathbb{R}\} < \eta.$$

Then for every $\epsilon > 0$ and $\delta > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{x \in [\delta, 1-\delta]^d, \, y \in \mathbb{R}} |\hat{F}_x(y) - F_x(y)| \geq \epsilon \right) = 0. \tag{19}$$

Assumption (i) requires that the covariates are sufficiently dense in $\mathcal{X}$, as is satisfied under strictly positive Lebesgue densities on $\mathcal{X}$. In order to derive rates of convergence, the size of the rectangles J in (i) would need to decrease with n, as in condition (A.2) of Mösching and Dümbgen (2020b); we leave this type of extension as a direction for future work. Assumption (iii) is the basic model assumption of IDR, while assumption (iv) requires uniform continuity of the conditional distributions, which is weaker than Hölder continuity in condition (A.1) of Mösching and Dümbgen (2020b).

Assumption (ii), which is always satisfied in the case of a total order, calls for a more detailed discussion. In words, the maximal number of mutually incomparable elements needs to grow at a rate slower than n^γ. Evidently, the easier elements can be ordered, the smaller the maximal antichain. Consequently, Theorem 2.3 continues to hold under the empirical stochastic order and the empirical increasing convex order on the covariates introduced in Section 3.3, and indeed under any preorder that is weaker than the componentwise order. The key to understanding the distributional implications of (ii) is Corollary 2 in Brightwell (1992), which states that for a sequence of independent random vectors from a uniform population on $[0,1]^d$ the size of a maximal antichain grows at a rate of $n^{1-1/d}$; see also the remark following the proof of Theorem 2.3 in Appendix B.

As comparability under the componentwise order is preserved under monotonic transformations, *any* covariate vector $X \in \mathbb{R}^d$ that can be obtained as a monotone transformation of a uniform random vector of arbitrary dimension guarantees (ii). This includes, e.g., all Gaussian random vectors with nonnegative correlation coefficients. In this light, assumption (ii) is rather weak, and well in line with the intuition that for multivariate isotonic (distributional) regression to work well, there ought be at least minor positive dependence between the covariates. In the context of our case study in Section 5, high positive correlations between the covariates are the rule, as exemplified by Table 3 in Raftery et al. (2005).

2.5. *Prediction*

As noted, the IDR solution $\hat{\boldsymbol{F}} = (\hat{F}_1, \dots, \hat{F}_n) \in \mathcal{P}^n_{\uparrow,\boldsymbol{x}}$ is defined at the covariate values $x_1, \dots, x_n \in \mathcal{X}$ only. Generally, if a (not necessarily optimal) distributional regression $\boldsymbol{F} = (F_1, \dots, F_n) \in \mathcal{P}^n_{\uparrow,\boldsymbol{x}}$ is available, a key task in practice is to make a prediction at a new covariate value $x \in \mathcal{X}$ where $x \notin \{x_1, \dots, x_n\}$. We denote the respective predictive CDF by F.

In the specific case $\mathcal{X} = \mathbb{R}$ of a single real-valued covariate there is a simple way of doing this, as frequently implemented in concert with the PAV algorithm. For simplicity we suppose that $x_1 < \cdots < x_n$. If $x < x_1$ we may let $F = F_1$; if $x \in (x_i, x_{i+1})$ for some $i \in \{1, \dots, n-1\}$ we may interpolate linearly, so that

$$F(z) = \frac{x - x_i}{x_{i+1} - x_i} F_i(z) + \frac{x_{i+1} - x}{x_{i+1} - x_i} F_{i+1}(z)$$

for $z \in \mathbb{R}$, and if $x > x_n$ we may set $F = F_n$. However, approaches that are based on interpolation do not extend to a generic covariate space, which may or may not be equipped with a metric.

In contrast, the method we describe now, which generalizes a proposal by Wilbur et al. (2005), solely uses information supplied by the partial order $\preceq$ on the covariate space $\mathcal{X}$. For a general covariate value $x \in \mathcal{X}$, the sets of the indices of direct predecessors and direct successors among the covariate values $x_1, \dots, x_n$ in the training data is defined as at (17) and (18), respectively with $X_1, \dots, X_n$ replaced by $x_1, \dots, x_n$. If the covariate space $\mathcal{X}$ is totally ordered, these sets contain at most one element. If the order is partial but not total, $p(x)$ and $s(x)$ may, and frequently do, contain more than one element. Assuming that $p(x)$ and $s(x)$ are non-empty, any predictive CDF F that is consistent with $\boldsymbol{F}$ must satisfy

$$\max_{i \in s(x)} F_i(z) \leq F(z) \leq \min_{i \in p(x)} F_i(z) \tag{20}$$

at all threshold values $z \in \mathbb{R}$. We now let F be the pointwise arithmetic average of these bounds, i.e.,

$$F(z) = \frac{1}{2}\left(\max_{i \in s(x)} F_i(z) + \min_{i \in p(x)} F_i(z)\right) \tag{21}$$

for $z \in \mathbb{R}$. If $s(x)$ is empty while $p(x)$ is non-empty, or vice-versa, a natural choice, which we employ hereinafter, is to let F equal the available bound given by the non-empty set. If x is not comparable to any of $x_1, \ldots, x_n$ the training data lack information about the conditional distribution at x, and a natural approach, which we adopt and implement, is to set F equal to the empirical distribution of the response values $y_1, \ldots, y_n$.

The difference between the bounds (if any) in (20) might be a useful measure of estimation uncertainty and could be explored as a promising avenue towards the quantification of ambiguity and generation of second-order probabilities (Ellsberg, 1961; Seo, 2009). In the context of ensemble weather forecasts, the assessment of ambiguity has been pioneered by Allen and Eckel (2012). Interesting links arise when the envelope in (20) is interpreted in the spirit of randomized predictive systems and conformal estimates as studied by Vovk et al. (2019); compare, e.g., their Figure 5 with our Figure 4b below.

3. Partial orders

The choice of a sufficiently informative partial order on the covariate space is critical to any successful application of IDR. In the extreme case of distinct, totally ordered covariate values $x_1, \ldots, x_n \in \mathcal{X}$ and a perfect monotonic relationship to the response values $y_1, \ldots, y_n$, the IDR distribution associated with x_i is simply the point measure in y_i, for $i = 1, \ldots, n$. The same happens in the other extreme, when there are no order relations at all. Hence, the partial order serves to regularize the IDR solution.

Thus far, we have simply assumed that the covariate space $\mathcal{X}$ is equipped with a partial order $\preceq$, without specifying how the order might be defined. If $\mathcal{X} \subseteq \mathbb{R}^d$, the usual componentwise order will be suitable in many applications, and we investigate it in Section 3.1. For covariates that are ordinal and admit a ranking in terms of importance, a lexicographic order may be suitable.

If groups of covariates are exchangeable, as in our case study on quantitative precipitation forecasts, other types of order relations need to be considered. In Sections 3.2 and 3.3 we study relations that are tailored to this setting, namely, the empirical stochastic order and empirical increasing convex order. Proofs are deferred to Appendix C.

3.1. Componentwise order

Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of the covariate space $\mathbb{R}^d$. The most commonly used partial order in multivariate isotonic regression is the *componentwise order* defined by

$$x \preceq x' \iff x_i \leq x'_i \text{ for } i = 1, \ldots, d.$$

This order becomes weaker as the dimension d of the covariate space increases: If $\tilde{x} = (x_1, \ldots, x_d, x_{d+1})$ and $\tilde{x}' = (x'_1, \ldots, x'_d, x'_{d+1})$ then $x \preceq x'$ is a necessary condition for $\tilde{x} \preceq \tilde{x}'$. The following result is an immediate consequence of this observation and the structure of the optimization problem in Definition 2.1.

Proposition 3.1. *Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)$ have components $x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$ and $x_i^* = (x_{i1}, \ldots, x_{id}, x_{i,d+1}) \in \mathbb{R}^{d+1}$ for $i = 1, \ldots, n$, and let S be a proper scoring rule.*

Then if $\mathbb{R}^d$ and $\mathbb{R}^{d+1}$ are equipped with the componentwise partial order, and $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{F}}^*$ denote S-based isotonic regressions of $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\boldsymbol{x}^*$, respectively, it is true that

$$\ell_{\mathrm{S}}(\hat{\boldsymbol{F}}^*) \leq \ell_{\mathrm{S}}(\hat{\boldsymbol{F}}).$$

In simple words, under the componentwise partial order, the inclusion of further covariates can only improve the in-sample fit. This behaviour resembles linear regression, where the addition of covariates can only improve the (unadjusted) R-square.

3.2. Empirical stochastic order

We now define a relation that is based on stochastic dominance and invariant under permutation.

Definition 3.1. Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$. Then x is smaller than or equal to x' in *empirical stochastic order*, for short $x \preceq_{\mathrm{st}} x'$, if the empirical distribution of $x_1, \ldots, x_d$ is smaller than the empirical distribution of $x'_1, \ldots, x'_d$ in the usual stochastic order.

This relation is tailored to groups of exchangeable, real-valued covariates. The following results summarizes its properties and compares to the componentwise order, which we denote by $\preceq$.

Proposition 3.2. *Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$ with order statistics $x_{(1)} \leq \cdots \leq x_{(d)}$ and $x'_{(1)} \leq \cdots \leq x'_{(d)}$.*

i) The relation $x \preceq_{\mathrm{st}} x'$ is equivalent to $x_{(i)} \leq x'_{(i)}$ for $i = 1, \ldots, d$.

ii) If $x \preceq x'$ then $x \preceq_{\mathrm{st}} x'$.

iii) If $x \preceq_{\mathrm{st}} x'$ and x and x' are comparable in the componentwise partial order, then $x \preceq x'$.

iv) If $x \preceq_{\mathrm{st}} x'$ and $x' \preceq_{\mathrm{st}} x$ then x and x' are permutations of each other. Consequently, the relation $\preceq_{\mathrm{st}}$ defines a partial order on $\mathbb{R}^d_{\uparrow}$.

In a nutshell, the empirical stochastic order is equivalent to the componentwise order on the sorted elements, and this relation is weaker than the componentwise order. However, unlike the componentwise order, the empirical stochastic order does not degenerate as further covariates are added. To the contrary, empirical distributions of larger numbers of exchangeable variables become more informative and more easily comparable.

3.3. Empirical increasing convex order

In applications, the empirical stochastic order might be too strong, in the sense that it does not generate sufficiently informative constraints. In this light, we now define a weaker partial order on $\mathbb{R}^d_{\uparrow}$, which also is based on a partial order for probability measures. Specifically, let X and X' be random variables with CDFs F and F'. Then F is smaller than F' in increasing convex order if $\mathbb{E}(\phi(X)) \leq \mathbb{E}(\phi(X'))$ for all increasing convex functions ϕ such that the expectations exist (Shaked and Shanthikumar, 2007, Section 4.A.1).

Definition 3.2. Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$. Then x is smaller than or equal to x' in *empirical increasing convex order*, for short $x \preceq_{\mathrm{icx}} x'$, if the empirical distribution of $x_1, \ldots, x_d$ is smaller than the empirical distribution of $x'_1, \ldots, x'_d$ in increasing convex order.

This notion provides another meaningful relation for groups of exchangeable covariates. The following result summarizes its properties and relates it to the empirical stochastic order.

Proposition 3.3. *Let $x = (x_1, \ldots, x_d)$ and $x' = (x'_1, \ldots, x'_d)$ denote elements of $\mathbb{R}^d$ with order statistics $x_{(1)} \leq \cdots \leq x_{(d)}$ and $x'_{(1)} \leq \cdots \leq x'_{(d)}$.*

i) The relation $x \preceq_{\mathrm{icx}} x'$ is equivalent to

$$\sum_{i=j}^{d} x_{(i)} \leq \sum_{i=j}^{d} x'_{(i)} \quad \textit{for} \;\; j = 1, \ldots, d.$$

ii) If $x \preceq_{\mathrm{st}} x'$ then $x \preceq_{\mathrm{icx}} x'$.

iii) If $x \preceq_{\mathrm{icx}} x'$ then

$$\frac{1}{d}\sum_{i=1}^{d} x_i + \frac{d-1}{2(d+1)} g(x) \;\leq\; \frac{1}{d}\sum_{i=1}^{d} x'_i + \frac{d-1}{2(d+1)} g(x'),$$

where g is the Gini mean difference,

$$g(x) = \frac{1}{d(d-1)} \sum_{i,j=1}^{d} |x_i - x_j|. \tag{22}$$

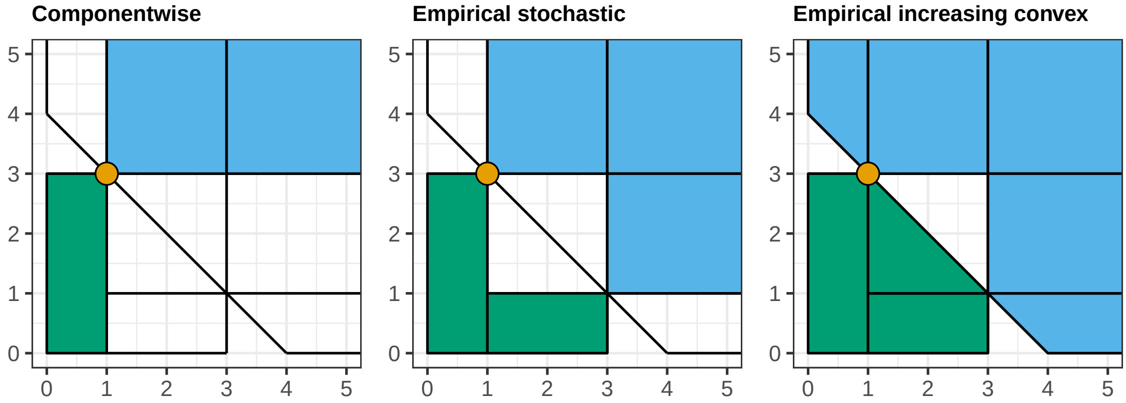


Fig. 3. Regions of smaller, greater and incomparable elements in the positive quadrant of $\mathbb{R}^2$, as compared to the point $(1, 3)$, for the (left) componentwise, (middle) empirical stochastic and (right) empirical increasing convex order. Coloured areas below (above) of $(1, 3)$ correspond to smaller (greater) elements, while blank areas contain elements incomparable to $(1, 3)$ in the given partial order.

iv) If $x \preceq_{\mathrm{icx}} x'$ and $x' \preceq_{\mathrm{icx}} x$ then x and x' are permutations of each other. Consequently, the relation $\preceq_{\mathrm{icx}}$ defines a partial order on $\mathbb{R}^d_{\uparrow}$.

Figure 3 illustrates the various types of relations for points in the positive quadrant of $\mathbb{R}^2$. As reflected by the nested character of the regions, the componentwise order is stronger than the empirical stochastic order, which in turn is stronger than the empirical increasing convex order. The latter is equivalent to *weak majorization* as studied by Marshall et al. (2011). In the special case of vectors with non-negative entries, their Corollary C.5 implies that $x \in \mathbb{R}^d$ is dominated by $x' \in \mathbb{R}^d$ in empirical increasing convex order if, and only if, it lies in the convex hull of the points of the form $(\xi_1 x'_{\pi(1)}, \dots, \xi_d x'_{\pi(d)})$, where π is a permutation and $\xi_i \in \{0, 1\}$ for $i = 1, \dots, d$.

4. Simulation study

Since we view IDR primarily as a tool for prediction, we compare it to other distributional regression methods in terms of predictive performance on continuous and discrete, univariate simulation examples, as measured by the CRPS. However, as noted below and formalized in Appendix D, the CRPS links asymptotically to L_2 estimation error, so under large validation samples prediction and estimation are assessed simultaneously. A detailed comparative study on mixed discrete-continuous data with a multivariate covariate vector is given in the case study in the next section.

Here, our simulation scenarios build on the illustrating example in the introduction. Specifically, we draw a covariate $X \sim \text{Unif}(0, 10)$ and then

$$
\begin{aligned}
Y_1 \mid X &\sim \text{Gamma}(\text{shape} = \sqrt{X}, \text{ scale} = \min\{\max\{X, 1\}, 6\}), && (23)\\
Y_2 \mid X &= Y_1 \mid X + 10 \cdot \mathbb{1}\{X \geq 5\}, && (24)\\
Y_3 \mid X &= Y_1 \mid X - 2 \cdot \mathbb{1}\{X \geq 7\}, && (25)\\
Y_4 \mid X &\sim \text{Poisson}(\lambda = \min\{\max\{X, 1\}, 6\})\}). && (26)
\end{aligned}
$$

Under each scenario we generate 500 training sets of size $n = 500$, 1 000, 2 000, and 4 000 each, fit distributional regression models, and validate on a test set of size $m = 5\,000$. For comparison with IDR, we use a nonparametric kernel (or nearest neighbor) smoothing technique (NP; Li and Racine, 2008), semiparametric quantile regression with monotone rearrangement (SQR; Koenker 2005; Chernozhukov et al. 2010), conditional transformation models (TRAM; Hothorn et al., 2014), and distributional or quantile random forests (QRF; Meinshausen 2006; Athey et al. 2019). These methods have been chosen as they are not subject to restrictive assumptions on the distribution of the response variable and have well established and well documented implementations in the statistical programming environment R (R Core Team, 2020). We also include the ideal forecast, i.e., the true conditional distribution of the response given the covariate, in the comparison.

Implementation details for the various methods are given in Table 3 in Appendix E. Here we only note that QRF uses the `grf` package (Tibshirani et al., 2020) with a splitting rule that is tailored to quantiles (Athey et al., 2019). We see that, unlike IDR, its competitors rely on manual intervention and tuning. For example, QRFs perform poorly under the default value of 5 for the tuning parameter `min.node.size`, which we have raised to 40. Further improvement may arise when tuning parameters, such as honesty fraction and node size, are judiciously adjusted to the specific scenario and training sample size at hand. In contrast, IDR is entirely free of implementation decisions, except for the subagging variant, IDR_{sbg}, where we average predictions based on estimates on 100 subsamples of size $n/2$ each.

Scenario (23) is the same as in the introduction and illustrated in Figure 1. It has a smooth covariate–response relationship, and NP, SQR, and even the misspecified TRAM technique, which are tailored to this type of setting, outperform QRF and IDR. However, the assumption of continuity in the response is crucial, as the results under the discontinuous scenario (24) demonstrate, where IDR and IDR_{sbg} perform best. In the non-isotonic scenario (25) IDR and IDR_{sbg} retain acceptable performance, even though the key assumption is violated. Not surprisingly, SQR faces challenges in the Poisson scenario (26), where the conditional quantile functions are piecewise constant, and IDR is outperformed only by TRAM. Throughout, the simplistic subagging variant of IDR has slightly lower mean CRPS than the default variant that is estimated on the full training set, and it would be interesting to explore the relation to the super-efficiency phenomenon described by Banerjee et al. (2019).

These results lend support to our belief that IDR can serve as a universal benchmark in probabilistic forecasting and distributional regression problems. For sufficiently large training samples, IDR offers competitive performance under any type

Table 1. Mean CRPS in smooth (23), discontinuous (24), non-isotonic (25), and discrete (26) simulation scenarios with training sets of size n.

	Smooth (23)				Discontinuous (24)			
n	500	1 000	2 000	4 000	500	1 000	2 000	4 000
NP	3.561	3.542	3.532	3.525	3.614	3.582	3.562	3.549
SQR	3.571	3.543	3.530	3.524	3.647	3.619	3.606	3.600
TRAM	3.560	3.543	3.535	3.531	3.642	3.625	3.616	3.612
QRF	3.589	3.561	3.555	3.553	3.614	3.576	3.561	3.556
IDR	3.604	3.568	3.548	3.535	3.628	3.581	3.555	3.540
IDR$_{\mathrm{sbg}}$	3.595	3.561	3.543	3.532	3.620	3.577	3.551	3.537
Ideal	3.516	3.516	3.516	3.516	3.516	3.516	3.516	3.516
	Non-isotonic (25)				Discrete (26)			
n	500	1 000	2 000	4 000	500	1 000	2 000	4 000
NP	3.564	3.544	3.534	3.527	1.136	1.131	1.128	1.126
SQR	3.574	3.546	3.533	3.527	1.129	1.121	1.116	1.114
TRAM	3.566	3.549	3.543	3.539	1.115	1.110	1.107	1.106
QRF	3.587	3.560	3.555	3.553	1.121	1.113	1.112	1.112
IDR	3.605	3.569	3.549	3.536	1.130	1.119	1.113	1.109
IDR$_{\mathrm{sbg}}$	3.597	3.564	3.545	3.534	1.128	1.118	1.112	1.109
Ideal	3.516	3.516	3.516	3.516	1.104	1.104	1.104	1.104

of type of linearly ordered outcome, without reliance on tuning parameters or other implementation choices, except when subsampling is employed.

5. Case study: Probabilistic quantitative precipitation forecasts

The past decades have witnessed tremendous progress in the science and practice of weather prediction (Bauer et al., 2015). Arguably, the most radical innovation consists in the operational implementation of ensemble systems and an accompanying culture change from point forecasts to distributional forecasts (Leutbecher and Palmer, 2008). An ensemble system comprises multiple runs of numerical weather prediction (NWP) models, where the runs or members differ from each other in initial conditions and numerical-physical representations of atmospheric processes.

Ideally, one would like to interpret an ensemble forecast as a random sample from the conditional distribution of future states of the atmosphere. However, this is rarely advisable in practice, as ensemble forecasts are subject to biases and dispersion errors, thereby calling for some form of statistical postprocessing (Gneiting and Raftery, 2005; Vannitsem et al., 2018). This is typically done by fitting a distributional regression model, with the weather variable of interest being the response variable, and the members of the forecast ensemble constituting the covariates, and applying this model to future NWP output, to obtain conditional predictive distributions for future weather quantities. State of the art techniques include Bayesian Model Averaging (BMA; Raftery et al., 2005; Sloughter et al., 2007), Ensemble

Table 2. Meteorological stations at airports, with International Air Transport Association (IATA) airport code, World Meteorological Organization (WMO) station ID, and data availability in days (years).

	IATA Code	WMO ID	Data Availability
Brussels, Belgium	BRU	06449	3406 (9.3)
Frankfurt, Germany	FRA	10637	3617 (9.9)
London, UK	LHR	03772	2256 (6.2)
Zurich, Switzerland	ZRH	06670	3241 (8.9)

Model Output Statistics (EMOS; Gneiting et al., 2005; Scheuerer, 2014), and Heteroscedastic Censored Logistic Regression (HCLR; Messner et al., 2014).

In this case study, we apply IDR to the statistical postprocessing of ensemble forecasts of accumulated precipitation, a variable that is notoriously difficult to handle, due to its mixed discrete-continuous character, which requires both a point mass at zero and a right skewed continuous component on the positive half-axis. As competitors to IDR, we implement the BMA technique of Sloughter et al. (2007), the EMOS method of Scheuerer (2014), and HCLR (Messner et al., 2014), which are widely used parametric approaches that have been developed specifically for the purposes of probabilistic quantitative precipitation forecasting. In contrast, IDR is a generic technique and fully automatic, once the partial order on the covariate space has been specified.

5.1. Data

The data in our case study comprise forecasts and observations of 24-hour accumulated precipitation from 06 January 2007 to 01 January 2017 at meteorological stations on airports in London, Brussels, Zurich and Frankfurt. As detailed in Table 2, data availability differs, and we remove days with missing entries station by station, so that all types of forecasts for a given station are trained and evaluated on the same data. Both forecasts and observations refer to the 24-hour period from 6:00 UTC to 6:00 UTC on the following day. The observations are in the unit of millimeter and constitute the response variable in distributional regression. They are typically, but not always, reported in integer multiples of a millimeter (mm).

As covariates, we use the 52 members of the leading NWP ensemble operated by the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al., 1996; Buizza et al., 2005). The ECMWF ensemble system comprises a high-resolution member (x_{HRES}), a control member at lower resolution (x_{CTR}) and 50 perturbed members ($x_1, \ldots, x_{50}$) at the same lower resolution but with perturbed initial conditions, and the perturbed members can be considered exchangeable (Leutbecher, 2019). To summarize, the covariate vector in distributional regression is

$$x = (x_1, \ldots, x_{50}, x_{\text{CTR}}, x_{\text{HRES}}) = (x_{\text{PTB}}, x_{\text{CTR}}, x_{\text{HRES}}) \in \mathbb{R}^{52}, \tag{27}$$

where $x_{\text{PTB}} = (x_1, \ldots, x_{50}) \in \mathbb{R}^{50}$. At each station, we use the forecasts for the corresponding latitude-longitude gridbox of size 0.25×0.25 degrees, and we consider prediction horizons of 1 to 5 days. For example, the two day forecast is initialized at 00:00 Universal Coordinated Time (UTC) and issued for 24-hour accumulated precipitation from 06:00 UTC on the next calendar day to 06:00 UTC on the day after. ECMWF forecast data are available online via the TIGGE system (Bougeault et al., 2010; Swinbank et al., 2016)

Statistical postprocessing is both a calibration and a downscaling problem: Forecasts and observations are at different spatial scales, whence the unprocessed forecasts are subject to representativeness error (Wilks, 2019, Chapter 8.9). Indeed, if we interpret the predictive distribution from the raw ensemble (27) as the empirical distribution of all 52 members — a customary approach, which we adopt hereinafter — there is a strong bias in probability of precipitation forecasts: Days with exactly zero precipitation are predicted much less often at the NWP model grid box scale than they occur at the point scale of the observations.

5.2. BMA, EMOS and HCLR

Before describing our IDR implementation, we review its leading competitors, namely, state of the art parametric distributional regression approaches that have been developed specifically for accumulated precipitation.

Techniques of ensemble model output statistics (EMOS; Gneiting et al., 2005) type can be interpreted as parametric instances of generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005). The specific variant of Scheuerer (2014) which we use here is based on the three-parameter family of left-censored generalized extreme value (GEV) distributions. The left-censoring generates a point mass at zero, corresponding to no precipitation, and the shape parameter allows for flexible skewness on the positive half-axis, associated with rain, hail or snow accumulations. The GEV location parameter is modeled as a linear function of x_{HRES}, x_{CTR}, $m_{\mathrm{PTB}} = \frac{1}{50}\sum_{i=1}^{50} x_i$ and

$$p_{\mathrm{ZERO}} = \frac{1}{52}\left(\mathbb{1}\{x_{\mathrm{HRES}} = 0\} + \mathbb{1}\{x_{\mathrm{CTR}} = 0\} + \sum_{i=1}^{50} \mathbb{1}\{x_i = 0\}\right),$$

and the GEV scale parameter is linear in the Gini mean difference (22) of the 52 individual forecasts in the covariate vector (27). While all parameters are estimated by minimizing the in-sample CRPS, the GEV shape parameter does not link to the covariates.

The general idea of the Bayesian model averaging (BMA; Raftery et al., 2005) approach is to employ a mixture distribution, where each mixture component is parametric and associated with an individual ensemble member forecast, with mixture weights that reflect the member's skill. Here we use the BMA implementation of Sloughter et al. (2007) for accumated precipitation in a variant that is based on x_{HRES}, x_{CTR}, $m_{\mathrm{PTB}} = \frac{1}{50}\sum_{i=1}^{50} x_i$ only, which we found to achieve more stable estimates and superior predictive scores than variants based on all members, as proposed by Fraley et al. (2010) in settings with smaller groups of exchangeable members. Hence, our BMA predictive CDF is of the form

$$F_x(y) = w_{\text{HRES}} G(y|x_{\text{HRES}}) + w_{\text{CTR}} G(y|x_{\text{CTR}}) + w_{\text{PTB}} G(y|m_{\text{PTB}})$$

for $y \in \mathbb{R}$, where the component CDFs $G(y|\cdot)$ are parametric, and the weights w_{HRES}, w_{CTR} and w_{PTB} are nonnegative and sum to one. Specifically, $G(y|x_{\text{HRES}})$ models the logit of the point mass at zero as a linear function of $\sqrt[3]{x_{\text{HRES}}}$ and $p_{\text{HRES}} = \mathbb{1}\{x_{\text{HRES}} = 0\}$, and the distribution for positive accumulations as a gamma density with mean and variance being linear in $\sqrt[3]{x_{\text{HRES}}}$ and x_{HRES}, respectively, and analogously for $G(y|x_{\text{CTR}})$ and $G(y|m_{\text{PTB}})$. Estimation relies on a two-step procedure, where the (component specific) logit and mean models are fitted first, followed by maximum likelihood estimation of the weight parameters and the (joint) variance model via the EM algorithm (Sloughter et al., 2007).

Heteroscedastic censored logistic regression (Messner et al., 2014) originates from the observation that conditional CDFs can be estimated by dichotomizing the random variable of interest at given thresholds and estimating the probability of threshold exceedance via logistic regression. The HCLR model used here assumes that square-root transformed precipitation follows a logistic distribution censored at zero, with location parameter linear in $\sqrt{x_{\text{HRES}}}$, $\sqrt{x_{\text{CTR}}}$ and the mean of the square-root transformed perturbed forecasts, and a scale parameter linear in the standard deviation of the square-root transformed perturbed forecasts. Like EMOS, HCLR can be interpreted within the GAMLSS framework of Rigby and Stasinopoulos (2005).

Code for BMA, EMOS and HCLR is available within the `ensembleBMA`, `ensembleMOS` and `crch` packages in R (Messner, 2018). Unless noted differently, we use default options in implementation decisions.

5.3. Choice of partial order for IDR

IDR applies readily in this setting, without any need for adaptations due to the mixed-discrete continuous character of precipitation accumulation, nor requiring data transformations or other types of implementation decisions. However, the partial order on the elements (27) of the covariate space $\mathcal{X} = \mathbb{R}^{52}$, or on a suitable derived space, needs to be selected thoughtfully, considering that the perturbed members $x_1, \ldots, x_{50}$ are exchangeable.

In the sequel, we apply IDR in three variants. Our first implementation is based on x_{HRES}, x_{CTR} and $m_{\text{PTB}} = \frac{1}{50}\sum_{i=1}^{50} x_i$ along with the componentwise order on $\mathbb{R}^3$, in that

$$x \preceq x' \iff m_{\text{PTB}} \le m'_{\text{PTB}},\ x_{\text{CTR}} \le x'_{\text{CTR}},\ x_{\text{HRES}} \le x'_{\text{HRES}}. \tag{28}$$

The second implementation uses the same variables and partial order, but combined with a simple subagging approach: Before applying IDR, the training data is split into the two disjoint subsamples of training observations with odd and even indices, and we average the predictions based on these two subsamples.

Our third implementation combines the empirical increasing convex order for x_{PTB} with the usual total order on $\mathbb{R}$ for x_{HRES}, whence

$$x \preceq x' \iff x_{\text{PTB}} \preceq_{\text{icx}} x'_{\text{PTB}},\ x_{\text{HRES}} \le x'_{\text{HRES}}. \tag{29}$$

Henceforth, we refer to the three implementations based on the partial orders in (28) and (29) as $\mathrm{IDR}_{\mathrm{cw}}$, $\mathrm{IDR}_{\mathrm{sbg}}$, and $\mathrm{IDR}_{\mathrm{icx}}$. With reference to the discussion preceding Theorem 2.1, the relations (28) and (29) define preorders on $\mathbb{R}^{52}$ and partial orders on $\mathbb{R}^3$ and $\mathbb{R}^{50}_{\uparrow} \times \mathbb{R}$, respectively.

We have experimented with other options as well, e.g., by incorporating the maximum $\max_{i=1,\ldots,50} x_i$ of the perturbed members in the componentwise order in (28), with the motivation that the maximum might serve as a proxy for the spread of the ensemble, or by using the empirical stochastic order $\preceq_{\mathrm{st}}$ in lieu of the empirical increasing convex order $\preceq_{\mathrm{icx}}$ in (29). IDR is robust to changes of this type, and the predictive performance remains stable, provided that the partial order honors the key substantive insights, in that the perturbed members $x_1, \ldots, x_{50}$ are exchangeable, while x_{HRES}, due to its higher native resolution, is able to capture local information that is not contained in x_{PTB} nor x_{CTR}. Hence, x_{HRES} ought to play a pivotal role in the partial order.

5.4. Selection of training periods

The selection of the training period is a crucial step in the statistical postprocessing of NWP output. Most postprocessing methods, including the ones used in this analysis, assume that there is a stationary relationship between the forecasts and the observations. As Hamill (2018) points out, this assumption is hardly ever satisfied in practice: NWP models are updated, instruments at observation stations get replaced, and forecast biases may vary seasonally. These problems are exacerbated by the fact that quantitative precipitation forecasts require large training datasets in order to include sufficient numbers of days with non-zero precipitation and extreme precipitation events.

For BMA and EMOS, a training period over a rolling window of the latest available 720 days at the time of forecasting is (close to) optimal at all stations. This resembles choices made by Scheuerer and Hamill (2015) who used a training sample of about 900 past instances. Scheuerer (2014) took shorter temporal windows, but merged instances from nearby stations into the training sets, which is not possible here. In general, it would be preferable to select training data seasonally (e.g., data from the same month), but in our case the positive effect of using seasonal training data does not outweigh the negative effect of a smaller sample size.

As a nonparametric technique, IDR requires larger sets of training data than BMA or EMOS. As training data for IDR, we used all data available at the time of forecasting, which is about 2 500 to 3 000 days for the stations Frankfurt, Brussels and Zurich, and 1 500 days for London Heathrow. The same training periods are also used for HCLR, where no positive effect of shorter, rolling training periods has been observed (Messner et al., 2014).

For evaluation, we use the years 2015 and 2016 (and 01 January 2017) for all postprocessing methods and the raw ensemble. This test dataset consists of roughly 700 instances for each station and lead time.

5.5. Results

Before comparing the BMA, EMOS, $\mathrm{IDR}_{\mathrm{cw}}$, $\mathrm{IDR}_{\mathrm{sbg}}$ and $\mathrm{IDR}_{\mathrm{icx}}$ techniques in terms of out-of-sample predictive performance over the test period, we exemplify them in Figure 4, where we show predictive CDFs for accumulated precipitation at Brussels on December 16, 2015, at a prediction horizon of 2 days. In panel (a) the marks at the bottom correspond to x_{HRES}, x_{CTR}, the perturbed members $x_1, \ldots, x_{50}$ and their mean m_{PTB}. The observation at 4 mm is indicated by the vertical line. Under all four techniques, the point mass at zero, which represents the probability of no precipitation, is vanishingly small. While the BMA, EMOS and HCLR CDFs are smooth and supported on the positive half-axis, the $\mathrm{IDR}_{\mathrm{cw}}$, $\mathrm{IDR}_{\mathrm{sbg}}$ and $\mathrm{IDR}_{\mathrm{icx}}$ CDFs are piecewise constant with jump points at observed values in the training period. Panel (b) illustrates the hard and soft constraints on the $\mathrm{IDR}_{\mathrm{cw}}$ CDF that arise from (20) under the order relation (28), with the thinner lines representing the $\mathrm{IDR}_{\mathrm{cw}}$ CDFs of direct successors and predecessors. In this example, the constraints are mostly hard, except for threshold values between 4 and 11 mm.

We now use the mean CRPS over the test period as an overall measure of out-of-sample predictive performance. Figure 5 shows the CRPS of the raw and post-processed forecasts for all stations and lead times, with the raw forecast denoted as ENS. While HCLR performs best in terms of the CRPS, the IDR variants show scores of a similar magnitude and outperform BMA in many instances. Figure 7 in Appendix E shows the difference of the empirical cumulative distribution function (ECDF) of the PIT defined at (8) to the bisector for the distributional forecasts. All three IDR variants show a PIT-distribution close to uniform, and so do BMA, EMOS and HCLR, as opposed to the raw ensemble, which is underdispersed.

In Figure 6 we evaluate probability of precipitation forecasts by means of the Brier score (Gneiting and Raftery, 2007), and Figure 8 in Appendix E shows reliability diagrams (Wilks, 2019; Dimitriadis et al., 2021). As opposed to the raw ensemble forecast, all distributional regression methods yield reliable probability forecasts. BMA, $\mathrm{IDR}_{\mathrm{cw}}$, $\mathrm{IDR}_{\mathrm{sbg}}$ and $\mathrm{IDR}_{\mathrm{icx}}$ separate the estimation of the point mass at zero, and of the distribution for positive accumulations, and the four methods perform ahead of EMOS. HCLR is outperformed by BMA and the IDR variants at lead times of one or two days, but achieves a lower Brier score at the longest lead time of five days.

Interestingly, IDR tends to outperform EMOS and HCLR for probability of precipitation forecasts, but not for precipitation accumulations. We attribute this to the fact that parametric techniques are capable of extrapolating beyond the range of the training responses, whereas IDR is not: The highest precipitation amount judged feasible by IDR equals the largest observation in the training set. Furthermore, unlike EMOS and HCLR, IDR does not use information about the spread of the raw ensemble, which is inconsequential for probability of precipitation forecasts, but may impede distributional forecasts of precipitation accumulations.

In all comparisons, the forecast performance of $\mathrm{IDR}_{\mathrm{cw}}$ and $\mathrm{IDR}_{\mathrm{sbg}}$ is similar. However, in our implementation, the simple subagging method used in $\mathrm{IDR}_{\mathrm{sbg}}$ reduced the computation time by up to one half.

To summarize, our results underscore the suitability of IDR as a benchmark

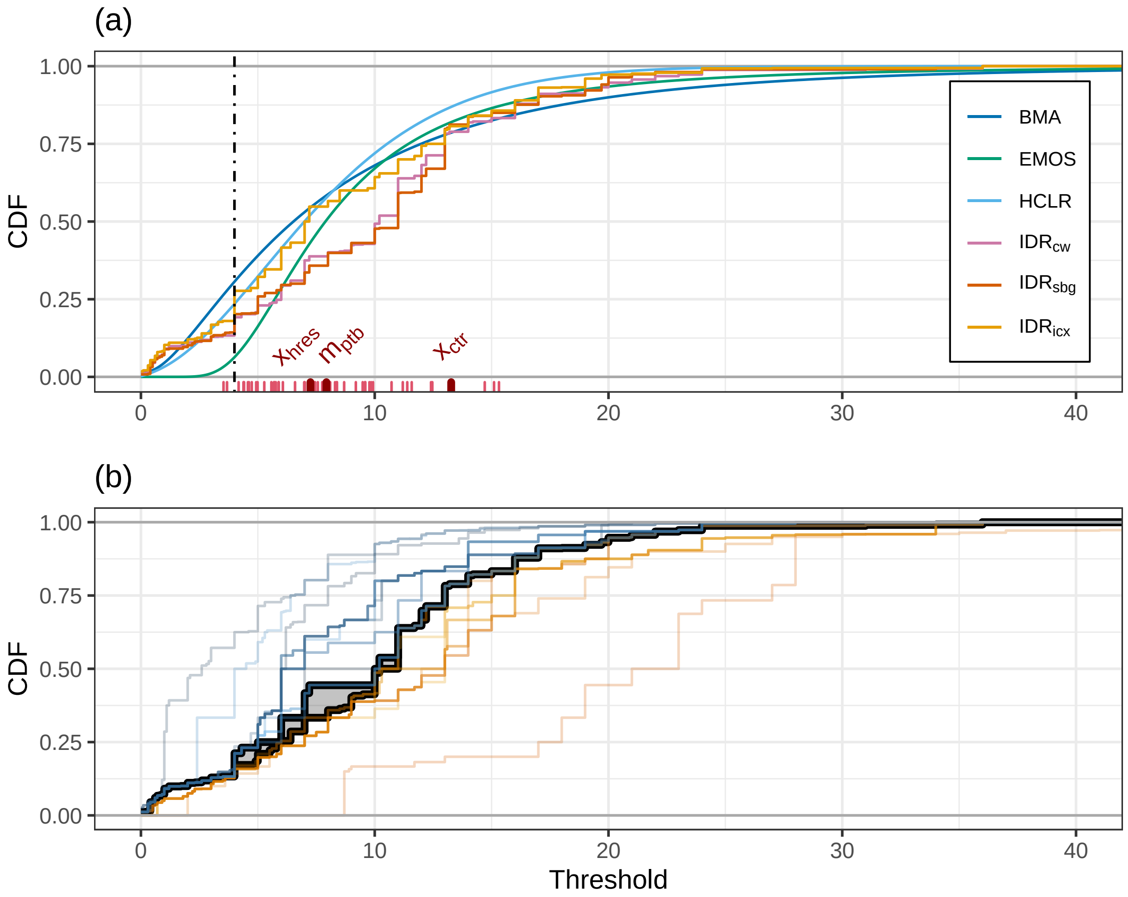


Fig. 4. Distributional forecasts for accumulated precipitation at Brussels, valid 16 December 2015 at a prediction horizon of 2 days. (a) BMA, EMOS, IDR$_{\text{cw}}$, IDR$_{\text{sbg}}$ and IDR$_{\text{icx}}$ predictive CDFs. The vertical line represents the observation. (b) IDR$_{\text{cw}}$ CDF along with the hard and soft constraints in (20) as induced by the order relation (28). The thin lines show the IDR$_{\text{cw}}$ CDFs at direct predecessors and successors.

technique in probabilistic forecasting problems. Despite being generic as well as fully automated, IDR is remarkably competitive relative to state of the art techniques that have been developed specifically for the purpose. In fact, in a wide range of applied problems that lack sophisticated, custom-made distributional regresssion solutions, IDR might well serve as a ready-to-use, top-performing method of choice.

6. Discussion

Stigler (1975) gives a lucid historical account of the 19th century transition from point estimation to distribution estimation. In regression analysis, we may be witnessing what future generations might refer to as the transition from conditional mean estimation to conditional distribution estimation, accompanied by a simultaneous transition from point forecasts to distributional forecasts (Gneiting and

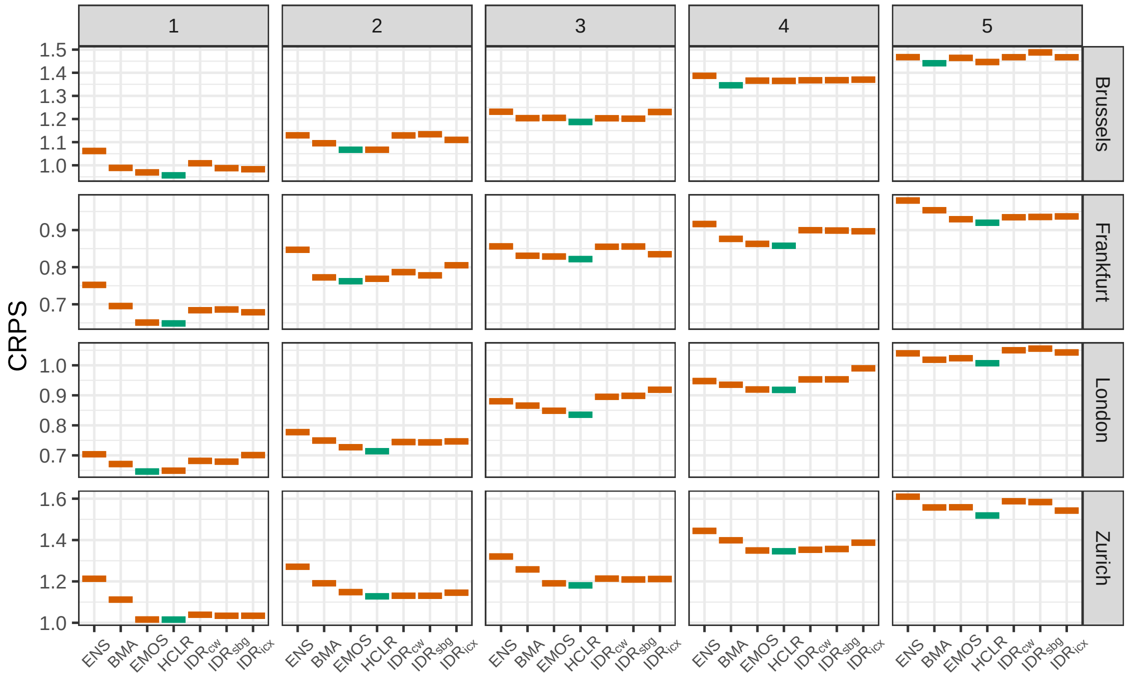


Fig. 5. Mean CRPS over the test period for raw and postprocessed ensemble forecasts of 24-hour accumulated precipitation at prediction horizons of 1, 2, 3, 4 and 5 days. The lowest mean score for a given lead time and station is indicated in green.

Katzfuss, 2014).

Isotonic distributional regression (IDR) is a nonparametric technique for estimating conditional distributions that takes advantage of partial order relations within the covariate space. It can be viewed as a far-reaching generalization of pool adjacent violators (PAV) algorithm based classical approaches to isotonic (non-distributional) regression, is entirely generic and fully automated, and provides for a unified treatment of continuous, discrete and mixed discrete-continuous real-valued response variables. Code for the implementation of IDR within R (R Core Team, 2020) and Python (`https://www.python.org/`) is available via the `isodistrreg` package at CRAN (`https://CRAN.R-project.org/package=isodistrreg`) and on github (`https://github.com/AlexanderHenzi/isodistrreg`; `https://github.com/evwalz/isodisreg`), with user-friendly functions for partial orders, estimation, prediction and evaluation.

IDR relies on information supplied by order constraints, and the choice of the partial order on the covariate space is a critical decision prior to the analysis. Only variables that contribute to the partial order need to be retained, and the order constraints serve to regularize the IDR solution. Weak orders lead to increased numbers of comparable pairs of training instances and predictive distributions that are more regular. The choice of the partial order is typically guided and informed by substantive expertise, as illustrated in our case study, and it is a challenge for

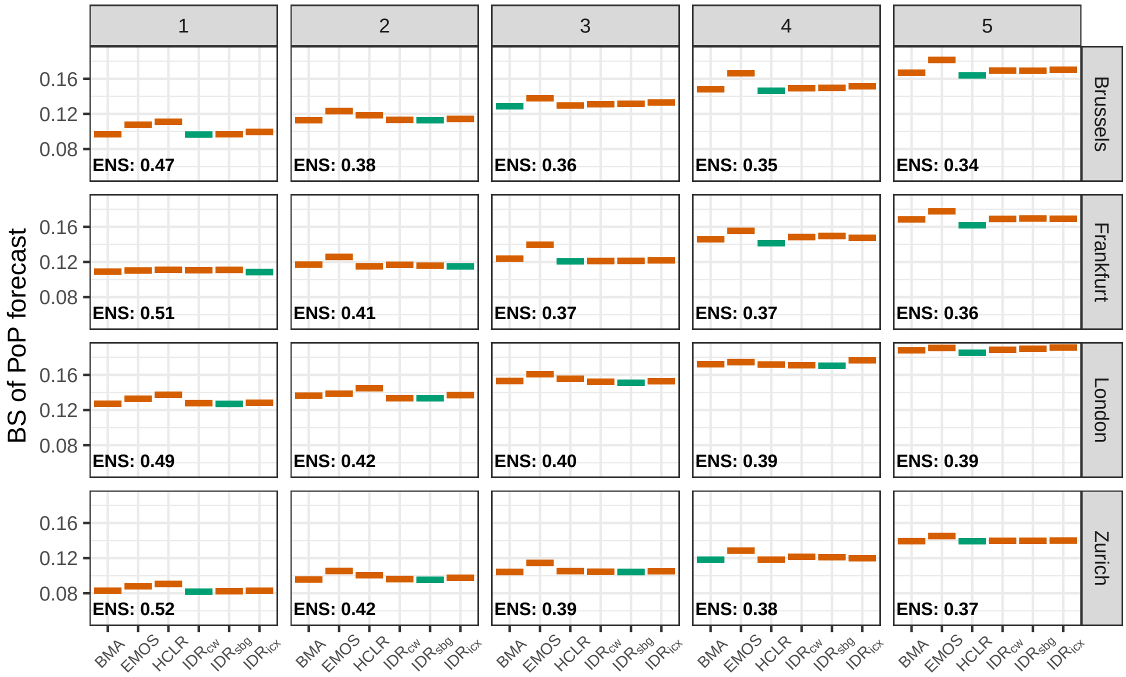


Fig. 6. Mean Brier score over the test period for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days. The lowest mean score for a given lead time and station is indicated in green.

future research to investigate whether the selection of the partial order could be automated. Given that IDR gains information through order constraints, it is a valid concern whether it is robust under misspecifications of the partial order. There is evidence that this is indeed the case: IDR has guaranteed in-sample threshold calibration (Theorem 2.2) and therefore satisfies a minimal requirement for reliable probabilistic forecasts under any (even misspecified) partial order. Moreover, El Barmi and Mukerjee (2005, Theorem 7) show that in the special case of a discrete, totally ordered covariate, isotonic regression asymptotically has smaller estimation error than non-isotonic alternatives even under mild violations of the monotonicity assumptions, akin to the performance of IDR in the non-isotonic setting (25) in our simulation study.

Unlike other methods for distributional regression, which require implementation decisions, such as the specification of parametric distributions, link functions, estimation procedures and convergence criteria, to be undertaken by users, IDR is fully automatic once the partial order and the training set have been identified. In this light, we recommend that IDR be used as a benchmark technique in distributional regression and probabilistic forecasting problems. With both computational efficiency and the avoidance of overfitting in mind, IDR can be combined with subsample aggregation (subagging) in the spirit of random forests. In our case study on quantitative precipitation forecasts, we used simplistic ad hoc choices for the size and number of subsamples. Future research on computationally efficient algorithmic implementations of IDR as well as optimal and automated choices of subsampling settings is highly desirable.

A limitation of IDR in its present form is that we only consider the usual stochastic order on the space $\mathcal{P}$ of the conditional distributions. Hence, IDR is unable to distinguish situations where the conditional distributions agree in location but differ in spread, shape or other regards. This restriction is of limited concern for response variables such as precipitation accumulation or income, which are bounded below and right skewed, but may impact the application of IDR to variables with symmetric distributions. In this light, we encourage future work on ramifications of IDR, in which $\mathcal{P}$ is equipped with partial orders other than the stochastic order, including but not limited to the likelihood ratio order (Mösching and Dümbgen, 2020a). Similarly, the "spiking" problem of traditional isotonic regression, which refers to unwarranted jumps of estimates at boundaries, arguably did not have adverse effects in our simulation and case studies. However, it might be of concern in other applications, where remedies of the type proposed by Wu et al. (2015) might yield improvement and warrant study.

Another promising direction for further research are generalizations of IDR to multivariate response variables. In weather prediction, this would allow simultaneous postprocessing of forecasts for several variables, and an open question is for suitable notions of multivariate stochastic dominance that allow efficient estimation in such settings.

Acknowledgements

The authors are grateful to the editor, the associate editor, four anonymous referees, Sebastian Arnold, Lutz Dümbgen, Ed George, Stephan Hemri, Alexander Jordan, Kristof Kraus, Alexandre Mösching, Anja Mühlemann, Aaditya Ramdas, Nicholas Reich, Benedikt Schulz, Peter Vogel, Jonas Wallin and Eva-Maria Walz for providing code, assistance with the data handling, comments and discussion. Tilmann Gneiting acknowledges support by the Klaus Tschira Foundation, by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 257899354 – TRR 165, and via the Fellowship programme operated by the European Centre for Medium-Range Weather Forecasts (ECMWF). Alexander Henzi and Johanna F. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation. Calculations were mostly performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

References

Allen, M. S. and F. A. Eckel (2012). Value from ambiguity in ensemble forecasts. *Weather and Forecasting 27*, 70–84.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics 37*, 1148–1178.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics 26*, 641–647.

Banerjee, M., C. Durot, and B. Sen (2019). Divide and conquer in nonstandard problems and the super efficiency phenomenon. *Annals of Statistics 47*, 720–757.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference Under Order Restrictions*. John Wiley & Sons.

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika 46*, 36–48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives II. *Biometrika 46*, 329–335.

Basel Committee on Banking Supervision (2016). Standard: Minimum Capital Requirements for Market Risk.

Bauer, P., A. Thorpe, and G. Brunet (2015). The quiet revolution of numerical weather prediction. *Nature 525*, 47–55.

Ben Bouallègue, Z., T. Haiden, and D. S. Richardson (2018). The diagonal score: Definition, properties, and interpretations. *Quarterly Journal of the Royal Meteorological Society 144*, 1463–1473.

Bougeault, P., Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, and S. Worley (2010). The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society 91*, 1059–1072.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth.

Brightwell, G. (1992). Random k-dimensional orders: Width and number of linear extensions. *Order 9*, 333–342.

Brunk, H. B. (1955). Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics 26*, 607–616.

Bühlmann, P. and B. Yu (2002). Analyzing bagging. *Annals of Statistics 30*, 927–961.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review 133*, 1076–1097.

Buja, A. and W. Stützle (2006). Observations on bagging. *Statistica Sinica 16*, 323–351.

Casady, R. J. and J. D. Cryer (1976). Monotone percentile regression. *Annals of Statistics 4*, 532–541.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*, 1093–1125.

Davidov, O. and G. Iliopoulos (2012). Estimating a distribution function subject to a stochastic ordering restriction: A comparative study. *Journal of Nonparametric Statistics 24*, 923–933.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A 147*, 278–290.

de Leeuw, J., K. Hornik, and P. Mair (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software 32(5)*, 1–19.

Dette, H., N. Neumeyer, and K. F. Pilz (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli 12*, 469–490.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

Dimitriadis, T., T. Gneiting, and A. I. Jordan (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences 118*, e2016191118.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society Series B 69*, 163–183.

Ehm, W., T. Gneiting, A. Jordan, and F. Krüger (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations, and forecast rankings (with discussion). *Journal of the Royal Statistical Society Series B 78*, 505–562.

El Barmi, H. and H. Mukerjee (2005). Inferences under a stochastic ordering constraint. *Journal of the American Statistical Association 100*, 252–261.

Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics 75*, 643–669.

Fraley, C., A. E. Raftery, and T. Gneiting (2010). Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review 138*, 190–202.

Gasthaus, J., K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski (2019). Probabilistic forecasting with spline quantile function RNNs. *Proceedings of Machine Learning Research 89*, 1901–1910.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*, 746–762.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B 69*, 243–268.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*, 125–151.

Gneiting, T. and A. E. Raftery (2005). Weather forecasting with ensemble methods. *Science 310*, 248–249.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*, 1098–1118.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*, 411–422.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

Groeneboom, P. and G. Jongbloed (2014). *Nonparametric Estimation under Shape Constraints*. Cambridge University Press.

Guntuboyina, A. and B. Sen (2018). Nonparametric shape-restricted regression. *Statistical Science 33*, 563–594.

Hall, P., R. C. L. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association 94*, 154–163.

Hamill, T. M. (2018). Practical aspects of statistical postprocessing. In S. Vannitsem, D. S. Wilks, and J. Messner (Eds.), *Statistical Postprocessing of Ensemble Forecasts*, pp. 187–217. Elsevier.

Han, Q., T. Wang, S. Chatterjee, and R. J. Samworth (2019). Isotonic regression in general dimensions. *Annals of Statistics 47*, 2440–2471.

Hayfield, T. and J. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5), 1–32.

Henzi, A., A. Mösching, and L. Dümbgen (2020). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. Preprint, arxiv.org/abs/2006.05527.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*, 559–570.

Hogg, R. V. (1965). On models and hypotheses with restricted alternatives. *Journal of the American Statistical Association 60*, 1153–1162.

Hothorn, T., T. Kneib, and P. Bühlmann (2014). Conditional transformation models. *Journal of the Royal Statistical Society Series B 76*, 3–27.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger (2004). Bagging survival trees. *Statistics in Medicine 23*, 77–91.

Hothorn, T. (2020). Most likely transformations: The mlt package. *Journal of Statistical Software 92*(1), 1–68.

Jordan, A. I., A. Mühlemann, and J. F. Ziegel (2021). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Annals of the Institute of Statistical Mathematics*, to appear.

Klein, N., T. Kneib, S. Lang, and A. Sohn (2015). Bayesian structured additive distributional forecasting with an application to regional income inequality in Germany. *Annals of Applied Statistics 9*, 1024–1052.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. (2020). *quantreg: Quantile Regression.* R package version 5.61.

Laio, F. and S. Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences 11*, 1267–1277.

Lee, C.-I. C. (1983). The min-max algorithm and isotonic regression. *Annals of Statistics 11*, 467–477.

Leutbecher, M. and T. N. Palmer (2008). Ensemble forecasting. *Journal of Computational Physics 227*, 3515–3539.

Leutbecher, M. (2019). Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society 145*, 107–128.

Li, Q. and J. Racine (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics 26*, 423–434.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Annals of Statistics 19*, 724–740.

Marshall, A. W., I. Olkin, and B. C. Arnold (2011). *Inequalities: Theory of Majorization and its Applications* (2nd ed.). Springer.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*, 1087–1096.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research 7*, 983–999.

Messner, J. W. (2018). Ensemble postprocessing with R. In S. Vannitsem, D. S. Wilks, and J. Messner (Eds.), *Statistical Postprocessing of Ensemble Forecasts*, pp. 291–326. Elsevier.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review 142*, 3003–3014.

Miles, R. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika 46*, 317–327.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society 122*, 73–119.

Mösching, A. and L. Dümbgen (2020a). Maximum likelihood estimation of a likelihood ratio ordered family of distributions. Preprint, arxiv.org/abs/2007.11521.

Mösching, A. and L. Dümbgen (2020b). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics 14*, 24–49.

Neelon, B. and D. B. Dunson (2004). Bayesian isotonic regression and trend analysis. *Biometrics 60*, 398–406.

Newey, W. K. and J. L. Powell (1987). Asymmetric least squares estimation and testing. *Econometrica 55*, 819–847.

Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salomon (2015). How do I know if my forecasts are better? Using benchmarks in hydrologic ensemble prediction. *Journal of Hydrology 522*, 697–713.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*, 1155–1174.

Rasp, S. and S. Lerch (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review 146*, 3885–3900.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society Series C (Applied Statistics) 54*, 507–554.

Robertson, T. and F. T. Wright (1975). Consistency in generalized isotonic regression. *Annals of Statistics 3*, 350–362.

Robertson, T. and F. T. Wright (1980). Algorithms in order restricted statistical inference and the Cauchy mean value property. *Annals of Statistics 8*, 645–651.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order Restricted Statistical Inference.* John Wiley & Sons.

Rojo, J. and H. El Barmi (2003). Estimation of distribution functions under second order stochastic dominance. *Statistica Sinica 13*, 903–926.

Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature 51*, 1063–1110.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science 28*, 616–640.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Annals of Statistics 17*, 1856–1879.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society 140*, 1086–1096.

Scheuerer, M. and T. M. Hamill (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review 143*, 4578–4596.

Schlosser, L., T. Hothorn, R. Stauffer, A. Zeileis, et al. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics 13*, 1564–1589.

Schulze Waltrup, L., F. Sobotka, T. Kneib, and G. Kauermann (2015). Expectile and quantile regression — David and Goliath? *Statistical Modelling 15*, 433–456.

Seo, K. (2009). Ambiguity and second-order belief. *Econometrica 77*, 1575–1605.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic Orders.* Springer, New York.

Shively, T. S., T. W. Sager, and S. G. Walker (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society Series B 71*, 159–175.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley (2007). Probabilistic quantitative precipitation forecasting using Baysian model averaging. *Monthly Weather Review 135*, 3209–3220.

Spady, R. H. and S. Stouli (2018). Dual regression. *Biometrika 105*, 1–18.

Stellato, B., G. Banjac, P. Goulart, A. Bemporad, and S. Boyd (2020). OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation 12*, 637–672.

Stellato, B., G. Banjac, P. Goulart, and S. Boyd (2019). *osqp: Quadratic Programming Solver using the 'OSQP' Library.* R package version 0.6.0.3.

Stigler, S. M. (1975). The transition from point to distribution estimation. *Bulletin of the International Statistical Institute 46(2)*, 332–340.

Swinbank, R., M. Kyouda, P. Buchanan, L. Froude, T. M. Hamill, T. D. Hewson, J. H. Keller, M. Matsueda, J. Methven, F. Pappenberger, M. Scheuerer, H. A. Titley, L. Wilson, and M. Yamaguchi (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society 97*, 49–67.

Taillardat, M., A. L. Fougères, P. Naveau, and O. Mestre (2019). Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting 34*, 617–634.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review 144*, 2375–2393.

Tibshirani, J., S. Athey, and S. Wager (2020). *grf: Generalized Random Forests.* R package version 1.2.0.

Umlauf, N. and T. Kneib (2018). A primer on Bayesian distributional regression. *Statistical Modelling 18*, 219–247.

van Eeden, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions.* Ph. D. thesis, University of Amsterdam.

Vannitsem, S., D. S. Wilks, and J. Messner (Eds.) (2018). *Statistical Postprocessing of Ensemble Forecasts.* Elsevier.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting 33*, 369–388.

Vovk, V., J. Shen, V. Manokhin, and M. Xie (2019). Nonparametric predictive distributions based on conformal prediction. *Machine Learning 108*, 445–474.

Wilbur, W. J., L. Yeganova, and W. Kim (2005). The synergy between PAV and AdaBoost. *Machine Learning 61*, 71–103.

Wilks, D. S. (2019). *Statistical Methods in the Atmospheric Sciences* (4th ed.). Elsevier.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association 67*, 187–191.

Wu, J., M. C. Meyer, and J. D. Opsomer (2015). Penalized isotonic regression. *Journal of Statistical Planning and Inference 161*, 13–24.

The appendices are available as Supplementary Material on the JRSSB website; see https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssb.12450.

A. Proofs for Section 2.2

Proof of Theorem 2.1. Let $\mathcal{A}$ be the lattice of all subsets of $\{1,\dots,n\}$ that yield admissible superlevel sets for an increasing function $\{x_1,\dots,x_n\} \to \mathbb{R}$. More precisely, a set $A \subseteq \{1,\dots,n\}$ belongs to $\mathcal{A}$ if and only if for any $i \in A$ and any x_j with $x_i \preceq x_j$ it follows that $j \in A$.

Let $z \in \mathbb{R}$. By Jordan et al. (2021, Theorem 1 and Lemma 4), the minimizer of the criterion

$$\frac{1}{n}\sum_{i=1}^{n}(p_i - \mathbb{1}\{z \geq y_i\})^2 \tag{30}$$

over all $\boldsymbol{p} = (p_1,\dots,p_n) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$ is uniquely determined and given by $\hat{\boldsymbol{F}}(z) = (\hat{F}_1(z),\dots,\hat{F}_n(z)) \in \mathbb{R}^n$ with

$$\hat{F}_i(z) = \min_{A\in\mathcal{A}:i\in A}\ \max_{A'\in\mathcal{A}:A'\subsetneq A} \frac{1}{\#(A\backslash A')}\sum_{j\in A\backslash A'}\mathbb{1}\{y_j \leq z\}, \tag{31}$$

for $i = 1,\dots,n$, where $\#B$ denotes the cardinality of a set B. From the definition of the CRPS it is clear that $\boldsymbol{F}$ minimizes $\ell_{\mathrm{CRPS}}(\boldsymbol{F})$ over all tuples of functions $\boldsymbol{F} = (F_1,\dots,F_n)$ with $F_i : \mathbb{R} \to \mathbb{R}$ such that for each $z \in \mathbb{R}$, $(F_1(z),\dots,F_n(z)) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$. It remains to show that for each $i = 1,\dots,n$, F_i is a valid CDF.

Let $i \in \{1,\dots,n\}$, $z \leq z'$, $B \subseteq \{1,\dots,n\}$. It is clear from (31) that the domain of F_i in $[0,1]$. Furthermore,

$$\frac{1}{\#B}\sum_{j\in B}\mathbb{1}\{y_j \leq z\} \leq \frac{1}{\#B}\sum_{j\in B}\mathbb{1}\{y_j \leq z'\}, \tag{32}$$

and therefore, by (31), $F_i(z) \leq F_i(z')$. The function F_i is also right-continuous because for $z' \downarrow z$, the right-hand side of (32) converges to the left-hand side. Finally, for $z \to \pm\infty$ the left-hand side of (32) converges to zero and one, respectively, which concludes the proof. □

Proof of Theorem 2.2. First, we show threshold calibration. Let (X, Y) be a random vector with distribution $(1/n)\sum_{i=1}^{n}\delta_{(x_i,y_i)}$ where $\delta_{(x_i,y_i)}$ denotes the Dirac measure at (x_i, y_i). Let $z \in \mathbb{R}$. By Lee (1983, Theorem 6.4), there exists a partition $\{B_m\}_{m=1}^{M}$ of $\{1,\dots,n\}$ such that

$$F_i(z) = F_{x_i}(z) = \sum_{m=1}^{M}\mathbb{1}\{i \in B_m\}\frac{1}{\#B_m}\sum_{j\in B_m}\mathbb{1}\{y_j \leq z\}.$$

Therefore, the σ-algebra generated by $F_X(z)$ is contained in the σ-algebra generated by $\{\bar{B}_m\}_{m=1}^M$ with $\bar{B}_m = \{(x_i, y_i) : \ i \in B_m\}$. Furthermore,

$$\begin{aligned} \mathbb{E}\left(\mathbb{1}\{Y \le z\}\mathbb{1}\{(X,Y) \in \bar{B}_m\}\right) &= \frac{1}{n}\sum_{j \in B_m} \mathbb{1}\{y_j \le z\} \\ &= \mathbb{E}\left(F_X(z)\mathbb{1}\{(X,Y) \in \bar{B}_m\}\right). \end{aligned}$$

Part i) for the scoring rules of type (12) follows directly from the arguments in the proof of Theorem 2.1. Let $z \in \mathbb{R}$. By Jordan et al. (2021, Theorem 1 and Lemma 4) the solution $\hat{\boldsymbol{F}}(z)$ at (31) is not only the unique minimizer of the criterion (30) but also the unique solution that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}\{c < p_i\} - \mathbb{1}\{y_i \le z\}\right)\left(c - \mathbb{1}\{y_i \le z\}\right) \tag{33}$$

over all $\boldsymbol{p} = (p_1, \dots, p_n) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$ simultaneously for all $c \in (0,1)$. As $\hat{\boldsymbol{F}} \in \mathcal{P}^n_{\downarrow,\boldsymbol{x}}$, and $(1/n)\sum_{i=1}^n \mathrm{S}_{z,c}(F_i, y_i)$ is equal to the expression at (33) with $p_i = F_i(z)$, we obtain the claim.

Part iii) is a direct consequence of the arguments for the second part of part i) and the representation theorem of Schervish (1989) for proper scoring rules of binary events.

Let $\alpha \in (0,1)$. Concerning part ii), observe that any function s_α satisfying the requirements of the theorem can be written as $\int \tilde{\mathrm{S}}^Q_{\alpha,\theta}(q,y)\,\mathrm{d}h(\theta)$ for some Borel measure h on $\mathbb{R}$; see Ehm et al. (2016, Theorem 1). Here,

$$\tilde{\mathrm{S}}^Q_{\alpha,\theta}(q,y) = \begin{cases} 1-\alpha, & y \le \theta < q, \\ \alpha, & q \le \theta < y, \\ 0, & \text{otherwise.} \end{cases}$$

By Jordan et al. (2021, Theorem 1 and Proposition 5) there exists a unique solution $\hat{\boldsymbol{q}}(\alpha) = (\hat{q}_1(\alpha), \dots, \hat{q}_n(\alpha)) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{\mathrm{S}}^Q_{\alpha,\theta}(q_i, y_i)$$

over all $\boldsymbol{q} = (q_1, \dots, q_n) \in \mathbb{R}^n_{\downarrow,\boldsymbol{x}}$ simultaneously over all $\theta \in \mathbb{R}$ such that for each $i \in \{1, \dots, n\}$, $\hat{q}_i(\alpha)$ is the lower α-sample-quantile of some subset of observations $B_i \subseteq \{y_1, \dots, y_n\}$. Indeed, the solution has a max-min representation as in (31) with the empirical mean of the indcators replaced by the lower α-sample quantile over all observations in $A \backslash A'$. The max-min representation for $\hat{q}_i(\alpha)$ yields that $\hat{q}_i(\cdot)$ is increasing and left-continuous because lower α-sample-quantiles are increasing and left-continuous as a function of α. Therefore, $\hat{q}_i(\cdot)$ is a valid quantile function for each $i = 1, \dots, n$, and the generalized inverse $\hat{\boldsymbol{q}}^{-1} = (\hat{q}_1^{-1}, \dots, \hat{q}_n^{-1})$ is a member of $\mathcal{P}^n_{\uparrow,\boldsymbol{x}}$.

Since $\mathrm{S}^Q_{\alpha,\theta}(F,y) = \tilde{\mathrm{S}}^Q_{\alpha,\theta}(F^{-1}(\alpha),y)$ for any CDF F, it follows from (3) that $\hat{\boldsymbol{q}}^{-1}$ is a CRPS-based isotonic regression of $\boldsymbol{y}$ on $\boldsymbol{x}$. To conclude the proof of part ii), it remains to note that $\hat{\boldsymbol{q}}^{-1} = \hat{\boldsymbol{F}}$ due to the uniqueness of $\hat{\boldsymbol{F}}$. The initial statement in part i) is now also immediate. □

B. Proofs and remarks for Section 2.4

The proof of Theorem 2.3 requires the following lemma, which is established in Mösching and Dümbgen (2020b, Theorem 4.6).

Lemma B.1. *Let $Z_1, Z_2, \ldots$ be independent random variables with distribution functions $G_1, G_2, \ldots$, respectively. For $k = 1, 2, \ldots$, let*

$$\hat{\mathbb{G}}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{Z_i \le \cdot\} \quad \text{and} \quad \bar{G}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} G_i(\cdot).$$

Then there exists a universal constant $M \le 2^{5/2}e$ such that for all $\eta \ge 0$,

$$\mathbb{P}\left(\sqrt{k}\,\|\hat{\mathbb{G}}_k - \bar{G}_k\|_\infty \ge \eta\right) \le M\exp(-2\eta^2),$$

where $\|\cdot\|_\infty$ denotes the usual supremum norm of functions.

Proof of Theorem 2.3. Let $\epsilon, \delta > 0$. By assumption (iv), there exists $r > 0$ such that

$$\sup\{|F_x(y) - F_{x'}(y)| : x, x' \in [0,1]^d, \|x - x'\| \le r, y \in \mathbb{R}\} < \frac{\epsilon}{4}. \tag{34}$$

Let $m = \max(\lceil 2/r \rceil, \lceil 2/\delta \rceil + 1)$ and define intervals $I_1 = [0, 1/m]$ and $I_j = ((j-1)/m, j/m]$ for $j = 2, \ldots, m$. For indices $j_1, \ldots, j_d \in \{1, \ldots, m\}$, let $I(j_1, \ldots, j_d) = \times_{k=1}^{d} I_{j_k} \subset [0,1]^d$. The collection of such rectangles, which we denote by $\mathcal{R}$, partitions $[0,1]^d$ into m^d disjoint subsets with $\sup_{x,x' \in I(j_1,\ldots,j_d)} \|x - x'\| \le r/2$.

By assumption (i), for each $J \in \mathcal{R}$, there exists $c_J > 0$ such that with asymptotic probability one, $\#(S_n \cap J) \ge nc_J$. Define $c = \min_{J \in \mathcal{R}} c_J > 0$, so that with asymptotic probability one, $\#(S_n \cap J) \ge nc > 0$. We assume in the following that for $(X_{n1}, Y_{nn}), \ldots, (X_{nn}, Y_{nn})$ the event in assumption (i) occurs for all $J \in \mathcal{R}$ as well as the event in assumption (ii). To ease notation, we drop the subscript n.

Let $x = (x_1, \ldots, x_d) \in [\delta, 1-\delta]^d$. Then $2/m < \delta \le \min_{i=1\ldots,d} x_i$ and $\max_{i=1,\ldots,d} x_i \le 1 - \delta < (m-2)/m$, and there exist indices $j_1, \ldots, j_d \in \{3, \ldots, m-2\}$ such that $x \in I(j_1, \ldots, j_d)$. Define

$$L(x) = I(j_1 - 1, \ldots, j_d - 1), \quad U(x) = I(j_1 + 1, \ldots, j_d + 1).$$

Then $v \preceq x \preceq w$ for all $v \in L(x)$ and $w \in U(x)$, and

$$\sup_{v \in L(x)} \|v - x\| \le r, \qquad \sup_{w \in U(x)} \|w - x\| \le r.$$

We see from (34) that

$$\sup_{v \in L(x) \cup U(x), y \in \mathbb{R}} |F_v(y) - F_x(y)| \le \frac{\epsilon}{4},$$

whereas the bounds in (16) give

$$\hat{F}_{X_u}(y) \le \hat{F}_x(y) \le \hat{F}_{X_l}(y), \quad y \in \mathbb{R},\ X_u \in U(x),\ X_l \in L(x).$$

Consequently, for $y \in \mathbb{R}$,

$$\begin{aligned} |\hat{F}_x(y) - F_x(y)| &\le \max_{j:X_j \in L(x)\cup U(x)} |\hat{F}_{X_j}(y) - F_{X_j}(y)| + \frac{\epsilon}{4} \\ &\le \sup_{j:X_j \in (1/m,(m-1)/m]^d, y\in\mathbb{R}} |\hat{F}_{X_j}(y) - F_{X_j}(y)| + \frac{\epsilon}{4}, \end{aligned}$$

and this upper bound does not depend on x. Therefore, it is sufficient to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{j:X_j \in (1/m,(m-1)/m]^d,\, y\in\mathbb{R}} |\hat{F}_{X_j}(y) - F_{X_j}(y)| \ge \frac{3\epsilon}{4} \right) = 0. \tag{35}$$

Let $\mathcal{A}_n$ be the collection of upper sets in S_n. By the min-max formula for antitonic regression, for $j = 1, \dots, n$ and $y \in \mathbb{R}$,

$$\hat{F}_{X_j}(y) = \min_{A\in\mathcal{A}_n : X_j \in A} \max_{A'\in\mathcal{A}_n : X_j \notin A'} \frac{1}{\#(A \setminus A')} \sum_{i:X_i \in A\setminus A'} \mathbb{1}\{Y_i \le y\}.$$

For $X_j \in (1/m,(m-1)/m]^d$, let $j_i = \max\{k : k/m < X_{j,i}\} - 1$ and $x_j = (j_1/m, \dots, j_d/m) \in \mathbb{R}^d$. Here, $X_{j,i}$ denotes the i-th component of X_j. Then, for all $v \in [x_j, X_j] := \{u \in [0,1]^d : x_j \preceq u \preceq X_j\}$ it holds that $\|v - X_j\| \le 2/m \le r$. Therefore, inequality (34) along with assumption (iii) imply that for all i in $\{1,\dots,n\}$ such that $X_i \succeq x_j$,

$$F_{X_i}(y) \le F_{x_j}(y) \le F_{X_j}(y) + \frac{\epsilon}{4}, \quad y \in \mathbb{R}.$$

Consequently, with $A_j = \{v \in [0,1]^d : v \succeq x_j\}$,

$$\hat{F}_{X_j}(y) - F_{X_j}(y) \le \max_{A'\in\mathcal{A}_n : X_j \notin A'} \frac{1}{\#(A_j \setminus A')} \sum_{i:X_i \in A_j\setminus A'} (\mathbb{1}\{Y_i \le y\} - F_{X_i}(y)) + \frac{\epsilon}{4}.$$

By the definition of $j_1, \dots, j_d$, $I(j_1 + 1, \dots, j_d + 1) \subseteq [x_j, X_j] \subseteq A_j \setminus A'$ for $A' \in \mathcal{A}_n$ with $X_j \notin A'$. Therefore, $\#(A_j \setminus A') \ge cn > 0$, where c is the constant introduced at the beginning of the proof. Lemma B.1 implies that for all $A' \subseteq A_j$ with $X_j \notin A'$, conditional on $X_1, \dots, X_n$,

$$\mathbb{P}\left(\sup_{y\in\mathbb{R}} \frac{1}{\#(A_j \setminus A')} \left| \sum_{i:X_i \in A_j\setminus A'} (\mathbb{1}\{Y_i \le y\} - F_{X_i}(y)) \right| \ge \frac{\epsilon}{2} \right) \le M \exp\Big(-\frac{c}{2}\epsilon^2 n\Big),$$

with a constant $M \le 2^{5/2}e$ that does not depend on j. In view of the Bonferroni inequality we get the upper bound

$$\begin{aligned} \mathbb{P}\left(\sup_{y\in\mathbb{R}} \left(\hat{F}_{X_j}(y) - F_{X_j}(y)\right) \ge \frac{3\epsilon}{4} \right) &\le \sum_{A'\in\mathcal{A}:X_j\notin A'} M \exp\left(-\frac{c}{2}\epsilon^2 n\right) \\ &\le \#(\mathcal{A}_n)\ M \exp\left(-\frac{c}{2}\epsilon^2 n\right), \end{aligned}$$

which does not depend on j.

For $A \in \mathcal{A}_n$, let $m(A) = \{x \in A : z \in A, z \preceq x \implies z = x\} \subseteq A$ be the associated set of minimal elements. Then $A = A' \iff m(A) = m(A')$ for $A, A' \in \mathcal{A}_n$, and so the number of upper sets in S_n equals the number of antichains. The size of a maximal antichain, which we denote by s_n, satisfies $s_n \geq 1$ and, by assumption (ii), $s_n \leq n^\gamma$. So if n is sufficiently large, $n^\gamma < n/2$ and

$$\#(\mathcal{A}_n) \leq \sum_{k=1}^{s_n} \binom{n}{k} \leq s_n \frac{n!}{(n-s_n)!\, s_n!} \leq \lceil n^\gamma \rceil \frac{n!}{(n - \lceil n^\gamma \rceil)!\, \lceil n^\gamma \rceil!}.$$

By Stirling's formula, the right hand side is asymptotically equivalent to

$$\begin{aligned}
& n^\gamma \frac{\sqrt{2\pi n}\,(n/e)^n}{\sqrt{2\pi(n-n^\gamma)}\,((n-n^\gamma)/e)^{n-n^\gamma}\sqrt{2\pi n^\gamma}\,(n^\gamma/e)^{n^\gamma}} \\
&= \frac{n^{-\gamma/2}}{\sqrt{2\pi(1-n^{\gamma-1})}} \frac{n^n}{(n-n^\gamma)^{n-n^\gamma} n^{\gamma n^\gamma}} \\
&= \frac{1}{\sqrt{2\pi(1-n^{\gamma-1})}} n^{-\gamma/2+n^\gamma(1-\gamma)} (1-n^{\gamma-1})^{n^\gamma - n} \\
&= \frac{1}{\sqrt{2\pi(1-n^{\gamma-1})}} \exp\left(\left(-\frac{\gamma}{2} + (1-\gamma)n^\gamma\right)\log n\right)(1-n^{\gamma-1})^{n^\gamma-n},
\end{aligned}$$

where the factor $(1-n^{\gamma-1})^{n^\gamma-n} = ((1-n^{\gamma-1})^{n^{1-\gamma}})^{-n^\gamma(1-n^{\gamma-1})}$ grows no faster than $\exp(n^\gamma)$, because $(1-1/x)^x \leq \exp(-1)$ for $x \geq 1$. Combining these results, we see that for n sufficiently large, $\#(\mathcal{A}_n) \leq \exp(C_1\, n^\gamma \log n)$, where C_1 is a constant that depends on γ. Hence, for n sufficiently large,

$$\begin{aligned}
\mathbb{P}\left(\sup_{y\in\mathbb{R}} \left(\hat{F}_{X_j}(y) - F_{X_j}(y)\right) \geq \frac{3\epsilon}{4}\right) &\leq \#(\mathcal{A}_n)\, M \exp\left(-\frac{c}{2}\epsilon^2 n\right) \\
&\leq M \exp\left(-\frac{c}{2}\epsilon^2 n + C_1 n^\gamma \log n\right) \\
&\leq M \exp\left(-C_2 n\right)
\end{aligned}$$

for some strictly positive constant C_2 that depends on γ. This upper bound does not depend on j, so

$$\mathbb{P}\left(\sup_{j: X_j \in (1/m,(m-1)/m]^p,\, y\in\mathbb{R}} \left(\hat{F}_{X_j}(y) - F_{X_j}(y)\right) \geq \frac{3\epsilon}{4}\right) \leq M \exp\left(-C_2 n\right) n$$

vanishes as $n \to \infty$. Analogous arguments yield the bound with F_{X_j} and $\hat{F}_{X_j}$ interchanged, which establishes (35) and completes the proof. □

As noted, the broad applicability of Theorem 2.3 rests on a powerful combinatorial result of Brightwell (1992, Corollary 2), which enables us to deduce consistency without having to check complex regularity conditions of the type in Robertson and

Wright (1975). The size of a maximal antichain also appears in the derivation of risk bounds for multiple isotonic regression for the mean in Han et al. (2019, p. 2447, and Lemma 4 in their Supplementary Material). Their Lemma 4 gives an asymptotic lower bound of $n^{1-1/d}$ for the size of a maximal antichain among n independent and identically distributed covariates $X_1, \dots, X_n \in \mathbb{R}^d$ with any Lebesgue density bounded from above, and might in fact also be derived from Brightwell (1992, Corollary 2). An intuitive explanation for the lower bound $n^{1-1/d}$ is that any distribution with bounded Lebesgue density can be restricted to a fixed subset where the density is positive, and asymptotically the maximum antichain of $X_1, \dots, X_n$ within this subset behaves as if $X_i \sim \text{Unif}[0,1]^d$, regardless of the dependence structure. This is an interesting result, because if the speed of convergence hinges on the maximal size of an antichain, as our proof and results in Han et al. (2019) suggest, then it may not be possible to improve the speed of convergence by assuming positively correlated components. Therefore, we believe that positive dependency between the components of the covariate vector does not affect convergence rates, though clearly it may have positive effects in finite sample settings.

C. Proofs for Sections 3.2 and 3.3

Proof of Proposition 3.2. Denote the CDF corresponding to the empirical distribution of $x_1, \dots, x_d$ and of $x'_1, \dots, x'_d$ by F and G, respectively. For part i), assume that $x_{(i)} \leq x'_{(i)}$ for $i = 1, \dots, d$, and let $z \in \mathbb{R}$. Then,

$$F(z) = \frac{\#\{i : x_{(i)} \leq z\}}{d} \geq \frac{\#\{i : x'_{(i)} \leq z\}}{d} = G(y),$$

hence F is smaller than G in the usual stochastic order. Conversely, if F is smaller then G, by choosing $z = x'_{(k)}$, $k = 1, \dots, d$, we obtain

$$\frac{\#\{i : x_{(i)} \leq x'_{(k)}\}}{d} = F(x'_{(k)}) \geq G(x'_{(k)}) = \frac{\#\{i : x'_{(i)} \leq x'_{(k)}\}}{d}.$$

By definition of the k-th order statistic, we know that $\#\{i : x'_{(i)} \leq x'_{(k)}\} \geq k$ (with equality if the x'_i are distinct). Therefore, $\#\{i : x_{(i)} \leq x'_{(k)}\} \geq k$. This can only be true if $x_{(k)} \leq x'_{(k)}$.

Concerning part ii), we can assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_d$, otherwise we reorder the pairs (x_i, y_i). Now apply part i): We know that $x_1 \leq x'_1$ and $x'_{(1)} \geq x_j$ for some j. But the components of x are sorted, hence $x'_{(1)} \geq x_j \geq x_1 = x_{(1)}$, and also $x'_1 \geq x'_{(1)} \geq x_j$. So we can think of the positions of x'_1 and $x'_{(1)}$ in x' to be exchanged, without violating the condition $x \preceq x'$. Now we can ignore the pair $(x'_1, x'_{(1)})$ and proceed in the same way for remaining components $(x_i)_{i=2}^d$ and $(x'_i)_{i=2}^d$.

For the proof of part iii), assume the opposite, i.e., $x_i \geq x'_i$ for $i = 1, \dots, d$. By ii), we know that $x \succeq_{\text{st}} x'$. By assumption $x \preceq_{\text{st}} x'$, hence x and x' are permutations of each other. But then either $x = x'$, or x and x' cannot be comparable in the componentwise order.

The last part is immediate from part i). □

Proof of Proposition 3.3. Part i) is a consequence of Theorem 4.A.3 of Shaked and Shanthikumar (2007). Part ii) follows from part i) and Proposition 3.2 i). For part iii) note that the Gini mean difference has the equivalent formula

$$g(x) = \frac{2}{d(d-1)} \sum_{i=1}^{d} x_{(i)}(2i - d - 1),$$

which can be rewritten as

$$g(x) = \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x_{(j)} - 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i.$$

Part i) implies that

$$\begin{aligned} g(x') + 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i' &= \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x'_{(j)} \\ &\geq \frac{4}{d(d-1)} \sum_{i=1}^{d} \sum_{j=i}^{d} x_{(j)} = g(x) + 2\frac{d+1}{d(d-1)} \sum_{i=1}^{d} x_i. \square \end{aligned}$$

D. Large sample equivalence of CRPS and L_2 measures

Here we show that the difference between the mean CRPS for the distributional regression method at hand and the mean CRPS for the ideal forecast is large sample equivalent to the (squared) L_2 error in conditional distribution estimation. This relates the CRPS, as introduced by Matheson and Winkler (1976) and arguably the most prevalent measure of predictive performance in distributional forecasting (Gneiting and Raftery, 2007), to traditional L_p measures, as used by Hall et al. (1999) and Spady and Stouli (2018) in the evaluation of conditional cumulative distribution function (CDF) estimation.

Specifically, suppose that the random variates $(x_1, y_1), \ldots, (x_m, y_m)$ are independent identically distributed from a population with bivariate law G. Let $F(Y|X)$ be any estimate of the conditional distributions of Y given X, and for $i = 1, \ldots, m$ let $F_i = F(Y \mid X = x_i)$ and $G_i = G(Y \mid X = x_i)$ denote the respective conditional CDFs for $x_1, \ldots, x_m$. Subject to the conditions of the bivariate strong law of large numbers,

$$\bar{\mathrm{S}}_m^F = \frac{1}{m} \sum_{i=1}^{m} \mathrm{CRPS}(F_i, y_i) \to \mathbb{E}_{(X,Y)\sim G} \left[\mathrm{CRPS}(F(Y|X), Y)\right]$$

and

$$\bar{\mathrm{S}}^G_m = \frac{1}{m}\sum_{i=1}^{m} \mathrm{CRPS}(G_i, y_i) \to \mathbb{E}_{(X,Y)\sim G}\left[\mathrm{CRPS}(G(Y|X), Y)\right]$$

almost surely. Therefore, subject to the conditions of the strong law and Fubini's theorem,

$$\begin{aligned}\bar{\mathrm{S}}^F_m - \bar{\mathrm{S}}^G_m &\to \mathbb{E}_{X\sim G}\,\mathbb{E}_{Y\sim G(Y|X)}\left[\mathrm{CRPS}(F(Y|X), Y) - \mathrm{CRPS}(G(Y|X), Y) \mid X\right] \\ &= \mathbb{E}_{X\sim G}\left[\int_{-\infty}^{\infty} (F(y \mid X) - G(y \mid X))^2 \,\mathrm{d}y\right] \\ &= \mathbb{E}_{X\sim G}\left[L_2^2\left(F(\cdot \mid X), G(\cdot \mid X)\right)\right]\end{aligned}$$

almost surely, where the first equality uses the analytic form of the CRPS divergence (Gneiting and Raftery, 2007, p. 367).

In the context of the simulation study in Section 4, the above setting corresponds to a single of the 500 Monte Carlo replicates, where F is an estimate on a training set of size n, and performance is evaluated on an independent test sample of size $m = 5\,000$. The large sample arguments remain valid when scores are averaged across Monte Carlo replicates.

E. Additional tables and figures

Table 3 provides implementation details for the distributional regression methods in the simulation study in Section 4.

Figure 7 assesses the probabilistic calibration of the postprocessing methods for precipitation forecasts in the case study in Section 5. Similarly, Figure 8 shows reliability diagrams for the postprocessed probability of precipitation forecasts, using the CORP approach developed by Dimitriadis et al. (2021).

Table 3. Implementation details for the distributional regression methods in the simulation study. We list the R packages and the specific functions used for estimation and prediction, along with choices for tuning parameters. For nonparametric kernel smoothing (NP) we use Gaussian kernels in (23), (24), and (25) and the `liracine` kernel in the Poisson scenario (26). To fit semiparametric quantile regression (SQR) and conditional transformation models (TRAM) we employ cubic B-splines with interior knots from 2 to 8 in steps of 2 and boundary knots 0 and 10 (`bs(x, ...)`). For TRAM, we use continuous outcome logistic regression (`Colr`) for (23), (24), and (25), and ordered categorical regression (`Polr`) in (26). For further detail, see the code, which is available at `https://github.com/AlexanderHenzi/isodistrreg`.

Package	
NP	`np` (Hayfield and Racine, 2008)
SQR	`quantreg` (Koenker, 2020)
TRAM	`tram` (Hothorn, 2020)
QRF	`grf` (Tibshirani et al., 2020)
IDR	`isodistrreg`
IDR_{sbg}	`isodistrreg`
Estimation	
NP	`npcdistbw(nmulti = 4, oykertype = "liracine", bwtype = adaptive_nn")`
SQR	`rq(y~., data = cbind(y = y, bs(x, ...)), tau = seq(0.005,0.995,0.001))`
TRAM	`Colr/Polr(y~., data = cbind(y = y, bs(x, ...)))`
QRF	`quantile_forest(min.node.size = 40, quantiles = seq(0.01,0.99,0.01))`
IDR	`idr()`
IDR_{sbg}	`idrbag(b = 100, digits = 6, p = 1/2)`
Prediction	
NP	`npcdist(eydat = grid)`
SQR	`predict.rqs()`
TRAM	`predict.ctm(K = 5000, type = "distribution")`
QRF	`predict.quantile_forest(quantiles = seq(0.005,0.995,0.001))`
IDR	`predict.idrfit(digits = 6)`
IDR_{sbg}	`idrbag(b = 100, digits = 6, p = 1/2)`

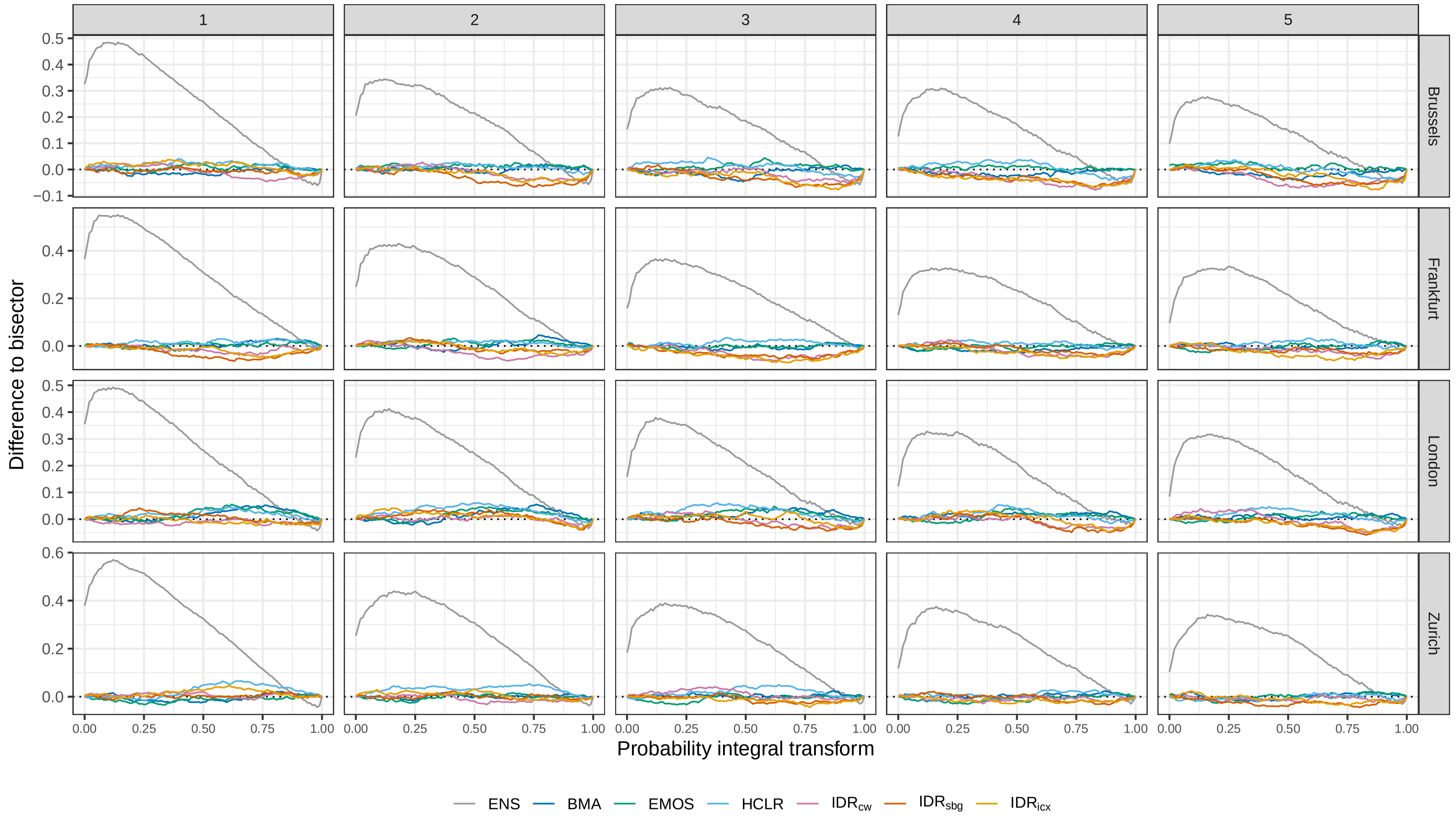


Fig. 7. Difference of the ECDF of the PIT to the bisector (positive: above bisector) for raw and postprocessed ensemble forecasts of 24-hour accumulated precipitation at prediction horizons of 1, 2, 3, 4 and 5 days, for the test period.

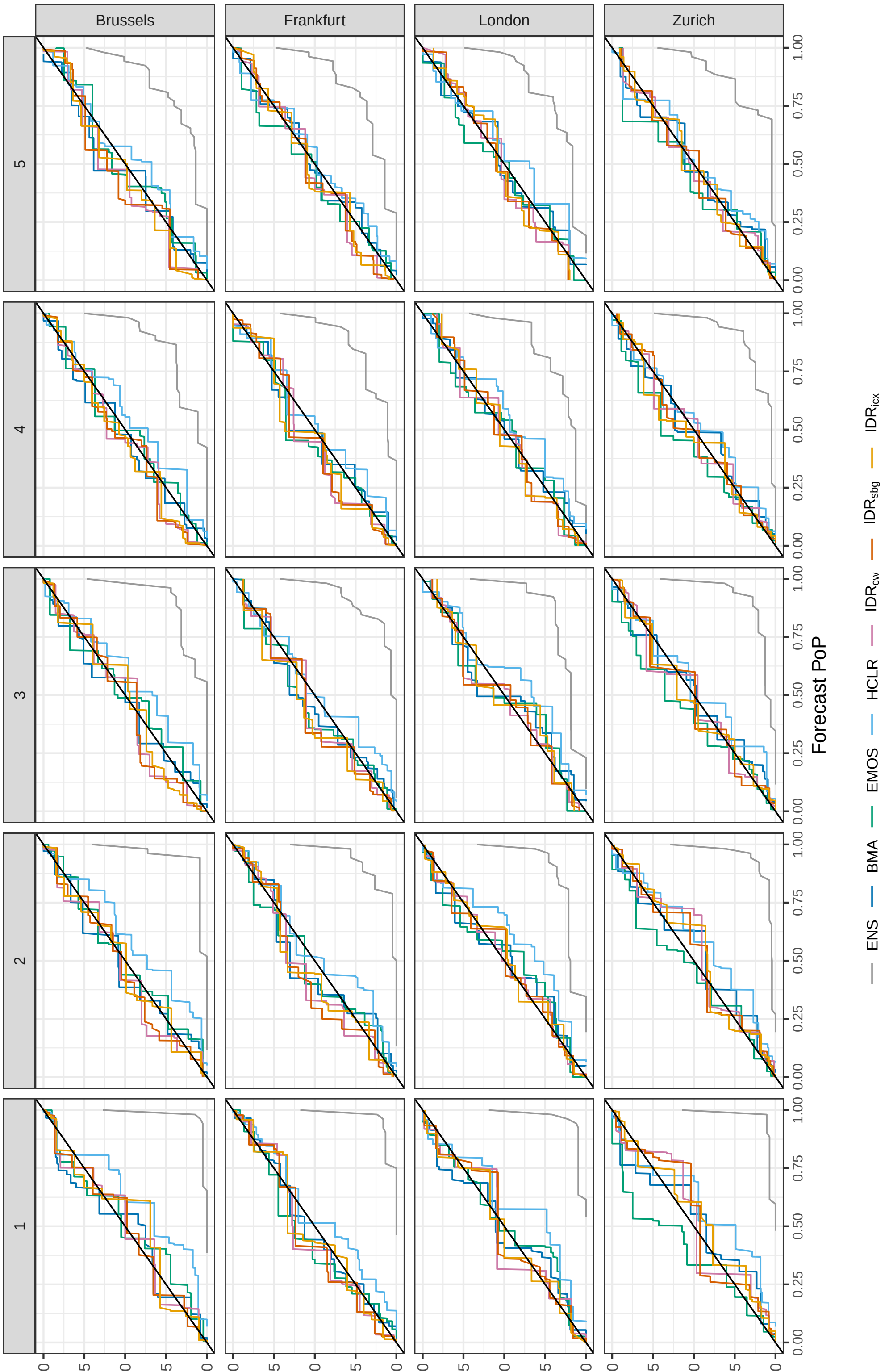


Fig. 8. CORP reliability diagrams for probability of precipitation forecasts at prediction horizons of 1, 2, 3, 4 and 5 days, for the test period.

2.2 Accelerating the Pool-Adjacent-Violators Algorithm for isotonic distributional regression

The content of this section is published as

HENZI, A., MÖSCHING, A. and DÜMBGEN, L. (2022+). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodology and Computing in Applied Probability*, to appear.

The version in this thesis is the arXiv preprint (identifier *arXiv:2006.05527*).

Accelerating the Pool-Adjacent-Violators Algorithm for Isotonic Distributional Regression

Alexander Henzi*[1], Alexandre Mösching†[2], and Lutz Dümbgen‡[1]

[1]University of Bern, Switzerland
[2]Georg-August-University of Göttingen, Germany

December 2, 2021

Abstract

In the context of estimating stochastically ordered distribution functions, the pool-adjacent-violators algorithm (PAVA) can be modified such that the computation times are reduced substantially. This is achieved by studying the dependence of antitonic weighted least squares fits on the response vector to be approximated.

Keywords: Monotone regression, sequential computation, weighted least squares

AMS 2000 subject classifications: 62G08, 62G30, 62-08

1 Introduction

Let $\mathcal{X}$ be a set equipped with a binary relation $\preceq$, for instance, some partial order. The general problem is as follows: For $m \geq 2$ pairs $(x_1, z_1), \ldots, (x_m, z_m) \in \mathcal{X} \times \mathbb{R}$ and weights $w_1, \ldots, w_m > 0$, let

$$A(\boldsymbol{z}) \ := \ \operatorname*{arg\,min}_{\boldsymbol{f} \in \mathbb{R}^m_{\downarrow,\boldsymbol{x}}} \sum_{j=1}^m w_j (z_j - f_j)^2, \tag{1}$$

where

$$\mathbb{R}^m_{\downarrow,\boldsymbol{x}} \ := \ \{\boldsymbol{f} \in \mathbb{R}^m : x_i \preceq x_j \text{ implies that } f_i \geq f_j\}.$$

Suppose that $\boldsymbol{z}^{(0)}, \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$ are vectors in $\mathbb{R}^m$ such that for $1 \leq t \leq n$, the two vectors $\boldsymbol{z}^{(t-1)}$ and $\boldsymbol{z}^{(t)}$ differ only in a few components, and our task is to compute all antitonic (i.e. monotone decreasing) approximations $A(\boldsymbol{z}^{(0)}), A(\boldsymbol{z}^{(1)}), \ldots, A(\boldsymbol{z}^{(n)})$. We show that $A(\boldsymbol{z}^{(t)})$ can be computed efficiently, provided we know already $A(\boldsymbol{z}^{(t-1)})$. Briefly speaking, this is achieved by noticing that $A(\boldsymbol{z}^{(t-1)})$ and $A(\boldsymbol{z}^{(t)})$ share some identical components, and that the remaining components of $A(\boldsymbol{z}^{(t)})$ can be determined directly from $A(\boldsymbol{z}^{(t-1)})$ and $\boldsymbol{z}^{(t)}$ with only a few operations.

[*] alexander.henzi@stat.unibe.ch
[†] alexandre.moesching@uni-goettingen.de
[‡] duembgen@stat.unibe.ch

The efficient computation of a sequence of antitonic approximations appears naturally in the context of isotonic distributional regression, see Henzi et al. (2021), Mösching and Dümbgen (2020) and Jordan et al. (2021). There, one observes random pairs (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ in $\mathcal{X} \times \mathbb{R}$ such that, conditional on $(X_i)_{i=1}^n$, the random variables $Y_1, Y_2, \ldots, Y_n$ are independent with distribution functions $F_{X_1}, F_{X_2}, \ldots, F_{X_n}$, where $(F_x)_{x \in \mathcal{X}}$ is an unknown family of distribution functions. Then the goal is to estimate the latter family under the sole assumption that $F_x \geq F_{x'}$ pointwise whenever $x \preceq x'$. This notion of ordering of distributions is known as stochastic ordering, or first order stochastic dominance. This isotonic distributional regression leads to the aforementioned least squares problem, where $x_1, \ldots, x_m$ denote the different elements of $\{X_1, X_2, \ldots, X_n\}$, and $\boldsymbol{z}^{(t)}$ has components

$$z_j^{(t)} \ := \ w_j^{-1} \sum_{i \,:\, X_i = x_j} 1_{[Y_i \leq Y_{(t)}]}$$

with $w_j := \#\{i \leq n : X_i = x_j\}$, $Y_{(0)} := -\infty$ and $Y_{(t)}$ is the t-th order statistic of the sample $\{Y_1, Y_2, \ldots, Y_n\}$.

Section 2 provides some facts about monotone least squares which are useful for the present task. For a complete account and derivations, we refer to Barlow et al. (1972) and Robertson et al. (1988). Then it is shown in Section 3 how to turn this into an efficient computation scheme in case of a total order $\preceq$. Finally, we discuss the specific application to isotonic distributional regression, and provide numerical experiments which show that computation times of the naive approach are decreased substantially with our procedure.

2 Some facts about antitonic least squares estimation

Since the sum on the right hand side of (1) is a strictly convex and coercive function of $\boldsymbol{f} \in \mathbb{R}^m$, and since $\mathbb{R}^m_{\downarrow,\boldsymbol{x}}$ is a closed and convex set, $A(\boldsymbol{z})$ is well-defined. It possesses several well-known characterizations, two of which are particularly useful for our considerations.

The first characterization uses local weighted averages. Let us first introduce some notations. In this article, upper, lower and level sets are seen as subsets of $\{1, \ldots, m\}$ inheriting the structure of $(\mathcal{X}, \preceq)$. More precisely, a set $U \subset \{1, \ldots, m\}$ is an upper set if $i \in U$ and $x_i \preceq x_j$ imply that $j \in U$. A set $L \subset \{1, \ldots, m\}$ is a lower set if $j \in L$ and $x_i \preceq x_j$ imply that $i \in L$. The families of all upper and all lower sets are denoted by $\mathcal{U}$ and $\mathcal{L}$, respectively. For a non-empty set $S \subset \{1, \ldots, m\}$, its weight and the weighted average of $\boldsymbol{z}$ over S are respectively defined as

$$w_S \ := \ \sum_{j \in S} w_j \quad \text{and} \quad M_S(\boldsymbol{z}) \ := \ w_S^{-1} \sum_{j \in S} w_j z_j.$$

Characterization I. For any index $1 \leq j \leq m$,

$$A_j(\boldsymbol{z}) \ = \ \min_{U \in \mathcal{U} :\, j \in U} \ \max_{L \in \mathcal{L} :\, j \in L} \ M_{U \cap L}(\boldsymbol{z}) \ = \ \max_{L \in \mathcal{L} :\, j \in L} \ \min_{U \in \mathcal{U} :\, j \in U} \ M_{L \cap U}(\boldsymbol{z}).$$

For all vectors $\boldsymbol{f} \in \mathbb{R}^m$, numbers $\xi \in \mathbb{R}$ and relations $\ltimes$ in $\{<, \leq, =, \geq, >\}$, let

$$[\boldsymbol{f} \ltimes \xi] \ := \ \{j \in \{1, \ldots, m\} : f_j \ltimes \xi\}.$$

For example, the family of sets $[\boldsymbol{f} = \xi]$ indexed by $\xi \in \{f_1, \ldots, f_m\}$ yields a partition of $\{1, \ldots, m\}$ such that two indices i and j belong to the same block if and only if $f_i = f_j$. In case of $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow,\boldsymbol{x}}$, $[\boldsymbol{f} < \xi]$ and $[\boldsymbol{f} \le \xi]$ are upper sets, whereas $[\boldsymbol{f} > \xi]$ and $[\boldsymbol{f} \ge \xi]$ are lower sets.

Characterization II. A vector $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow,\boldsymbol{x}}$ equals $A(\boldsymbol{z})$ if and only if for any number $\xi \in \{f_1, \ldots, f_m\}$,

$$\begin{aligned} M_{U\cap[\boldsymbol{f}=\xi]}(\boldsymbol{z}) &\ \ge\ \xi \quad \text{for } U \in \mathcal{U} \text{ such that } U \cap [\boldsymbol{f} = \xi] \neq \emptyset, \qquad (2)\\ M_{L\cap[\boldsymbol{f}=\xi]}(\boldsymbol{z}) &\ \le\ \xi \quad \text{for } L \in \mathcal{L} \text{ such that } L \cap [\boldsymbol{f} = \xi] \neq \emptyset. \qquad (3)\end{aligned}$$

In particular, $\xi = M_{[\boldsymbol{f}=\xi]}(\boldsymbol{z})$.

One possible reference for Characterizations I and II is Domínguez-Menchero and González-Rodríguez (2007). They treat the case of a quasi-order $\preceq$ and more general target functions $\sum_{j=1}^m h_j(f_j)$ to be minimized over $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow,\boldsymbol{x}}$. For the present setting with an arbitrary binary relation $\preceq$ and weighted least squares, a relatively short and self-contained derivation of these two characterizations is available from the authors upon request.

The next lemma summarizes some facts about changes in $A(\boldsymbol{z})$ if some components of $\boldsymbol{z}$ are increased.

Lemma 2.1. *Let $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^m$ such that $\tilde{\boldsymbol{z}} \ge \boldsymbol{z}$ component-wise. Then the following conclusions hold true for $\boldsymbol{f} := A(\boldsymbol{z})$, $\tilde{\boldsymbol{f}} := A(\tilde{\boldsymbol{z}})$ and $K := \{k : \tilde{z}_k > z_k\}$:*

(i) *$\boldsymbol{f} \le \tilde{\boldsymbol{f}}$ component-wise.*

(ii) *$\tilde{f}_i = f_i$ whenever $f_i < \min_{k\in K} f_k$.*

(iii) *$\tilde{f}_i = f_i$ whenever $\tilde{f}_i > \max_{k\in K} \tilde{f}_k$.*

(iv) *$\tilde{f}_i = \tilde{f}_j$ whenever $f_i = f_j$ and $x_i, x_j \preceq x_k$ for all $k \in K$.*

Figure 1 illustrates the statements of Lemma 2.1 on $\mathbb{R}^2$ equipped with the componentwise order in case of $K = \{j_o\}$. The colored areas show level sets of a hypothetical antitonic regression $\boldsymbol{f}$, and x_{j_o} is the point where $\tilde{z}_{j_o} > z_{j_o}$. By part (ii) of Lemma 2.1, we know that $\tilde{f}_i = f_i$ if $f_i < f_{j_o}$, so the values of $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ are equal on the orange and yellow regions in the top right corner, which is indicated by saturated colors. Furthermore, when passing from $\boldsymbol{z}$ to $\tilde{\boldsymbol{z}}$, the slightly transparent pink, blue and green level sets on the bottom left (including the point x_{j_o}) can only be merged, but never be split. This follows from part (iv) of Lemma 2.1. Finally, for all points in the faded pink, blue and green areas, there is no statement about the behavior of the antitonic regression when passing from $\boldsymbol{z}$ to $\tilde{\boldsymbol{z}}$.

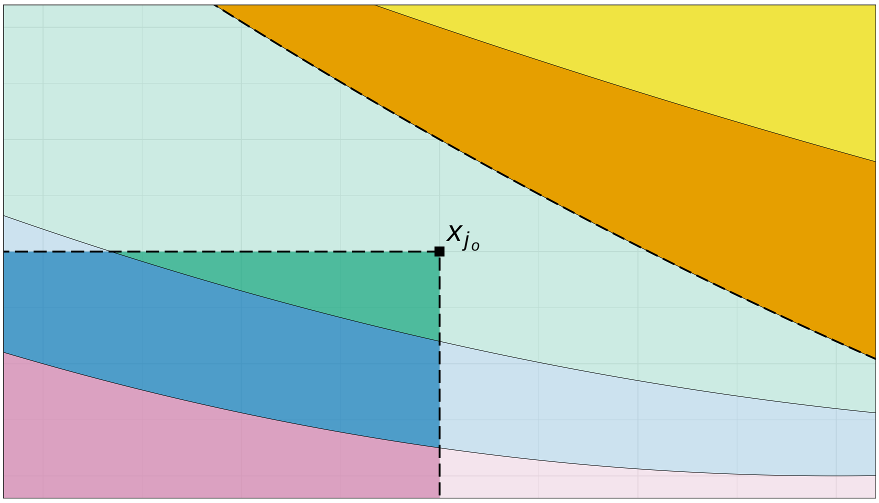


Figure 1: Illustration of the statements of Lemma 2.1 on $\mathbb{R}^2$.

Proof of Lemma 2.1. Part (i) is a direct consequence of Characterization I.

As to part (ii), if $f_i < \min_{k\in K} f_k$, then $K \subset [\boldsymbol{f} > f_i]$, whence

$$\begin{aligned}
\tilde{f}_i &= \max_{L\in\mathcal{L}:\, i\in L}\ \min_{U\in\mathcal{U}:\, i\in U} M_{L\cap U}(\tilde{\boldsymbol{z}}) && \text{(Char. I)}\\
&\le \max_{L\in\mathcal{L}:\, i\in L} M_{L\cap[\boldsymbol{f}\le f_i]}(\tilde{\boldsymbol{z}}) && (i \in [\boldsymbol{f} \le f_i] \in \mathcal{U})\\
&= \max_{L\in\mathcal{L}:\, i\in L} M_{L\cap[\boldsymbol{f}\le f_i]}(\boldsymbol{z}) && (K \cap [\boldsymbol{f} \le f_i] = \emptyset)\\
&= \max_{L\in\mathcal{L}:\, i\in L} \sum_{\xi\le f_i:\, L\cap[\boldsymbol{f}=\xi]\ne\emptyset} \frac{w_{L\cap[\boldsymbol{f}=\xi]}}{w_{L\cap[\boldsymbol{f}\le f_i]}} M_{L\cap[\boldsymbol{f}=\xi]}(\boldsymbol{z})\\
&\le \max_{L\in\mathcal{L}:\, i\in L} \sum_{\xi\le f_i:\, L\cap[\boldsymbol{f}=\xi]\ne\emptyset} \frac{w_{L\cap[\boldsymbol{f}=\xi]}}{w_{L\cap[\boldsymbol{f}\le f_i]}}\, \xi && \text{(Char. II)}\\
&\le f_i.
\end{aligned}$$

This inequality and part (i) show that $\tilde{f}_i = f_i$.

Part (iii) is proved analogously. If $\tilde{f}_i > \max_{k\in K} \tilde{f}_k$, then $K \subset [\tilde{\boldsymbol{f}} < \tilde{f}_i]$, whence

$$\begin{aligned}
f_i &= \min_{U\in\mathcal{U}:\, i\in U}\ \max_{L\in\mathcal{L}:\, i\in L} M_{U\cap L}(\boldsymbol{z}) && \text{(Char. I)}\\
&\ge \min_{U\in\mathcal{U}:\, i\in U} M_{U\cap[\tilde{\boldsymbol{f}}\ge \tilde{f}_i]}(\boldsymbol{z}) && (i \in [\tilde{\boldsymbol{f}} \ge \tilde{f}_i] \in \mathcal{L})\\
&= \min_{U\in\mathcal{U}:\, i\in U} M_{U\cap[\tilde{\boldsymbol{f}}\ge \tilde{f}_i]}(\tilde{\boldsymbol{z}}) && (K \cap [\tilde{\boldsymbol{f}} \ge \tilde{f}_i] = \emptyset)\\
&= \min_{U\in\mathcal{U}:\, i\in U} \sum_{\xi\ge \tilde{f}_i:\, U\cap[\tilde{\boldsymbol{f}}=\xi]\ne\emptyset} \frac{w_{U\cap[\tilde{\boldsymbol{f}}=\xi]}}{w_{U\cap[\tilde{\boldsymbol{f}}\ge \tilde{f}_i]}} M_{U\cap[\tilde{\boldsymbol{f}}=\xi]}(\tilde{\boldsymbol{z}})\\
&\ge \min_{U\in\mathcal{U}:\, i\in U} \sum_{\xi\ge \tilde{f}_i:\, U\cap[\tilde{\boldsymbol{f}}=\xi]\ne\emptyset} \frac{w_{U\cap[\tilde{\boldsymbol{f}}=\xi]}}{w_{U\cap[\tilde{\boldsymbol{f}}\le \tilde{f}_i]}}\, \xi && \text{(Char. II)}\\
&\ge \tilde{f}_i.
\end{aligned}$$

This inequality and part (i) show that $\tilde{f}_i = f_i$.

Part (iv) follows directly from parts (i) and (iii). Let i and j be different indices such that $f_i = f_j$ and $x_i, x_j \preceq x_k$ for all $k \in K$. It follows from $\tilde{\boldsymbol{f}} \in \mathbb{R}^m_{\downarrow,\boldsymbol{x}}$ that $\tilde{f}_i, \tilde{f}_j \geq \max_{k\in K} \tilde{f}_k$. Consequently, if $\tilde{f}_j > \tilde{f}_i$, then $\tilde{f}_j > \max_{k\in K} \tilde{f}_k$, so parts (i) and (iii) would imply that

$$\tilde{f}_i \ \geq \ f_i = f_j \ = \ \tilde{f}_j,$$

contradicting $\tilde{f}_j > \tilde{f}_i$. □

The special case of a total order. If one replaces the preorder $\preceq$ by a total order $\leq$ on $\mathcal{X}$, as for example in the case of the usual total order on a subset of $\mathbb{R}$, the conclusions of Lemma 2.1 take a simpler form. In case of a total order, we assume that the covariates are ordered as follows

$$x_1 \ \leq \ x_2 \ \leq \ \cdots \ \leq \ x_m,$$

so that $i \leq j$ implies that $x_i \leq x_j$, while $x_i < x_j$ implies that $i < j$.

Corollary 2.2. *Let $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^m$ such that $\boldsymbol{z} \leq \tilde{\boldsymbol{z}}$ component-wise. Then the following conclusions hold true for $\boldsymbol{f} := A(\boldsymbol{z})$ and $\tilde{\boldsymbol{f}} := A(\tilde{\boldsymbol{z}})$:*

(i) *$\boldsymbol{f} \leq \tilde{\boldsymbol{f}}$ component-wise.*

(ii) *Let $k \in \{1, \ldots, m-1\}$ such that $f_k > f_{k+1}$ and $(\tilde{z}_j)_{j>k} = (z_j)_{j>k}$. Then*

$$(\tilde{f}_j)_{j>k} \ = \ (f_j)_{j>k}.$$

(iii) *Let $k \in \{2, \ldots, m\}$ such that $\tilde{f}_{k-1} > \tilde{f}_k$ and $(\tilde{z}_j)_{j<k} = (z_j)_{j<k}$. Then*

$$(\tilde{f}_j)_{j<k} \ = \ (f_j)_{j<k}.$$

(iv) *Let $k \in \{2, \ldots, m\}$ such that $(\tilde{z}_j)_{j<k} = (z_j)_{j<k}$. Then*

$$\{j < k : \tilde{f}_j > \tilde{f}_{j+1}\} \ \subset \ \{j < k : f_j > f_{j+1}\}.$$

3 A sequential algorithm for total orders

Lemma 2.1 is potentially useful to accelerate algorithms for isotonic distributional regression with arbitrary partial orders, possibly in conjunction with the recursive partitioning algorithm by Luss and Rosset (2014), but this will require additional research. Now we focus on improvements of the well-known pool-adjacent-violators algorithm (PAVA) for a total order.

3.1 General considerations

In what follows, we assume that $x_1 < \cdots < x_m$, so $\mathbb{R}^m_{\downarrow,\boldsymbol{x}}$ coincides with $\mathbb{R}^m_{\downarrow} = \{\boldsymbol{f} \in \mathbb{R}^m : f_1 \geq \cdots \geq f_m\}$. To understand the different variants of the PAVA, let us recall two basic facts about $A(\boldsymbol{z})$. Let $\mathcal{P} = (P_1, \ldots, P_d)$ be a partition of $\{1, \ldots, m\}$ into blocks $P_s = \{b_{s-1}+1, \ldots, b_s\}$, where $0 = b_0 < b_1 < \cdots < b_d = m$, and let $\mathbb{R}^m_{\mathcal{P}}$ be the set of vectors $\boldsymbol{f} \in \mathbb{R}^m$ such that $f_i = f_j$ whenever i, j belong to the same block of $\mathcal{P}$.

Fact 1. Let $r_1 > \cdots > r_d$ be the sorted elements of $\{A_i(\boldsymbol{z}) : 1 \le i \le m\}$, and let $\mathcal{P}$ consist of the blocks $P_s = \{i : A_i(\boldsymbol{z}) = r_s\}$. Then $r_s = M_{P_s}(\boldsymbol{z})$ for $1 \le s \le d$.

Fact 2. Suppose that $A(\boldsymbol{z}) \in \mathbb{R}^m_{\mathcal{P}}$ for a given partition $\mathcal{P}$ with $d \ge 2$ blocks. If $s \in \{1, \ldots, d-1\}$ such that $M_{P_s}(\boldsymbol{z}) \le M_{P_{s+1}}(\boldsymbol{z})$, then $A_i(\boldsymbol{z})$ is constant in $i \in P_s \cup P_{s+1}$. That means, one may replace $\mathcal{P}$ with a coarser partition by pooling P_s and P_{s+1} and still, $A(\boldsymbol{z}) \in \mathbb{R}^m_{\mathcal{P}}$.

Fact 1 is a direct consequence of Characterization II. To verify Fact 2, suppose that $\boldsymbol{f} \in \mathbb{R}^m_{\downarrow} \cap \mathbb{R}^m_{\mathcal{P}}$ such that $f_i = r_s$ for $i \in P_s$, $f_i = r_{s+1}$ for $i \in P_{s+1}$, and $r_s > r_{s+1}$. Now we show that $\boldsymbol{f}$ cannot be equal to $A(\boldsymbol{z})$. For $t \ge 0$ let $\boldsymbol{f}(t) \in \mathbb{R}^m_{\mathcal{P}}$ be given by

$$f_i(t) \;=\; f_i - 1_{[i \in P_s]} t w_{P_s}^{-1} + 1_{[i \in P_{s+1}]} t w_{P_{s+1}}^{-1}.$$

Then $\boldsymbol{f}(0) = \boldsymbol{f}$, and $\boldsymbol{f}(t) \in \mathbb{R}^m_{\downarrow}$ if $t \le (r_s - r_{s+1})/(w_{P_{s+1}}^{-1} + w_{P_s}^{-1})$. But

$$\frac{d}{dt}\Big|_{t=0} \sum_{i=1}^m w_i (f_i(t) - z_i)^2 \;=\; 2(r_{s+1} - r_s) - 2\big(M_{P_{s+1}}(\boldsymbol{z}) - M_{P_s}(\boldsymbol{z})\big) \;<\; 0,$$

so for sufficiently small $t > 0$, $\boldsymbol{f}(t) \in \mathbb{R}^m_{\downarrow}$ and is superior to $\boldsymbol{f}(0)$. Hence $\boldsymbol{f} \ne A(\boldsymbol{z})$.

Facts 1 and 2 indicate already a general PAV strategy to compute $A(\boldsymbol{z})$. One starts with the finest partition $\mathcal{P} = (\{1\}, \ldots, \{m\})$. As long as $\mathcal{P}$ contains two neighboring blocks P_s and P_{s+1} such that $M_{P_s}(\boldsymbol{z}) \ge M_{P_{s+1}}(\boldsymbol{z})$, the partition $\mathcal{P}$ is coarsened by replacing P_s and P_{s+1} with the block $P_s \cup P_{s+1}$.

Standard PAVA. Specifically, one works with three tuples: $\mathcal{P} = (P_1, \ldots, P_d)$ is a partition of $\{1, \ldots, b_d\}$ into blocks $P_s = \{b_{s-1}+1, \ldots, b_s\}$, where $0 = b_0 < b_1 < \cdots < b_d$. The number b_d is running from 1 to m, and the number $d \ge 1$ changes during the algorithm, too. The tuples $\mathcal{W} = (W_1, \ldots, W_d)$ and $\mathcal{M} = (M_1, \ldots, M_d)$ contain the corresponding weights $W_s = w_{P_s}$ and weighted means $M_s = M_{P_s}(\boldsymbol{z})$. Before increasing b_d, the tuples $\mathcal{P}$, $\mathcal{W}$ and $\mathcal{M}$ describe the minimizer of $\sum_{i=1}^{b_d} w_i (f_i - z_i)^2$ over all $\boldsymbol{f} \in \mathbb{R}^{b_d}_{\downarrow}$. Here is the complete algorithm:

Initialization: We set $\mathcal{P} \leftarrow (\{1\})$, $\mathcal{W} \leftarrow (w_1)$, $\mathcal{M} \leftarrow (z_1)$, and $d \leftarrow 1$.

Induction step: If $b_d < m$, we add a new block by setting

$$\mathcal{P} \;\leftarrow\; (\mathcal{P}, \{b_d + 1\}), \quad \mathcal{W} \;\leftarrow\; (\mathcal{W}, w_{b_d+1}), \quad \mathcal{M} \;\leftarrow\; (\mathcal{M}, z_{b_d+1}),$$

and $d \leftarrow d + 1$. Then, while $d > 1$ and $M_{d-1} \le M_d$, we pool the "violators" P_{d-1} and P_d by setting

$$\begin{aligned} \mathcal{P} \;&\leftarrow\; \big((P_j)_{j<d-1}, P_{d-1} \cup P_d\big), \\ \mathcal{M} \;&\leftarrow\; \Big((W_j)_{j<d-1}, \frac{W_{d-1}M_{d-1} + W_d M_d}{W_{d-1} + W_d}\Big), \\ \mathcal{W} \;&\leftarrow\; \big((W_j)_{j<d-1}, W_{d-1} + W_d\big), \end{aligned}$$

and $d \leftarrow d - 1$.

Finalization: Eventually, $\mathcal{P}$ is a partition of $\{1, \ldots, m\}$ into blocks such that $M_1 > \cdots > M_d$ and

$$A_j(\boldsymbol{z}) \;=\; M_s \quad \text{for } j \in P_s \text{ and } 1 \le s \le d.$$

Modified PAVA. In our specific applications of the PAVA, we are dealing with vectors $\boldsymbol{z}$ containing larger blocks $\{a, \ldots, b\}$ on which $i \mapsto z_i$ is constant. Indeed, in regression settings with continuously distributed covariates and responses, $\boldsymbol{z}$ will always be a $\{0,1\}$-valued vector. Then it is worthwhile to utilize fact 2 and modify the initialization as well as the very beginning of the induction step as follows:

For the initialization, we determine the largest index b_1 such that $z_1 = \cdots = z_{b_1}$ and the corresponding weight W_{P_1} with $P_1 = \{1, \ldots, b_1\}$. Then we set $\mathcal{P} \leftarrow (P_1)$, $\mathcal{W} \leftarrow (w_{P_1})$ and $\mathcal{M} \leftarrow (z_{b_1})$, where $P_1 = \{1, \ldots, b_1\}$.

At the beginning of the induction step, we determine the largest index $b_{d+1} > b_d$ such that $z_{b_d+1} = \cdots = z_{b_{d+1}}$ and the corresponding weight $W_{P_{d+1}}$ with $P_{d+1} = \{b_d + 1, \ldots, b_{d+1}\}$. Then we set $\mathcal{P} \leftarrow (\mathcal{P}, P_{d+1})$, $\mathcal{W} \leftarrow (\mathcal{W}, W_{P_{d+1}})$, $\mathcal{M} \leftarrow (\mathcal{M}, z_{b_{d+1}})$, and $d \leftarrow d + 1$.

Abridged PAVA. Suppose that we have computed $A(\boldsymbol{z})$ with corresponding tuples $\mathcal{P} = (P_1, \ldots, P_d)$, $\mathcal{W} = (W_1, \ldots, W_d)$ and $\mathcal{M} = (M_1, \ldots, M_d)$ via the PAVA. Now let $\tilde{\boldsymbol{z}} \in \mathbb{R}^m$ such that $\tilde{z}_{j_o} > z_{j_o}$ for one index $j_o \in \{1, \ldots, m\}$, while $(\tilde{z}_j)_{j \neq j_o} = (z_j)_{j \neq j_o}$. Let $j_o \in P_{s_o}$ with $s_o \in \{1, \ldots, d\}$. By parts (ii) and (iv) of Corollary 2.2, the partition corresponding to $A(\tilde{\boldsymbol{z}})$ will be a coarsening of the partition with the following blocks:

$$P_s \ \text{ for } 1 \leq s < s_o, \quad \{b_{s_o-1} + 1, \ldots, j_o\}, \quad \{j\} \ \text{ for } j_o < j \leq b_{s_o}, \quad P_s \ \text{ for } s_o < s \leq d.$$

Moreover, $A_i(\tilde{\boldsymbol{z}}) = A_i(\boldsymbol{z})$ for $i > b_{s_o}$. This allows us to compute $A(\tilde{\boldsymbol{z}})$ as follows, keeping copies of the auxiliary objects for $A(\boldsymbol{z})$ and indicating this with a superscript $\boldsymbol{z}$:

Initialization: We determine $s_o \in \{1, \ldots, d^{\boldsymbol{z}}\}$ such that $j_o \in P^{\boldsymbol{z}}_{s_o}$. Then we set

$$\begin{aligned} \mathcal{P} &\leftarrow \big((P^{\boldsymbol{z}}_s)_{s<s_o}, \{b^{\boldsymbol{z}}_{s_o-1} + 1, \ldots, j_o\}\big), \\ \mathcal{M} &\leftarrow \big((M^{\boldsymbol{z}}_s)_{s<s_o}, M_{P_{s_o}}(\tilde{\boldsymbol{z}})\big), \\ \mathcal{W} &\leftarrow \big((W^{\boldsymbol{z}}_s)_{s<s_o}, w_{P_{s_o}}\big) \end{aligned}$$

and $d \leftarrow s_o$. While $d > 1$ and $M_{d-1} \leq M_d$, we pool the violators P_{d-1} and P_d as in the induction step of PAVA. (This initialization is justified by part (iv) of Corollary 2.2.)

Induction step: If $j_o < b^{\boldsymbol{z}}_{s_o}$, we run the induction step of PAVA for b_d running from $j_o + 1$ to $b^{\boldsymbol{z}}_{s_o}$ with $\tilde{\boldsymbol{z}}$ in place of $\boldsymbol{z}$.

Finalization: If $b^{\boldsymbol{z}}_{s_o} < m$, we set

$$\begin{aligned} \mathcal{P} &\leftarrow \big(\mathcal{P}, (P^{\boldsymbol{z}}_s)_{s_o<s\leq d^{\boldsymbol{z}}}\big), \\ \mathcal{M} &\leftarrow \big(\mathcal{M}, (M^{\boldsymbol{z}}_s)_{s_o<s\leq d^{\boldsymbol{z}}}\big), \\ \mathcal{W} &\leftarrow \big(\mathcal{W}, (W^{\boldsymbol{z}}_s)_{s_o<s\leq d^{\boldsymbol{z}}}\big) \end{aligned}$$

and $d \leftarrow d + d^{\boldsymbol{z}} - s_o$. The new pair $(\mathcal{P}, \mathcal{M})$ yields the vector $A(\tilde{\boldsymbol{z}})$. This finalization is justified by part (ii) of Corollary 2.2.

Computational complexity. It directly follows from the algorithmic description that when $A(\boldsymbol{z})$ is available, the abridged PAVA for computing $A(\tilde{\boldsymbol{z}})$ requires not more operations than the standard PAVA. Its computational complexity is therefore at most of order $O(m)$ if $x_1, \ldots, x_m$ are already sorted. More precisely, the number of averaging operations

$\boldsymbol{z}$	1	3	2	0	-1	1	$1/2$	-1	1	
$b_d = 1$	1									$d = 1$
$b_d = 2$	1	3								$d = 2$
	2	2								$d = 1$
$b_d = 3$	2	2	2							$d = 2$
	2	2	2							$d = 1$
$b_d = 4$	2	2	2	0						$d = 2$
$b_d = 5$	2	2	2	0	-1					$d = 3$
$b_d = 6$	2	2	2	0	-1	1				$d = 4$
	2	2	2	0	0	0				$d = 3$
	2	2	2	0	0	0				$d = 2$
$b_d = 7$	2	2	2	0	0	0	$1/2$			$d = 3$
	2	2	2	$1/8$	$1/8$	$1/8$	$1/8$			$d = 2$
$b_d = 8$	2	2	2	$1/8$	$1/8$	$1/8$	$1/8$	-1		$d = 3$
$b_d = 9$	2	2	2	$1/8$	$1/8$	$1/8$	$1/8$	-1	1	$d = 4$
	2	2	2	$1/8$	$1/8$	$1/8$	$1/8$	0	0	$d = 3$

Table 1: Running the PAVA for a vector $\boldsymbol{z}$.

in the abridged PAVA is bounded from above by $d^{\boldsymbol{z}} + (b^{\boldsymbol{z}}_{s_o} - b^{\boldsymbol{z}}_{s_o - 1})$, where $d^{\boldsymbol{z}}$ is the partition size of the antitonic regression $A(\boldsymbol{z})$ and $b^{\boldsymbol{z}}_{s_o} - b^{\boldsymbol{z}}_{s_o-1}$ is the number of elements in the set $P^{\boldsymbol{z}}_{s_o}$ containing the index j_o where the value of $\boldsymbol{z}$ changes. In many practical applications this number is much smaller than m, but in the worst case it may equal exactly m; for example, let $w_i = 1$ and $z_i = m - i$ for $i = 1, \ldots, m$, $j_o = m$, and $\tilde{z}_m = m^2$.

Numerical example. We illustrate the previous procedures with two vectors $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^9$ and $\boldsymbol{w} = (1)_{j=1}^9$. Table 1 shows the main steps of the PAVA for $\boldsymbol{z}$. The first line shows the components of $\boldsymbol{z}$, the other lines contain the current candidate for $(f_j)_{j=1}^{b_d}$, where $\boldsymbol{f} = A(\boldsymbol{z})$ eventually, and the current partition $\mathcal{P}$ is indicated by extra vertical bars. Table 2 shows the abridged PAVA for two different vectors $\tilde{\boldsymbol{z}}$.

3.2 Application to isotonic distributional regression

Now we consider a regression framework similar to the one discussed in Mösching and Dümbgen (2020), Henzi et al. (2021) and Jordan et al. (2021). We observe pairs (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ consisting of numbers $X_i \in \mathcal{X}$ (covariate) and $Y_i \in \mathbb{R}$ (response), where $\mathcal{X}$ is a given real interval. Conditional on $(X_i)_{i=1}^n$, the observations $Y_1, Y_2, \ldots, Y_n$ are viewed as independent random variables such that for $x \in \mathcal{X}$ and $y \in \mathbb{R}$,

$$\mathbb{P}(Y_i \le y) \ = \ F_x(y) \quad \text{if } X_i = x.$$

Here $(F_x)_{x \in \mathcal{X}}$ is an unknown family of distribution functions. We only assume that $F_x(y)$ is non-increasing in $x \in \mathcal{X}$ for any fixed $y \in \mathbb{R}$. That means, the family $(F_x)_{x \in \mathcal{X}}$ is increasing with respect to stochastic order.

$\boldsymbol{z}$	1	3	2	0	-1	1	1/2	-1	1	
$A(\boldsymbol{z})$	2	2	2	1/8	1/8	1/8	1/8	0	0	
$\tilde{\boldsymbol{z}}$	1	3	2	0	**1**	1	1/2	-1	1	
$b_d = 5$	2	2	2	1/2	1/2					$d = 2$
$b_d = 6$	2	2	2	1/2	1/2	1				$d = 3$
	2	2	2	2/3	2/3	2/3				$d = 2$
$b_d = 7$	2	2	2	2/3	2/3	2/3	1/2			$d = 3$
$b_d = 9$	2	2	2	2/3	2/3	2/3	1/2	0	0	$d = 4$

$\boldsymbol{z}$	1	3	2	0	-1	1	1/2	-1	1	
$A(\boldsymbol{z})$	2	2	2	1/8	1/8	1/8	1/8	0	0	
$\tilde{\boldsymbol{z}}$	1	3	2	**2**	-1	1	1/2	-1	1	
$b_d = 4$	2	2	2	2						$d = 2$
	2	2	2	2						$d = 1$
$b_d = 5$	2	2	2	2	-1					$d = 2$
$b_d = 6$	2	2	2	2	-1	1				$d = 3$
	2	2	2	2	0	0				$d = 2$
$b_d = 7$	2	2	2	2	0	0	1/2			$d = 3$
	2	2	2	2	1/6	1/6	1/6			$d = 2$
$b_d = 9$	2	2	2	2	1/6	1/6	1/6	0	0	$d = 3$

Table 2: Running the abridged PAVA for two vectors $\tilde{\boldsymbol{z}} \approx \boldsymbol{z}$.

Variant of PAVA	mean (sd) of T_j		
Standard T_1	6.0394 (1.5257)	mean (sd) of T_1/T_j	
Modified T_2	1.7482 (0.4224)	3.4618 (0.3816)	mean (sd) of T_2/T_3
Abridged T_3	0.2080 (0.1052)	30.8308 (6.1209)	8.9012 (1.4469)

Table 3: Computation times in seconds and ratios of running times.

Let $x_1 < x_2 < \cdots < x_m$ be the elements of $\{X_1, X_2, \ldots, X_n\}$, and let

$$w_j \ := \ \#\{i : X_i = x_j\}, \quad 1 \le j \le m.$$

Then one can estimate $\boldsymbol{F}(y) := (F_{x_j}(y))_{j=1}^m$ by

$$\widehat{\boldsymbol{F}}(y) \ := \ A(\boldsymbol{z}(y)),$$

where $\boldsymbol{z}(y)$ has components

$$z_j(y) \ := \ w_j^{-1} \sum_{i:\, X_i = x_j} 1_{[Y_i \le y]}, \quad 1 \le j \le m.$$

Suppose we have rearranged the observations such that $Y_1 \le Y_2 \le \cdots \le Y_n$. Let $\boldsymbol{z}^{(0)} := \boldsymbol{0}$ and

$$\boldsymbol{z}^{(t)} \ := \ \Bigl(w_j^{-1} \sum_{i \le t:\, X_i = x_j} 1_{[Y_i \le Y_t]} \Bigr)_{j=1}^m$$

for $1 \le t \le n$. Note that $\boldsymbol{z}^{(t-1)}$ and $\boldsymbol{z}^{(t)}$ differ in precisely one component, and that

$$\boldsymbol{z}(y) \ = \ \begin{cases} \boldsymbol{z}^{(0)} & \text{if } y < Y_1, \\ \boldsymbol{z}^{(t)} & \text{if } Y_t \le y < Y_{t+1}, \ 1 \le t < n, \\ \boldsymbol{z}^{(n)} & \text{if } y \ge Y_n. \end{cases}$$

Thus it suffices to compute $A(\boldsymbol{z}^{(t)})$ for $t = 0, 1, \ldots, n$. But $A(\boldsymbol{z}^{(0)}) = \boldsymbol{0}$, $A(\boldsymbol{z}^{(n)}) = \boldsymbol{1}$, and for $1 \le t < n$, one may apply the abridged PAVA to the vectors $\boldsymbol{z} := \boldsymbol{z}^{(t-1)}$ and $\tilde{\boldsymbol{z}} := \boldsymbol{z}^{(t)}$. This leads to an efficient algorithm to compute all vectors $A(\boldsymbol{z}^{(t)})$, $0 \le t \le n$, if implemented properly.

Numerical experiment 1. We generated data sets with $n = 1000$ independent observation pairs (X_i, Y_i), $1 \le i \le n$, where X_i is uniformly distributed on $[0, 10]$ while $\mathcal{L}(Y_i \,|\, X_i = x)$ is the gamma distribution with shape parameter $\sqrt{x}$ and scale parameter $2 + (x-5)/\sqrt{2 + (x-5)^2}$. Figure 2 shows one such data set. In addition, one sees estimated β-quantile curves for levels $\beta \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, resulting from the estimator $\widehat{\boldsymbol{F}}$.

Now we simulated 1000 such data sets and measured the times T_1, T_2, T_3 for computing the estimator $\widehat{\boldsymbol{F}}$ via the standard, the modified and the abridged PAVA, respectively. Table 3 reports the sample means and standard deviations of these computation times in the 1000 simulations. In addition, one sees the averages and standard deviations of the ratios T_i/T_j, for $1 \le i < j \le 3$. It turned out that using the modified instead of

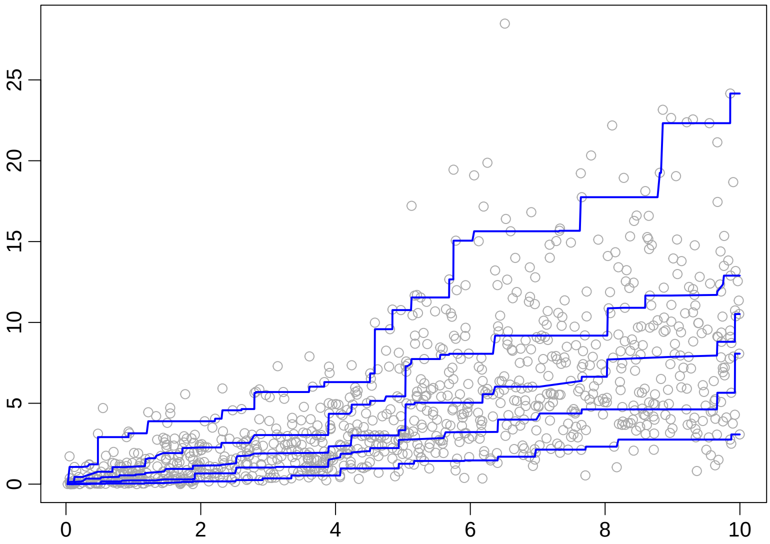


Figure 2: A data set with estimated quantile curves.

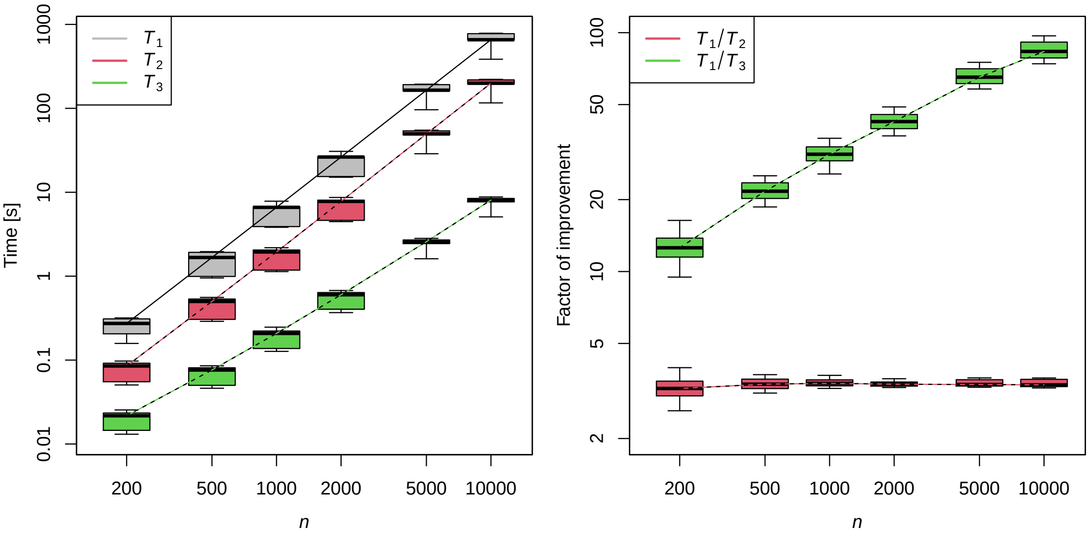


Figure 3: Boxplots of computation times and ratios of running times for varying sample sizes. The whiskers indicate the 10% and 90% sample quantiles. The other elements of the boxplots are standard. A logarithmic scale was used for both axes.

the standard PAVA reduced the computation time by a factor of 3.46 already. Using the abridged PAVA yielded a further improvement by a factor of 8.90.

Figure 3 displays the result of simulation experiments for sample sizes ranging from 200 to 10 000, where the data were generated using the procedure mentioned earlier. The simulations indicate that the improvement due to using modified instead of standard PAVA is almost constant in n, whereas the improvement due to abridged instead of modified PAVA increases with n. Presumably, the complexity of the abridged PAVA for computing the isotonic distributional regression remains quadratic in n. But our numerical experiments show that the constant is substantially smaller than the one resulting from applying the usual PAVA with complexity $O(n)$ for $n-1$ different levels of the response.

Numerical experiment 2. The goal of this experiment is to study the influence of the strength of the monotone association between X and Y on the efficiency gain of the abridged PAVA for isotonic distributional regression. The gains of abridged PAVA are expected to be milder when Y is independent of X, and to become larger as the monotone association strengthens. The reason behind it is that, while the standard PAVA proceeds independently of the stochastic order, the abridged PAVA relies on the index j_o indicating the component increasing in $\boldsymbol{z}(t-1)$ and on the nature of the partition corresponding to $A(\boldsymbol{z}(t-1))$, at a certain state $t \in \{1, \ldots, n\}$ of the procedure. If the monotone association is weak, then the partition corresponding to $A(\boldsymbol{z}(t-1))$ tends to contain fewer blocks in total and relatively large blocks in the middle of $\{1, \ldots, n\}$. If the index j_o happens to lie in a block containing many indices to the right of j_o, even the abridged PAVA will have to inspect all of these.

To demonstrate this claim, we simulated n independent bivariate Gaussian random vectors $(X, Y)^\top$ with correlation $\mathrm{Corr}(X, Y) = \rho \geq 0$. Note that the respective means and variances of X and Y have no influence on the results of the experiment. Indeed, the running times are invariant under strictly isotonic transformations of X and of Y. In particular, the simulations for $\rho = 0$ cover all situations in which X and Y are stochastically independent with continuous distribution functions. The stochastic order between $\mathcal{L}(Y|X = x_1)$ and $\mathcal{L}(Y|X = x_2)$ for $x_1 < x_2$ becomes stronger as the correlation $\rho \in [0, 1)$ increases, from an equality in distribution when $\rho = 0$ to a deterministic ordering when ρ approaches 1. Now, for sample sizes n ranging from 200 to 10 000 and for each correlation $\rho \in \{0, 0.5, 0.9\}$, the mean and standard deviation of the time ratio T_3/T_1 were estimated from 1 000 repetitions. The results are summarized in Table 4. As expected, the efficiency gain is smallest for $\rho = 0$. But even then, it is larger than 6 for $n \geq 200$ and larger than 9 for $n \geq 1\,000$.

Declarations

Acknowledgments. The authors are grateful to a reviewer for constructive comments.

Funding. This work was supported by Swiss National Science Foundation.

Conflict of interest/Competing interests. Not applicable.

Availability of data and material. Not applicable.

Code availability. R code is available at
https://github.com/AlexanderHenzi/abridgedPava.

n	$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
200	6.5337	(2.3496)	10.7581	(3.6704)	13.5695	(4.6390)
500	8.3029	(2.6393)	18.7010	(5.7806)	26.1813	(8.0763)
1 000	9.1351	(3.1800)	27.6290	(7.5007)	41.4116	(11.2161)
2 000	9.7559	(3.3532)	39.3180	(10.0337)	62.8382	(16.3293)
5 000	10.7495	(4.0525)	62.4600	(18.2002)	108.4198	(31.1414)
10 000	12.5190	(5.6193)	91.9084	(33.4657)	168.5030	(58.7712)

Table 4: Means (and standard deviations) of the factor of improvement T_3/T_1 for different correlation values ρ between X and Y and sample sizes n.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Domínguez-Menchero, J. S. and González-Rodríguez, G. (2007). Analyzing an extension of the isotonic regression problem. *Metrika* **66** 19–30.

Henzi, A., Ziegel, J. F. and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83** 963–993.

Jordan, A. I., Mühlemann, A. and Ziegel, J. F. (2021). Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals. *Annals of the Institute of Statistical Mathematics* to appear.

Luss, R. and Rosset, S. (2014). Generalized isotonic regression. *J. Comput. Graph. Statist.* **23** 192–210.
URL https://doi.org/10.1080/10618600.2012.741550

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electron. J. Stat.* **14** 24–49.
URL https://doi.org/10.1214/19-EJS1659

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order restricted statistical inference.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Ltd., Chichester.

2.3 Consistent estimation of distribution functions under increasing concave and convex stochastic ordering

The content of this section is published as arXiv preprint,

Henzi, A. (2021). Consistent estimation of distribution functions under increasing concave and convex stochastic ordering. *arXiv preprint arXiv:2105.03101.*

Consistent estimation of distribution functions under increasing concave and convex stochastic ordering

Alexander Henzi

University of Bern, Switzerland
alexander.henzi@stat.unibe.ch

March 21, 2022

Abstract

A random variable Y_1 is said to be smaller than Y_2 in the increasing concave stochastic order if $\mathbb{E}[\phi(Y_1)] \leq \mathbb{E}[\phi(Y_2)]$ for all increasing concave functions ϕ for which the expected values exist, and smaller than Y_2 in the increasing convex order if $\mathbb{E}[\psi(Y_1)] \leq \mathbb{E}[\psi(Y_2)]$ for all increasing convex ψ. This article develops nonparametric estimators for the conditional cumulative distribution functions $F_x(y) = \mathbb{P}(Y \leq y \mid X = x)$ of a response variable Y given a covariate X, solely under the assumption that the conditional distributions are increasing in x in the increasing concave or increasing convex order. Uniform consistency and rates of convergence are established both for the K-sample case $X \in \{1, \ldots, K\}$ and for continuously distributed X.

1 Introduction

The nonparametric estimation of distribution functions under stochastic order restrictions is a classical problem in statistics. It can be formulated very generally as the task to estimate the conditional distributions of a random variable Y given a covariate X, solely under the assumption that these distributions are increasing in a certain stochastic order. The classical and best understood order is first order stochastic dominance, requiring that the conditional cumulative distribution functions (CDFs) $F_x(y) = \mathbb{P}(Y \leq y \mid X = x)$ are decreasing in x for every fixed $y \in \mathbb{R}$. Brunk et al. (1966) were the first to consider this constrained estimation problem in the two sample case $X \in \{1, 2\}$. Almost 40 years later, El Barmi and Mukerjee (2005) have described an estimator for the K-sample case $X \in \{1, \ldots, K\}$, and again after more than a decade, Mösching and Dümbgen (2020b) extended it to continuously distributed X. In a further leap of complexity, Henzi et al. (2021c) have shown that consistent estimation under first order stochastic dominance is even possible with partially ordered covariates $X \in \mathbb{R}^d$. Stronger orders considered in the literature are the uniform stochastic ordering and the likelihood ratio order, see El Barmi and Mukerjee (2016) and Mösching and Dümbgen (2020a) and the references therein. A weaker constraint is stochastic precedence (Arcones et al., 2002), and a structurally different stochastic order is the peakedness order, where the variability of the conditional distributions of Y around a center is increasing in the covariate (Rojo and Batún-Cutz, 2007; El Barmi and Mukerjee, 2012; El Barmi and Wu, 2017).

So far, the main efforts in developing estimators under stochastic order restrictions have been focused on first order stochastic dominance and stronger orders, and consistency results in the case of continuously distributed X have only been derived for first order stochastic dominance. This is a limitation insofar as these orders require the conditional CDFs $F_x(y)$ to be decreasing in x for all fixed y. In particular, the CDFs for different values of x are not allowed to cross, which in practice often happens in the tails when the variability of Y increases (or decreases) with x. The purpose of this article is to develop consistent estimators under the increasing concave and increasing convex stochastic order, which are weaker orders applicable in situations where first order stochastic dominance is not appropriate. Estimation under the increasing concave order has been studied before by Rojo and El Barmi (2003) and El Barmi and Marchev (2009) in the two-sample case $X \in \{1, 2\}$. In this article, uniform consistency and rates of convergence are established both in the K-sample case and for continuously distributed X.

For two random variables Y_1 and Y_2 with finite expected values, Y_1 is said to be smaller in the increasing concave order than Y_2 if $\mathbb{E}[\phi(Y_1)] \leq \mathbb{E}[\phi(Y_2)]$ for all increasing concave functions ϕ for which the expectations exist. Similarly, Y_1 is smaller than Y_2 in the increasing convex order if $\mathbb{E}[\psi(Y_1)] \leq \mathbb{E}[\psi(Y_2)]$ for all increasing convex functions ψ, which is equivalent to $-Y_2$ being smaller than $-Y_1$ in the increasing concave order. In the following, these orders are abbreviated as $Y_1 \preceq_{\text{icv}} Y_2$ and $Y_1 \preceq_{\text{icx}} Y_2$, respectively, and $\preceq_{\text{icx}}$ and $\preceq_{\text{icv}}$ are both used as orders on random variables and on their CDFs. Another characterization (see Shaked and Shanthikumar, 2007, Chapter 4) for the increasing concave order is

$$\mathbb{E}[(y - Y_1)_+] = \int_{-\infty}^{y} F_1(t)\,dt \geq \int_{-\infty}^{y} F_2(t)\,dt = \mathbb{E}[(y - Y_2)_+],\ y \in \mathbb{R},$$

where $(z)_+ = \max(z, 0)$, and F_1 and F_2 are the CDFs of Y_1 and Y_2, respectively. For the increasing convex order, the analogous condition reads as

$$\mathbb{E}[(Y_1 - y)_+] = \int_{y}^{\infty} 1 - F_1(t)\,dt \leq \int_{y}^{\infty} 1 - F_2(t)\,dt = \mathbb{E}[(Y_2 - y)_+],\ y \in \mathbb{R}.$$

A useful sufficient condition for the increasing concave order is that the CDFs F_1 and F_2 cross at a single point y_0 with $F_1(y) \leq F_2(y)$ for $y \leq y_0$ and $F_1(y) \geq F_2(y)$ for $y \geq y_0$, or with the reverse inequalities for the CDFs in case of the increasing convex order. The increasing concave order is well-known in economics as second order stochastic dominance, with ‘second order’ referring to the fact that monotonicity is required for the integrated CDFs and not for the CDFs themselves. If Y_1 and Y_2 are portfolio returns, then $Y_1 \preceq_{\text{icv}} Y_2$ means that all individuals with increasing concave utility functions ϕ, i.e. all risk-averse utility maximizers, prefer Y_2 over Y_1. In the literature on finance and insurance, the increasing convex order appears under the name stop-loss order, a term introduced by Goovaerts et al. (1982) referring to the characterization $\mathbb{E}[(Y_1 - y)_+] \leq \mathbb{E}[(Y_2 - y)_+]$, which states that the expected stop-loss of Y_2 over any retention limit y exceeds the stop-loss of Y_1. This suggests using $\preceq_{\text{icx}}$ as an order for comparing risks. Detached from its economic interpretation, the increasing concave (convex) order can be seen as an order relation where the central tendency of Y increases with X, but variability decreases (increases). In the special case that the expected values $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_2]$ are equal, the orders reduce to the convex order (Shaked and Shanthikumar, 2007, Chapter 3), which is a prominent example for variability orders.

From a practical point of view, the estimation under stochastic order restrictions is nothing but another method for estimating conditional distributions. The paradigm has proven to be useful various applications where order restrictions appear naturally and estimators under stochastic order constraints are attractive because they usually require little tuning and no assumptions on the shape of the conditional distributions. For example, Henzi et al. (2021c) have derived a competitive benchmark for postprocessing numerical weather predictions, assuming that the conditional distribution of the actual observation given the forecast(s) increases in first order stochastic dominance. In the particular case of precipitation forecasts, their method can directly estimate the point mass at zero (probability of no precipitation) and the continuous distribution of positive precipitation amounts without any need for specific adaptations, whereas other postprocessing methods have to be specially tailored to this mixed discrete-continuous distribution. Further applications are the estimation of growth curves (Mösching and Dümbgen, 2020b) and of survival times depending on the severity of a carcinoma (El Barmi and Mukerjee, 2016), and the probabilistic prediction of the length of stay of intensive care unit patients (Henzi et al., 2021b,a)

The structure of the article is as follows. Section 2 describes the estimator, and it is shown that it generalizes the one by El Barmi and Marchev (2009) for the two-sample case. This connection sheds light on the selection of a tuning parameter in their method. As often in estimation under stochastic order restrictions, constructing a proper estimator which satisfies the order constraints is not trivial, and the proposed method relies heavily on tools and results from the monotone regression literature. In Section 3, uniform consistency is proven both for discrete and continuous covariates X, and rates of convergences are derived in different settings. All proofs deferred to the appendix. Simulation examples in Section 4 illustrate the performance of the estimator in comparison with extant methods, and Section 5 presents an application to conditional distribution estimation and calibration testing in economic surveys.

2 Estimation

To avoid redundancy, only the estimation for the increasing concave order is presented here; the necessary adaptations for the increasing convex order are straightforward. Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R} \times \mathbb{R}$ be covariate-observation pairs based on which the conditional distributions are to be estimated. In the literature on estimation under stochastic order restrictions, the CDFs $F_x(y)$ are often only estimated at the distinct values $x_1 < \cdots < x_d$ of $X_1, \dots, X_n$ and $y_1 < \cdots < y_m$ and of $Y_1, \dots, Y_n$, and frequently used estimation methods are nonparametric maximum likelihood estimation (NPLME) (e.g. in Dykstra et al., 1991; Mösching and Dümbgen, 2020a) and monotone least squares regression (e.g. in El Barmi and Mukerjee, 2005; Mösching and Dümbgen, 2020b). However, these two approaches turn out to be unrewarding in the case of the increasing concave order. Firstly, they lead to a constrained optimization problem with $\mathcal{O}(n^2)$ variables in general, namely the estimators $\hat{F}_{x_i}(y_j)$ for $F_{x_i}(y_j)$, which is not efficiently solvable for large n. And secondly, for the $\preceq_{\text{icv}}$-constrained estimator, proving consistency using the definition of the estimator as maximizer of the likelihood or least squares estimator seems intractable. The construction here is therefore an indirect approach. For $x, y \in \mathbb{R}$, define

$$M_x(y) = \int_{-\infty}^{y} F_x(t)\, dt = \mathbb{E}[(y - Y)_+ \mid X = x].$$

Under the assumption that $F_x \preceq_{\mathrm{icv}} F_{x'}$ if $x \leq x'$, the quantities $M_x(y)$ should be decreasing in x for all fixed y, and they satisfy

$$M_x'(y+) = \lim_{h \to 0,\, h > 0} \frac{M_x(y+h) - M_x(y)}{h} = F_x(y).$$

This suggests that an estimator $\hat{M}_x$ for M_x may yield, under some conditions, an estimator for F_x. We restrict the estimation of $F_x(y)$ to $x \in \{x_1, \ldots, x_d\}$; in Section 3, it is shown that under a continuity assumption, any interpolation method to obtain estimates for $x \notin \{x_1, \ldots, x_d\}$ is sufficient for uniform consistency.

Since $M_{X_i}(y)$ equals the expected value $\mathbb{E}[(y-Y)_+ \mid X = X_i]$, a reasonable estimator for it is the antitonic least squares regression $\tilde{M}_{X_1}(y), \ldots, \tilde{M}_{X_n}(y)$ of $(y-Y_1)_+, \ldots, (y-Y_n)_+$ with covariates $X_1, \ldots, X_n$, that is,

$$[\tilde{M}_{X_1}(y), \ldots, \tilde{M}_{X_n}(y)] \;=\; \operatorname*{argmin}_{\eta \in \mathbb{R}^n:\, \eta_i \geq \eta_j \text{ if } X_i \leq X_j} \sum_{i=1}^{n} [\eta_i - (y - Y_i)_+]^2.$$

The order constraints enforce $\tilde{M}_{X_i}(y) = \tilde{M}_{X_j}(y)$ if $X_i = X_j$, so the above problem is equivalent to the reduced, weighted antitonic regression

$$[\tilde{M}_{x_1}(y), \ldots, \tilde{M}_{x_d}(y)] \;=\; \operatorname*{argmin}_{\eta \in \mathbb{R}^d:\, \eta_1 \geq \cdots \geq \eta_d} \sum_{i=1}^{d} w_i [\eta_i - h_i(y)]^2,$$

where $w_i = \#\{j \leq n : \, X_j = x_i\}$, $i = 1, \ldots, d$, and

$$h_i(y) = \frac{1}{w_i} \sum_{j:\, X_j = x_i} (y - Y_j)_+.$$

This antitonic regression has the min-max-representation

$$\tilde{M}_{x_i}(y) = \min_{k=1,\ldots,i} \max_{j=k,\ldots,d} \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s h_s(y), \tag{1}$$

see Equations (1.9)-(1.13) of Barlow et al. (1972). In principle, one could now try to estimate $F_{x_i}(y)$ by taking the right-sided derivative of $\tilde{M}_{x_i}(\cdot)$ at y. However, $\tilde{M}_{x_i}$ is not necessarily convex and therefore its derivative may be decreasing and not a CDF. To correct this, let $\hat{M}_{x_i}$ be the greatest convex minorant to the function $y \mapsto \tilde{M}_{x_i}(y)$, which is the pointwise greatest convex function bounded by $\tilde{M}_{x_i}$ from above, and define $\hat{F}_{x_i}(y)$ as the right-hand slope of $\hat{M}_{x_i}(\cdot)$ at y. The following proposition, which is a consequence of the above min-max-formula and properties of greatest convex minorants (see Appendix A), shows that this is a valid strategy.

Proposition 2.1.

(i) The functions $\tilde{M}_{x_i}(y)$ and $\hat{M}_{x_i}(y)$ are increasing and piecewise linear in y for fixed $i \in \{1, \ldots, d\}$, and decreasing in i for fixed $y \in \mathbb{R}$.

(ii) The functions $\hat{F}_{x_i}(y)$ for fixed $i \in \{1, \ldots, d\}$ are piecewise constant CDFs with $F_{x_i}(y) = 0$ for $y < y_1$ and $F_{x_i}(y) = 1$ for $y \geq y_m$.

In practice, it is not possible to compute $\tilde{M}_{x_i}(y)$ and $\hat{M}_{x_i}(y)$ at all $y \in \mathbb{R}$. Although these functions are piecewise linear, there is no efficient procedure to identify the knots where their slope changes. A pragmatic solution is to evaluate $\tilde{M}_x$ and $\hat{M}_x$ on a fine grid $t_1 < \cdots < t_k$ with $t_1 = y_1$ and $t_k = y_m$, and interpolate linearly in between. This has the consequence that the CDFs $\hat{F}_{x_i}$ can only put mass on $t_1, \dots, t_k$. By a standard result about isotonic regression (see Appendix A), the right-sided slope of the greatest convex minorant to the interpolation of $(t_1, \tilde{M}_{x_i}(t_1)), \dots, (t_k, \tilde{M}_{x_i}(t_k))$ equals the isotonic regression of the slopes

$$\tilde{F}_{x_i}(t_j) = \frac{\tilde{M}_{x_i}(t_{j+1}) - \tilde{M}_{x_i}(t_j)}{t_{j+1} - t_j}$$

with weights $t_{j+1} - t_j$, $j = 1, \dots, k-1$. This isotonic regression directly yields the estimators for the conditional CDFs,

$$[\hat{F}_{x_i}(t_1), \dots, \hat{F}_{x_i}(t_{k-1})] \;=\; \operatorname*{argmin}_{\xi \in \mathbb{R}^{k-1}:\, \xi_1 \le \cdots \le \xi_{k-1}} \sum_{j=1}^{k-1} (t_{j+1} - t_j)[\xi_j - \tilde{F}_{x_i}(y_j)]^2,$$

and $\hat{F}_{x_i}(t_k) = 1$ by Proposition 2.1 (ii) if $t_k = y_m$. To summarize, the estimation procedure consists of two series of monotone regressions, informally speaking one in the X-direction for fixed threshold y to obtain $\preceq_{\text{icv}}$-ordered distributions, and another in the Y-direction for fixed covariate x_i to ensure that the CDFs are increasing. It is not necessary to compute the functions $\hat{M}_{x_i}$ explicitly, since the computation of the greatest convex minorant is indirect via its right-hand slope. The exact solution of monotone regression problems can be obtained efficiently with the Pool-Adjacent Violators Algorithm (PAVA), which has complexity $\mathcal{O}(N)$ with sorted covariate and sample size N. Hence the overall complexity of the estimation procedure is $\mathcal{O}(n^2)$ if the number of distinct values in $X_1, \dots, X_n$ or $Y_1, \dots, Y_n$ grows at the rate $\mathcal{O}(n)$.

If the distinct values $y_1, \dots, y_m$ of $Y_1, \dots, Y_n$ are taken as the grid for computation, then the estimated distributions $\hat{F}_{x_i}$ can only put mass on the actual observations in the data, and they are equal to the conditional empirical cumulative distribution functions (ECDF) if these already satisfy the increasing concave order condition. That is, if $\check{F}_{x_j}$ is the ECDF of all Y_i with $X_i = x_j$ and if $\check{F}_{x_i} \preceq_{\text{icv}} \dots \preceq_{\text{icv}} \check{F}_{x_d}$, then $\hat{F}_{x_j} = \check{F}_{x_j}$ for $j = 1, \dots, d$. If in addition $X_1, \dots, X_n$ are pairwise distinct, $\hat{F}_{X_i}$ is the Dirac measure at Y_i for $i = 1, \dots, n$. The estimators under first order stochastic dominance (El Barmi and Mukerjee, 2005; Mösching and Dümbgen, 2020b) also have this property. However, with the increasing concave order, even if the grid $\{t_1, \dots, t_k\}$ contains $\{y_1, \dots, y_m\}$, the estimator can put probability mass on points outside of $\{y_1, \dots, y_m\}$. In particular, if the response variable is known to take values in a discrete set, say $\mathbb{Z}$, then the grid should be chosen within this set to avoid positive estimated probabilities outside of the actual support.

The increasing concave order is preserved under pointwise convex combinations of CDFs, i.e. if $F_1 \preceq_{\text{icv}} F_2$ and $G_1 \preceq_{\text{icv}} G_2$, then also $\lambda F_1 + (1-\lambda) G_1 \preceq_{\text{icv}} \lambda F_2 + (1-\lambda) G_2$ for $\lambda \in (0, 1)$. This fact opens the possibility to combine the estimation procedure with sample splitting as suggested in Henzi et al. (2021c) for first order stochastic dominance. Instead of estimating the conditional distributions with the complete dataset, one may draw random subsamples from the data and aggregate the estimated conditional CDFs from each run by their pointwise average. This subsample aggregation (subagging) yields smoother estimated CDFs and prevents overfitting. Alternatively, the data can also be partitioned into several disjoint subsets instead of drawing subsamples, and Banerjee et al. (2019) have proved that this divide and conquer strategy may lead to better convergence rates of isotonic mean regression. While partitioning of the data is a valid strategy for very large datasets, it seems more desirable to apply subagging in smaller datasets to avoid that the estimator depends too strongly on the chosen partition. Note that in principle, the averaging step in subagging or sample splitting could also be done on the level of the estimators $\tilde{M}_x$ instead of the CDFs $\hat{F}_x$, but if the goal is to obtain smoother CDFs, it is more natural to perform averaging of $\hat{F}_x$.

Finally, we show that the estimator proposed here generalizes the one by El Barmi and Marchev (2009) for the case $X_i \in \{1, 2\}$. Their estimator depends on a parameter $\alpha \in [0, 1]$, and equality holds for $\alpha = \#\{i \leq n : X_i = 1\}/n$. This follows from the fact that with $\check{M}_j(y) = \int_{-\infty}^{y} \check{F}_j(t)\,dt$, one can write $\tilde{M}_j(y)$ as

$$\tilde{M}_j(y) = \mathbb{1}\{\check{M}_j(y) \geq \check{M}_2(y)\}\check{M}_1(y) + \mathbb{1}\{\check{M}_1(y) < \check{M}_2(y)\}[\alpha\check{M}_1(y) + (1-\alpha)\check{M}_1(y)],$$

for $j = 1, 2$, where $\mathbb{1}$ is the indicator function. Taking the right-hand slope of the greatest convex minorant of the above functions then yields Equation (8) from El Barmi and Marchev (2009). The choice $\alpha = \#\{i \leq n : X_i = 1\}/n$ was already suggested in their article, and it corresponds to the natural weight for which $\tilde{M}_j$, $j = 1, 2$, are the antitonic regression estimators.

3 Uniform consistency

The following notation and assumptions are required for stating the theorems about uniform consistency. Let (X_{ni}, Y_{ni}), $i = 1, \dots, n$, $n \in \mathbb{N}$, be a triangular array defined on a measurable space $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}$. For a sequence of events $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$, the statement '$A_n$ holds with asymptotic probability one' means $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$. The covariates $X_{n1}, \dots, X_{nn}$ are assumed to be independent and have distinct values $x_1 < \cdots < x_d$, and the response variables $Y_{n1}, \dots, Y_{nn}$ are independent conditional on $X_{n1}, \dots, X_{nn}$ such that $\mathbb{P}(Y_{ni} \leq y \mid X_{ni}) = F_{X_{ni}}$, with the CDFs F_x increasing in x in the increasing concave order. The distinct values of $Y_{n1}, \dots, Y_{nn}$ are again denoted by $y_1 < \cdots < y_m$. A subscript n in $\tilde{M}_{n;x}(y)$, $\hat{M}_{n;x}(y)$, and $\hat{F}_{n;x}(y)$ will be used to indicate that these quantities depend on the sample size n, but the dependency of m and d on n is not written explicitly to lighten the notation. If $x \notin \{x_1, \dots, x_d\}$, it is only assumed that $\tilde{M}_{n;x_i}(y) \geq \tilde{M}_{n;x}(y) \geq \tilde{M}_{n;x_{i+1}}(y)$ for all $y \in \mathbb{R}$ if $x \in [x_i, x_{i+1})$, and $\tilde{M}_{n;x}(y) = \tilde{M}_{n;x_1}(y)$ if $x < x_1$ or $\tilde{M}_{n;x}(y) = \tilde{M}_{n;x_d}(y)$ if $x \geq x_d$. The same property then also holds for $\hat{M}_{n;x}$.

The key condition for proving consistency of $\hat{F}_{n;x}(y)$ is the following.

(A) There exists $(c_n)_{n\in\mathbb{N}} \subset [0, \infty)$ and a sequence of sets $(I_n)_{n\in\mathbb{N}}$, $I_n \subset \mathbb{R}$, such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{y\in\mathbb{R},\, x\in I_n} |\tilde{M}_{n;x}(y) - M_x(y)| \geq c_n \right) = 0.$$

Sufficient conditions for (A) will be given below, and the convergence rate c_n depends on whether X is discrete or continuously distributed and on the tail properties of F_x. If $X_{n1}, \dots, X_{nn} \in \{1, \dots, K\}$, one can simply set $I_n = \{1, \dots, K\}$. For continuously distributed covariates on an interval I, I_n will be of the form $I_n = \{x \in I : x \pm \delta_n \in I\}$ with $\delta_n \to 0$, that is, it is in general not possible to show consistency at the boundary of the covariate domain. This is also the case in isotonic mean regression and estimation under first order stochastic dominance (Mösching and Dümbgen, 2020b).

The following proposition establishes the connection between the uniform consistency of $\tilde{M}_{n;x}(y)$ and $\hat{F}_{n;x}(y)$.

Proposition 3.1. *Assume that (A) holds and $I \subseteq \mathbb{R}$ is a set such that $I_n \subseteq I$, $n \in \mathbb{N}$.*

(i) If there exist $J \subseteq \mathbb{R}$ and constants $C \geq 0$, $\beta > 0$ such that $|F_x(y) - F_x(z)| \leq C|y - z|^\beta$ for all $y, z \in J$, $x \in I$, then with $J_n = \{y \in J : y \pm c_n^{1/(1+\beta)} \in J\}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{y\in J_n,\, x\in I_n} |\hat{F}_{n;x}(y) - F_x(y)| \geq (2 + C)c_n^{\beta/(1+\beta)}\right) = 0.$$

(ii) If the distribution functions F_x, $x \in I$, have support in $\mathbb{Z}$ and if $\tilde{M}_{n;x}$ is computed with grid $\{y_1, y_1 + 1, \ldots, y_m - 1, y_m\}$, then

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{y\in \mathbb{R},\, x\in I_n} |\hat{F}_{n;x}(y) - F_x(y)| \geq 2c_n\right) = 0.$$

Proposition 3.1 shows that if $\tilde{M}_{n;x}(y)$ is uniformly consistent in x and y at a rate c_n, then the estimator $\hat{F}_{n;x}(y)$ is also uniformly consistent. When the response variable is integer-valued, $\hat{F}_{n;x}$ is consistent at the same rate. Otherwise, if the distribution functions F_x are Hölder continuous with index β, the corresponding rate for $\hat{F}_{n;x}(y)$ is $c_n^{\beta/(1+\beta)}$, for example $c_n^{1/2}$ if the F_x are Lipschitz continuous. Note that in the case $J = \mathbb{R}$, the sets J_n in Proposition 3.1 (i) are also equal to $\mathbb{R}$.

We proceed to state conditions under which (A) holds. For the K-sample case, the assumption on the covariate is the following.

(K) The covariates take values in $I = \{1, \ldots, K\}$, and $\min_{j=1,\ldots,K} \mathbb{P}(X_{ni} = j) = p$ for some $p > 0$.

In the continuous case, the assumptions are analogous to (A.1) and (A.2) in Mösching and Dümbgen (2020b).

(C1) The covariates $X_{n1}, \ldots, X_{nn}$ admit a Lebesgue density bounded away from zero by $p > 0$ on an interval I.

(C2) There exists a constant $L > 0$ such that for all $u, v \in I$ and $y \in \mathbb{R}$,

$$|M_u(y) - M_v(y)| \leq L|u - v|.$$

Note that the set I and the constants p in (K) and (C1) and L in (C2) do not depend on n. Condition (C1) could be replaced by the weaker assumption that the number of points in every subinterval of I of a certain size grows sufficiently fast, like in (A.2) of Mösching and Dümbgen (2020b, see also their Remark 3.2). In particular, it is not necessary to assume that the covariates $X_{n1}, \ldots, X_{nn}$ are pairwise distinct or independent. However, this more general condition would require to introduce additional notation and constants. Similarly, in (K), it is sufficient that each value $j \in \{1, \ldots, K\}$ is attained at least $n\delta$ times with asymptotic probability one for some $\delta > 0$. The Lipschitz assumption (C2) in the continuous case is standard in isotonic regression (see e.g. Yang et al., 2019; Dai et al., 2020), and it could be replaced by Hölder continuity with index $\alpha \in (0, 1)$ at the cost of a slower convergence rate.

Since the goal is to prove consistency for an estimator of the expected values $\mathbb{E}[(y - Y)_+ \mid X = x]$, it is natural that some additional assumptions on the tail behaviour of the distributions F_x are required. In the two cases below, the set I is assumed to be the one from (K) or from (C1), (C2).

(P) There exist $\lambda > 2$ and $y_0 \geq 0$ such that for all $y \geq y_0$ and $x \in I$,

$$\mathbb{P}(|Y| \geq y \mid X = x) \leq y^{-\lambda}.$$

(E) There exist $\lambda > 0$ and $y_0 \geq 0$ such that for all $y \geq y_0$ and $x \in I$,

$$\mathbb{P}(|Y| \geq y \mid X = x) \leq \exp(-\lambda y).$$

Theorem 3.2. *Condition (A) holds with*

$$c_n = \begin{cases} 4p^{-1/2} n^{-1/2+1/\lambda} \log(n)^{1/2+1/\lambda}, & \textit{under (K) and (P)}, \\ 8p^{-1/2} \lambda^{-1} n^{-1/2} \log(n)^{3/2}, & \textit{under (K) and (E)}, \\ [4p^{-1/2} + L] n^{-1/3+2/(3\lambda)} \log(n)^{1/3+2/(3\lambda)}, & \textit{under (C1), (C2), and (P)}, \\ [4p^{-1/2} + L](2/\lambda)^{2/3} n^{-1/3} \log(n), & \textit{under (C1), (C2), and (E)}, \end{cases}$$

and

$$I_n = \begin{cases} \{1, \ldots, K\}, & \textit{under (K)}, \\ \{x \in I : x \pm n^{-1/3+2/(3\lambda)} \log(n)^{1/3+2/(3\lambda)} \in I\}, & \textit{under (C1), (C2), and (P)}, \\ \{x \in I : x \pm (2/\lambda)^{2/3} n^{-1/3} \log(n) \in I\}, & \textit{under (C1), (C2), and (E)}. \end{cases}$$

In the K-sample case, Theorem 3.2 implies uniform consistency at a rate of at least $(\log(n)/n)^{1/4}$ if the distribution functions $F_x(y)$ are Lipschitz continuous in y and have exponential tails. This is slower than the $n^{-1/2}$-rate of the empirical distribution functions stratified by the K covariate values, and suggests that this lower bound is not always tight. Indeed, if the conditional CDFs F_j, $j = 1, \ldots, K$, have support on disjoint, pointwise increasing intervals, then $\hat{F}_{n;j}$ are equal to the ECDFs of the corresponding subsamples and hence known to converge at the faster rate. Nevertheless, the result extends the ones from the current literature. In the two-sample case $K = 2$, Rojo and El Barmi (2003) establish strong uniform convergence and pointwise but not uniform root-n convergence for their estimator, while El Barmi and Marchev (2009) only prove strong uniform consistency, but do not derive rates of convergence.

For a continuously distributed covariate and exponential tails, $\tilde{M}_x(y)$ converges uniformly in x and y at a rate of $n^{-1/3}$ up to a logarithmic factor, which is known to be the global rate of convergence of the isotonic regression estimator. When the conditional distributions have power tails with exponent λ, the rate becomes slower by a factor of $n^{2/(3\lambda)}$. In general, the global $n^{-1/3}$-rate of convergence for isotonic regression does not require the assumption of exponential tails, but the results across the literature are not directly comparable. For example, Zhang (2002) shows that with bounded second moments, the risk of the isotonic mean regression estimator, i.e. the root mean squared error at the design points, scales at a rate of $n^{-1/3}$, whereas Yang et al. (2019) prove uniform consistency with the same rate (up to logarithmic factors) in the supremum norm under sub-gaussianity of the error terms. Theorem 3.2 yields a stronger statement since convergence is also uniform in the parameter y, and with exponential tails it still matches the optimal global rate up to the logarithmic factor. For $\hat{F}_x(y)$, Theorem 3.2 implies a rate of at least $n^{-1/6}$ under the favorable assumption (E) and Lipschitz continuity of $F_x(y)$ in y.

4 Simulations

In the following simulation examples, the $\preceq_{\mathrm{icv}}$- and $\preceq_{\mathrm{icx}}$-order constrained estimators are compared to competitors in terms of the L_1 distance between the estimated and the true CDFs, and in terms of the mean absolute error (MAE) of quantile estimates,

$$\mathrm{L}_1(\hat{F}_n, F) = \mathbb{E}\left(\int_{-\infty}^{\infty} |\hat{F}_n(y) - F(y)|\, dy\right), \quad dq_\gamma(\hat{F}_n, F) = \mathbb{E}\Big(|\hat{F}_n^{-1}(\gamma) - F^{-1}(\gamma)|\Big),$$

where $\hat{F}_n$ denotes an estimator for F and the expected value is taken over the sampling distribution of $\hat{F}_n$ for the given sample size. The expected value in the definition of L_1 and dq_γ is approximated by the empirical mean over $10'000$ simulations for the examples with discrete and $5'000$ simulations for those with continuously distributed covariates.

With covariate values $X \in [1, 4]$, the following three settings are considered:

$$Y_1 = X^{1/2} + \big[1 + (X-2)/(1+(X-2)^2)^{1/2}\big]\varepsilon,\ \varepsilon \sim \mathrm{Student}(\mathrm{df} = 10), \tag{2}$$

$$Y_2 \sim \mathrm{Gamma}(\mathrm{shape} = X,\ \mathrm{rate} = X^{9/10}), \tag{3}$$

$$Y_3 \sim \text{Beta-binomial}(n = 50,\ \alpha = X^3,\ \beta = 1 + X^3). \tag{4}$$

The conditional distributions of Y_1 given X are ordered in the increasing convex order, and the those of Y_2 and Y_3 with respect to the increasing concave order; see Figure 1 (a) for an illustration. In the K-sample case, X takes values in $\{1, 4\}$, $\{1, 2, 3, 4\}$, and $\{1, 1.5, \ldots, 3.5, 4\}$, i.e. $K = 2, 4, 7$, which allows comparing the change in estimation error at previously available values of X when the number of samples increases. For simulation examples with continuous covariate, the sample of X is generated independent and uniformly distributed on $[1, 4]$. In all simulations the distinct observed values of the response variable are taken as grid for the computation of the $\preceq_{\mathrm{icv}}$- and $\preceq_{\mathrm{icx}}$-order constrained estimators.

Table 1 shows the performance order restricted estimators compared to the ECDF in the K-sample case, with fixed group sizes $n = 30, 50$ as in El Barmi and Marchev (2009). For $K = 2$ only few corrections are required to obtain conditional distributions satisfying the order constraints, whence the order restricted estimator brings no improvement over the ECDF stratified by X. However, as the number of groups increases, the order restricted estimators benefit from the larger total sample size and achieve a lower estimation error both globally, i.e. in L_1-distance, and for most quantiles considered.

In the continuous case, the estimator under first order stochastic dominance by Mösching and Dümbgen (2020b) is chosen as competitor. As Figure 1 (a) shows, for the simulations (3) and (4) the conditional quantile curves up to the seventh decile are all increasing in the covariate X, and so are the conditional quantile curves above the third decile in (2). Therefore, although first order stochastic dominance is violated, it serves as a reasonable approximation in these problems. Figure 2 shows the relative performance of the estimators for $n = 500$. The estimator by Mösching and Dümbgen (2020b) achieves a lower absolute error for the median, for the 0.1-quantile in (3) and (4), and for the 0.9-quantile in (2), uniformly over all values of X. This has to be expected, since the corresponding quantile curves are monotone and estimation under this correct constraint is more efficient than with the weaker $\preceq_{\mathrm{icv}}$- and $\preceq_{\mathrm{icx}}$-constraints. The picture is different for the low quantiles in (2) and the high quantiles in (3) and (4), where the conditional quantile curves are antitonic and the best isotonic approximation is constant, which generally provides a poor fit. Figure 2 also compares the errors of

Figure 1: (a) Deciles of the conditional distributions in the simulation examples (2), (3), (4). The median is depicted as a dashed line. (b) Quantile curves (levels $0.1, 0.3, 0.5, 0.7, 0.9$) for simulation example (3) together with $\preceq_{\mathrm{icv}}$-ordered estimator ($n = 500$; ICV and subagging variant $\mathrm{ICV}_{\mathrm{sbg}}$ with 50 subsamples of size $250 = n/2$).

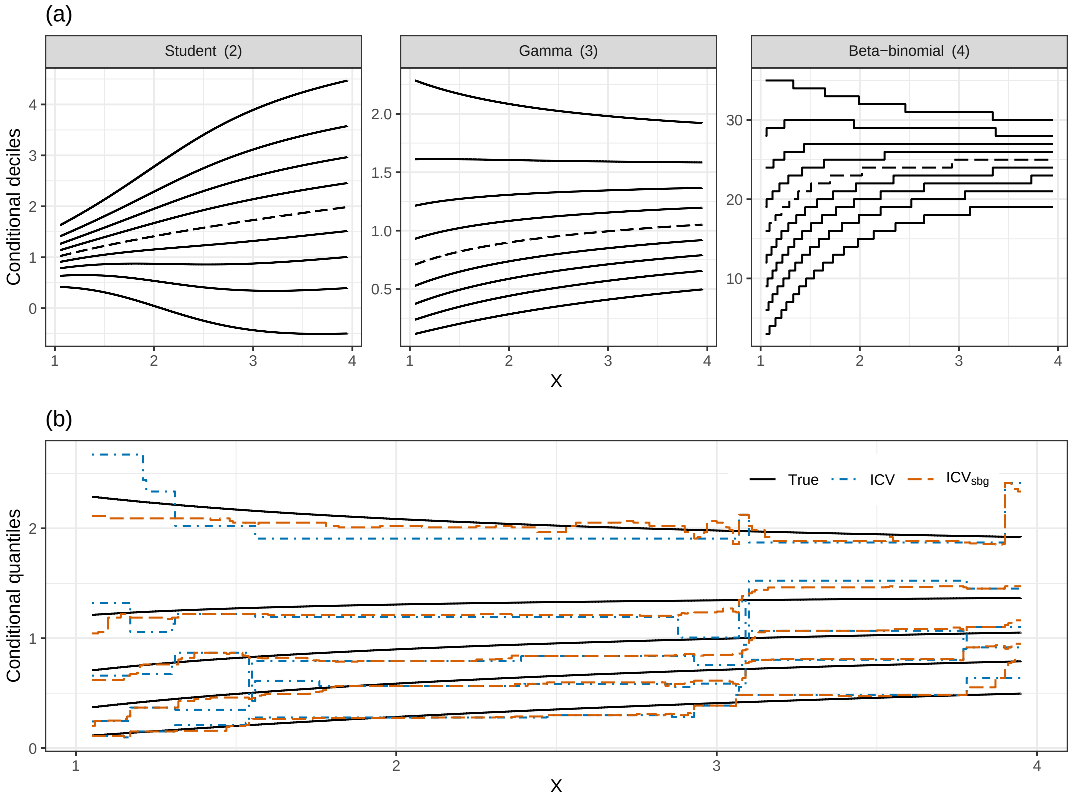
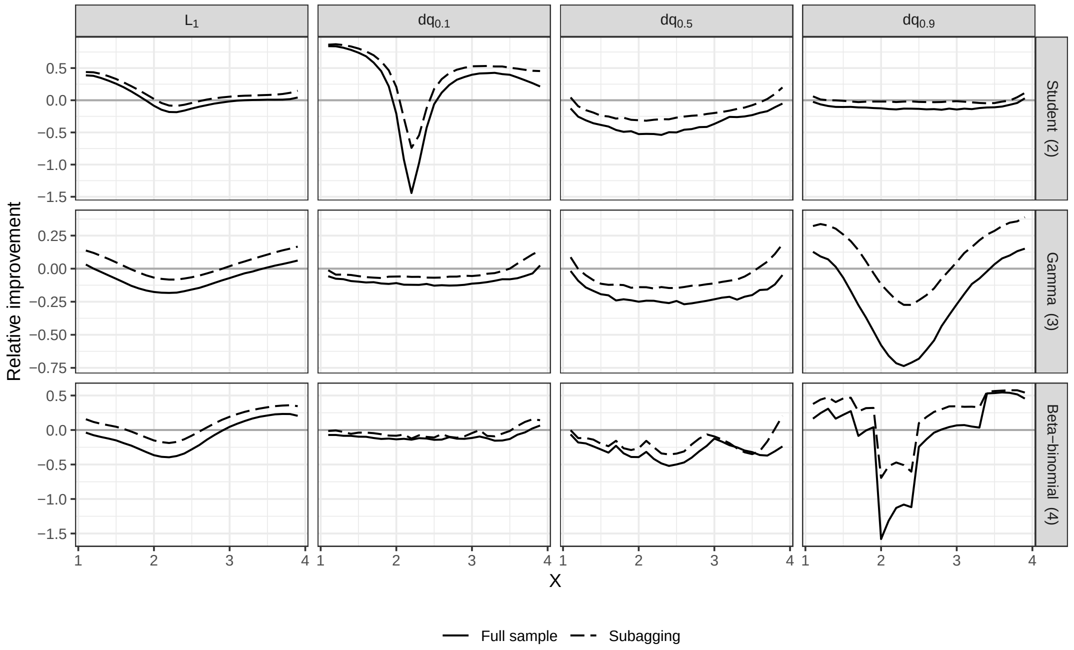


Table 1: Relative improvement in mean L_1 distance, mean absolute error of quantile estimates of $\preceq_{\text{icv}}$- and $\preceq_{\text{icx}}$-order constrained estimator compared to ECDF stratified by the value of X, for $K = 2$, 4, 7 and group sizes of $n = 30$, 50.

$n = 30$		Student (2)				Gamma (3)				Beta-binomial (4)			
K	X	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$
2	1.0	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.07	0.00	0.00	0.00	0.00
	4.0	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	-0.04	0.00	0.00	0.00	0.00
4	1.0	0.00	0.00	0.00	0.00	0.06	0.00	0.05	0.12	0.00	0.00	0.00	0.00
	2.0	-0.01	-0.10	0.01	-0.01	0.08	0.05	0.12	0.19	0.05	0.00	0.05	0.09
	3.0	0.06	0.19	0.08	0.05	0.09	0.13	0.18	0.21	0.13	0.09	0.09	0.12
	4.0	0.05	0.16	0.10	0.10	0.03	0.11	0.11	0.14	0.08	0.09	0.15	0.10
7	1.0	0.00	-0.01	0.00	0.00	0.08	0.04	0.10	0.13	0.01	0.00	0.00	0.01
	1.5	-0.02	-0.10	0.00	-0.01	0.14	0.13	0.21	0.29	0.05	0.00	0.04	0.05
	2.0	0.01	0.01	0.05	0.00	0.16	0.20	0.26	0.35	0.12	0.05	0.12	0.14
	2.5	0.08	0.23	0.15	0.08	0.18	0.25	0.31	0.37	0.21	0.14	0.22	0.24
	3.0	0.13	0.33	0.23	0.19	0.17	0.28	0.32	0.36	0.28	0.24	0.22	0.29
	3.5	0.14	0.33	0.26	0.25	0.15	0.28	0.30	0.32	0.27	0.29	0.29	0.20
	4.0	0.09	0.23	0.19	0.21	0.07	0.19	0.19	0.22	0.15	0.15	0.26	0.19

$n = 50$		Student (2)				Gamma (3)				Beta-binomial (4)			
K	X	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$	L$_1$	$dq_{0.1}$	$dq_{0.5}$	$dq_{0.9}$
2	1.0	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.05	0.00	0.00	0.00	0.00
	4.0	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	-0.02	0.00	0.00	0.00	0.00
4	1.0	0.00	0.00	0.00	0.00	0.04	0.00	0.02	0.10	0.00	0.00	0.00	0.00
	2.0	-0.01	-0.07	0.00	0.00	0.05	0.02	0.08	0.14	0.03	0.00	0.02	0.08
	3.0	0.04	0.13	0.05	0.04	0.07	0.08	0.14	0.20	0.09	0.05	0.04	0.07
	4.0	0.04	0.13	0.07	0.07	0.03	0.07	0.09	0.11	0.06	0.06	0.12	0.05
7	1.0	0.00	0.00	0.00	0.00	0.07	0.02	0.07	0.13	0.00	0.00	0.00	0.00
	1.5	-0.01	-0.07	0.00	0.00	0.11	0.09	0.16	0.25	0.03	0.00	0.02	0.03
	2.0	0.00	-0.04	0.02	-0.01	0.13	0.14	0.21	0.29	0.08	0.02	0.08	0.12
	2.5	0.05	0.15	0.09	0.03	0.15	0.19	0.26	0.32	0.16	0.09	0.19	0.20
	3.0	0.11	0.28	0.19	0.15	0.16	0.23	0.28	0.34	0.23	0.17	0.16	0.25
	3.5	0.13	0.27	0.23	0.22	0.14	0.24	0.27	0.31	0.25	0.26	0.26	0.16
	4.0	0.08	0.20	0.16	0.18	0.07	0.16	0.16	0.18	0.14	0.11	0.26	0.12

Figure 2: Relative improvement in L_1 distance and mean absolute error of quantile estimates of the $\preceq_{\text{icv}}$- and $\preceq_{\text{icx}}$-order constrained estimator compared to the estimator under first order stochastic dominance, for $n = 500$. The solid lines show the improvement when the estimators are computed on the full sample, and the dashed lines for a subagging variant with 50 subamples of size 250.



subagging variants of the estimators; see Figure 1 (b) for an illustration of the estimated quantile curves in the Gamma example. For both estimators, 50 random subsamples of size $250 = n/2$ are drawn from the data, and the conditional CDFs from each fit to the subsamples are averaged pointwise. It can be seen that the $\preceq_{\text{icv}}$- and $\preceq_{\text{icx}}$-order constrained estimators benefit more from subagging than the estimator with first order stochastic dominance. A more detailed comparison of different subagging variants and results for other sample sizes are given in Appendix D.

5 Case study

It is well known that in the the evaluation of point forecasts, a wrongly specified loss function, such as the absolute error for comparing mean forecasts, may lead to counterintuitive results and distorted forecast rankings (Gneiting, 2011). This causes problems in the interpretation of economic surveys, where respondents are often asked to issue point predictions for future quantities, but it is unspecified what functional of their (hypothetical) predictive distribution is meant. As a remedy, various tests of forecast rationality, or forecast calibration, have been proposed in the literature. A recent contribution is by Dimitriadis et al. (2019), who develop tests for the hypothesis that a given point forecast is the mean, median, or mode functional, or a convex combination of the three. The case study in this section demonstrates that the estimation of conditional distributions can complement such tests to gain additional information for the interpretation of point forecasts.

If X denotes a point forecast and Y the observation, the hypothesis of forecast rationality with respect to a functional T can be defined as $X = T[\mathcal{L}(Y \mid X)]$, where $\mathcal{L}(Y \mid X)$ denotes the conditional law of Y given the forecast X. This formulation is a special but important case of equation (2.1) in Dimitriadis et al. (2019), which allows including additional information available to the forecaster for conditioning. If T is the mean functional, then forecast rationality is equivalent to the moment condition $\mathbb{E}(Y - X \mid X) = 0$. For the median, the corresponding condition is $\mathbb{E}(\mathbb{1}\{Y \geq X\} \mid X) = 0.5$, provided that $\mathcal{L}(Y \mid X)$ is a continuous distribution. Based on such conditions, Dimitriadis et al. (2019) developed asymptotic tests for forecast rationality.

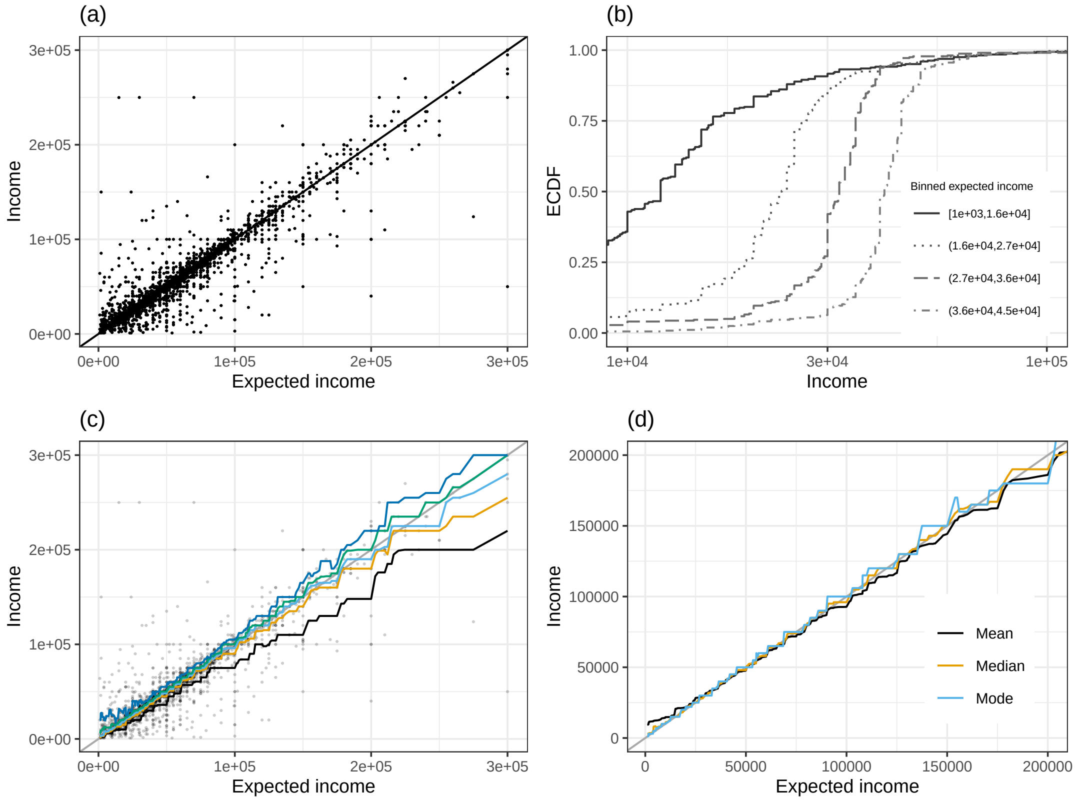
Distributional regression provides a different, more qualitative approach to this problem. If the conditional distributions $\mathcal{L}(Y \mid X = x)$ were known, one could easily derive the functional of interest $T(x) = T[\mathcal{L}(Y \mid X = x)]$ and detect violations of forecast rationality by directly comparing $T(x)$ and x. Estimators with stochastic order restrictions allow to mimic this ideal situation, without having to impose restrictive or implausible assumptions on the conditional distributions. For almost any sufficiently precise point forecast, it is a reasonable assumption that the distributions $\mathcal{L}(Y \mid X = x)$ are ordered in the increasing concave or convex order or even in first order stochastic dominance. Moreover, estimating $\mathcal{L}(Y \mid X)$ under stochastic order constraints only requires the ranks of the forecasts $X_1, \dots, X_n$ in a sample, but not their values. This makes a comparison of X_i and $T(X_i)$ sensible, because $X_1, \dots, X_n$ themselves have not been provided to the model. Other estimation methods, such as kernel regression (Li and Racine, 2008), generally do not have this property.

To illustrate the approach, consider the data example from Section 5.1 of Dimitriadis et al. (2019). In the Labor Market Survey by the Federal Reserve Bank of New York[1], respondents are asked three times per year to report their annualized income in four months. The sample analysed here ranges from March 2015 to November 2019. Some respondents participate in several rounds of the survey, and only the first round is included for those individuals which occur several times to obtain independent observations. Additionally, like in Dimitriadis et al. (2019), observations with very high or low expected income (above 300'000 or below 1000, 4.0% of the sample; an upper bound of 1 million was used in Dimitriadis et al. (2019)) are removed since the data is very sparse and uninformative for such values, as are cases when the ratio of expectation and income or the inverse ratio is between 9 and 13 (27 instances), which might be due to misplaced decimal points or erroneously reporting monthly instead of annualized income. The remaining sample consists of 3161 observations.

Panels (a) and (b) of Figure 3 illustrate the joint distribution of the income expectations and realizations. There is a strong monotone relationship, and for income expectations below 50'000, the variability in the realized income decreases as the expected income increases, in a similar fashion as in simulation example (3) from Section 4. It is therefore questionable if the distributions are increasing with respect to first order stochastic dominance, but the weaker increasing concave order is an appropriate constraint. To estimate the conditional distributions, a subagging version of the $\preceq_{\mathrm{icv}}$-order restricted estimator with 50 subsamples of half of the total sample size is applied. From the estimated distributions, the mean, median, and mode functional are

[1] Source: Survey of Consumer Expectations, © 2013-2020 Federal Reserve Bank of New York. The SCE data are available without charge at https://www.newyorkfed.org/microeconomics/sce, and may be used subject to license terms posted there. The New York Fed disclaims any responsibility for this analysis and interpretation of Survey of Consumer Expectations data.

Figure 3: (a) Expected and realized income in the case study. (b) ECDF of the realized income for binned expectations. The boundaries of the bins are the 0.1 to 0.4-quantile of the income expectations. (c) Estimated quantile curves (levels $0.1, 0.3, 0.5, 0.7, 0.9$). (d) Mean, median and mode functional computed from the estimated conditional distributions (for expectations and incomes below 200'000).



then computed, with the mode taken as the location of the largest jump of the conditional CDFs, which are piecewise constant stepfunctions. Panels (c) and (d) of Figure 3 display estimated quantile curves and the three functionals depending on the income expectation.

For the mean functional, the forecast rationality test of Dimitriadis et al. (2019) yields a p-value of $1.7 \cdot 10^{-12}$, computed with the R package `fcrat` available on https://github.com/Schmidtpk/fcrat. As can be seen in Figure 3, the conditional mean curve lies above the bisector for expectations below 25'000, and below the bisector when the expectation exceeds 75'000, so there is indeed a systematic deviation of the income expectation from the estimated mean. For the median and the mode functional, the p-values of the rationality test are $4.5 \cdot 10^{-8}$ and 0.93, respectively. This huge difference in the p-values is in contrast to the curves in Figure 3 (d), where the expected income does not seem to deviate systematically from either functional. A simulation reveals that in this particular application, the p-value for the median should indeed be interpreted with care. By taking the estimated medians as new income expectation and simulating new observations from the estimated conditional distributions, one obtains datasets which look similar to the original data, but the income expectation equals the median of the underlying distribution by construction. Over 10'000 simulations, the rejection rate for the median rationality test is 0.03, 0.11 and 0.19 at the levels 0.01, 0.05, and 0.10 – the test is anticonservative. The reason for the non-validity of the median rationality test is likely to be the discreteness in the data: The realized incomes only take 526 distinct values with a sample size of $n = 3161$, and in 22% of the cases the income expectation is exactly equal to the realized income. Hence the condition $\mathbb{E}(\mathbb{1}\{Y \geq X\} \mid X) = 0.5$ may be violated even if X is equal to the conditional median due a point mass of the conditional distributions at the expected income X.

In conclusion, the $\preceq_{\text{icv}}$-constrained estimator suggests that both median and mode could rationalize the income expectations, and it confirms that the income expectations should not be interpreted as a mean forecast.

6 Discussion

In this article, the estimator for conditional distributions under increasing concave order constraints by El Barmi and Marchev (2009) has been generalized to the K-sample case and continuous covariates, and uniform rates of convergence have been established. This augments the current literature on estimation under stochastic order constraints by a general estimator under a weaker order than first order stochastic dominance.

There are several potential avenues for future work. A natural generalization is to consider partially ordered instead of real-valued covariates X, like in Henzi et al. (2021c) for first order stochastic dominance. A careful look at the construction of the estimator in Section 2 reveals that this is indeed possible, by applying antitonic regression with respect to the given partial order in the first estimation step. However, a proof of consistency with partially ordered covariates remains an open task.

For practical applications, the simulation examples in this article suggest that the $\preceq_{\text{icv}}$- and $\preceq_{\text{icx}}$-order constrained estimators benefit from smoothing. Motivated by the fact that isotonic regression for the mean may be improved by data splitting and averaging, subsample aggregating was suggested as method to smooth the estimated conditional CDFs. A theoretical analysis of the superefficiency problem studied by Banerjee et al. (2019) for the case of conditional distribution estimation is desirable, both for first order stochastic dominance and for the increasing concave and convex order.

A different approach for smoothing the conditional CDFs would be kernel smoothing. If K is any CDF, $h > 0$ a bandwidth and $(\hat{F}_x)_{x\in\mathbb{R}}$ are the estimators under the increasing concave or convex order, then

$$\hat{K}_x(y) = \int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right) d\hat{F}_x(t)$$

is again a CDF. These distributions are increasing in x in the given order by Theorem 4.A.18 of Shaked and Shanthikumar (2007), and smooth in y if the functions K are smooth. In this approach, the conditional CDFs $\hat{K}_x$ are a weighted average of (integrated) kernel functions, with the weights chosen in such a way that the stochastic order constraints are satisfied. Such a smoothing procedure is also applicable with first order stochastic dominance or more generally any other stochastic order which is preserved under convex mixing of distribution functions. The analysis of consistency and optimal bandwidth selection for such estimators would be a valuable contribution to the literature on estimation under stochastic order restrictions.

Acknowledgements

This work was supported by the Swiss National Science Foundation. The author is grateful to Johanna Ziegel and Timo Dimitriadis for helpful comments and discussions.

References

Arcones, M. A., Kvam, P. H., and Samaniego, F. J. (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association*, 97:170–182.

Banerjee, M., Durot, C., Sen, B., et al. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *Annals of Statistics*, 47:720–757.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Brunk, H., Franck, W., Hanson, D., and Hogg, R. (1966). Maximum likelihood estimation of the distributions of two stochastically ordered random variables. *Journal of the American Statistical Association*, 61:1067–1080.

Dai, R., Song, H., Barber, R. F., and Raskutti, G. (2020). The bias of isotonic regression. *Electronic Journal of Statistics*, 14:801.

Dimitriadis, T., Patton, A. J., and Schmidt, P. (2019). Testing forecast rationality for measures of central tendency. *arXiv preprint arXiv:1910.12545*.

Dümbgen, L., Freitag, S., and Jongbloed, G. (2004). Consistency of concave regression with an application to current-status data. *Mathematical Methods of Statistics*, 13:69–81.

Dykstra, R., Kochar, S., Robertson, T., et al. (1991). Statistical inference for uniform stochastic ordering in several populations. *Annals of Statistics*, 19:870–888.

El Barmi, H. and Marchev, D. (2009). New and improved estimators of distribution functions under second-order stochastic dominance. *Journal of Nonparametric Statistics*, 21:143–153.

El Barmi, H. and Mukerjee, H. (2005). Inferences under a stochastic ordering constraint: the k-sample case. *Journal of the American Statistical Association*, 100:252–261.

El Barmi, H. and Mukerjee, H. (2012). Peakedness and peakedness ordering. *Journal of Multivariate Analysis*, 111:222–233.

El Barmi, H. and Mukerjee, H. (2016). Consistent estimation of survival functions under uniform stochastic ordering; the k-sample case. *Journal of Multivariate Analysis*, 144:99–109.

El Barmi, H. and Wu, R. (2017). On estimation of peakedness-ordered distributions. *Communications in Statistics-Theory and Methods*, 46:4855–4869.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106:746–762.

Goovaerts, M., De Vylder, F., and Haezendonck, J. (1982). Ordering of risks: a review. *Insurance: Mathematics and Economics*, 1:131–161.

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, on behalf of RISC-19-ICU Investigators for Switzerland, P. D., and Ziegel, J. F. (2021a). Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units. *PLOS ONE*, 16:1–14.

Henzi, A., Kleger, G.-R., and Ziegel, J. F. (2021b). Distributional (single) index models. *Journal of the American Statistical Association*. Forthcoming.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021c). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:963–993.

Li, Q. and Racine, J. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26:423–434.

Mösching, A. and Dümbgen, L. (2020a). Estimation of a likelihood ratio ordered family of distributions–with a connection to total positivity. *arXiv preprint arXiv:2007.11521*.

Mösching, A. and Dümbgen, L. (2020b). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.

Rojo, J. and Batún-Cutz, J. (2007). Estimation of symmetric distributions subject to a peakedness order. In *Advances In Statistical Modeling And Inference: Essays in Honor of Kjell A Doksum*, pages 649–669. World Scientific.

Rojo, J. and El Barmi, H. (2003). Estimation of distribution functions under second order stochastic dominance. *Statistica Sinica*, 13:903–926.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer Series in Statistics. Springer, New York.

Yang, F., Barber, R. F., et al. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, 13:646–677.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Annals of Statistics*, 30:528–555.

A Greatest convex minorants

Let $I \subseteq \mathbb{R}$ be an interval and $f : I \to \mathbb{R}$ a function. The greatest convex minorant of f is the pointwise greatest convex function g such that $g(x) \leq f(x)$ for all $x \in I$. It exists if and only if f can be bounded from below by an affine linear function, and if the greatest convex minorant exists, it is unique since the pointwise supremum of convex functions is again convex. By the same reason, if f_1 and f_2 are functions with greatest convex minorants g_1 and g_2, then $f_1(x) \geq f_2(x)$ for all x implies that also $g_1 \geq g_2$.

A standard result about isotonic regression (see e.g. Robertson et al., 1988, Theorem 1.2.1) states that the isotonic regression of $z_1, \dots, z_r$ with weights $w_1, \dots, w_r > 0$, that is, the minimizer of

$$\sum_{i=1}^{r} w_i(\theta_i - z_i)^2$$

over all $\theta_1 \leq \dots \leq \theta_r$, equals the left-hand slope of the greatest convex minorant to the function that results from linearly interpolating

$$(0, 0), \ \left(\sum_{i=1}^{k} w_i, \sum_{i=1}^{k} w_k z_k \right), \ k = 1, \dots, r.$$

This result allows to describe right-hand slope of the greatest convex minorant of any piecewise linear function with finitely many knots.

Lemma A.1. *Let $f : [t_1, t_k] \to \mathbb{R}$ be piecewise linear with knots at $t_1 < \dots < t_k$ and let g be its greatest convex minorant. Then the right-hand slope of g at $t_1, \dots, t_{k-1}$ is given by the isotonic regression of $[f(t_{i+1}) - f(t_i)]/[t_{i+1} - t_i]$ with weights $t_{i+1} - t_i$, $i = 1, \dots, k - 1$.*

The following lemma is known as Marshall's Inequality.

Lemma A.2. *Let $\mathcal{I} \subseteq \mathbb{R}$ be an interval and $f : \mathcal{I} \to \mathbb{R}$ a function, and let g be the greatest convex minorant of f and $h : \mathcal{I} \to \mathbb{R}$ any convex function. Assume that $\|f - h\|_\infty < \infty$, where $\| \cdot \|_\infty$ is the usual supremum norm of functions. Then,*

$$\|g - h\|_\infty \leq \|f - h\|_\infty.$$

Proof. Let $\varepsilon = \|f - h\|_\infty$. The function $\tilde{h}(x) = h(x) - \varepsilon$ is convex and satisfies $f(x) \geq \tilde{h}(x)$ for all $x \in \mathcal{I}$ by definition of ε. This and the definition of g imply that

$$f(x) \geq g(x) \geq h(x) - \varepsilon, \, x \in \mathcal{I}.$$

Since also $f(x) - h(x) \leq \varepsilon$ by the definition of ε, this yields

$$-\varepsilon \leq g(x) - h(x) \leq f(x) - h(x) \leq \epsilon,$$

and so

$$\|g - h\|_\infty \leq \varepsilon = \|f - h\|_\infty.$$

□

B Proofs for Section 2

Proof of Proposition 2.1. Formula (1) shows that $\tilde{M}_{x_i}(y)$ is decreasing in i and increasing in y when the respective other argument is fixed, and

$$\tilde{M}_{x_i}(y) = 0,\ y \le y_1, \quad \tilde{M}_{x_i}(y_m + t) = \tilde{M}_{x_i}(y_m) + t,\ t > 0. \tag{5}$$

In particular, it follows that the greatest convex minorant $\hat{M}_{x_i}$ of $\tilde{M}_{x_i}$ exists. For $k, j \in \{1, \ldots, d\}$ with $k \le j$, the functions

$$y \mapsto \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s h_s(y).$$

are piecewise linear with finitely many knots, a property which is preserved when taking pointwise maxima and minima of finitely many functions. Therefore, the $\tilde{M}_{x_i}$ are also piecewise linear. For any $i \in \{1, \ldots, d\}$, $y \in \mathbb{R}$ and $t > 0$,

$$\begin{aligned} 0 \le \tilde{M}_{x_i}(y+t) &= \min_{k=1,\ldots,i} \max_{j=k,\ldots,d} \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s h_s(y+t) \\ &\le \min_{k=1,\ldots,i} \max_{j=k,\ldots,d} \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s [h_s(y) + t] = \tilde{M}_{x_i}(y) + t, \end{aligned}$$

so $0 \le [\tilde{M}_{x_i}(y+t) - \tilde{M}_{x_i}(y)]/t$, and hence $\hat{M}_{x_i}(y)$ is increasing in y. Lemma A.1 and (5) together with the inequality $[\tilde{M}_{x_i}(y+t) - \tilde{M}_{x_i}(y)]/t \le 1$ imply that $\hat{F}_{x_i} \in [0, 1]$ with $\hat{F}_{x_i}(y) = 0$ for $y < y_1$ and $\hat{F}_{x_i}(y) = 1$ for $y \ge y_m$, and $\hat{F}_{x_i}$ is continuous from the right and increasing because is is the right-hand derivative of a convex function. Finally, $\hat{M}_{x_i}(y)$ is decreasing in i because $\tilde{M}_{x_i}(y)$ is pointwise decreasing in i for all fixed y; see Appendix A. □

C Proofs for Section 3

Proof of Proposition 3.1. The proof is similar to the proof of Corollary 1 in Dümbgen et al. (2004). With $(c_n)_{n\in\mathbb{N}}$ from (A), define

$$A_n = \left\{ \sup_{y \in \mathbb{R},\, x \in I_n} |\tilde{M}_{n;x}(y) - M_x(y)| < c_n \right\}.$$

Then $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$, and in the following derivations, assume that the inequality in A_n holds. In case (i), let $v_n = c_n^{1/(1+\beta)}$. For $x \in I_n$, by convexity of $\hat{M}_x(\cdot)$,

$$\frac{\hat{M}_{n;x}(y) - \hat{M}_{n;x}(y - v_n)}{v_n} \le \hat{F}_{n;x}(y) \le \frac{\hat{M}_{n;x}(y + v_n) - \hat{M}_{n;x}(y)}{v_n},$$

and the same property holds for F_x and M_x instead of $\hat{F}_{n;x}$ and $\hat{M}_{n;x}$. The function $M_x(\cdot)$ is convex, so due to Lemma A.2,

$$\sup_{y\in\mathbb{R}} |\hat{M}_{n;x}(y) - M_x(y)| \le \sup_{y\in\mathbb{R}} |\tilde{M}_{n;x}(y) - M_x(y)|.$$

Combining these facts yields, for any $y \in J_n$,

$$\begin{aligned}\hat{F}_{n;x}(y) &\geq \frac{\hat{M}_{n;x}(y) - \hat{M}_{n;x}(y - v_n)}{v_n} \\ &\geq \frac{M_x(y) - |\hat{M}_{n;x}(y) - M_x(y)| - M_x(y - v_n) - |\hat{M}_{n;x}(y - v_n) - M_x(y - v_n)|}{v_n} \\ &\geq F_x(y - v_n) - 2c_n/v_n \\ &\geq F_x(y) - Cv_n^{\beta} - 2c_n/v_n = F_x(y) - (2 + C)c_n^{\beta/(1+\beta)},\end{aligned}$$

and similarly

$$\hat{F}_{n;x}(y) \leq \frac{\hat{M}_{n;x}(y + v_n) - \hat{M}_{n;x}(y)}{v_n} \leq F_x(y) + (2 + C)c_n^{\beta/(1+\beta)}.$$

Thus $|\hat{F}_{n;x}(y) - F_x(y)| \leq (2 + C)c_n^{\beta/(1+\beta)}$ with on A_n for $x \in I_n$ and $y \in J_n$, for each $n \in \mathbb{N}$. Under (ii), for $y \in \mathbb{Z}$ and $x \in I_n$,

$$\hat{F}_{n;x}(y) = \hat{M}_{n;x}(y + 1) - \hat{M}_{n;y}(y) \leq M_x(y + 1) - M_x(y) + 2c_n = F_x(y) + 2c_n,$$

and analogously $\hat{F}_{n;x}(y) \geq F_x(y) - 2c_n$, which gives $|\hat{F}_{n;x}(y) - F_x(y)| \leq 2c_n$ For $y \in \mathbb{R} \setminus \mathbb{Z}$, the same bound is valid since $F_x(y) = F_x(\lfloor y \rfloor)$ and $\hat{F}_{n;x}(y) = \hat{F}_{n;x}(\lfloor y \rfloor)$, where the latter holds if $\tilde{M}_{n;x}(y)$ and $\hat{M}_{n;x}(y)$ are only computed at $y \in \mathbb{Z}$ and interpolated linearly. □

The proof of Theorem 3.2 requires several auxiliary results.

Proposition C.1. *Let $Z_1, \ldots, Z_k$ be random variables with values in a non-degenerate interval $[a, b] \subset \mathbb{R}$. Then there exists a universal constant $M \leq 2^{5/2}e$ such that for all $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{z \in \mathbb{R}} \frac{1}{\sqrt{k}} \Big| \sum_{i=1}^{k} (z - Z_i)_+ - \mathbb{E}[(z - Z_i)_+] \Big| \geq \varepsilon \right) \leq M \exp\left(\frac{-2\varepsilon^2}{(b - a)^2}\right)$$

Proof. Let F_i be the cumulative distribution function of Z_i. The assumption $F_i(z) = 0$ for $s < a$ implies that $\mathbb{E}[(z - Z_i)_+] = \int_a^z F_i(z)\, ds$, so

$$\begin{aligned}\frac{1}{\sqrt{k}} \Big| \sum_{i=1}^{k} (z - Z_i)_+ - \mathbb{E}[(z - Z_i)_+] \Big| &= \frac{1}{\sqrt{k}} \Big| \sum_{i=1}^{k} \int_a^z \mathbb{1}\{Z_i \leq s\} - F_i(s)\, ds \Big| \\ &\leq \frac{1}{\sqrt{k}} \int_a^z \Big| \sum_{i=1}^{k} \mathbb{1}\{Z_i \leq s\} - F_i(s) \Big|\, ds \\ &\leq \frac{1}{\sqrt{k}} \int_a^z \sup_{u \in \mathbb{R}} \Big| \sum_{i=1}^{k} \mathbb{1}\{Z_i \leq u\} - F_i(u) \Big|\, ds \\ &= \frac{1}{\sqrt{k}} (b - a) \sup_{u \in \mathbb{R}} \Big| \sum_{i=1}^{k} \mathbb{1}\{Z_i \leq u\} - F_i(u) \Big|.\end{aligned}$$

Theorem 4.6 of Mösching and Dümbgen (2020b) now yields the result. □

For $\gamma > 0$ and $z \in \mathbb{R}$, let $t_\gamma(z) = \min(\max(-\gamma, z), \gamma)$. The following inequality, which follows by simple case distinctions, will be applied several times: For all $y, z \in \mathbb{R}$,

$$|(y - z)_+ - (y - t_\gamma(z))_+| \le (\gamma + z)_- + (z - \gamma)_+, \tag{6}$$

where $(x)_- = \max(0, -x)$ and $(x)_+ = \max(0, x)$ for $x \in \mathbb{R}$.

Lemma C.2. *Let Z be a random variable such that for some $z_0 > 0$ and all $z \ge z_0$,*

$$\mathbb{P}(|Z| \ge z) \le \begin{cases} z^{-\lambda}, & \text{for some } \lambda > 1, \text{ or} \\ \exp(-\lambda z), & \text{for some } \lambda > 0. \end{cases}$$

Then for $\gamma \ge z_0$,

$$\mathbb{E}\left(\sup_{z \in \mathbb{R}} |(z - Z)_+ - (z - t_\gamma(Z))_+|\right) \le \begin{cases} \gamma^{-\lambda+1}/(\lambda - 1), & \text{or} \\ \exp(-\lambda\gamma)/\lambda. \end{cases}$$

Proof. Replacing z by the random variable Z in (6) implies that for all $\gamma \ge 0$,

$$\mathbb{E}\Big(\sup_{z \in \mathbb{R}} |(z - Z)_+ - (z - t_\gamma(Z))_+|\Big) \le \mathbb{E}[(\gamma + Z)_- + (Z - \gamma)_+].$$

To compute the expected value in the upper bound, let F denote the cumulative distribution function of Z. Then,

$$\mathbb{E}[(\gamma + Z)_-] = \int_{-\infty}^{-\gamma} F(z)\, ds, \quad \mathbb{E}[(Z - \gamma)_+] = \int_\gamma^\infty 1 - F(z)\, ds.$$

This implies

$$\mathbb{E}\Big(\sup_{z \in \mathbb{R}} |(z - Z)_+ - (z - t_\gamma(Z))_+|\Big) \le \int_\gamma^\infty F(-s) + (1 - F(s))\, ds = \int_\gamma^\infty \mathbb{P}(|Z| \ge s)\, ds.$$

In the first case, for $\gamma \ge z_0$, it holds $\int_\gamma^\infty \mathbb{P}(|Z| \ge s)\, ds \le \gamma^{-\lambda+1}/(\lambda - 1)$. In the second case, the upper bound is $\exp(-\lambda\gamma)/\lambda$. □

Proposition C.1 and Lemma C.2 allow to derive an analogous result to Corollary 4.7 of Mösching and Dümbgen (2020b), for which some additional notation is required. For $y \in \mathbb{R}$ and $r, s \in \{1, \ldots, n\}$, $r \le s$, define $w_{rs} = s - r + 1$ and

$$\mathbb{M}_{rs}(y) = \frac{1}{w_{rs}} \sum_{i=r}^{s} (y - Y_{ni})_+, \quad \bar{M}_{rs}(y) = \frac{1}{w_{rs}} \sum_{i=r}^{s} \mathbb{E}[(y - Y_{ni})_+].$$

Recall that the estimator $\tilde{M}_{n;x_i}$ has the representation

$$\tilde{M}_{n;x_i}(y) = \min_{k=1,\ldots,i} \max_{j=k,\ldots,d} \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s h_s(y),$$

for the distinct values $x_1 < \cdots < x_d$ of $X_{n1}, \ldots, X_{nn}$, $w_i = \#\{j \le n : X_{nj} = x_i\}$, and

$$h_i(y) = \frac{1}{w_i} \sum_{j\colon X_j = x_i} (y - Y_{nj})_+, \; i = 1, \ldots, d.$$

For fixed $i \in \{1, \dots, d\}$, let $1 \le r(i) \le s(i) \le d$ be indices such that

$$\tilde{M}_{n;x_i}(y) = \frac{1}{\sum_{k=r(i)}^{s(i)} w_k} \sum_{k=r(i)}^{s(i)} w_k h_k(y).$$

Assuming $X_{n1} \le \cdots \le X_{nn}$, with $\tilde{r}(x) = \min\{j \le n : X_{nj} = x_{r(i)}\}$, $\tilde{s}(x) = \max\{j \le n : X_{nj} = x_{r(i)}\}$, the estimator $\tilde{M}_{n;x_i}(y)$ equals

$$\tilde{M}_{n;x_i}(y) = \frac{1}{\tilde{s}(i) - \tilde{r}(i) + 1} \sum_{k=\tilde{r}(i)}^{\tilde{s}(i)} (y - Y_{nk})_+.$$

This implies that

$$\max_{1 \le r \le s \le d} \left\| \frac{1}{\sum_{k=r}^{s} w_k} \sum_{k=r}^{s} w_k (h_k - M_{x_k}) \right\|_\infty \le \max_{1 \le r \le s \le n} \|\mathbb{M}_{rs} - \bar{M}_{rs}\|_\infty,$$

and an asymptotic upper bound for $\max_{1 \le r \le s \le n} \|\mathbb{M}_{rs} - \bar{M}_{rs}\|_\infty$ is derived below.

Proposition C.3. *Let $R_n = \max_{1 \le r \le s \le n} w_{rs}^{1/2} \|M_{rs} - \bar{M}_{rs}\|_\infty$. Then for any $D > 2$,*

$$\lim_{n \to \infty} \mathbb{P}\left(R_n \le D \log(n)^{1/2} \gamma_n\right) = 1,$$

where

$$\gamma_n = \begin{cases} (n \log(n))^{1/\lambda}, & \textit{under (P)}, \\ 2\log(n)/\lambda, & \textit{under (E)}. \end{cases}$$

Proof. For $\gamma > 0$, define

$$u(\gamma) = \begin{cases} \gamma^{-\lambda+1}/(\lambda - 1), & \text{under (P)}, \\ \exp(-\lambda\gamma)/\lambda, & \text{under (E)}, \end{cases} \qquad p(\gamma) = \begin{cases} \gamma^{-\lambda}, & \text{under (P)}, \\ \exp(-\lambda\gamma), & \text{under (E)}. \end{cases}$$

By Lemma C.2, for any $y \in \mathbb{R}$ and $\gamma \ge y_0$,

$$\frac{1}{w_{rs}} \left| \sum_{i=r}^{s} \mathbb{E}[(y - Y_{ni})_+] - \mathbb{E}[(y - t_\gamma(Y_{ni}))_+] \right| \le u(\gamma).$$

Also by (P) or (E) and by (6),

$$\mathbb{P}\left(\sup_{y \in \mathbb{R}} |(y - Y_{ni})_+ - (y - t_\gamma(Y_{ni}))_+| > 0\right) \le \mathbb{P}(|Y_{ni}| \ge \gamma) \le p(\gamma).$$

This implies that the events

$$B_n = \left\{ \sup_{y \in \mathbb{R}, i=1,\dots,n} |(y - Y_{ni})_+ - (y - t_\gamma(Y_{ni}))_+| = 0 \right\}$$

satisfy $\mathbb{P}(B_n) \ge 1 - np(\gamma)$. Let ${}_\gamma\mathbb{M}_{rs}$ and ${}_\gamma\bar{M}_{rs}$ be defined as $\mathbb{M}_{rs}$ and $\bar{M}_{rs}$ but with the truncated variables $t_\gamma(Y_{ni})$ instead of Y_{ni}. By the above considerations, conditional on B_n, for any $1 \le r \le s \le n$,

$$\|\mathbb{M}_{rs} - \bar{M}_{rs}\|_\infty = \sup_{y \in \mathbb{R}} \frac{1}{w_{rs}} \left| \sum_{i=r}^{s} (y - Y_{ni})_+ - \mathbb{E}[(y - Y_{ni})_+] \right| \le \|{}_\gamma\mathbb{M}_{rs} - {}_\gamma\bar{M}_{rs}\|_\infty + u(\gamma)$$

Proposition C.1 implies that

$$\mathbb{P}\left(\sup_{y\in\mathbb{R}} w_{rs}^{1/2}|\,{}_{\gamma}\mathbb{M}_{rs}(y) - {}_{\gamma}\bar{M}_{rs}(y)| \geq \varepsilon\right) \leq M \exp\left(\frac{-2\varepsilon^2}{(2\gamma)^2}\right).$$

Replace now γ by

$$\gamma_n = \begin{cases} [n\log(n)]^{1/\lambda}, & \text{under (P)}, \\ 2\log(n)/\lambda, & \text{under (E)}. \end{cases}$$

This yields

$$n \cdot p(\gamma_n) = \begin{cases} n[n\log(n)]^{-\lambda/\lambda} = \log(n)^{-1}, \\ n\exp(-2\lambda\log(n)/\lambda) = n^{-1}, \end{cases}$$

and therefore $\lim_{n\to\infty}\mathbb{P}(B_n) = 1$. Also,

$$n^{1/2}\cdot u(\gamma_n) = \begin{cases} n^{1/2}[n\log(n)]^{(1-\lambda)/\lambda}/(\lambda-1) = n^{-1/2+1/\lambda}\log(n)^{1/\lambda-1}/(\lambda-1), \\ n^{1/2}\exp(-2\lambda\log(n)/\lambda)/\lambda = n^{-3/2}/\lambda, \end{cases}$$

which gives $\lim_{n\to\infty} n^{1/2}\cdot u_n = 0$, using $\lambda > 2$ in the first case. For $\delta > 0$, define $\varepsilon_n = 2(1+\delta)\log(n)^{1/2}\gamma_n$. Then, for n large enough such that $n^{1/2}u(\gamma_n) \leq \delta\log(n)^{1/2}\gamma_n$, and by conditioning on B_n,

$$\begin{aligned}
\mathbb{P}(R_n \geq \varepsilon_n) &\leq \sum_{1\leq r\leq s\leq n} \mathbb{P}(w_{rs}^{1/2}\|\mathbb{M}_{rs} - \bar{M}_{rs}\|_\infty \geq \varepsilon_n) \\
&\leq \sum_{1\leq r\leq s\leq n} \mathbb{P}\left(w_{rs}^{1/2}\|\,{}_{\gamma_n}\mathbb{M}_{rs} - {}_{\gamma_n}\bar{M}_{rs}\|_\infty + w_{rs}^{1/2}u(\gamma_n) \geq \varepsilon_n\right) \\
&\leq \sum_{1\leq r\leq s\leq n} \mathbb{P}\left(w_{rs}^{1/2}\|\,{}_{\gamma_n}\mathbb{M}_{rs} - {}_{\gamma_n}\bar{M}_{rs}\|_\infty + n^{1/2}u(\gamma_n) \geq \varepsilon_n\right) \\
&\leq \sum_{1\leq r\leq s\leq n} \mathbb{P}\left(w_{rs}^{1/2}\|\,{}_{\gamma_n}\mathbb{M}_{rs} - {}_{\gamma_n}\bar{M}_{rs}\|_\infty \geq 2(1+\delta/2)\log(n)^{1/2}\gamma_n\right) \\
&\leq \frac{Mn(n+1)}{2}\exp\left(-\frac{8(1+\delta/2)^2\log(n)\gamma_n^2}{(2\gamma_n)^2}\right) \\
&\leq \frac{M}{2}\exp(2\log(n+1) - 2(1+\delta/2)^2\log(n)) \to 0,\ n\to\infty. \qquad \square
\end{aligned}$$

Proof of Theorem 3.2, discrete setting (K). For $j = 1,\ldots,K$, let $A_j = \{i \in \{1,\ldots,n\} : X_{ni} = j\}$, and define $\check{M}_{n;j} = \sum_{i\in A_i}(y - Y_{ni})_+/\#A_i$. Recall that $\tilde{M}_{n;j}(y)$ is the antitonic regression of $(X_{ni}, (y-Y_{ni})_+)$, $i = 1,\ldots,n$. Corollary B of Robertson et al. (1988, p. 42) implies that for all $y \in \mathbb{R}$,

$$\max_{j=1,\ldots,K}|M_j(y) - \tilde{M}_{n,j}(y)| \leq \max_{j=1,\ldots,K}|M_j(y) - \check{M}_{n,j}(y)|$$

This gives

$$\max_{j=1,\ldots,K}\|M_j - \tilde{M}_{n,j}\|_\infty \leq \max_{j=1,\ldots,K}\|M_j - \check{M}_{n,j}\|_\infty.$$

Assume that $X_{n1} \leq \cdots \leq X_{nn}$, and define $k(j) = \max\{k \in \{1,\ldots,n\} : X_{nk} = j\}$ for $j = 1,\ldots,K$, and $k(0) = 0$. Then $\#A_j = k(j) - k(j-1)$, and by assumption (K),

$$\min_{j=1,\ldots,K}\frac{k(j)-k(j-1)}{n} = \frac{\#A_j}{n} \geq p/2.$$

with asymptotic probability one. Since $\check{M}_{n;j}(y) = \mathbb{M}_{(k(j-1)+1),k(j)}(y)$ and $w_{(k(j-1)+1),k(j)} = \#A_j$, Proposition C.3 implies that, with asymptotic probability one for any $D > 2$ and $j = 1, \dots, K$,

$$\|\check{M}_{n;j} - M_j\|_\infty \leq (w_{(k(j-1)+1),k(j)})^{-1/2} R_n \leq \left(\frac{np}{2}\right)^{-1/2} R_n \leq D\gamma_n \left(\frac{2}{p}\right)^{1/2} \left(\frac{\log(n)}{n}\right)^{1/2}.$$

With $D = \sqrt{8} > 2$, the upper bound equals

$$c_n = \begin{cases} 4p^{-1/2} n^{-1/2+1/\lambda} \log(n)^{1/2+1/\lambda}, & \text{under (P)}, \\ 8p^{-1/2}\lambda^{-1} n^{-1/2} \log(n)^{3/2}, & \text{under (E)}. \quad \square \end{cases}$$

Proof of Theorem 3.2, continuous setting (C1), (C2). With Proposition C.3, one can apply the same strategy of proof as for Theorem 3.3 in Mösching and Dümbgen (2020b). Let δ_n be a sequence such that $\lim_{n\to\infty} \delta_n = 0$ and $\lim_{n\to\infty} n\delta_n/\log(n) = \infty$. By assumption (C1) and by the result in Section 4.3 of Mösching and Dümbgen (2020b), for all subintervals $\mathcal{I} \subseteq I$ of length at least δ_n and any $q \in (0, p)$, the inequality $\{i \leq n : X_{ni} \in \mathcal{I}\} \geq qn\delta_n$ holds with asymptotic probability one. Let $x \in I$ such that $x - \delta_n \in I$, and define

$$r(x) = \min\{i \leq n : X_{ni} \geq x - \delta_n\}, \ j(x) = \max\{i \leq n : X_{ni} \leq x\}.$$

By the above considerations, with asymptotic probability one, $r(x)$ and $j(x)$ are well-defined, satisfy $r(x) \leq j(x)$, $x - \delta_n \leq X_{nr(x)} \leq X_{nj(x)} \leq x$, and $\#\{j \leq n : X_{nj} \in [x - \delta_n, x]\} \geq qn\delta_n$. Therefore, for any $y \in \mathbb{R}$,

$$\begin{aligned} \tilde{M}_{n;x}(y) - M_x(y) &\leq \tilde{M}_{n;x_{j(x)}}(y) - M_x(y) \\ &= \min_{k=1,\dots,i} \max_{j=k,\dots,d} \frac{1}{\sum_{s=k}^{j} w_s} \sum_{s=k}^{j} w_s h_s(y) - M_x(y) \\ &\leq \max_{n \geq s \geq j(x)} \mathbb{M}_{r(x)s}(y) - M_x(y) \\ &\leq (qn\delta_n)^{-1/2} R_n + \max_{n \geq s \geq j(x)} \bar{M}_{r(x)s}(y) - M_x(y) \\ &\leq (qn\delta_n)^{-1/2} R_n + M_{x_{r(x)}}(y) - M_x(y) \qquad (7) \\ &\leq (qn\delta_n)^{-1/2} R_n + L\delta_n, \qquad (8) \end{aligned}$$

using antitonicity of $t \mapsto \tilde{M}_t(y)$ in the first line, equation (1) in the second line, and antitonicity of $t \mapsto M_t(y)$ in the second-last step. An analogous argument for $M_x(y) - \tilde{M}_{n;x}(y)$ and the asymptotic bound for R_n in Proposition C.3 yield

$$|\tilde{M}_{n;x}(y) - M_{n,x}(y)| \leq (qn\delta_n)^{-1/2} \cdot D \log(n)^{1/2} \gamma_n + L\delta_n.$$

for $D > 2$. The convergence rates of these two summands are balanced if $\delta_n = (\log(n)/n)^{1/3} \gamma_n^{2/3}$, and for $D = \sqrt{8}$ and $q = p/2$, the upper bound equals

$$c_n = \begin{cases} [4p^{-1/2} + L] n^{-1/3+2/(3\lambda)} \log(n)^{1/3+2/(3\lambda)}, & \text{under (P)}, \\ [4p^{-1/2} + L](2/\lambda)^{2/3} n^{-1/3} \log(n), & \text{under (E)}. \end{cases}$$

□

In Section 2, it is suggested to estimate $\tilde{M}_x(y)$ and $\hat{M}_x(y)$ only on a finite grid $t_1, \ldots, t_k$. Below is a proof that this indeed does not influence the convergence rates, provided that $t_1 = y_1$, $t_k = y_m$, and that the grid is fine enough.

Proof that convergence rates are valid under interpolation. Assume that (A) holds, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{y\in J_n, x\in I_n} |\tilde{M}_{n;x}(y) - M_x(y)| \geq c_n \right) = 0$$

for some sequences of sets $I_n, J_n \subseteq \mathbb{R}$. Let $\tilde{m}_{n;x}$ be the linear interpolation of $\tilde{M}_{n;x}$ computed on this grid. That is, for $y \in (t_i, t_{i+1}]$, set $\tilde{m}_{n;x}(y) = \lambda \tilde{M}_x(t_i) + (1-\lambda)\tilde{M}_{n;x}(t_{i+1})$ with $\lambda = (t_{i+1} - y)/(t_{i+1} - t_i)$, and $\tilde{m}_{n;x}(y) = 0 = \tilde{M}_{n;x}(y)$ for $y \leq y_1 = t_1$ and $\tilde{m}_{n;x}(y) = \tilde{M}_{n;x}(y_m) + (y - y_m) = \tilde{M}_{n;x}(t_k) + (y - t_k)$ for $y \geq y_m = t_k$. Then, since $M_x(\cdot)$ is Lipschitz continuous with Lipschitz constant 1,

$$\begin{aligned}&|\tilde{m}_{n;x}(y) - M_x(y)| \leq \\ &\quad \max\Big(|\tilde{M}_{n;x}(t_i) - M_x(t_i)| + |t_i - y|, |\tilde{M}_{n;x}(t_{i+1}) - M_x(t_{i+1})| + |t_{i+1} - y| \Big)\end{aligned}$$

for all $y \in \mathbb{R}$. Provided that $\sup_{i=1,\ldots,k-1} |t_i - t_{i+1}| \leq c_n$, this implies $\sup_{y\in J_n} |\tilde{m}_{n;x}(y) - M_x(y)| \leq 2c_n$, so the same convergence rate applies if $\tilde{M}_x(y)$ and $\hat{M}_x(y)$ are evaluated on a sufficiently fine grid. If $Y_{n1} < \cdots < Y_{nn}$ are independent and admit a density bounded away from zero on $J \supseteq J_n$, then the results of Section 4.3 in Mösching and Dümbgen (2020b) imply that $\sup_{i=1,\ldots,n-1} |Y_{ni} - Y_{n(i+1)}| \leq c_n$ holds with asymptotic probability one for the c_n from Theorem 3.2, so it is admissible in this case to take the observed values $y_1, \ldots, y_m$ as the grid. □

D Additional figures for Section 4

Figure 4 shows the same comparison as Figure 2 for $n = 1000$ and $n = 1500$. In Figures 5 and 6, different variants of subagging are compared. Using more than $n/2$ of the total data in subsamples is generally not better than $n/2$ or less. A higher number of subsamples improves the subagging variants of the estimators, but the effect diminishes as the number of subsamples increases.

Figure 4: Relative improvement in L_1 distance and mean absolute error of quantile estimates of the $\preceq_{\mathrm{icv}}$- and $\preceq_{\mathrm{icx}}$-order constrained estimator compared to the estimator under first order stochastic dominance, for $n = 1000$ and $n = 1500$. The solid lines show the improvement when the estimators are computed on the full sample, and the dashed lines for a subagging variant with 50 subamples of size $n/2$.

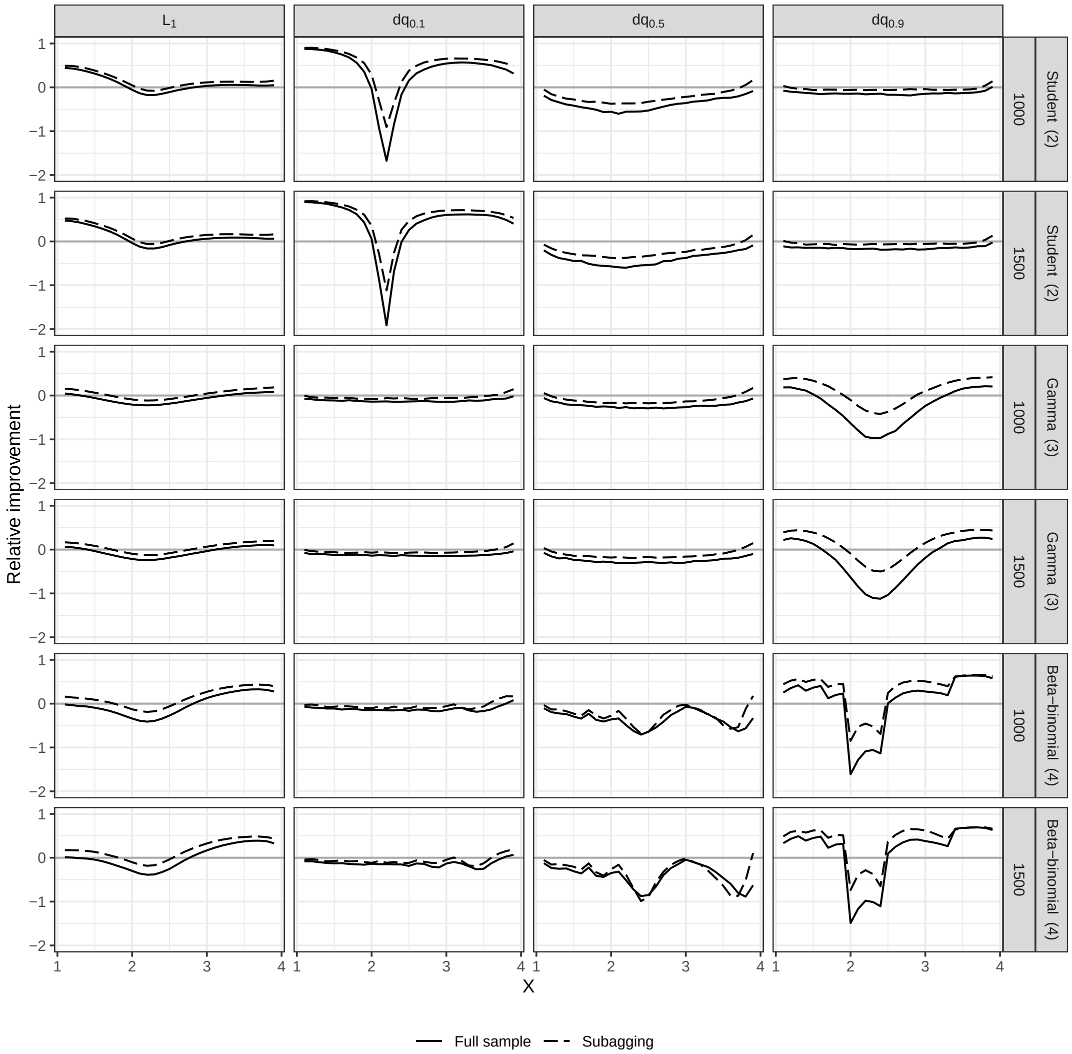


Figure 5: Relative improvement of subagging variants of the $\preceq_{\mathrm{icv}}$- and $\preceq_{\mathrm{icx}}$-constrained estimators (ICV/ICX) and of the estimator with first order stochastic dominance constraints (FSD) compared to the version without subagging. The sample size is $n = 1000$ fraction of data in each subsample is $n/2 = 500$.

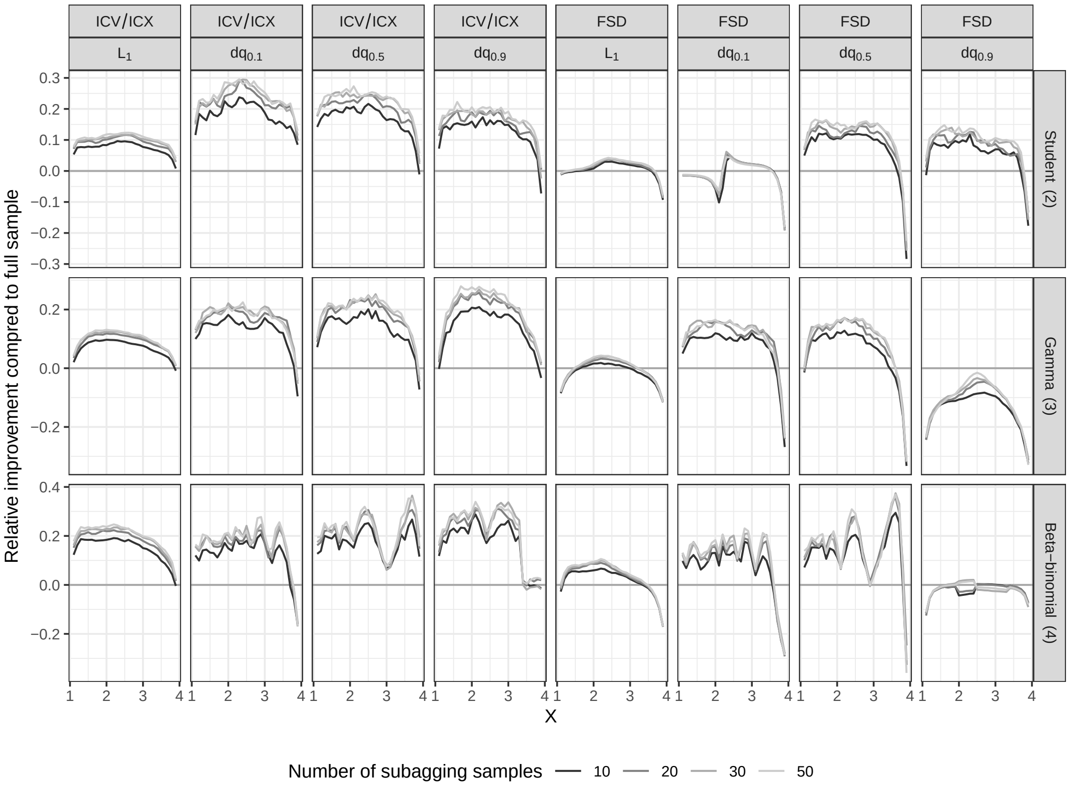
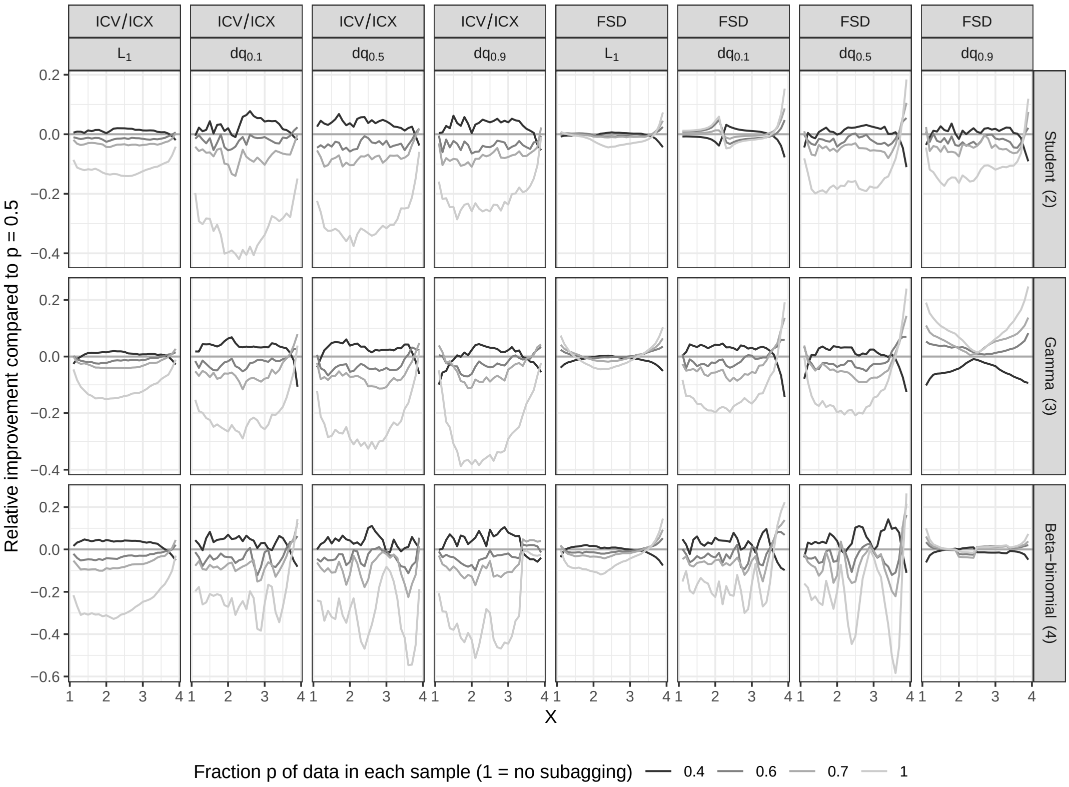


Figure 6: Relative improvement of subagging versions of $\preceq_{\text{icv}}$- and $\preceq_{\text{icx}}$-constrained estimators (ICV/ICX) and of the estimator with first order stochastic dominance constraints (FSD), compared to the variant with subsamples of size $n/2$. The sample size is $n = 1000$ and the number of subsamples is 50 for the variants with subagging.



Chapter 3

Distributional index models

3.1 Distributional (single) index models

The content of this section is published as

Henzi, A., Kleger, G.-R. and Ziegel, J. F. (2021+). Distributional (single) index models. *Journal of the American Statistical Association*, to appear.

The article is followed by its supplementary material. Both are available on `https://doi.org/10.1080/01621459.2021.1938582`.

Distributional (Single) Index Models

Alexander Henzi, Gian-Reto Kleger, Johanna F. Ziegel*

June 13, 2022

Abstract

A Distributional (Single) Index Model (DIM) is a semi-parametric model for distributional regression, that is, estimation of conditional distributions given covariates. The method is a combination of classical single index models for the estimation of the conditional mean of a response given covariates, and isotonic distributional regression. The model for the index is parametric, whereas the conditional distributions are estimated non-parametrically under a stochastic ordering constraint. We show consistency of our estimators and apply them to a highly challenging data set on the length of stay (LoS) of patients in intensive care units. We use the model to provide skillful and calibrated probabilistic predictions for the LoS of individual patients, that outperform the available methods in the literature.

Keywords: Distributional regression, intensive care unit length of stay, probabilistic forecast, single index model, stochastic ordering constraint

*Alexander Henzi is PhD student in Statistics, Johanna F. Ziegel is Professor of Applied Stochastics, Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland (e-mail: alexander.henzi@stat.unibe.ch, johanna.ziegel@stat.unibe.ch); Gian-Reto Kleger, MD, is Head the Division of Intensive Care Medicine, Cantonal Hospital, St. Gallen, Rorschacherstrasse 95, 9007 St.Gallen, Switzerland (e-mail: gian-reto.kleger@kssg.ch). A. Henzi and J. F. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation. The authors thank the Swiss Society of Intensive Care Medicine for providing the data. The work has greatly benefitted from discussions with Lutz Dümbgen, Tilmann Gneiting, Alexander Jordan, and Alexandre Mösching. This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of the American Statistical Association on 19 July 2021, available online: https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1938582

1 Introduction

Regression approaches for the full conditional distribution of an outcome given covariates are gaining momentum in the literature (Hothorn et al., 2014, and the references therein). They have already become an indispensable tool in probabilistic weather forecasting (Gneiting and Katzfuss, 2014; Vannitsem et al., 2018) but also find numerous applications in other fields such as economics, social sciences and medicine; see e.g. Machado and Mata (2000), Chernozhukov et al. (2013), Klein et al. (2015), Duarte et al. (2017) and Silbersdorff et al. (2018).

If the outcome is real-valued, then conditional distributions can be characterized in terms of their cumulative distribution function (CDF) or quantile function, and various techniques for the estimation of these objects have been proposed. Foresi and Peracchi (1995) and Peracchi (2002) build on the extant methods for the estimation of single quantiles or probabilities (Koenker, 2005), and suggest to approximate the conditional distribution by a cascade of regressions for quantiles or for the CDF evaluated at certain thresholds. A drawback of this approach is that the resulting estimates are not necessarily isotonic (the so-called 'quantile crossing problem') and thus require correction, for which remedies have already been developed, see e.g. Dette and Volgushev (2008); Chernozhukov et al. (2010).

A broad class of methods that directly yield well-defined probability distributions are generalized additive models for location, shape and scale (Rigby and Stasinopoulos, 2005, GAMLSS). They build on generalized linear models (McCullagh and Nelder, 1989, GLM) and generalized additive models for the mean (Hastie and Tibshirani, 1990, GAM) but also allow to model shape and scale parameters as functions of covariates. The GAMLSS framework has has been extended to Bayesian statistics (Umlauf et al., 2018) and combined with popular machine learning techniques such as boosting (Thomas et al., 2018), neural networks (Rasp and Lerch, 2018) and regression forests (Schlosser et al., 2019).

Finally, there are also powerful semi-parametric and nonparametric techniques for the estimation of conditional distributions. Fully nonparametric methods estimate the conditional distribution functions locally, for example by kernel functions (Hall et al., 1999; Dunson et al., 2007; Li and Racine, 2008), or by partitioning of the covariate space, as in quantile random forests (Meinshausen, 2006; Athey et al., 2019). A frequently used semi-parametric distributional regression method is Cox regression (Cox, 1972), which models the hazard rate of the outcome but also allows to derive its survival function. Conditional transformation models (Hothorn et al., 2014) assume a parametric distribution for an unknown monotone transformation of the response, which is estimated along with the model parameters. Hall and Yao (2005); Zhang et al. (2017) propose semi-parametric methods that reduce the dimension of the covariate space by a suitable projection, and then estimate the conditional distributions non-parametrically given the projections by kernel methods.

We introduce a new approach to distributional regression that can be seen as a combination of a single index model with isotonic distributional regression (IDR, Henzi et al., 2019). The dimension reduction of the covariate space achieved by the single index assumption is in the spirit of Hall and Yao (2005); Zhang et al. (2017) but the combination with IDR is new, and has the advantage to be free of any implementation choices or tuning parameters.

Let Y be a real-valued response and X a covariate in some covariate space $\mathcal{X}$. We want to estimate the conditional distribution of Y given X, that is, $\mathcal{L}(Y \mid X)$. To expose the main idea, suppose that $\mathcal{X} = \mathbb{R}^d$. Then, a Distributional (Single) Index Model (DIM) could be

$$\mathbb{P}(Y \leqslant y \mid X) = F_{\alpha_0^\top X}(y), \quad \text{for all } y \in \mathbb{R}, \tag{1}$$

where $\alpha_0 \in \mathbb{R}^d$, $\alpha_0^\top X$ denotes the scalar product between α_0 and X, and $(F_u)_{u \in \mathbb{R}}$ is a family of CDFs such that

$$F_u \preceq_{\mathrm{st}} F_v \quad \text{if } u \leqslant v, \tag{2}$$

where $\preceq_{\mathrm{st}}$ denotes the usual stochastic order, that is $F_u \preceq_{\mathrm{st}} F_v$ if $F_u(y) \geqslant F_v(y)$ for all $y \in \mathbb{R}$. We call $\theta(x) = \alpha_0^T x$ in representation (1) the index (function).

If the parameter α_0 in the previous example (1) is known, then a natural method to estimate the unknown family $(F_u)_u$ of stochastically ordered CDFs is IDR as introduced by Henzi et al. (2019), see also Mösching and Dümbgen (2020). IDR is a nonparametric technique to estimate conditional distributions under stochastic ordering constraints. In brief, IDR works as follows. Given training data $(\vartheta_1, y_1), \ldots, (\vartheta_n, y_n)$, where $\vartheta_i \in \Theta$ for some partially ordered set Θ, IDR yields the unique optimal vector $\hat{\mathbf{F}} = (\hat{F}_1, \ldots, \hat{F}_n)$ of CDFs that minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(F_i, y_i),$$

over all vectors $(F_1, \ldots, F_n)$ of CDFs that respect the stochastic ordering constraints $F_i \preceq_{\mathrm{st}} F_j$ if $\vartheta_i \preceq \vartheta_j$, $i, j = 1, \ldots, n$. Here, for any CDF F and $y \in \mathbb{R}$,

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leqslant z\})^2 \, \mathrm{d}z \tag{3}$$

is the widely applied proper scoring rule called the continuous ranked probability score (CRPS, Matheson and Winkler, 1976; Gneiting et al., 2007). If we have a sample (x_1, y_1), $\ldots$, (x_n, y_n) from $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, we can apply IDR to the training data $(\alpha_0^\top x_1, y_1)$, $\ldots$, $(\alpha_0^\top x_n, y_n)$, that is, we set $\vartheta_i = \alpha_0^\top x_i$, $i = 1, \ldots, n$ and $\Theta = \mathbb{R}$. This yields a distributional regression model for (X, Y) that may be used to provide probabilistic predictions for Y given X, see Henzi et al. (2019, Section 2.5) and Section 4.

DIMs are closely related to generalized linear models, which assume that the conditional distributions $(F_u)_u$ belong to a known exponential family of distributions with mean $\mathbb{E}(Y \mid X = x) = g(\alpha_0^T x)$, where g is a fixed, strictly monotone link function. In fact, the Gaussian, Poisson, Gamma and Binomial GLM can be subsumed under the DIM, since they also satisfy the stochastic ordering constraint on the conditional distributions. Our approach, to leave the conditional distributions $(F_u)_u$ unspecified, is already widely applied in classical regression for the mean, where models of the type $\mathbb{E}(Y \mid X = x) = g(\alpha_0^T x)$ with unknown link function g are called single index models. Typically, g is assumed to be a smooth function and estimated by kernel regression or local polynomial approximation (Härdle et al., 1993) or local polynomial approximation (Carroll et al., 1997; Zou and Zhu, 2014). More recently, shape constrained single index models have been considered with monotone (Balabdaoui et al., 2019a) and convex (Kuchibhotla et al., 2017) link functions. DIMs directly extend monotone single index models for the mean, since the stochastic ordering assumption on the conditional distributions implies an isotonic conditional mean function.

There is a vast literature on the estimation of the index in single index models, and we refer to Lanteri et al. (2020) for a comprehensive overview. In Section 3, we discuss estimators for the index and the distribution functions in DIMs. Briefly, when IDR is used to estimate the conditional distribution functions, then it is sufficient to know the index function up to isotonic transformations, i.e. to find a pseudo index function that approximates the *ordering* implied by the true index. This approach is supported by the asymptotic analysis in Section 5, which shows that when a monotone transformation of the estimated index function is consistent at the parametric rate, then a DIM with that index estimator is consistent.

A major application of distributional regression techniques is forecasting. It has been recognized in many problems, such as weather prediction or economic forecasting, that point forecasts are unable to account for the full forecast uncertainty and should be replaced by probabilistic forecasts (Gneiting and Katzfuss, 2014). Distributional regression methods are statistical tools to provide such probabilistic forecasts. One fundamental contribution of DIMs is that they allow to associate a natural distributional prediction to point forecasts: If a point forecast from a statistical model is taken as the index in a DIM, for example the estimated conditional expected value, then the DIM naturally extends this deterministic forecast to a probabilistic one. Moreover, the only prerequisite is an isotonic relationship between the point forecast and the outcome in a stochastic ordering sense, which is often a natural and intuitive assumption for reasonable point forecasts.

In Section 6, we use a DIM for predictions in a highly challenging dataset on the length of stay (LoS) of intensive care unit (ICU) patients. Accurate LoS predictions could serve as a tool for ICU physicians, for example to plan the number of available beds, or to identify potential long stay patients at an early stage. Moreover, the same models that are used for prediction may also be used for risk-adjustment and benchmarking across different ICUs. In the last twenty years, there have been many approaches to find appropriate regression models for LoS, see Zimmerman et al. (2006); Moran and Solomon (2012); Verburg et al. (2014) for some examples and Verburg et al. (2014); Kramer (2017) for literature reviews. The extant methods typically model the conditional mean and are unsatisfactory when applied for single patient predictions, since the distribution of LoS is strongly right-skewed with a large variance even after conditioning on covariates. We therefore argue that LoS predictions should be probabilistic. In Section 6, we derive calibrated and informative probabilistic forecasts for LoS, and show that the DIM outperforms existing distributional regression methods in terms of predictive accuracy.

2 Distributional index models

In this section, we define the DIM in its most general form. Let Y be a real-valued response, and let X be covariates in some general space $\mathcal{X}$. The link between X and Y is the index function $\theta\colon \mathcal{X} \to \mathbb{R}^d$, where $\mathbb{R}^d$ is equipped with some partial order $\preceq$. Let further $(F_u)_{u\in\mathbb{R}^d}$ be a family of CDFs such that $F_u \preceq_{\text{st}} F_v$ if $u \preceq v$. The DIM then assumes that

$$\mathbb{P}(Y \leqslant y \mid X) = F_{\theta(X)}(y). \tag{4}$$

Due to the stochastic ordering assumption, it directly follows that the conditional distributions are ordered in the index, that is, $\theta(x) \preceq \theta(x')$ implies $F_{\theta(x)} \preceq_{\text{st}} F_{\theta(x')}$.

We assume further that the function θ belongs to a finite dimensional vector space $\mathcal{F}$, i.e. a parametric model for θ. If $\theta_1, \dots, \theta_p$ are a basis of $\mathcal{F}$ and if $d = 1$, then we recover the form $\mathbb{P}(Y \leqslant y \mid \tilde{X} = \tilde{x}) = F_{\alpha_0^T\tilde{x}}(y)$, where $\tilde{x} = (\theta_1(x) \dots, \theta_p(x))$, and hence, the analogy to single index models. However, the estimation procedure suggested in the next section can be applied with any dimension d and any partial order $\preceq$ on $\mathbb{R}^d$.

3 Estimation

Having motivated and formalized the DIM, we propose a method for estimation. Assume that a training dataset (x_i, y_i), $i = 1, \dots, n$, of independent realizations of (X, Y) satisfying the model assumption (4) is available.

In principle, it would be desirable to have a simultaneous estimator for both the index and the distribution functions. In Section 5, we show that simultaneous estimation is possible theoretically, but computationally infeasible. The method we propose here, and for which we provide asymptotic results, is a two-stage estimation in which first the index θ is estimated, say by $\hat{\theta}$, and then the conditional CDFs based on pairs $(\hat{\theta}(x_i), y_i)$. This is inspired by the 'plug-in estimators' for monotone single index models suggested in Balabdaoui et al. (2019a). The estimation procedure is straightforward and reads as follows:

1. Estimate θ with some estimator $\hat{\theta}$ on the data $(x_i, y_i)_{i=1}^n$,
2. compute the in-sample predictions $\vartheta_i = \hat{\theta}(x_i)$, $i = 1, \dots, n$,
3. estimate the distribution functions $\hat{F}_u, u \in \mathbb{R}^d$, using $(\vartheta_i, y_i)_{i=1}^n$.

In the next two subsections, we reverse the order of the estimation procedure and first suggest our method for Step 3, because this has important implications for the choice of the index estimators in Step 1.

3.1 Isotonic distributional regression

Because of model assumption (4), we seek an estimator $\hat{F}_u, u \in \mathbb{R}^d$, such that $\hat{F}_u \preceq_{\mathrm{st}} \hat{F}_v$ if $u \preceq v$, i.e. $\hat{F}_u(y) \geqslant \hat{F}_v(y)$ for all $y \in \mathbb{R}$ and given u, v. For fixed y, this suggests to define $\hat{\boldsymbol{F}} = (\hat{F}_{\vartheta_1}, \ldots, \hat{F}_{\vartheta_n})$ as

$$\hat{\boldsymbol{F}}(y) = \underset{\eta_k \geqslant \eta_l \text{ if } \vartheta_k \preceq \vartheta_l}{\operatorname{argmin}} \sum_{i=1}^{n} (\eta_i - \mathbb{1}\{y_i \leqslant y\})^2. \tag{5}$$

It turns out that (5) indeed yields a collection of well-defined conditional CDFs, and this estimator is called the IDR in Henzi et al. (2019). By Henzi et al. (2019, Theorem 2.2), IDR can equivalently be defined in terms of conditional quantile functions, $\hat{\boldsymbol{q}} = (\hat{q}_{\vartheta_1}, \ldots, \hat{q}_{\vartheta_n})$, where

$$\hat{\boldsymbol{q}}(\alpha) = \underset{\beta_k \leqslant \beta_l \text{ if } \vartheta_k \preceq \vartheta_l}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbb{1}\{y_i \leqslant \beta_i\} - \alpha)(\beta_i - y_i) \tag{6}$$

for any $\alpha \in (0, 1)$, and the argmin is defined as the componentwise smallest minimizer if it is not unique. IDR estimates the conditional distributions non-parametrically under the stochastic order constraints. For IDR, the index u can take values in any partially ordered set Θ. The particular choice of the loss functions, i.e. the squared error for the estimation of probabilities in (5) and the classical quantile loss function in (6), is in fact irrelevant here: Any other consistent loss function for the expectation or quantiles would yield the same result (Henzi et al., 2019; Jordan et al., 2019).

The above estimators are defined when the index u (in $\hat{F}_u$ or $\hat{q}_u$) is in $\{\vartheta_1, \ldots, \vartheta_n\} \subseteq \Theta$. The CDFs or quantile functions for an arbitrary u can be derived by interpolation of $\hat{F}_{\vartheta_1}, \ldots, \hat{F}_{\vartheta_n}$ or $\hat{q}_{\vartheta_1}, \ldots, \hat{q}_{\vartheta_n}$ for $\Theta = \mathbb{R}$, and a suitable generalization thereof for general partially ordered Θ (Henzi et al., 2019, Section 2.5).

The following proposition is a direct consequence of the above formulas. It shows invariance properties of IDR, which make it a suitable method for estimating the conditional distributions in DIMs. We use the notation $\hat{F}_u(y; \boldsymbol{\vartheta}, \boldsymbol{y})$ and $\hat{q}_u(\alpha; \boldsymbol{\vartheta}, \boldsymbol{y})$ for the IDR CDFs and quantile functions estimated with training data $\boldsymbol{\vartheta} = (\vartheta_k)_{k=1}^m$ and $\boldsymbol{y} = (y_k)_{k=1}^m$.

Proposition 3.1 (Invariance of IDR)**.** *Let $\boldsymbol{y} = (y_k)_{k=1}^m \in \mathbb{R}^m$ and $\boldsymbol{\vartheta} = (\vartheta_k)_{k=1}^m \in \Theta^m$, and let Θ' be a partially ordered set with order $\preceq'$. Let further $g : \Theta \to \Theta'$ be such that $\vartheta_k \preceq \vartheta_l$ if and only if $g(\vartheta_k) \preceq' g(\vartheta_l)$ and $h : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Define $g(\boldsymbol{\vartheta}) = (g(\vartheta_k))_{k=1}^m$. Then, for $j = 1, \ldots, m$, $y \in \mathbb{R}$, $\alpha \in (0, 1)$,*

$$\hat{q}_{g(\vartheta_j)}(\alpha;\, g(\boldsymbol{\vartheta}), h(\boldsymbol{y})) = h(\hat{q}_{\vartheta_j}(\alpha; \boldsymbol{\vartheta}, \boldsymbol{y})), \quad \hat{F}_{g(\vartheta_j)}(h(y);\, g(\boldsymbol{\vartheta}), h(\boldsymbol{y})) = \hat{F}_{\vartheta_j}(y; \boldsymbol{\vartheta}, \boldsymbol{y}).$$

Proposition 3.1 shows that when IDR is used to estimate the conditional distributions in Step 3, then it is sufficient to know the index θ up to increasing transformations. Moreover, any isotonic transformation can be applied to the response Y to simplify the estimation of θ in Step 1, and then reverted by its inverse, without affecting the estimation of the conditional distributions. Hence, the task of estimating the index function θ is simplified to finding an estimator for a pseudo index that induces the same ordering on $\theta(x_i)$, $i = 1, \ldots, n$.

3.2 Index estimators

A simple but effective way to estimate the index in DIMs are classical generalized linear models. This might be surprising, because it seems that a parametric assumption has to be imposed on the distribution functions $(F_u)_u$ for this approach. However, due to the invariance of DIMs under monotone transformations (Proposition 3.1), it is sufficient that such a parametric assumption holds only approximately, in the sense that a monotone transformation of the index estimator converges to the index function; see Assumption (A4) in Section 5. The only requirement is that the linear predictor of the GLM exhibits an isotonic relationship with the outcome. This can be verified by the rank correlation between the index and the outcome, or by plots of the empirical distribution of the outcome stratified according to the index. A further advantage of this approach is that GLMs are well-understood, implemented efficiently in nearly every statistical software, and one can directly build on extant literature from non-distributional regression to find a suitable index estimator. The effectiveness of GLMs in the context of DIMs is demonstrated in the data application in Section 6.

Another powerful tool for index estimation in DIMs is quantile regression (Koenker, 2005). The stochastic ordering of the conditional distributions in DIMs is equivalent to the assumption that the conditional quantile functions $q_{\theta(x)}(\alpha)$ are increasing in the index $\theta(x)$ for every $\alpha \in (0, 1)$. One can thus estimate one or several quantiles by quantile regression, e.g. the median and/or the 90% quantile, and obtain estimates of the complete distribution by taking this (these) quantile(s) as the index (vector) in a DIM. Compared to the direct application of quantile regression for the estimation of conditional distributions, one does not need to specify a grid of quantiles over the whole unit interval and correct quantile crossings, but can focus on the estimation of a small number of quantiles that reveal the ordering of the conditional distributions.

In the case of a distributional single index model $F_{\theta(X)}(y) = F_{\alpha_0^T X}(y)$, that is a DIM with $d = 1$, one might estimate the index α_0 via methods for single index models. For the monotone single index model, efficient estimators have been developed recently (Balabdaoui et al., 2019b; Balabdaoui and Groeneboom, 2020). Index estimators for the single index model, such the one proposed in Lanteri et al. (2020), also allow for non-monotone relationships between the index function $\alpha_0^T x$ and the response, and hence monotonicity should be checked carefully. Compared to GLMs as a pseudo index, single index models gain flexibility by not assuming any fixed functional form of the relationship between $\alpha_0^T X$ and the outcome Y. The drawbacks are that it is more difficult to accommodate high dimensional categorical variables and to let numeric covariates enter the index-function in a non-linear fashion, e.g. via polynomial or spline expansions, which is essential in our data application on ICU LoS. Since the DIM is already invariant under monotone transformations of the index function, it is questionable whether the benefits of using single index methods surpass these drawbacks. The same concerns are also valid for estimation methods for distributional single index models in the spirit of Hall and Yao (2005), which requires a notion of distance on the covariate space and is hence not directly applicable when categorical covariates are present.

3.3 Extension: Sample splitting and bagging

The estimation procedure suggested so far uses in-sample predictions with the estimated index function, $\hat{\theta}(x_i)$, as covariates for distributional regression with IDR. Depending on the index estimator, this strategy may be prone to overfitting. As a remedy, we propose a procedure in the spirit of (sub)sample aggregation (bagging).

Instead of estimating both the index function and the conditional distributions on the whole dataset, one may split the data (randomly) into two separate parts for these tasks, say $D_1 = \{1, \ldots, \lfloor n\xi \rfloor\}$ and $D_2 = \{\lfloor n\xi \rfloor + 1, \ldots, n\}$ for some $\xi \in (0, 1)$. The index function is estimated with (x_i, y_i), $i \in D_1$, and the second part of the data with the *out-of-sample* predictions $\hat{\theta}(x_j)$, $j \in D_2$, serves as training data for IDR. To avoid that the estimated distribution functions depend on the random split of the training data, this procedure should be repeated several times, every time with a different split of the training data, and the conditional distribution functions are averaged in the end. The application of (sub-)sample aggregating ((sub-)bagging) has already been suggested in Henzi et al. (2019) in conjunction with IDR, where it yields smoother distribution functions and (in the case of subagging) reduces the computation time for larger datasets with multivariate covariates ($d \geqslant 2$). These advantages can also be expected for the DIM. In addition, the consistency result (Theorem 5.1) still holds under sample splitting when the data is split into D_1 and D_2 at a constant fraction $\xi \in (0, 1)$.

4 Prediction

This section reviews basic tools for the evaluation of probabilistic forecasts, and related properties of DIMs when used for forecasting. We denote by F a generic, random probabilistic forecast for a random variable Y, and all probability statements are understood with respect to the joint distribution of F and Y, which we denote by $\mathbb{P}$. For the distributional index model, the randomness of $F = F_{\theta(X)}$ is fully captured in the index $\theta(X)$.

As argued in Gneiting et al. (2007), *calibration* is a minimal requirement for probabilistic forecasts, meaning that the forecast should be statistically compatible with the distribution of the response. Of particular interest for DIMs is threshold calibration, requiring

$$\mathbb{P}(Y \leqslant y \mid F(y)) = F(y), \quad y \in \mathbb{R}. \tag{7}$$

It is shown in Henzi et al. (2019) that IDR, and hence also the DIM, is always in-sample threshold calibrated, that is, (7) holds when $\mathbb{P}$ is the empirical distribution of the training data used to estimate the distribution functions. Threshold calibration can be assessed by reliability diagrams (Wilks, 2011), in which estimated forecast probabilities $\hat{F}(y)$ are binned and compared to the observed event frequencies in each bin. Another prominent tool for calibration checks is the probability integral transform (PIT)

$$Z = F(Y-) + V\left(F(Y) - F(Y-)\right), \tag{8}$$

where V is uniformly distributed on $[0,1]$ and independent of F and Y, and $F(y-) = \lim_{z\uparrow y} F(z)$. If Z is uniformly distributed, then the forecast F is said to be probabilistically calibrated. The PIT can be used to identify forecast biases as well as underdispersion and overdispersion (Diebold et al., 1998; Gneiting et al., 2007).

Among different calibrated probabilistic forecasts, the most informative forecast is arguably the one with the narrowest prediction intervals. This property, which only concerns the forecast distribution F, is referred to as *sharpness* (Gneiting et al., 2007). Sharpness and calibration are often assessed jointly by means of proper scoring rules (Gneiting and Raftery, 2007), which map probabilistic forecasts and observations to a numerical score. An important example is the CRPS defined at (3). IDR enjoys in-sample optimality among all stochastically ordered forecasts with respect to a broad class of proper scoring rules, including the CRPS and weighted versions of it, that is,

$$\mathrm{CRPS}_\mu(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leqslant z\})^2 \,\mathrm{d}\mu(z),$$

where μ is a locally-finite measure. This emphasizes that IDR is a natural way to estimate the probability distributions in DIMs, since it is not tailored to a specific loss function.

5 Consistency

5.1 Two stage estimation

We work with a triangular array of random elements $(X_{ni}, Y_{ni}) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, and assume that for all n, the following hold:

(A1) The random elements X_{ni}, $i = 1, \dots, n$, are independent and identically distributed, and Y_{ni}, $i = 1, \dots, n$, are independent conditional on $(X_{ni})_{i=1}^n$ with

$$\mathbb{P}(Y_{ni} \leqslant y \mid X_{ni}) = F_{\theta(X_{ni})}(y),$$

where $\theta : \mathcal{X} \to \mathbb{R}$ is a function and $(F_u)_{u\in\mathbb{R}}$ is a family of distributions such that $F_u \preceq_{\mathrm{st}} F_v$ if $u \leqslant v$.

(A2) There exists a constant $L > 0$ such that for all $u, v, y \in \mathbb{R}$,

$$|F_u(y) - F_v(y)| \leqslant L|u - v|.$$

(A3) On an interval I, the random variables $\theta(X_{ni})$ admit a density with respect to the Lebesgue measure which is bounded from below by $C_1 > 0$ and from above by C_2.

(A4) There exist a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$ and a constant $C_0 > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathcal{X}} |g(\hat{\theta}_n(x)) - \theta(x)| \geqslant C_0(\log(n)/n)^{1/2}\right) = 0.$$

We denote by $\hat{F}_{n;u}$ the IDR estimator computed with training data $(\hat{\theta}_n(X_{nj}), Y_{nj})_{j=1}^n$, i.e.

$$\hat{F}_{n;u}(y) = \hat{F}_u(y; (\hat{\theta}_n(X_{nj}))_{j=1}^n, (Y_{nj})_{j=1}^n),$$

with the notation of Section 3.1.

Theorem 5.1 (Consistency of DIM)**.** *Under assumptions (A1)-(A4), there exists a constant $C > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{y\in\mathbb{R}, x\in\mathcal{X}_n} |\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y)| \geqslant C\left(\frac{\log n}{n}\right)^{1/6}\right) = 0,$$

where $\mathcal{X}_n = \{x \in \mathcal{X} : [\theta(x) \pm (\log n/n)^{1/6}] \subseteq I\}$.

An analogous result to Theorem 5.1 can be shown for the variant of the DIM with sample splitting described in Section 3.3. The requirements under sample splitting are slightly weaker, namely, the density of $\theta(X_{ni})$ does not have to be bounded from above in (A2), and in (A4), it is sufficient that the index estimator $\hat{\theta}_n$ converges at a rate of $o((\log(n)/n)^{1/3})$ instead of $n^{-1/2}$. The resulting convergence rate of the DIM with sample splitting is of order at least $(\log(n)/n)^{1/3}$. The proofs of Theorem 5.1, both, with and without sample splitting, rely on the consistency results about the monotonic least squares estimator in Mösching and Dümbgen (2020), and are given in Appendix A.

Assumption (A1) is the basic model assumption of DIMs. The Lipschitz-continuity in (A2) also appears in the monotone single index model for the mean (Balabdaoui et al., 2019a). Since the distributional single index model and the monotone single index model are equivalent when Y is binary, the Lipschitz assumption (A2) is natural in this context; also (A3) can be derived from the assumptions in Balabdaoui et al. (2019a). Assumptions (A2) and (A3) are required for the consistency of the monotone least squares estimator, with (A3) ensuring that the 'design points' $\theta(X_{nj})$ are dense enough in a region of interest, c.f. Mösching and Dümbgen (2020). A parametric model $\theta = \alpha_1\theta_1 + \cdots + \alpha_p\theta_p$ satisfies this assumption when at at least one of the summands $\alpha_i\theta_i$ admits a continuous distribution on I with density bounded away from zero. In (A4), we require uniform consistency of a monotone transformation of the index estimator at a rate of $n^{-1/2}$, i.e. not necessarily consistency of the index estimator itself. In a parametric model $\theta = \alpha_1\theta_1 + \cdots + \alpha_p\theta_p$, uniform consistency is satisfied for any $\sqrt{n}$-consistent estimator of the coefficients $\alpha_1, \ldots, \alpha_p$, when the functions $\theta_1, \ldots, \theta_p$ are bounded. All estimators suggested in Section 3.2 are consistent at the rate $n^{-1/2}$ under suitable conditions.

5.2 Simultaneous estimation

In this subsection, we treat the question to what extent simultaneous estimation of the index and the distribution functions is possible and sensible in the DIM. Currently, the results are of theoretical interest only.

It has been shown in Balabdaoui et al. (2019a) that for the monotone single index model, there exists a simultaneous minimizer (ψ_0, α_0) of the squared error

$$\sum_{i=1}^{n} (\psi_0(\alpha_0^T x_i) - y_i)^2$$

where $\psi_0 : \mathbb{R} \to \mathbb{R}$ is an increasing function, $\alpha_0 \in \{x \in \mathbb{R}^p : \|x\| = 1\}$ is the index, and $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$. The minimizer is in general not unique.

A similar result also holds in the distributional index model, when the loss function is defined as

$$l(\hat{\theta}, \hat{\boldsymbol{F}}) = \sum_{i=1}^{n} \mathrm{CRPS}(\hat{F}_{\hat{\theta}(x_i)}, y_i). \tag{9}$$

For basis functions $\theta_1, \ldots, \theta_p$ of the vector space $\mathcal{F}$ containing the true index function θ, every index estimator $\hat{\theta} : \mathcal{X} \to \mathbb{R}^d$ can be written as $\hat{\theta} = \hat{\alpha}_1\theta_1 + \cdots + \hat{\alpha}_p\theta_p$. The loss (9) has a unique minimizer $\hat{\boldsymbol{F}} = (\hat{F}_{\hat{\theta}(x_i)}, \ldots, \hat{F}_{\hat{\theta}(x_n)})$ for fixed $\hat{\theta}$, namely the IDR. This minimizer only depends on $\hat{\theta}$ via the partial order on the points $\hat{\theta}(x_i)$, $i = 1, \ldots, n$. But the number of partial orders on n points is finite, and so there exists a minimizer of (9).

In general, the number of partial orders induced by index functions $\hat{\theta}$ is too large for a direct minimization of (9) to be possible: When $\mathcal{X} = \mathbb{R}^p$ and $\theta_1, \ldots, \theta_p$ are the coordinate projections, then the number of total orders grows at a rate of $n^{2(p-1)}$ (Balabdaoui et al., 2019a). Moreover, when the index space is partially but not totally ordered, trivial solutions (a perfect fit to the training data) may appear, namely if the points $\hat{\theta}(x_i)$, $i = 1, \ldots, n$, are all incomparable in the partial order. Hence, the simultaneous estimation of the index and the distribution functions in DIMs is generally not feasible. A related interesting question for further research is to find a way to directly parametrize and estimate partial orders for isotonic or isotonic distributional regression, instead of indirectly via an index function.

6 Data application

We apply a DIM to derive probabilistic forecasts for intensive care unit (ICU) length of stay (LoS) based on patient information available 24 hours after admission. The main difficulty of such predictions is that, even conditional on many demographic and physiologic patient specific covariates, there is often great uncertainty in the LoS. In addition to unknown factors (e.g. frailty status, patient or family wishes), the LoS also depends on non-patient-related information such as ICU organization and resources. We therefore model the LoS using data of single ICUs rather than a merged dataset, thus keeping the ICU-related variables fixed. This allows forecasts within each single ICU as well as the comparison of the forecasted LoS of patients across ICUs. The same methodology can also be used on a joint dataset of several ICUs, giving a reference LoS forecast on the combined case-mix. Using these predictions for risk-adjustment and benchmarking is promising but goes beyond the scope of this paper.

All computations in this application were performed in R 4.0 (R Core Team, 2020) using the packages `mgcv` (Wood, 2017) for the estimation of index models and Cox proportional hazards regression, `quantreg` (Koenker, 2020) for quantile regression, and `isodistrreg` (Henzi et al., 2019, https://github.com/AlexanderHenzi/isodistrreg) for IDR.

6.1 Data and variables

Since 2005, the Swiss Society of Intensive Care Medicine collects ICU key figures and information on patient admissions in the Minimal Dataset of the Swiss Society of Intensive Care Medicine (MSDi). Our analysis is based on a part of this dataset suitable for LoS predictions, namely, we include 18 out of 86 ICUs which, after the application of selection criteria described below, include more than 10'000 patient admissions. The codes used as identifiers for the ICUs were generated randomly. The sample sizes range from 10'041 to 36'865 with an average of 17'181 observations per ICU. The cutoff of 10'000 is based on our experience with IDR and probabilistic forecasts in general, which require sufficiently large datasets for a meaningful and stable evaluation, especially when the models involve large numbers of covariates and a skewed response variable, as it is the case here. However, the prediction methods can also be applied to smaller datasets.

Based on literature review, we identified the variables described in Table 1 as relevant for LoS forecasts (Zimmerman et al. (2006); Verburg et al. (2014, Table S1); Niskanen et al. (2009)). We exclude patients that were transferred from or to another ICU, because their LoS is incomplete. As in Zimmerman et al. (2006), we also remove patients younger than 16 years and patients admitted after transplant operations or because of burns. Patients with missing values in the variables in Table 1 are excluded, too.

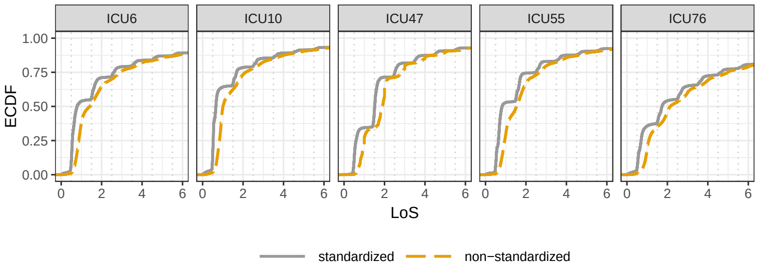
Table 1 documents at what time after admission the relevant covariates for LoS predictions are available. While all variables are available 24 hours after patient admission, the information is completed also for patients staying at the ICU less than one day. For example, ICU interventions within the first 24 hours are then only interventions performed until patient discharge, and the SAPS II is computed based on the worst physiological values until discharge instead of the worst values in the first 24 hours at the ICU.

In preliminary tests, we found that for probabilistic LoS forecasts, the usual definition of LoS as the time between patient admission and discharge is problematic, because most ICUs discharge patients during specific time windows, but the admission times are spread throughout the day. As a consequence, it may happen that the predicted LoS for certain patients does not conform with the discharge practice of a ICU, e.g. there might be a high predicted probability for a patient being discharged around midnight but the ICU actually discharges patients in the early afternoon. To circumvent this problem, we decided to measure the LoS as the *time between the next midnight after patient admission until discharge*, thereby standardizing all admission to the same (day)time and revealing the true pattern in the patient discharge times; see Figure 1. All results in this section use this definition of the LoS. Patients who do not stay over at least two calendar days are excluded, which is

Table 1: Covariates used for ICU length of stay predictions. Availability of the variables is given by 'admission' (at patient admission) or by the number of hours after admission.

Variable	**Availability**	**Description**
Age	admission	patient age at admission
Sex	admission	male, female
Planned	admission	admission is announced at least 12h in advance (true/false)
Readmission	admission	patient was discharged from the same ICU at most 48 hours ago (true/false)
Admission source	admission	admission source (emergency room; intermediate care unit, high dependency unit, recovery room; hospital ward; surgery; others)
Location before hospital admission	admission	location before *hospital* admission (home; other hospital; others)
Diagnosis	24h	main diagnosis on first day (structured into: cardiovascular, respiratory, gastrointestinal, neurological, metabolic, trauma, others; in total 36 different specific ICU relevant diagnoses)
NEMS	8h	NEMS (Miranda et al., 1997) over first shift after patient admission (8-12h)
SAPS II	24h	Le Gall et al. (1993)
Interventions	24h	interventions 24 hours before until 24 hours after admission (13 categories of interventions, e.g. surgeries, interventions in respiratory system, cardiovascular interventions)

Figure 1: Empirical distribution functions of the standardized and non-standardized LoS for selected ICUs. The standardized LoS is defined as $Y - 1 + h/24$, where h is the admission hour of a patient and Y is the non-standardized LoS, i.e. the time between patient admission and discharge. Only patients with positive standardized LoS are included.



unproblematic since in practice, the data required for predictions is only available 24 hours after admission and the forecast should be conditioned on the event that the patient already stayed at the ICU for 24 hours. Forecasts for the non-standardized LoS, i.e. the time between admission and discharge, can be derived via the relation

$$\mathbb{P}(Y > 1 + t | Y > 1) = \frac{\mathbb{P}(\tilde{Y} > t + h/24 | \tilde{Y} > 0)}{\mathbb{P}(\tilde{Y} > h/24 | \tilde{Y} > 0)},$$

where Y and $\tilde{Y} = Y - 1 + h/24$ denote the LoS and the standardized LoS measured in days, respectively, and h the admission hour of a given patient. Since only patients staying at least until midnight of the admission day are used as training data, our LoS forecasts are conditioned on the event $\{\tilde{Y} > 0\}$ in the above equation.

We select the most recent 20% of the observations in each ICU for model validation, thereby mimicking a realistic situation in which past data are used to predict the LoS of present and future patients. This implies that forecasts might be inaccurate if the relationship between the covariates and LoS changes over time, and it is part of our analysis to check to what extent past data can be reasonably used to predict the LoS of future patients. Of the remaining data, randomly selected 75% are used for model fitting and 25% for model selection via out-of-sample predictions. All comparisons of different variants of a distributional regression model were performed by such out-of-sample predictions.

6.2 Derivation of DIM

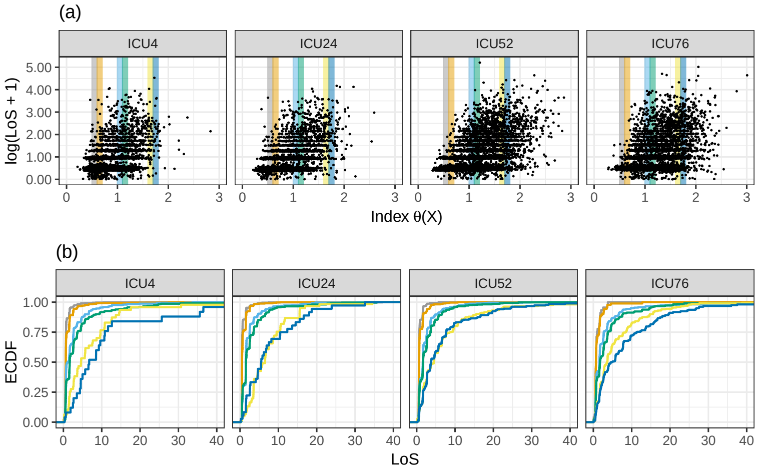
To derive an index estimator for the DIM, we can benefit from the comparisons of regression models for point forecasts for LoS in the extant literature. Moran and Solomon (2012) and Verburg et al. (2014) found that a Gaussian linear regression for the expected log-LoS is suitable for point forecasts, and we use this as our candidate for the index estimator and will refer to it as the 'lognormal index model'. We use the transformation $y \mapsto \log(y+1)$, which results in more symmetric distributions than the logarithm. All variables from Table 1 were included in the model, and the effects of the continuous variables age, SAPS and NEMS were modeled by cubic regression splines. Interactions of variables were explored but not included in the final model. We also tested whether merging factor levels with few observations improved the model, but the untransformed covariates yielded the best forecasts in out-of-sample predictions on the part of the data used for model selection.

We tested two other index estimators for the expected LoS to investigate the robustness of the DIM with respect to the index. The first one estimates the expected log-LoS under the assumption of a scaled t-distribution. The mean is modeled as a function of the covariates, with the same specification as for the lognormal index model, and the degrees of freedom are estimated, with a minimal threshold of 5 to ensure stability. This model is structurally similar to the lognormal index model, but more robust with respect to outliers, which occur even after the log-transformation. The second alternative is a gamma regression for the untransformed LoS with logarithm as the link function. While the three index models yield different predictions on the scale of the LoS, they largely agree when only the *ordering* of the predictions is considered: Over the 18 ICUs, the rank correlation between predictions by two of the models is 0.98 on average with a minimum of 0.86. As a consequence, there is no significant difference between the corresponding DIM forecasts: Evaluated on the dataset for model selections, the average CRPS over all ICUs of DIM forecasts based on different models only differs by up to 0.01, while the averages are around 1.40. The predictions based on the lognormal index model achieved the best results in most ICUs and were therefore selected for the predictions on the validation data.

Due to the large training datasets, splitting of the training data as described in Section 3.3 only has a marginal effect on the predictions. Estimating the index function on the full training data and the conditional distributions on in-sample predictions only increased the average CRPS by 0.01 (on 1.40), compared to a bagging approach with 100 random splits of the training data into equally sized parts for the estimation of the index and the CDFs. For the final evaluation, we show the results of the simpler variant without bagging.

Figure 2 illustrates how to perform a check of the stochastic ordering assumption of the DIM: We bin the observed LoS according to the index value, and plot the empirical cumulative distribution functions (ECDFs) of the LoS in each bin. By varying the positions and sizes of the bins, it can be seen that the empirical distributions are indeed sufficiently well ordered. The Spearman correlation between the index and the observed LoS is 0.53 on average over all ICUs (range $0.40-0.65$), which confirms that there is an isotonic relationship between the index and the actual LoS for most ICUs, taking into account the high uncertainty in the LoS of ICU patients even conditional on patient information collected at the first day.

Figure 2: (a) Index function and log(LoS + 1) for selected ICUs. (b) ECDFs of the LoS stratified into the bins given by the vertical shaded stripes in panel (a).



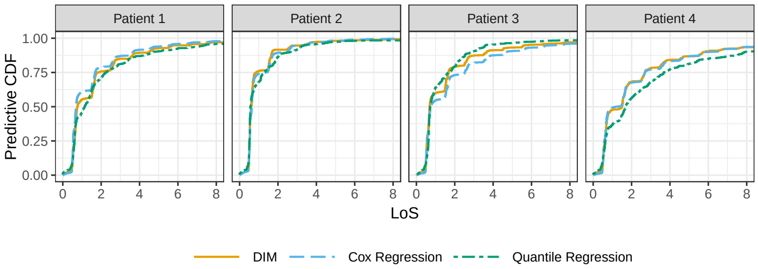
6.3 Alternative regression methods

We compare the DIM to two other distributional regression methods: A Cox proportional hazards model (Cox, 1972) and quantile regression with monotone rearrangement (Koenker, 2005; Chernozhukov et al., 2010). For both, we use the same variables and specifications as in the index estimator for the DIM, which was superior compared to other variants tested; detailed results are provided in the supplementary material.

A Cox proportional hazards model is a classical choice for modeling survival times, and it shares some similarities with a DIM. Both models are semi-parametric and based on stochastic order restrictions on the conditional distributions, namely the usual stochastic ordering in the DIM and the hazard rate order in Cox regression, which is stronger than the usual stochastic order (Shaked and Shanthikumar, 2007, Theorem 1.B.1). While the distribution functions are estimated non-parametrically in Cox regression, the relationship between different conditional distributions is modeled parametrically via the hazard ratio, as opposed to the DIM, where only the ordering on the conditional distributions is modeled parametrically by the index function.

Quantile regression, on the other hand, imposes less assumptions on the conditional distributions. The conditional quantiles are modeled separately and satisfy no stochastic order constraints. In particular, if there are strong violations of the stochastic order assumptions of the DIM or Cox regression, we would expect that the more flexible quantile regression achieves better forecasts by fitting crossing quantile curves for different patients. This allows an informal check of the underlying assumptions of Cox regression and the DIM (see Figure

Figure 3: Predictive CDFs for four selected patients based on the training data of the ICU the patients were admitted to.



S1 in the supplementary material). We use a grid of quantiles from 0.005 to 0.995 with steps of 0.001, which gave better results than a coarser grid with steps of 0.01.

We also tested fully parametric models of GAMLSS type, and kernel methods as implemented in the `np` package in R (Hayfield and Racine, 2008). Unfortunately, we could not find a sufficiently flexible parametric family for a GAMLSS, and the application of kernel methods was not feasible due to computational problems with the large datasets and high numbers of covariates. As for the DIM, computation is obviously more demanding than for fully parametric methods, but still fast thanks to the sequential implementation of IDR described in Henzi et al. (2020). On a personal computer with Intel(R) Core i7-8650 CPU, computation with the lognormal index model without bagging takes 3 seconds for the smallest ICU (6'024 observations in training dataset) and 25 seconds for the largest ICU (22'219 observations). Estimation and prediction on the total dataset (all 18 ICUs) require about 2.5 minutes.

6.4 Results

Figure 3 illustrates the probabilistic forecasts for different patients based on the training data of the ICU the patients were admitted to. Patient 1, male, 32 years old, was admitted because of a severe sepsis or septic shock. Patient 2 is a 67 years old female with aortic aneurysm or aortic dissection, Patient 3 is 58 years old, male with a metabolic decompensation, and Patient 4 is a 78 old female admitted from a high dependency unit with subarachnoidal hemorrhage. Patient 2 has the shortest predicted LoS: The DIM and Cox regression predict that she leaves the ICU at the first day after admission with a probability of almost 75%. For the remaining patients, the predictive CDFs are more skewed, and a LoS of more than three days is not unlikely. It is immediately visible that the DIM and Cox regression are able to recover the pattern in the ICU discharge times, with flat pieces of the CDFs around midnight. Quantile regression, on the other hand, merely interpolates this pattern.

Figure 4: Reliability diagrams of probabilistic forecasts for the predicted probability that the LoS exceeds 1, 5, 9, 13 days. The forecast probability is grouped into the bins $[0, 0.1], (0.1, 0.2], \ldots, (0.9, 1]$ and the observed frequencies are drawn at the midpoints of the bins. Only bins with more than two observations are included.

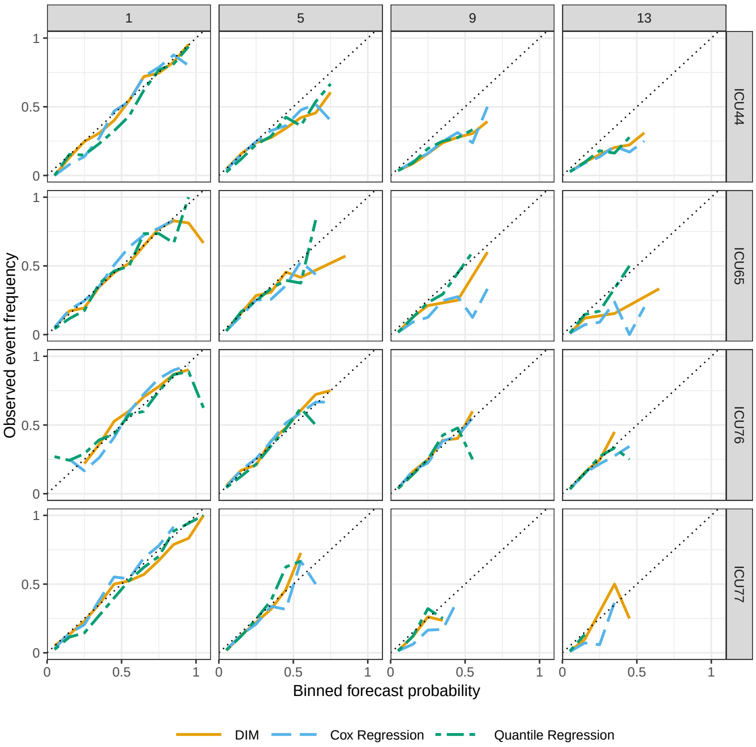
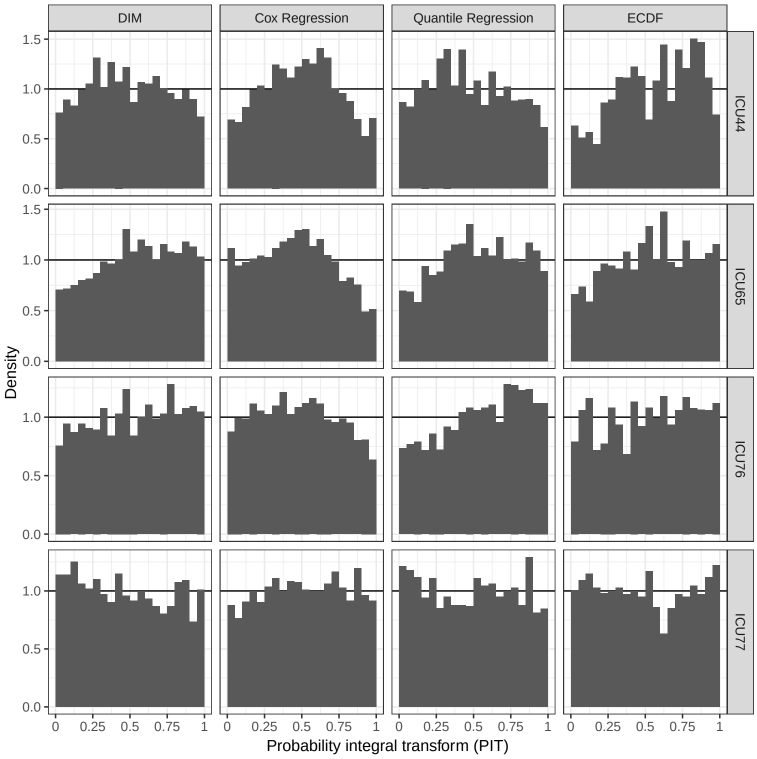


Table 2: Summary statistics (mean, median and standard deviation) of numeric variables in the dataset.

ICU	LoS			Age			NEMS			SAPS		
	mean	med.	sd	mean	med.	sd	mean	med.	sd	mean	med.	sd
ICU44	3.9	1.5	7.8	59.0	61	17.6	27.1	27	8.5	34.0	31	18.9
ICU65	1.8	0.6	4.3	67.2	69	13.9	25.5	25	7.9	28.7	28	12.5
ICU76	4.3	1.7	7.2	63.2	66	15.6	30.3	30	8.3	41.2	40	17.2
ICU77	1.8	0.6	3.2	65.0	68	15.9	21.9	18	8.0	31.1	28	16.1

Figure 5: PIT histograms of the probabilistic forecasts with bins of width 1/20.



Here, detailed results are only shown for the best and worst two ICUs with respect to the CRPS of the DIM forecasts; see the supplementary material for tables and figures for all ICUs. Summary statistics of the LoS and other numeric variables for the patients of these ICUs are given in Table 2. All probabilistic regression methods can reliably predict the probability that the LoS exceeds $k = 1, 5, 9, 13$ days; see Figure 4. Figure 5 shows that the forecasts achieve a better probabilistic calibration than the ECDF of the LoS in the training data, which is uninformative as a forecast and does not take into account changes in the ICU-case mix that are reflected in the covariates. Further improvements of calibration may be possible by selecting a tailored training dataset, taking into account organizational changes, and developments in treatments that have an influence on the LoS or on the relationship between covariates and the LoS. Such information is not available in our dataset.

While all three distributional regression methods yield similar results in terms of calibration, there is a clear ranking with respect to forecast accuracy: In all ICUs, the DIM achieves the lowest CRPS, followed by quantile regression in second and Cox regression in third place. For comparison, Table 3 also shows the CRPS of the ECDF forecast, and of the deterministic point forecast of the lognormal index model, which is its mean absolute error. Interestingly, the ECDF forecast achieves a lower mean CRPS in all ICUs (average improvement of 13%) than the point forecast, although it does not take any covariate information into account. This highlights the superiority of even simple probabilistic forecast over point forecasts in the context of ICU LoS. A further average improvement of 13% in the mean CRPS is achieved when going from the uninformative ECDF forecast to the worst of the probabilistic regression methods in terms of CRPS, which is Cox regression. The differences in the CRPS of the forecasts using distributional regression methods are smaller, but consistent over the ICUs: In terms of average CRPS, quantile regression outperforms Cox regression in 15 out of 18 ICUs, and the DIM outperforms Cox regression in all and quantile regression in all except 2 ICUs. The difference in CRPS between the DIM and quantile regression is highly significant when tested with Wilcoxon's signed rank test except for the ICUs with identifiers 19 and 33, where the p-values are 0.101 and 0.219 and quantile regression achieves lower average scores. Wilcoxon's signed rank test was applied because the CRPS differences are heavy-tailed, so a t-test is not appropriate (see Figure S6 in the supplementary material).

In conclusion, with distributional regression methods and especially the DIM, it is possible to obtain reliable, reasonably well calibrated, and informative probabilistic forecasts for ICU LoS in a realistic setting. These forecasts are not only more informative than point forecasts, but also reduce the forecast error by more than 25%.

7 Discussion

In this paper, we have introduced DIMs as intuitive and flexible models for distributional regression. Distributional regression approaches provide full conditional distributions of the outcome given covariate information, and are thus more informative than classical regression approaches for the conditional mean, median or specific quantiles. However, specifying a good distributional regression model is usually less intuitive than specifying a regression model for, say, the conditional mean. An appealing feature of DIMs is that for the modeling of the index function classical approaches and intuition for modeling a conditional mean or median can be used. Given the index function, the shape of the full conditional distribution is then learned from training data using IDR, that is, distributional regression under stochastic ordering constraints. The second step does not involve any parameter tuning or implementation choices.

The idea of reducing the complexity of a potentially high-dimensional covariate space by using an index function in distributional regression has also been used in the work of Hall and Yao (2005); Zhang et al. (2017). In these works, the index function has to be univariate and parametrizes a distance on the covariate space that is then used for kernel methods to estimate the conditional distributions. In contrast, the index function in a DIM parametrizes partial orders on the covariate space allowing for stochastic order constrained distributional

Table 3: CRPS of probabilistic forecasts. The column 'Point' shows the mean absolute error of the point forecast obtained from the lognormal index model, and p-values of Wilcoxon's signed rank test for the difference in CRPS between DIM and quantile regression are given in the column labelled p. P-values smaller than 10^{-16} are written as 0.

ICU	p	DIM	Quantile reg.	Cox reg.	ECDF	Point
ICU4	$1.18 \cdot 10^{-11}$	1.074	1.076	1.089	1.191	1.399
ICU6	$3.81 \cdot 10^{-12}$	1.360	1.385	1.386	1.605	1.830
ICU10	0	1.194	1.221	1.209	1.312	1.553
ICU19	$1.01 \cdot 10^{-1}$	1.041	1.032	1.048	1.189	1.350
ICU20	$5.13 \cdot 10^{-6}$	2.216	2.223	2.241	2.505	2.859
ICU24	0	1.099	1.111	1.141	1.265	1.416
ICU33	$2.19 \cdot 10^{-1}$	0.975	0.974	0.983	1.090	1.363
ICU39	$1.38 \cdot 10^{-16}$	1.332	1.352	1.383	1.697	1.872
ICU44	$1.06 \cdot 10^{-3}$	2.256	2.259	2.328	2.480	2.952
ICU47	$3.69 \cdot 10^{-5}$	0.977	0.980	1.036	1.231	1.363
ICU52	$7.40 \cdot 10^{-5}$	1.845	1.866	1.868	2.121	2.580
ICU55	0	1.062	1.085	1.055	1.253	1.445
ICU58	$1.25 \cdot 10^{-15}$	1.393	1.409	1.442	1.763	1.970
ICU65	0	0.908	0.914	0.981	1.062	1.194
ICU76	0	2.420	2.448	2.458	2.783	3.468
ICU77	$1.76 \cdot 10^{-16}$	0.921	0.936	0.938	1.117	1.260
ICU79	$1.86 \cdot 10^{-11}$	1.446	1.457	1.512	2.172	2.228
ICU80	0	0.942	0.971	0.949	1.094	1.253
Mean		1.359	1.372	1.392	1.607	1.853

regression in the second step.

Finding an informative index function is critical and usually requires expertise of the problem at hand. However, in many cases, existing models for the conditional mean or median can be used directly, as demonstrated in the application on ICU LoS. Indeed, it may even happen that a poorly fitting conditional mean model works well for a DIM since it is sufficient that the model is correct up to monotone transformations, or, in other words, that it is a good model for a pseudo index.

The distributional regression approach in Chernozhukov et al. (2020) allows to accomodate continuous, discrete and mixed discrete-continuous outcomes. The same is true for IDR, and thus for DIM models. While the case study in this paper concerns a continuous outcome, IDR has been successfully applied to a mixed discrete-continuous outcome in Henzi et al. (2019). It would be interesting to investigate the different benefits and drawbacks of DIM models versus the methods of Chernozhukov et al. (2020) in particular in the case of discrete outcomes.

Since IDR can be combined well with (sub-)bagging, the same also holds for DIMs. (Sub-)bagging is useful to avoid overfitting, may increase computational efficiency, and lead to smoother estimated conditional CDFs. We have explored bagging in our data application in Section 6 with relatively at hoc choices for the number of random splits of the training data. A systematic study of optimal choices for subsample sizes and/or iterations is desirable.

A promising future extension of DIMs is to replace the IDR step by distributional regression under a stronger stochastic ordering constraint such as a likelihood ratio ordering constraint, or by a weaker one such as second order stochastic dominance. However, this requires fundamental advances concerning the estimation of distributions under these constraints.

A Proof of Theorem 5.1

The following lemma is Theorem 4.6 in Mösching and Dümbgen (2020), which we state for completeness.

Lemma A.1. *Let $Z_1, Z_2, Z_3, \ldots$ be independent random variables with respective distribution functions $G_1, G_2, G_3, \ldots$. For $k \in \mathbb{N}$, let*

$$\hat{\mathbb{G}}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\{Z_i \leqslant \cdot\} \quad \textit{and} \quad \bar{G}_k(\cdot) = \frac{1}{k}\sum_{i=1}^{k} G_i(\cdot).$$

Then there exists a universal constant $M \leqslant 2^{5/2}e$ such that for all $\eta \geqslant 0$,

$$\mathbb{P}\left(\sqrt{k}\|\hat{\mathbb{G}}_k - \bar{G}_k\|_\infty \geqslant \eta\right) \leqslant M\exp(-2\eta^2),$$

where $\|\cdot\|_\infty$ denotes the usual supremum norm of functions.

The results and proofs below use the following definitions. We denote by $\lambda(J)$ the Lebesgue measure of a measurable set $J \subset \mathbb{R}$, and define the events

$$B_n = \left\{ \sup_{x \in \mathcal{X}} |g(\hat{\theta}_n(x)) - \theta(x)| < C_0(\log(n)/n)^{1/2} \right\}. \tag{10}$$

For $1 \leqslant r \leqslant s \leqslant n$ and a permutation σ of $\{1, \ldots, n\}$, let

$$w_{rs} = s - r + 1, \qquad \hat{\mathbb{F}}^{\sigma}_{rs} = \frac{1}{w_{rs}} \sum_{i=r}^{s} \mathbb{1}\{Y_{n\sigma(i)} \leqslant \cdot\},$$

$$\bar{F}^{\sigma}_{\theta;rs}(\cdot) = \frac{1}{w_{rs}} \sum_{i=r}^{s} F_{\theta(X_{n\sigma(i)})}(\cdot), \qquad \bar{F}^{\sigma}_{\hat{\theta};rs}(\cdot) = \frac{1}{w_{rs}} \sum_{i=r}^{s} F_{\hat{\theta}_n(X_{n\sigma(i)})}(\cdot).$$

We use π to denote a permutation such that $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \cdots \leqslant \hat{\theta}_n(X_{n\pi(n)})$. The permutation π is a function of $(X_{ni}, Y_{ni})_{i=1}^n$ via $(X_{ni})_{i=1}^n$ and $\hat{\theta}_n$. Let

$$M^{\pi}_n = \max_{1 \leqslant r \leqslant s \leqslant n} w_{rs}^{1/2} \| \hat{\mathbb{F}}^{\pi}_{rs} - \bar{F}^{\pi}_{\theta;rs} \|_{\infty}. \tag{11}$$

Lemma A.2. *Under (A3) and (A4), there exists a constant $s = s(C_0, C_2) > 0$ such that*

$$\lim_{n \to \infty} \mathbb{P}(M^{\pi}_n \geqslant s n^{1/4} \log(n)^{1/4}) = 0.$$

Proof. Define $m = m(n) = \max(1, \lfloor \lambda(I)/(2c_n) \rfloor)$ with $c_n = C_0(\log(n)/n)^{1/2}$, where C_0 is from assumption (A4). Then, for n large enough such that $c_n \leqslant \lambda(I)/4$,

$$2c_n \leqslant \frac{\lambda(I)}{m} \leqslant 4c_n. \tag{12}$$

Slice the interval I from (A3) into m equally sized, disjoint intervals $J_1, \ldots, J_m$ (ordered increasingly). Let $\mathcal{I}_k = \{i \in \{1, \ldots, n\} : \theta(X_{ni}) \in J_k\}$, $n_k = \#\mathcal{I}_k$ for $k = 1, \ldots, m$, and $N_n = \max_{k=1,\ldots,m} n_k$. Define also $\mathcal{I}_j = \varnothing$ for $j \notin \{1, \ldots, m\}$ and $\bigcup_{i=a}^{b} A_i = \varnothing$ for any sets A_i and $a > b$.

Let $r, s \in \{1, \ldots, n\}$, $r \leqslant s$, be indices that attain the maximum in (11), and define the index set $\mathcal{I}^* = \pi(\{r, \ldots, s\})$, so that

$$M^{\pi}_n = \left\| \frac{1}{(\#\mathcal{I}^*)^{1/2}} \sum_{i \in \mathcal{I}^*} \left(\mathbb{1}\{Y_{ni} \leqslant \cdot\} - F_{\theta(X_{ni})}(\cdot) \right) \right\|_{\infty}.$$

Note that the indices r and s are (complicated but measurable) functions of (X_i, Y_i), $i = 1, \ldots, n$, and thus random variables. Therefore, the set $\mathcal{I}^*$ is also a random set of indices.

If $i, j \in \mathcal{I}^*$ and $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nj}))$, with g from (A4), then $k \in \mathcal{I}^*$ for all k such that $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nk})) < g(\hat{\theta}_n(X_{nj}))$. This follows from $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \cdots \leqslant \hat{\theta}_n(X_{n\pi(n)})$, because if $i = \pi(i_0)$, $j = \pi(j_0)$ and $k = \pi(k_0)$, then $g(\hat{\theta}_n(X_{ni})) < g(\hat{\theta}_n(X_{nk})) < g(\hat{\theta}_n(X_{nj}))$ implies that $i_0 < k_0 < j_0$, and $k_0 \in \{i_0, \dots, j_0\} \subseteq \{r, \dots, s\}$ gives $k = \pi(k_0) \in \pi(\{r, \dots, s\}) = \mathcal{I}^*$.

Under the event B_n defined at (10), $i \in \mathcal{I}_k$ and (12) imply that $g(\hat{\theta}(X_{ni})) \in J_t$ for some $t \in \{k-1, k, k+1\}$. Therefore, for $l, k \in \{1, \dots, m\}$ with $l - k > 2$, it follows $g(\hat{\theta}(X_{ni})) < g(\hat{\theta}(X_{nj}))$ for all $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_l$. So if $\mathcal{I}^*$ contains indices $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_l$ with $l - k > 2$, then $\mathcal{I}^*$ must also contain all elements of the sets $\mathcal{I}_t$ for $k + 2 < t < l - 2$. Let $\kappa = \min\{j \in \{1, \dots, m\} : \mathcal{I}_j \cap \mathcal{I}^* \neq \varnothing\}$, $\ell = \max\{j \in \{1, \dots, m\} : \mathcal{I}_j \cap \mathcal{I}^* \neq \varnothing\}$. By the previous considerations, $\mathcal{I}^*$ may contain arbitrary elements of $\mathcal{I}_t$ with $t \in \{\kappa, \kappa+1, \kappa+2, \ell-2, \ell-1, \ell\}$, and it must contain all indices in $\mathcal{I}_j$ for $\kappa + 3 \leqslant j \leqslant \ell - 3$. In conclusion, under B_n, $\mathcal{I}^*$ is almost surely contained in the collection of index sets defined by

$$S_n = \bigcup_{1 \leqslant k \leqslant l \leqslant m} \left\{ \mathcal{J} \cup \left(\bigcup_{t=k+3}^{l-3} \mathcal{I}_t \right) : \mathcal{J} \subseteq \left(\bigcup_{t=k}^{k+2} \mathcal{I}_t \right) \cup \left(\bigcup_{t=l-2}^{l} \mathcal{I}_t \right) \right\}.$$

Indeed, on the event B_n, we know that $\mathcal{I}^*$ must contain all elements of $\mathcal{I}_j$ for $\kappa+3 \leqslant t \leqslant \ell-3$. This explains the part $\bigcup_{t=k+3}^{l-3} \mathcal{I}_t$ in the definition of S_n. As for the $\mathcal{I}_k$ with subscript not in $\{\kappa+3, \dots, \ell-3\}$, $\mathcal{I}^*$ may contain any arbitrary selection from their elements. This arbitrary selection is $\mathcal{J} \subseteq \left(\bigcup_{t=k}^{k+2} \mathcal{I}_t \right) \cup \left(\bigcup_{t=l-2}^{l} \mathcal{I}_t \right)$. For κ and ℓ, all pairs (k, l) with $k \leqslant l$ are possible, which gives the union over $1 \leqslant k \leqslant l \leqslant n$.

Because $\#\mathcal{I}_t \leqslant N_n$ for all t, one can derive from the definition of S_n that

$$\#S_n \leqslant m^2 \cdot 2^{6N_n} = m^2 \exp(6 \log(2) N_n).$$

We now compute an upper bound for N_n, which is a function of $\theta(X_{n1}), \dots, \theta(X_{nn})$ only. Denote by P and G the distribution and the CDF of $\theta(X_{n1})$, and by $\hat{P}$ and $\hat{G}$ the empirical distribution and the empirical CDF of $\theta(X_{n1}), \dots, \theta(X_{nn})$. For any $c \geqslant 0$,

$$\begin{aligned}
\mathbb{P}(N_n \geqslant c) &\leqslant \sum_{k=1}^{m} \mathbb{P}(n_k \geqslant c) \\
&\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(\hat{P}(J_k) - P(J_k) \geqslant \frac{c}{n} - P(J_k) \Big) \\
&\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G - \hat{G}\|_\infty \geqslant \frac{c}{n} - P(J_k) \Big).
\end{aligned}$$

For n sufficiently large, $P(J_k) \leqslant 4C_2C_0c_n = 4C_2C_0(\log(n)/n)^{1/2}$ by (12) and by (A3). Replacing c by $d_n = R\log(n)^{1/2}n^{1/2}$ with $R = \max(2, 8C_2C_0)$ and applying Lemma A.1 and

(12) yields

$$\begin{aligned}\mathbb{P}(N_n \geqslant d_n) &\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G-\hat{G}\|_\infty \geqslant \frac{d_n}{n} - 4C_2C_0(\log(n)/n)^{1/2}\Big)\\ &\leqslant \sum_{k=1}^{m} \mathbb{P}\Big(2\|G-\hat{G}\|_\infty \geqslant \frac{d_n}{2n}\Big)\\ &\leqslant mM\exp\left(-2n\Big(\frac{d_n}{4n}\Big)^2\right)\\ &\leqslant \frac{\lambda(I)M}{2(\log(n)/n)^{1/2}}\exp\left(-\log(n)/2\right)\\ &\leqslant \frac{\lambda(I)M}{2\log(n)^{1/2}}\exp\left(-\log(n)/2+\log(n)/2\right) \to 0,\ n\to\infty.\end{aligned}$$

So with asymptotic probability one,

$$\begin{aligned}\#S_n \leqslant m^2\exp\left(6\log(2)R\log(n)^{1/2}n^{1/2}\right) &\leqslant \frac{\lambda(I)^2}{4c_n^2}\exp\left(6R\log(2)\log(n)^{1/2}n^{1/2}\right)\\ &\leqslant r_0\exp\left(r_1\log(n)^{1/2}n^{1/2}\right),\end{aligned}$$

with $r_0 = \lambda(I)^2/(4C_0)$ and $r_1 = 6R\log(2)+1$. Define $D_n = \{\#S_n \leqslant r_0\exp(r_1\log(n)^{1/2}n^{1/2})\}$, let $\mathfrak{S}_n$ be the power set of $\{1,\dots,n\}$, and, for $\mathcal{J}\in\mathfrak{S}_n$,

$$M_n^{\mathcal{J}} = \left\|\frac{1}{(\#\mathcal{J})^{1/2}}\sum_{i\in\mathcal{J}}\left(\mathbb{1}\{Y_{ni}\leqslant\cdot\} - F_{\theta(X_{ni})}(\cdot)\right)\right\|_\infty.$$

Then, for $z_n = s\log(n)^{1/4}n^{1/4}$ with an arbitrary $s>0$,

$$\begin{aligned}&\mathbb{P}(M_n^\pi \geqslant z_n) = \mathbb{E}\left(\mathbb{1}\left\{M_n^{\mathcal{I}^*}\geqslant z_n\right\}\right)\\ &= \mathbb{E}\left(\sum_{\mathcal{J}\in\mathfrak{S}_n}\mathbb{1}\{\mathcal{I}^*=\mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}}\geqslant z_n\right\}\right)\\ &\leqslant \mathbb{P}(B_n^c) + \mathbb{E}\left(\mathbb{1}B_n\sum_{\mathcal{J}\in\mathfrak{S}_n}\mathbb{1}\{\mathcal{I}^*=\mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}}\geqslant z_n\right\}\right)\\ &= \mathbb{P}(B_n^c) + \mathbb{E}\left(\mathbb{1}B_n\sum_{\mathcal{J}\in S_n}\mathbb{1}\{\mathcal{I}^*=\mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}}\geqslant z_n\right\}\right)\\ &\leqslant \mathbb{P}(B_n^c) + \mathbb{E}\left(\sum_{\mathcal{J}\in S_n}\mathbb{1}\{\mathcal{I}^*=\mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}}\geqslant z_n\right\}\right)\\ &\leqslant \mathbb{P}(B_n^c) + \mathbb{P}(D_n^c) + \mathbb{E}\left(\mathbb{1}D_n\mathbb{E}\left[\sum_{\mathcal{J}\in S_n}\mathbb{1}\{\mathcal{I}^*=\mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}}\geqslant z_n\right\}\middle| X_{n1},\dots,X_{nn}\right]\right).\end{aligned}$$

In the last inequality we use the fact that $\mathbb{1}D_n$ is a function of $X_{n1}, \ldots, X_{nn}$ and

$$\mathbb{E}\left[\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\middle| X_{n1}, \ldots, X_{nn}\right] \leqslant 1 \text{ a.s.},$$

since $\mathcal{I}^* = \mathcal{J}$ may only hold for exactly one index set $\mathcal{J}$. Finally,

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{1}D_n\mathbb{E}\left[\sum_{\mathcal{J}\in S_n} \mathbb{1}\{\mathcal{I}^* = \mathcal{J}\}\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\}\middle| X_{n1}, \ldots, X_{nn}\right]\right)\\
&\leqslant \mathbb{E}\left(\mathbb{1}D_n \sum_{\mathcal{J}\in S_n} \mathbb{E}\left[\mathbb{1}\left\{M_n^{\mathcal{J}} \geqslant z_n\right\} \mid X_{n1}, \ldots, X_{nn}\right]\right).\\
&= \mathbb{E}\left(\mathbb{1}D_n \sum_{\mathcal{J}\in S_n} \mathbb{P}\left[M_n^{\mathcal{J}} \geqslant z_n \mid X_{n1}, \ldots, X_{nn}\right]\right).\\
&\leqslant \mathbb{E}\left(\mathbb{1}D_n(\#S_n)M\exp(-2z_n^2)\right)\\
&\leqslant r_0 M \exp\left(-(2s^2 - r_1)\log(n)^{1/2}n^{1/2}\right) \to 0,\ n \to \infty,
\end{aligned}$$

for $s > \sqrt{r_1/2}$, using Lemma A.1 in the second-last inequality. □

Lemma A.3 shows that for suitable constants D and sequences $(\delta_n)_{n\in\mathbb{N}}$ with limit zero, all subintervals of I with length at least δ_n contain at least $Dn\delta_n$ elements of $\{g(\hat{\theta}_n(X_{nj})) : j = 1, \ldots, n\}$. That is, the pseudo-covariates $g(\hat{\theta}_n(X_{nj}))$ are asymptotically dense in I.

Lemma A.3. *Under (A3) and (A4), with $\hat{w}(B) = \#\{j \in \{1, \ldots, n\} : g(\hat{\theta}_n(X_{nj})) \in B\}$, for any sequence $(\delta_n)_{n\in\mathbb{N}}$ such that $\delta_n \geqslant 4C_0(\log(n)/n)^{1/2}$, the event*

$$\left\{\inf\left\{\frac{\hat{w}(I_n)}{n\lambda(I_n)} : \textit{intervals } I_n \subset I \textit{ with } \lambda(I_n) \geqslant \delta_n\right\} \geqslant D\right\} \tag{13}$$

has asymptotic probability one for any $D < C_1/2$.

Proof of Lemma A.3. Similarly to the definition of $\hat{w}$, let $w(B) = \#\{j \in \{1, \ldots, n\} : \theta(X_{nj}) \in B\}$ for $B \subseteq I$. Define $c_n = C_0(\log(n)/n)^{1/2}$ with C_0 from (A4). Then on the event B_n defined at (10), for any interval $J \subseteq I$ with $\lambda(J) \geqslant 2c_n$,

$$\begin{aligned}
\hat{w}(J) - w(J) &\geqslant -\#\{j \in \{1, \ldots, n\} :\ \hat{\theta}_n(X_{nj}) \notin J,\ \theta(X_{nj}) \in J\}\\
&\geqslant -w(\{z \in J :\ z + c_n \notin J \text{ or } z - c_n \notin J\}).
\end{aligned}$$

This gives $\hat{w}(J) \geqslant\ w(J\backslash\{z \in J :\ z + c_n \notin J \text{ or } z - c_n \notin J\})$. The assumption $\delta_n \geqslant 4c_n$ implies that $\delta_n - 2c_n \geqslant \delta_n/2$. For any interval $I_n \subseteq I$ of length at least δ_n, the set $\tilde{I}_n = I_n\backslash\{z \in I_n : z + c_n \notin I_n \text{ or } z - c_n \notin I_n\}$ is an interval of length

$$\lambda(\tilde{I}_n) = \lambda(I_n) - 2c_n \geqslant \lambda(I_n) - \delta_n/2 \geqslant \lambda(I_n) - \lambda(I_n)/2 = \lambda(I_n)/2.$$

This and $\hat{w}(I_n) \geqslant w(\tilde{I}_n)$ yield

$$\begin{aligned}\hat{m}_n &:= \inf\left\{\frac{\hat{w}(I_n)}{n\lambda(I_n)} : \text{intervals } I_n \subset I \text{ with } \lambda(I_n) \geqslant \delta_n\right\}\\ &\geqslant \inf\left\{\frac{w(\tilde{I}_n)}{n\lambda(\tilde{I}_n)} : \text{intervals } \tilde{I}_n \subset I \text{ with } \lambda(\tilde{I}_n) \geqslant \delta_n/2\right\}/2 =: m_n.\end{aligned}$$

Define $A_n = \{\hat{m}_n \geqslant D\}$ and $\tilde{A}_n = \{m_n \geqslant D\}$ for $D < C_1/2$. Then $\tilde{A}_n \subseteq A_n$ and

$$\mathbb{P}(A_n) \geqslant \mathbb{P}(A_n \cap B_n) \geqslant \mathbb{P}(\tilde{A}_n \cap B_n) = \mathbb{P}(\tilde{A}_n) + \mathbb{P}(B_n) - \mathbb{P}(\tilde{A}_n \cup B_n) \to 1,\ n \to \infty,$$

since $\lim_{n\to\infty} \mathbb{P}(B_n) = 1$ by (A4) and $\lim_{n\to\infty} \mathbb{P}(\tilde{A}_n) = 1$ by (A3) and by Equation 4.6 of Mösching and Dümbgen (2020, Section 4.3). □

Proof of Theorem 5.1. Proposition 3.1 implies that for all $u \in \mathbb{R}$,

$$\hat{F}_u(y; (\hat{\theta}_n(X_{nj}))_{j=1}^n, (Y_{nj})_{j=1}^n) = \hat{F}_{g(u)}(y; (g(\hat{\theta}_n(X_{nj})))_{j=1}^n, (Y_{nj})_{j=1}^n).$$

To lighten the notation, we can therefore drop g from (A4) and simply write $\hat{\theta}_n(\cdot)$ instead of $g(\hat{\theta}_n(\cdot))$. Assume that $\hat{\theta}_n(X_{n\pi(1)}) \leqslant \hat{\theta}_n(X_{n\pi(2)}) \leqslant \ldots \leqslant \hat{\theta}_n(X_{n\pi(n)})$ and define $\delta_n = (\log n/n)^{1/6}/2$. Lemma A.3 and (A4) imply that for all $x \in \mathcal{X}_n = \{x \in \mathcal{X} : [\theta(x) \pm 2\delta_n] \subseteq I\}$, the indices

$$\begin{aligned}r(x) &= \min\{j \in \{1,\ldots,n\} : \hat{\theta}_n(X_{n\pi(j)}) \geqslant \hat{\theta}_n(x) - \delta_n\}\\ j(x) &= \max\{j \in \{1,\ldots,n\} : \hat{\theta}_n(X_{n\pi(j)}) \leqslant \hat{\theta}_n(x)\}\end{aligned}$$

are well defined with asymptotic probability one, because $[\hat{\theta}_n(x)-\delta_n, \hat{\theta}_n(x)]$ is of length δ_n and contained in I since $\theta(x) + (\log n/n)^{1/6} \geqslant \hat{\theta}_n(x) \geqslant \hat{\theta}_n(x) - \delta_n \geqslant \theta(x) - \delta_n - C_0 n^{-1/2} > \theta(x) - (\log n/n)^{1/6}$ for n sufficiently large, on the event B_n defined at (10). They satisfy $r(x) \leqslant j(x)$ and $\hat{\theta}_n(x) - \delta_n \leqslant \hat{\theta}_n(X_{nr(x)}) \leqslant \hat{\theta}_n(X_{nj(x)}) \leqslant \hat{\theta}_n(x)$ and, with asymptotic probability one due to Lemma A.3, $w_{r(x)j(x)} = \#\{j \in \{1,\ldots,n\} : \hat{\theta}_n(x) - \delta_n \leqslant \hat{\theta}_n(X_{n\pi(j)}) \leqslant \hat{\theta}_n(x)\} \geqslant Dn\delta_n$ for $0 < D < C_1/2$. Therefore, almost surely with respect to the joint law of (X_{ni}, Y_{ni}),

$i = 1, \ldots, n$, for any $y \in \mathbb{R}$,

$$\begin{aligned}
\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y) &\leqslant \hat{F}_{n;\hat{\theta}_n(X_{nj(x)})}(y) - F_{\theta(x)}(y) \\
&= \min_{r \leqslant j(x)} \max_{s \geqslant j(x)} \hat{\mathbb{F}}^{\pi}_{rs}(y) - F_{\theta(x)}(y) \\
&\leqslant \max_{s \geqslant j(x)} \hat{\mathbb{F}}^{\pi}_{r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant w_{r(x)j(x)}^{-1/2} M_n^{\pi} + \max_{s \geqslant j(x)} \bar{F}^{\pi}_{\theta;r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M_n^{\pi} \\
&\qquad + \max_{s \geqslant j(x)} \left(\bar{F}^{\pi}_{\theta;r(x)s}(y) - \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) + \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) \right) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M_n^{\pi} + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + \max_{s \geqslant j(x)} \bar{F}^{\pi}_{\hat{\theta};r(x)s}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M_n^{\pi} + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + F_{\hat{\theta}_n(X_{nr(x)})}(y) - F_{\theta(x)}(y) \\
&\leqslant (Dn\delta_n)^{-1/2} M_n^{\pi} + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + L|\hat{\theta}_n(X_{nr(x)}) - \theta(x)| \\
&\leqslant (Dn\delta_n)^{-1/2} M_n^{\pi} + L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + L\delta_n.
\end{aligned}$$

The equality in the second line is the classical min-max formula for monotone regression, see e.g. Equation (2.2) in Mösching and Dümbgen (2020), and the first and the third last inequality use antitonicity of $u \mapsto F_u(y)$. By assumption (A4) and with the constant $s > 0$ from Lemma A.2, the event

$$\{M_n^{\pi} \leqslant s(n \log(n))^{1/4}\} \cap \left\{ \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| < \delta_n \right\}$$

has asymptotic probability one. On this event, the previous considerations imply

$$\sup_{x \in \mathcal{X}_n, y \in \mathbb{R}} (\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)})(y) \leqslant s(Dn\delta_n)^{-1/2}(n \log(n))^{1/4} + 2L\delta_n \leqslant C \left(\frac{\log(n)}{n} \right)^{1/6},$$

with $C = [s(2D^{-1})^{1/2} + L]$. To finish the proof, we show that $F_{\theta(x)}(y) - \hat{F}_{n;\hat{\theta}_n(x)}(y)$ can be bounded in the same way.

Similar to before, define the indices $r'(x) = \min\{j \in \{1, \ldots, n\} : \hat{\theta}_n(X_{nj}) \geqslant \hat{\theta}_n(x)\}$, $j'(x) = \max\{j \in \{1, \ldots, n\} : \hat{\theta}_n(X_{nj}) \leqslant \hat{\theta}_n(x) + \delta_n\}$. Then with asymptotic probability one, also $r'(x) \leqslant j'(x)$ and $\hat{\theta}_n(x) \leqslant \hat{\theta}_n(X_{nr'(x)}) \leqslant \hat{\theta}_n(X_{nj'(x)}) \leqslant \hat{\theta}_n(x) + \delta_n$, $w_{r'(x)j'(x)} \geqslant Dn\delta_n$.

Thus,

$$\begin{aligned}
\hat{F}_{n;\hat{\theta}_n(x)}(y) - F_{\theta(x)}(y) &\geqslant \hat{F}_{n;\hat{\theta}_n(X_{nr'(x)})}(y) - F_{\theta(x)} \\
&= \min_{r \leqslant r'(x)} \max_{s \geqslant r'(x)} \hat{\mathbb{F}}^{\pi}_{rs}(y) - F_{\theta(x)}(y) \\
&\geqslant \min_{r \leqslant r'(x)} \hat{\mathbb{F}}^{\pi}_{rj'(x)}(y) - F_{\theta(x)}(y) \\
&\geqslant -w^{-1/2}_{r'(x)j'(x)} M^{\pi}_n + \min_{r \leqslant r'(x)} \bar{F}^{\pi}_{\theta;rj'(x)}(y) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n \\
&\qquad + \min_{r \leqslant r'(x)} \big(\bar{F}^{\pi}_{\theta;rj'(x)}(y) - \bar{F}^{\pi}_{\hat{\theta};rj'(x)}(y) + \bar{F}^{\pi}_{\hat{\theta};rj'(x)}(y)\big) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| + F_{\hat{\theta}_n(X_{nj'(x)})}(y) - F_{\theta(x)}(y) \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| - L|\hat{\theta}_n(X_{nj'(x)}) - \theta(x)| \\
&\geqslant -(Dn\delta_n)^{-1/2} M^{\pi}_n - L \sup_{x \in \mathcal{X}} |\hat{\theta}_n(x) - \theta(x)| - L\delta_n. \qquad \square
\end{aligned}$$

Proof of Theorem 5.1 with sample splitting. Assume that the index estimator $\hat{\theta}_n$ is computed with data $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$ and the distribution functions with $(\hat{\theta}_n(X_{ni}), Y_{ni})_{i=\lfloor n\xi \rfloor+1}^{n}$. The statement of Lemma A.3 also holds when $C_0(\log(n)/n)^{1/2}$ is replaced by $(\log(n)/n)^{1/3}$. By conditioning on $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$ and on X_{ni}, $i = \lfloor n\xi \rfloor + 1, \ldots, n$, Corollary 4.7 of Mösching and Dümbgen (2020) implies that M^{π}_n (computed with the data $(\hat{\theta}_n(X_{ni}), Y_{ni})_{i=\lfloor n\xi \rfloor+1}^{n}$) satisfies $\mathbb{P}(M^{\pi}_n \geqslant (R\log(n(1-\xi)))^{1/2}) \to 0$, $n \to \infty$, for any $R > 1$. This requires the fact that the permutation π is constant when conditioned on $(X_{ni}, Y_{ni})_{i=1}^{\lfloor n\xi \rfloor}$. One may now follow exactly the same steps as in the proof for the theorem without sample splitting, but with sample size $\lfloor n(1-\xi) \rfloor$ instead of n, $\delta_n = (n(1-\xi)/\log(n(1-\xi)))^{1/3}/2$ instead of $(n/\log(n))^{1/6}$ and $\{M^{\pi}_n \leqslant (R\log(n(1-\xi)))^{1/2}\}$ instead of $\{M^{\pi}_n \leqslant s(n\log(n))^{1/4}\}$, obtaining an upper bound of $C'(\log(n)/n)^{1/3}$ for the error, where $C' > 0$ also depends on ξ. □

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 37:1148–1178.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019a). Least squares estimation in the monotone single index model. *Bernoulli*, 25:3276–3310.

Balabdaoui, F. and Groeneboom, P. (2020). Profile least squares estimators in the monotone single index model. *arXiv e-prints*, page arXiv:2001.05454.

Balabdaoui, F., Groeneboom, P., and Hendrickx, K. (2019b). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46:517–544.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78:1093–1125.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81:2205–2268.

Chernozhukov, V., Fernández-Val, I., Melly, B., and Wüthrich, K. (2020). Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 115:123–137.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34:187–202.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, 70:609–627.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883.

Duarte, E., de Sousa, B., Cadarso-Suárez, C., Klein, N., Kneib, T., and Rodrigues, V. (2017). Studying the relationship between a woman's reproductive lifespan and age at menarche using a Bayesian multivariate structured additive distributional regression model. *Biometrical Journal*, 59:1232–1246.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B*, 69:163–183.

Foresi, S. and Peracchi, F. (1995). The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90:451–466.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69:243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

Gneiting, T. and Walz, E.-M. (2019). Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *arXiv e-prints*.

Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94:154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics*, 33:1404–1421.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27:1–32.

Henzi, A., Mösching, A., and Dümbgen, L. (2020). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. Preprint, arxiv.org/abs/2006.05527.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2019). Isotonic distributional regression. *arXiv e-prints*, page arXiv:1909.03725.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B*, 76:3–27.

Jordan, A. I., Mühlemann, A., and Ziegel, J. F. (2019). Optimal solutions to the isotonic regression problem. *arXiv e-prints*, page arXiv:1904.04761.

Klein, N., Kneib, T., Lang, S., and Sohn, A. (2015). Bayesian structured additive distributional forecasting with an application to regional income inequality in Germany. *Annals of Applied Statistics*, 9:1024–1052.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2020). *quantreg: Quantile Regression*. R package version 5.55.

Kramer, A. A. (2017). Are ICU length of stay predictions worthwhile? *Critical Care Medicine*, 45:379–380.

Kuchibhotla, A. K., Patra, R. K., and Sen, B. (2017). Least squares estimation in a single index model with convex Lipschitz link. *arXiv e-prints*, page arXiv:1708.00145.

Lanteri, A., Maggioni, M., and Vigogna, S. (2020). Conditional regression for single-index models. *arXiv e-prints*, page arXiv:2002.10008.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*, 270:2957–2963.

Li, Q. and Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26:423–434.

Machado, J. A. F. and Mata, J. (2000). Box–Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics*, 15:253–274.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22:1087–1096.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 2nd edition.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Miranda, D. R., Moreno, R., and Iapichino, G. (1997). Nine equivalents of nursing manpower use score (NEMS). *Intensive Care Medicine*, 23:760–765.

Moran, J. L. and Solomon, P. J. (2012). A review of statistical estimators for risk-adjusted length of stay: analysis of the Australian and new Zealand intensive care adult patient data-base, 2008–2009. *BMC Medical Research Methodology*, 12:68.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

Niskanen, M., Reinikainen, M., and Pettilä, V. (2009). Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. *Intensive Care Medicine*, 35:1060–1067.

Peracchi, F. (2002). On estimating conditional quantiles and distribution functions. *Computational Statistics & Data Analysis*, 38:433–447.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C*, 54:507–554.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics*, 13:1564–1589.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer, New York.

Silbersdorff, A., Lynch, J., Klasen, S., and Kneib, T. (2018). Reconsidering the income-health relationship using distributional regression. *Health Economics*, 27:1074–1088.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28:673–687.

Umlauf, N., Klein, N., and Zeileis, A. (2018). Bamlss: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, 27:612–627.

Vannitsem, S., Wilks, D. S., and Messner, J., editors (2018). *Statistical Postprocessing of Ensemble Forecasts*. Elsevier.

Verburg, I. W., de Keizer, N. F., de Jonge, E., and Peek, N. (2014). Comparison of regression methods for modeling intensive care length of stay. *PloS one*, 9(10):e109684.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Elsevier, 3rd edition.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition.

Zhang, J., Chen, Q., Lin, B., and Zhou, Y. (2017). On the single-index model estimate of the conditional density function: Consistency and implementation. *Journal of Statistical Planning and Inference*, 187:56–66.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., Malila, F. M., and Shaffer, V. L. (2006). Intensive care unit length of stay: Benchmarking based on acute physiology and chronic health evaluation (APACHE) IV. *Critical care medicine*, 34:2517–2529.

Zou, Q. and Zhu, Z. (2014). M-estimators for single-index model using B-spline. *Metrika*, 77:225–246.

B Supplementary Material

B.1 Model selection

All models in the data application (Section 6 in the article) have been fine-tuned, and different model variants were evaluated via out-of-sample predictions on the part of the data left for model selection. Table S1 and Table S2 provide detailed results and show that the performance of the methods is robust in terms of CRPS-ranking and consistent with the findings in the article. The key steps in model tuning are summarized below.

Response transformations: The outcome variable, LoS, is strongly right skewed, which suggests a log-transformation.

- The index estimation for DIM may benefit from response transformations, but the transformation does not directly have an impact on the estimation of the conditional CDFs. Index models with (lognormal, scaled-t) and without (gamma) log-transformation of the LoS have been considered, c.f. Section 6.2 in the article.

- Cox regression is invariant under strictly isotonic transformations of the response, so no response transformations need to be considered.

- Quantile regression is more robust to outliers than regression models for the mean, and it does not necessarily require transformations with skew response variables. Nevertheless, we verified if transformation $y \mapsto \log(y+1)$ as used in the DIM index estimation improves the results. (The transformation $\log(y)$ was also checked but clearly inferior.) The transformed model gave only a minor improvement on average over the ICUs, and diverging, meaningless distributions for some patients (removed for the computation of the averages in Table S1), and has therefore been discarded.

Covariate selection: The choice of covariates, including modelling effects of continuous variables with splines, can be expected to have similar effects for all methods.

- In all models, cubic splines were used to model the effects of the continuous variables age, NEMS and SAPS II. For Cox regression and for the index in DIM, determining a suitable dimension of the spline basis was done by using `k.check` of the `mgcv` package and by graphical tools for checking the robustness of the fit. The dimension parameter `k` was finally fixed at 12 for both regression methods.

- For quantile regression, cubic splines with equispaced knots or with knots at quantiles of the respective variables in the training data have been compared. The equispaced knots yielded better results, with a spline space dimension similar to the one for DIM and Cox regression. Additive quantile regression smoothing (`rqss`) in the `quantreg` package has been explored, but it only offers estimation at single quantiles for each fit and up to two continuous covariates, so the standard method `rq` has been selected.

- We have explored whether merging factor levels with few observations (less than 30 or 50 per category) improves the predictions. The effect was clearly negative for point forecasts for the mean LoS, as judged with the coefficient of predictive ability (Gneiting and Walz, 2019), and has not been further pursued.

Model-specific parameters:

- DIM: The influence of different parametric families for the index function is discussed in Section 6.2 in the article, see also Table S1. Detailed results on the CRPS differences with and without bagging are in Table S2.

- Cox regression: A possibility to make Cox regression more flexible is stratification by categorical variables. We did not pursue this approach because it may drastically reduce the number of observations for the baseline hazard estimation and thus for the CDF estimation for some groups of patients. (This may be less a problem if the hazard rate and not the distribution functions are the object of interest.)

- Quantile regression: Quantile regression is estimated on a grid of quantiles. As mentioned in the article, grids with spacing of 0.01 and 0.001 have been compared. In principle, the function `rq` in the `quantreg` package offers estimation of the full quantile regression process, but the resulting grid was too fine and led to computational difficulties. As can be seen by comparing the sixth and seventh column in Table S1, a finer grid consistently reduces the CRPS over the ICUs. But given that the improvement by moving from a spacing of 0.01 to 0.001 is rather small, we expect only minor benefits from estimating the whole quantile regression process.

B.2 Discreteness of LoS

Chernozhukov et al. (2013, Appendix SB) demonstrate that discreteness in the outcome variable influences the performance of quantile regression relative to other distributional regression techniques. Table S3 and Figure S1 summarize the cumulative proportion of the most frequent LoS values for each ICU as a measure of discreteness. Compared with Figure SB.1. in the supplementary material of Chernozhukov et al. (2013), the discreteness in the outcome variable is substantially lower in our study. Moreover, there is no relationship between the performance of DIM and Cox regression relative to quantile regression (Table 3 in the article) and the degree of discreteness as summarized in Table S3. As mentioned in the first paragraph of Section 6.4 and visible in Figure 3 in the article, quantile regression indeed has difficulties in fitting the pattern in the ICU discharge times with marked peaks before noon and in the afternoon. Nevertheless, it clearly outperforms Cox regression, which is able to correctly recognize this pattern. Based on these two observations, we argue that the disadvantage of quantile regression due to discreteness of the LoS is at most of limited extent and not decisive in our study.

Table S1: Mean CRPS on data for model selection for different variants of distributional regression methods. Asterisks (*, **, ***) indicate the three models with the lowest CRPS for each ICU. The DIM models are abbreviated as `logn`, `scat` and `gamma` for the variants with lognormal, scaled-t and gamma parametric families for index estimation, without bagging. The codes for quantile regression represent models with equispaced knots for splines (`e`) or with knots at quantiles of the respective variables (`q`), untransformed response variable (`u`) or with the transformation $\log(1 + y)$ (`log`). The first quantile regression model (sixth column in table) is fitted on a grid with spacing 0.001 (`0.001`), the others on a grid with spacing 0.01.

	Cox. reg.	DIM			Quantile regression				
Tuning		`logn`	`scat`	`gamma`	`e_u_0.001`	`e_u`	`e_log`	`q_log`	`q_u`
ICU4	1.212	1.188*	1.193***	1.190**	1.202	1.203	1.200	1.203	1.204
ICU6	1.632	1.606*	1.610**	1.622***	1.628	1.631	1.623	1.628	1.644
ICU10	1.094	1.076**	1.081***	1.075*	1.098	1.099	1.090	1.092	1.103
ICU19	1.253	1.248	1.252	1.262	1.241***	1.242	1.238**	1.237*	1.243
ICU20	1.880	1.839*	1.865***	1.853**	1.904	1.908	1.885	1.882	1.917
ICU24	0.972	0.937*	0.946	0.945***	0.948	0.948	0.951	0.945**	0.955
ICU33	0.903	0.895	0.897	0.897	0.893**	0.894	0.893*	0.893***	0.895
ICU39	1.907	1.865*	1.879***	1.870**	1.884	1.885	1.883	1.885	1.891
ICU44	2.266	2.232*	2.239***	2.238**	2.298	2.301	2.263	2.263	2.307
ICU47	1.306	1.233	1.255	1.245	1.220*	1.221**	1.234	1.227***	1.232
ICU52	2.034	1.998*	1.999**	2.012	2.002***	2.004	2.010	2.011	2.007
ICU55	1.196	1.178*	1.210	1.187	1.184	1.185	1.182**	1.182***	1.186
ICU58	1.344	1.312*	1.317**	1.320***	1.329	1.330	1.344	1.330	1.328
ICU65	1.069	1.004*	1.007**	1.010***	1.011	1.012	1.029	1.040	1.024
ICU76	2.552	2.521**	2.532***	2.517*	2.551	2.552	2.543	2.549	2.558
ICU77	0.838	0.832**	0.835***	0.825*	0.842	0.843	0.837	0.837	0.845
ICU79	1.266	1.211*	1.215**	1.233***	1.263	1.263	1.267	1.285	1.257
ICU80	0.996	0.983**	0.999	0.981*	0.998	0.998	0.992***	0.997	1.002
Mean	1.429	1.398*	1.407***	1.405**	1.416	1.418	1.415	1.416	1.422

Table S2: Increase in CRPS of the DIM when in-sample predictions on the training data are used for the estimation of the conditional CDFs instead of the bagging approach with 100 subsamples, for the lognormal, scaled-t and gamma index models. See Table S1 for the average CRPS without bagging. Positive values correspond to higher CRPS (worse predictions) of the variant without bagging.

	Lognormal	Scaled-t	Gamma
ICU4	0.0030	−0.001	0.0020
ICU6	0.0060	0.0050	0.0130
ICU10	0.0010	0.0010	0.0010
ICU19	0.0100	0.0060	0.0230
ICU20	0.0030	0.0240	0.0160
ICU24	0.0020	0.0040	0.0050
ICU33	0.0040	0.0010	0.0030
ICU39	0.0100	0.0070	0.0120
ICU44	−0.002	−0.002	0.0050
ICU47	0.0030	0.0020	0.0110
ICU52	0.0070	0.0040	0.0060
ICU55	0.0020	0.0190	0.0150
ICU58	0.0060	0.0040	0.0100
ICU65	0.0040	0.0030	0.0030
ICU76	000000	0.0030	000000
ICU77	0.0070	0.0070	0.0020
ICU79	0.0070	−0.001	0.0110
ICU80	0.0040	0.0080	0.0020
Mean	0.0040	0.0050	0.0080

Figure S1: Cumulative probabilities of LoS attaining one of the k most frequent values, $k = 1, 2, \ldots, 25$, stratified by ICU (identifiers omitted).

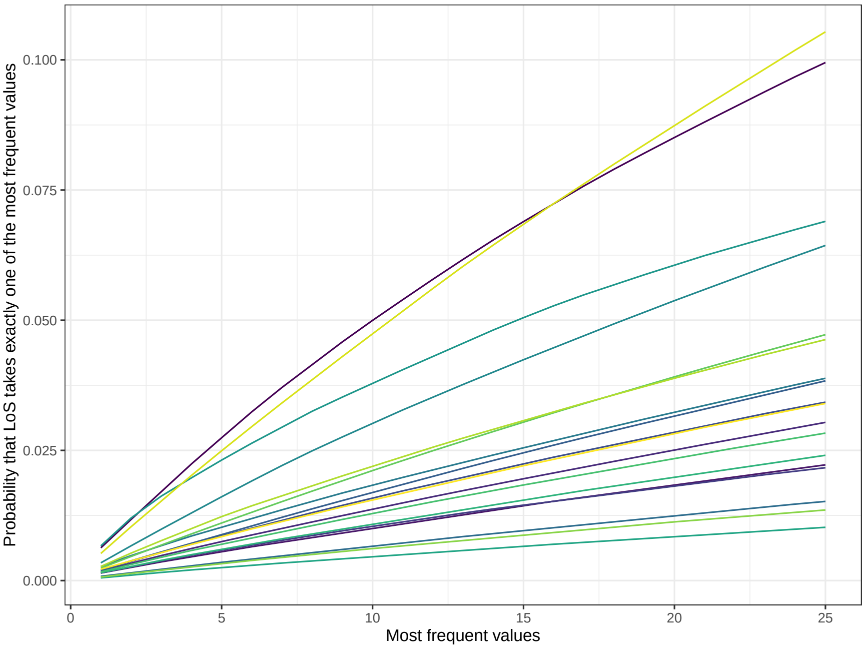


Table S3: Cumulative probabilities of LoS attaining one of the k most frequent values, $k = 1, 2, 10, 25$, stratified by ICU.

ICU	1	2	3	4	5	10	25
ICU4	0.006	0.012	0.017	0.022	0.027	0.050	0.099
ICU6	0.001	0.003	0.004	0.005	0.006	0.010	0.022
ICU10	0.002	0.003	0.005	0.006	0.007	0.014	0.030
ICU19	0.001	0.003	0.004	0.005	0.006	0.010	0.022
ICU20	0.002	0.004	0.005	0.007	0.009	0.016	0.034
ICU24	0.002	0.004	0.005	0.007	0.009	0.017	0.038
ICU33	0.001	0.001	0.002	0.003	0.004	0.007	0.015
ICU39	0.003	0.005	0.007	0.009	0.010	0.018	0.039
ICU44	0.003	0.007	0.010	0.013	0.016	0.030	0.064
ICU47	0.007	0.012	0.016	0.020	0.023	0.038	0.069
ICU52	0.001	0.001	0.002	0.002	0.002	0.005	0.010
ICU55	0.002	0.003	0.004	0.005	0.006	0.011	0.024
ICU58	0.002	0.003	0.004	0.006	0.007	0.013	0.028
ICU65	0.002	0.005	0.007	0.009	0.011	0.021	0.047
ICU76	0.001	0.001	0.002	0.003	0.003	0.006	0.014
ICU77	0.003	0.005	0.008	0.010	0.012	0.022	0.046
ICU79	0.005	0.010	0.015	0.020	0.025	0.047	0.105
ICU80	0.002	0.004	0.005	0.007	0.008	0.015	0.034

Table S4: Summary statistics (mean, median and standard deviation) of numeric variables in the dataset.

ICU identifier	LoS			Age			NEMS			SAPS		
	mean	median	sd	mean	median	sd	mean	median	sd	mean	median	sd
ICU4	2.1	0.7	4.4	65.7	68	14.5	25.4	25	9.6	29.0	26	14.5
ICU6	2.8	0.8	5.5	65.2	69	16.7	23.3	21	7.9	35.8	33	18.1
ICU10	2.0	0.7	3.9	62.9	66	15.7	27.8	27	8.8	41.6	39	18.2
ICU19	2.1	0.9	4.7	64.7	67	15.3	20.1	18	7.4	29.9	25	16.6
ICU20	3.7	0.7	8.0	64.2	67	15.3	25.5	24	8.3	31.6	27	17.5
ICU24	2.0	0.6	4.5	63.4	66	15.4	24.1	18	7.7	29.5	26	16.8
ICU33	2.0	1.0	3.3	66.1	69	15.8	19.9	18	7.5	36.5	33	17.4
ICU39	2.9	1.0	6.2	62.6	65	16.5	23.2	18	7.1	28.8	26	15.9
ICU44	3.9	1.5	7.8	59.0	61	17.6	27.1	27	8.5	34.0	31	18.9
ICU47	2.5	1.5	5.1	67.6	69	12.8	25.9	25	7.4	27.7	26	12.7
ICU52	3.7	1.6	6.3	60.5	63	17.3	26.2	27	10.3	40.8	39	18.5
ICU55	2.4	0.8	4.4	64.6	67	16.1	20.6	18	7.8	30.8	27	16.4
ICU58	2.6	0.7	4.8	61.7	64	16.4	22.5	18	7.3	28.5	26	15.0
ICU65	1.8	0.6	4.3	67.2	69	13.9	25.5	25	7.9	28.7	28	12.5
ICU76	4.3	1.7	7.2	63.2	66	15.6	30.3	30	8.3	41.2	40	17.2
ICU77	1.8	0.6	3.2	65.0	68	15.9	21.9	18	8.0	31.1	28	16.1
ICU79	2.7	0.5	5.9	55.8	57	17.0	22.4	18	7.1	19.1	15	15.3
ICU80	1.8	0.6	3.7	65.3	68	16.1	19.4	18	7.3	29.0	27	13.1

B.3 Additional figures and tables

Table S4 shows summary statistics of the ICU LoS, patient age, SAPS II and NEMS for all ICUs.

Figure S2 shows probabilistic LoS forecasts obtained by quantile regression, for eight randomly selected patients per ICU. While there are some crossings in the CDFs (e.g. in ICUs 47 and 52), the CDFs for most patients do not cross and are hence comparable with respect to stochastic dominance, suggesting that the model assumption of the DIM is reasonable for ICU LoS.

Figures S3 and S4 show reliability diagrams for the predicted probability that the LoS exceeds $k = 1, 2, \ldots, 14$ days for all forecasting methods and ICUs. PIT histograms are shown in Figures S5 and S6.

Figure S7 shows the difference in CRPS between the quantile regression forecasts and the DIM forecasts. For all ICUs, there is a considerable number of outliers (defined as points outside the 25% (75%) quantile minus (plus) 1.5 times the interquartile range), so Wilcoxon's signed rank test was applied to compare the CRPS, instead of a t-test.

Figure S2: Predictive CDFs obtained by quantile regression, for randomly selected patients.

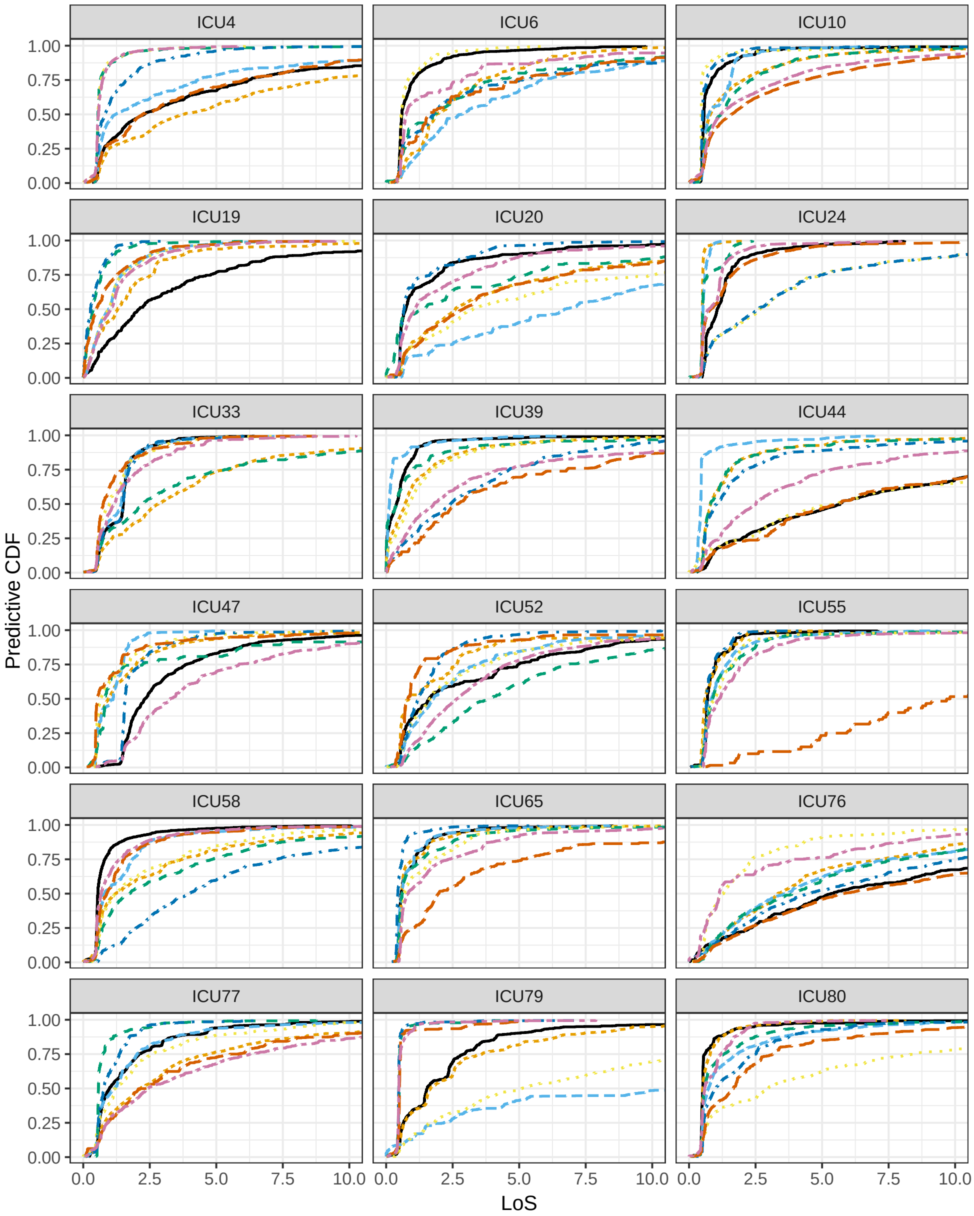


Figure S3: Reliability diagrams of probabilistic forecasts for the predicted probability that the LoS exceeds 1, 2, . . . , 7 days. The forecast probability is grouped into the bins $[0, 0.1], (0.1, 0.2], \ldots, (0.9, 1]$. Only bins with more than two observations are included.

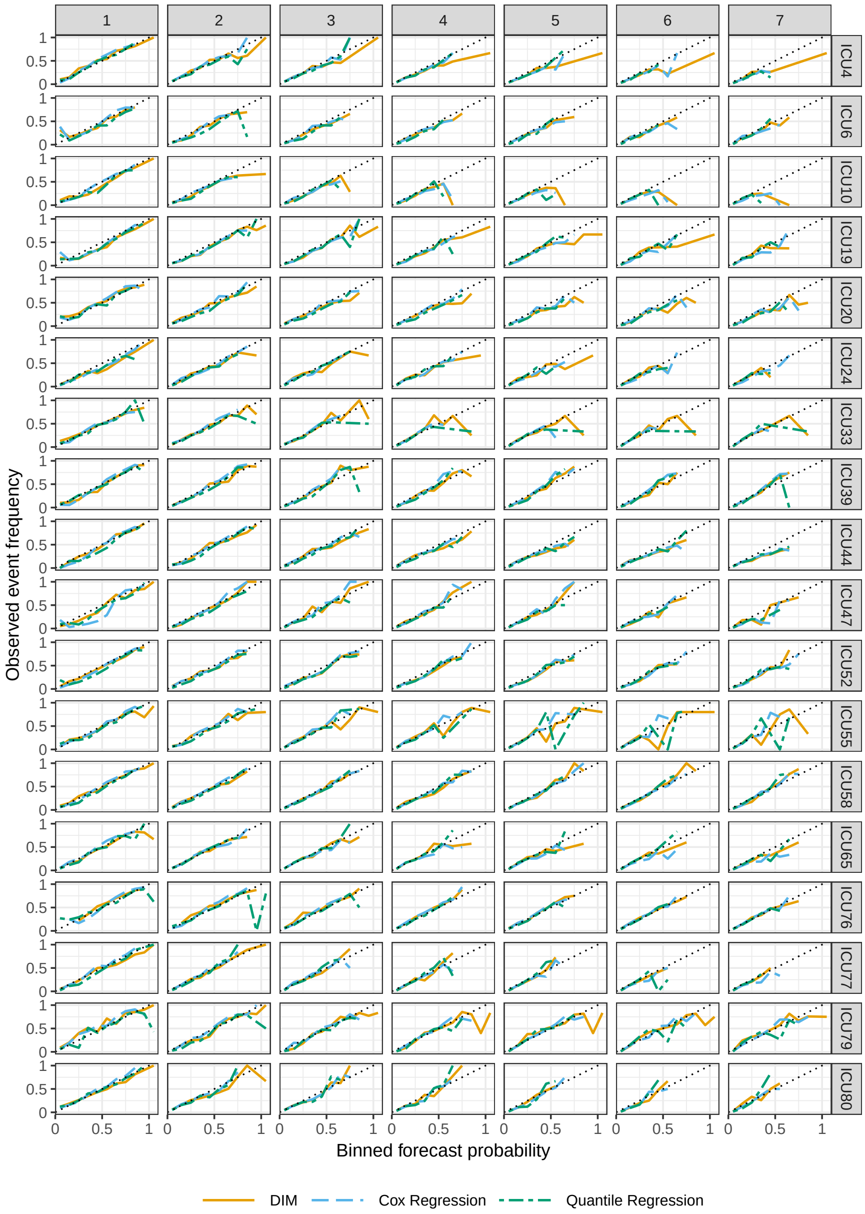


Figure S4: Reliability diagrams of probabilistic forecasts for the predicted probability that the LoS exceeds $8, 9, \ldots, 14$ days. The curves are as specified in Figure S3.

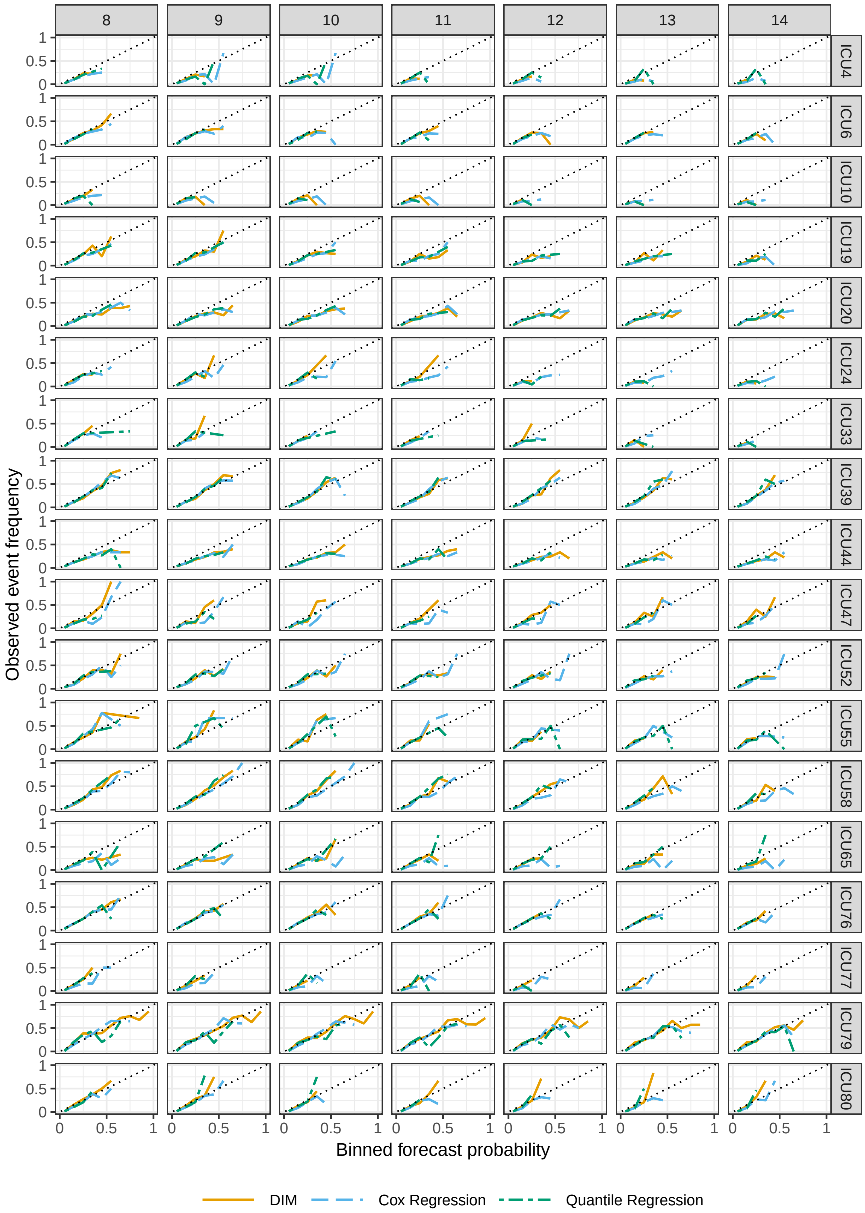


Figure S5: PIT histograms of the probabilistic forecasts with bins of width 1/20 (first nine ICUs).

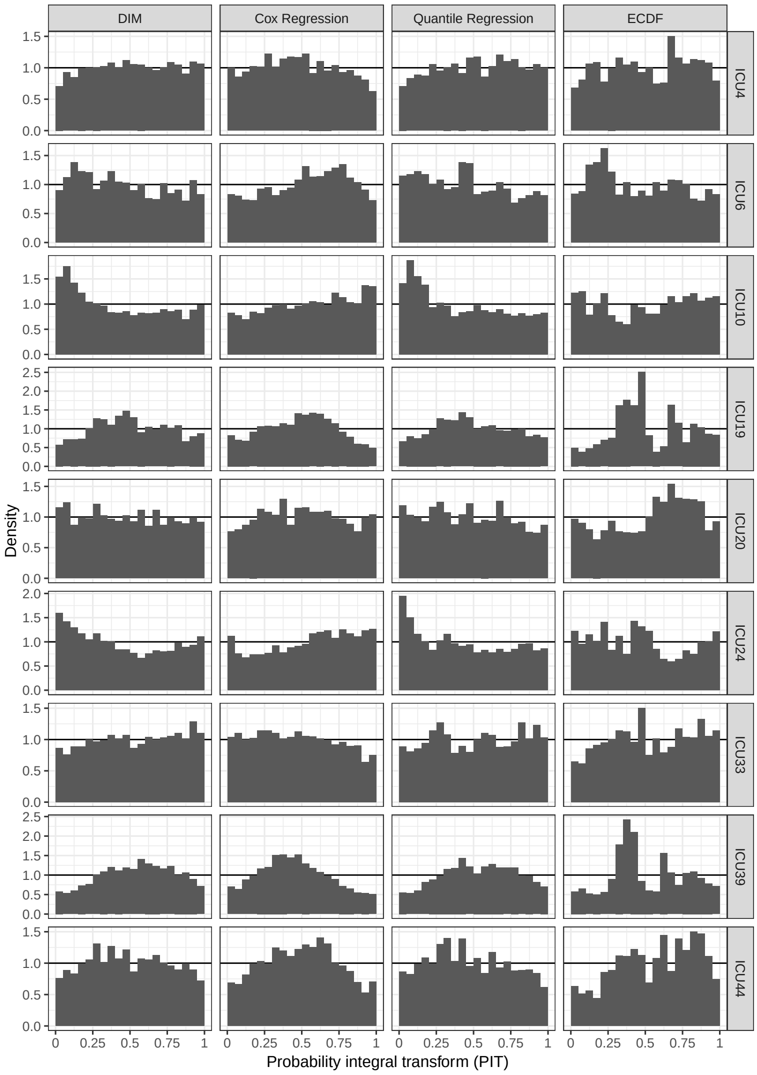


Figure S6: PIT histograms of the probabilistic forecasts with bins of width 1/20 (second half of the ICUs).

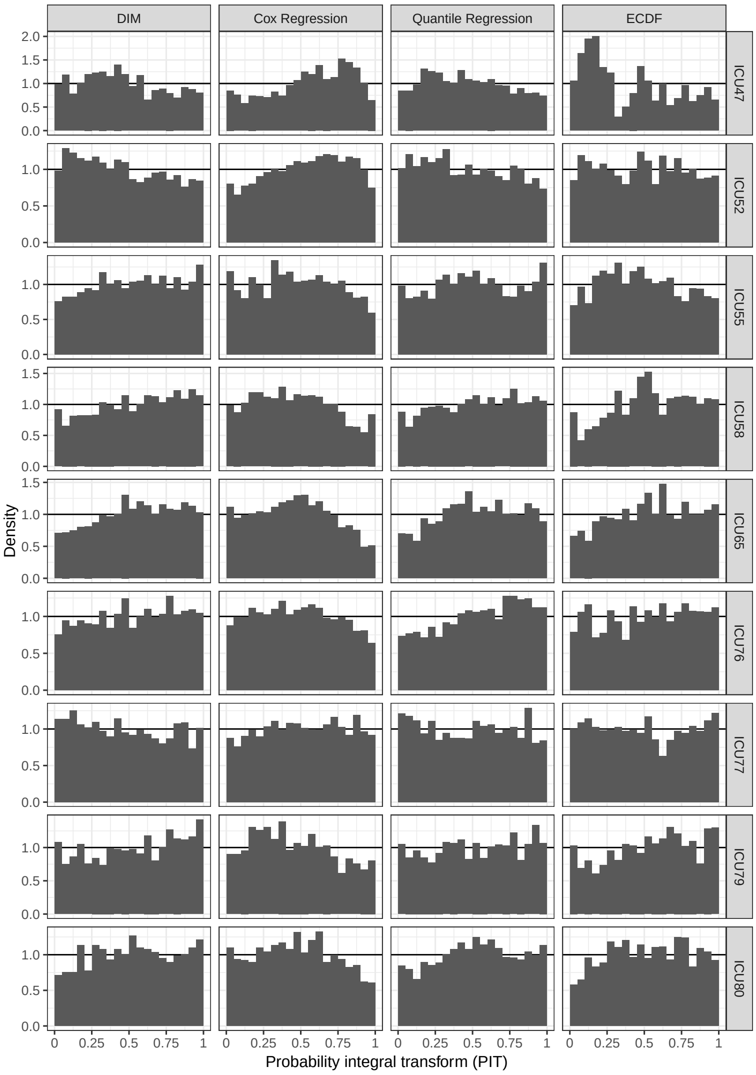
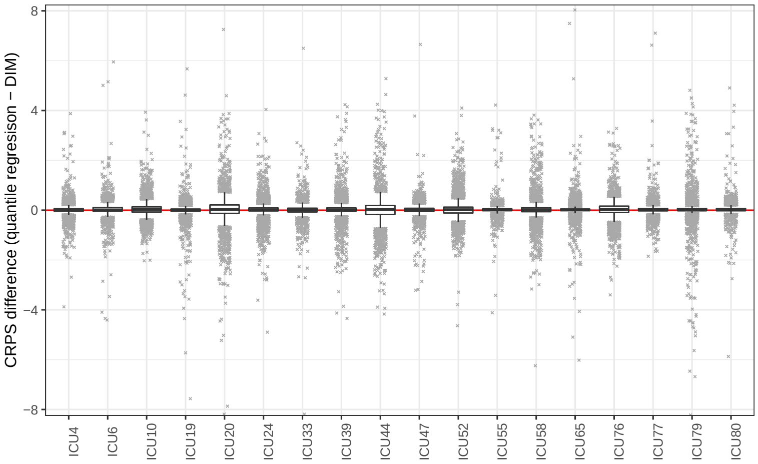


Figure S7: Boxplot of the difference in the CRPS of the quantile regression forecasts and of the DIM forecasts. Outliers are displayed as crosses (with horizontal jitter).



3.2 Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units

The content of this section is published as

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, P. D., Ziegel, J. F., on behalf of the RISC-19-ICU Investigators for Switzerland (2021). Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units. *PLOS ONE* **16** e0247265.

The article is followed by its supplementary material, which is also available on `https://doi.org/10.1371/journal.pone.0247265`.

RESEARCH ARTICLE

Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units

Alexander Henzi[1], Gian-Reto Kleger[2], Matthias P. Hilty[3,4], Pedro D. Wendel Garcia[3,4], Johanna F. Ziegel[1]*, on behalf of RISC-19-ICU Investigators for Switzerland[¶]

1 Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland, **2** Division of Intensive Care Medicine, Cantonal Hospital, St.Gallen, Switzerland, **3** The RISC-19-ICU Registry Board, University of Zurich, Zürich, Switzerland, **4** Institute of Intensive Care Medicine, University Hospital of Zürich, Zürich, Switzerland

¶ Membership of the RISC-19-ICU Investigators for Switzerland is provided in the Acknowledgments.
* johanna.ziegel@stat.unibe.ch

Abstract

Rationale

The COVID-19 pandemic induces considerable strain on intensive care unit resources.

Objectives

We aim to provide early predictions of individual patients' intensive care unit length of stay, which might improve resource allocation and patient care during the on-going pandemic.

Methods

We developed a new semiparametric distributional index model depending on covariates which are available within 24h after intensive care unit admission. The model was trained on a large cohort of acute respiratory distress syndrome patients out of the Minimal Dataset of the Swiss Society of Intensive Care Medicine. Then, we predict individual length of stay of patients in the RISC-19-ICU registry.

Measurements

The RISC-19-ICU Investigators for Switzerland collected data of 557 critically ill patients with COVID-19.

Main results

The model gives probabilistically and marginally calibrated predictions which are more informative than the empirical length of stay distribution of the training data. However, marginal calibration was worse after approximately 20 days in the whole cohort and in different subgroups. Long staying COVID-19 patients have shorter length of stay than regular acute respiratory distress syndrome patients. We found differences in LoS with respect to age categories and gender but not in regions of Switzerland with different stress of intensive care unit resources.

OPEN ACCESS

Citation: Henzi A, Kleger G-R, Hilty MP, Wendel Garcia PD, Ziegel JF, on behalf of RISC-19-ICU Investigators for Switzerland (2021) Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units. PLoS ONE 16(2): e0247265. https://doi.org/10.1371/journal.pone.0247265

Editor: Martina Crivellari, Vita Salute University of Milan, ITALY

Received: September 24, 2020

Accepted: February 3, 2021

Published: February 19, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0247265

Data Availability Statement: Any intensive care unit or other centre treating critically ill COVID-19 patients is invited to join the RISC-19-ICU registry at https://www.risc-19-icu.net. While the registry protocol prevents the deposition of the full registry dataset in a third-party repository, analyses on the full dataset may be requested by any collaborating centre after approval of the study protocol by the registry board. Reproducibility of the results in the present study is ensured by providing underlying code and an adapted dataset to exemplarily test the later. The registry protocol and data dictionary is publicly accessible at https://www.risc-19-icu.net.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Conclusion

A new probabilistic model permits calibrated and informative probabilistic prediction of LoS of individual patients with COVID-19. Long staying patients could be discovered early. The model may be the basis to simulate stochastic models for bed occupation in intensive care units under different casemix scenarios.

1 Introduction

During the COVID-19 pandemic, governments worldwide imposed severe restrictions on public life in order to limit the spread of the SARS-CoV-2 virus. A critical point in the decision making process was the limitation of beds in intensive care units (ICU) in order to adequately treat all severe cases of COVID-19. Many countries increased the number of ICU beds substantially at the onset of the crisis. A critical issue with severe COVID-19 disease is the frequent need for prolonged ICU treatment. For informed decision making it is important to quantitatively assess how long the patients are expected to be in an ICU.

At the example of Switzerland, we propose a prediction method for the individual length of stay (LoS) of patients in ICUs, and apply it to COVID-19 patients. The predictions are given for each patient based on covariates available within 24 hours after ICU admission. The method generates probabilistic predictions, that is, for each patient that enters the ICU, we provide a predictive cumulative distribution function (CDF) that comprehensively quantifies the uncertainty of the LoS at the time point of prediction. In particular, the predictive CDF allows to give prediction intervals with nay desired coverage probability. More precisely, the predictive CDF is an estimate of the conditional distribution of the LoS of the patient given covariates, which include age, gender, Simplified Acute Physiology Score (SAPS II) [1], and Nine Equivalents of nursing Manpower use Score (NEMS) (first shift) [2]. Fig 1 shows some predictive CDFs for randomly selected COVID-19 patients black, and true LoS as vertical lines. For each possible value t of the LoS, the value of the predictive CDF, $F(t)$, gives the probability that the patient stays at most t days in the ICU. Conversely, $1 - F(t)$ gives the probability that the patient stays longer than t days in the ICU. For example, patient 1 had an LoS of 20 days. The predicted probability that the patient stays at most 20 days was 0.91, and the probability for a stay of at least 10 days was 0.26 (or 0.74 for at most 10 days). Patient 4 stays longest in the ICU. This is in agreement with the predictive CDFs, since for all possible t, the probability of staying longer than t is highest for patient 4. The waves in the curves are explained by the fact that patients have a higher possibility to leave the ICU at certain times of the day, and a lower at others.

Probabilistic predictions allow to assess the uncertainty of the LoS comprehensively. Therefore they are preferable to forecasts for the mean or median LoS only. Their usefulness is illustrated by the following examples. The probabilistic LoS predictions allow to derive probabilistic forecasts for the number of patients who are still at the ICU at a certain day in future. This may be useful for planning purposes. For a single patient admitted today with predictive LoS distribution F, the probability that the patient is still at the ICU after t days equals $1 - F(t)$. From the probabilities for single patients, one may compute (with statistical software) the probability that any given number of patients is still at the ICU after t days. This allows to answer questions like 'How likely is it that there are at least two free beds in five days?' or 'What is the smallest number of patients we expect to stay until next week with a high probability (say, 90%)?'. The LoS forecasts, and so also the answers to such questions, take into account the individual characteristics of the patients currently at the ICU. The probabilistic LoS

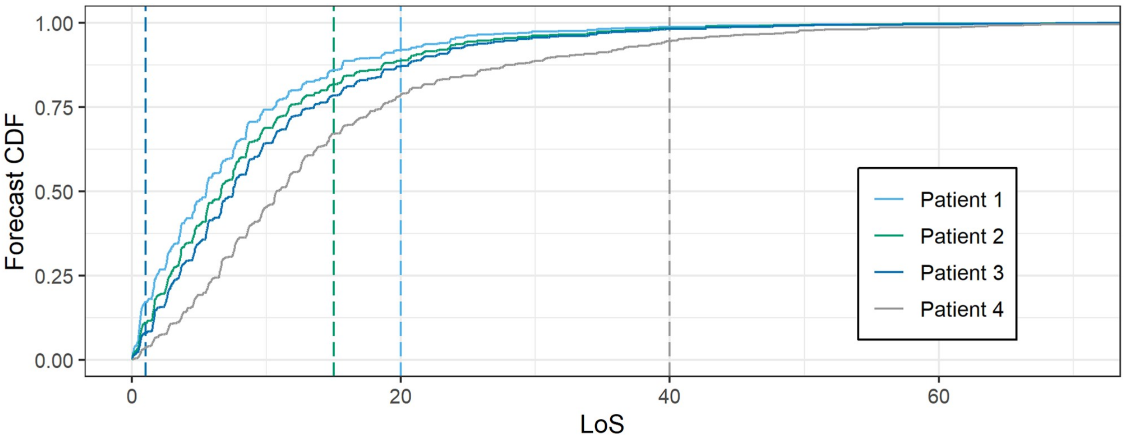


Fig 1. Predictive CDFs for the LoS of some COVID-19 patients with corresponding realizations as a vertical line. Four patients were drawn at random. The four wavy lines represent their predictive CDFs for the LoS based on covariates that are available at most 24 hours after ICU admission, that is, for each value *t* of the LoS on the horizontal axis, the curve gives the probability that the respective patient stays at most *t* days in the ICU. The vertical dashed lines represent the actually observed values of the LoS of the patients, which are unknown at the time of prediction. The larger the increase of the CDF on a given interval on the horizontal axis, the higher the probability of observing an LoS in this interval. For example, the predicted probability for the LoS of patient 1 being between 0 and 5 days is 0.47, whereas this probability is 0.40 for patient 2, 0.35 for patient 3, and only 0.19 for patient 4. The CDF of patient 4 lies substantially below the CDFs of the other patients which is in agreement with patient 4 having the longest realized LoS.

https://doi.org/10.1371/journal.pone.0247265.g001

predictions also allow to give alerts for patients that are likely to stay unusually long in the LOS. For example, fix a threshold of x days, say $x = 25$, and give an alert if the probability that the patient stays longer than x days exceeds, say, 90%. That is, if $1 - F(x) > 90\%$, where F is the predictive LoS distribution of a specific patient.

For planning of normal ward and intermediate care unit to ICU patient flows, such information is key to allow optimized resource allocation. On a larger scale, one could plan regional patient allocations to multiple hospitals based on such algorithms. The current health care crisis has emphasized the importance of patient flow logistics, and informative predictions of LoS are essential for this purpose.

It is documented in the literature that the prediction of the LoS at the patient level is difficult, and none of the available prediction models is providing satisfactory forecasts [3] with a possible exception being the complex models presented in [4, 5] for the purpose of benchmarking. Furthermore, the focus has almost exclusively been on only point predictions for the mean LoS, which is not ideal given that the LoS distribution is heavily skewed.

Recently, methodological progress has been made by Ziegel's group [6]: Based on data in the format of the Minimal Dataset of the Swiss Society of Intensive Care Medicine (MDSi), it is possible to give skillful and calibrated probabilistic predictions for the LoS of patients in ICUs 24h after their admission. In particular, the predictions for the probability of exceedance of the LOS over a certain threshold is shown to be reliable. The proposed method is semi-parametric, which makes it highly adaptive to the shape of the conditional LoS distributions. However, it requires large training data sets. The currently available data on COVID-19 patients in Swiss ICUs is (fortunately) not sufficient. Therefore, we suggest to borrow strength from the MDSi in order to predict the conditional LoS of COVID-19 patients.

The LoS of a patient in an ICU does not only depend on their physical condition but also on the characteristics and policies of the ICU. Even within a small country such as Switzerland such differences can be observed [6]. We restrict the analysis in this paper to Switzerland but the methodology can be adapted to other countries given sufficient data is available. We use the prediction method for the LoS to analyze the characteristics of the LoS of COVID-19 patients with respect to age differences, and gender differences. Since some parts of Switzerland were hit harder by the pandemic than others, we also use the predictions to analyse regional differences in the LoS.

2 Patients and methods

2.1 RISC-19-ICU and MDSi

Risk Stratification in COVID-19 patients in the Intensive Care Unit (RISC-19-ICU) registry, is a collaborative effort with the participation of a majority of the Swiss ICUs to provide a basis for decision support during the ongoing public health crisis, endorsed by the Swiss Society of Intensive Care Medicine (https://www.risc-19-icu.net/) [7, 8]. ICU data were reported on a daily basis, including near real-time data on LoS. The registry was deemed exempt from the need for additional ethics approval and patient informed consent by the ethics committee of the University of Zurich (KEK 2020-00322, ClinicalTrials.gov Identifier: NCT04357275). Fully anonymized datasets, in regard to Swiss law, were collected using a secure REDCap infrastructure provided by the Swiss Society of Intensive Care Medicine.

557 critically ill patients with COVID-19 that have been admitted to an ICU in Switzerland have entered the registry as of the snapshot date, June15, 2020, 481 of which have already been dismissed from the ICU or have died, that is for 86.36% of the patients the LoS is available. There are 18 patients for which one or more of the covariates are not available. Overall, covariates and LoS observations are available for 473 patients, and we call these the COVID-19 dataset. Censoring is a non-trivial problem in the COVID-19 dataset and we address this issue in detail in Section A of S1 Appendix.

The Minimal Dataset of the Swiss Society of Intensive Care Medicine (MDSi) has been introduced in 2005 and contains fully anonymized key data of the entire number of ICU patients in certified Swiss ICU's (https://www.sgi-ssmi.ch/de/datensatz.html). In addition to demographic data, the MDSi includes SAPS II as initial illness severity score and NEMS per nursing shift as a workload score.

Because almost any patient with severe COVID-19 disease presents chiefly like acute respiratory distress syndrome (ARDS), the training data consists of all patients in the MDSi with the diagnosis of ARDS which were admitted to Swiss ICUs in the years 2012 to 2018. Of the 2411 admissions, 856 were excluded because they satisfy one or more of the following criteria: missing or implausible values for SAPS II or NEMS (135), age younger than 16 (5), admitted with burns as initial diagnosis (3) or undergoing transplant operations 24 hours before or after ICU admission (8), readmissions (132), and patients admitted from ICUs or transferred to other ICUs (580). The exclusion of patients transferred from or to ICUs is because their LoS is incomplete and therefore not suitable for prediction. For the LoS predictions, admissions are standardized to a common admission time at midnight, in order to recover patterns in the ICU discharge times [6]. As a consequence, 99 patients had to be excluded because they did not stay in the ICU at least until midnight of the admission date. After exclusions, the training dataset consists of 1555 observations.

Concerning the covariates that are available for prediction, the possibilities are limited to covariates that are available in the COVID-19 dataset and the training data in the same format. Clear choices are the gender and age of patients. Furthermore, we have included SAPS II and the NEMS of the first ICU shift as covariates since they are informative for the LoS [9–11].

2.2 Statistical methods

Distributional Index Models (DIMs) have been introduced in [6]. They are semi-parametric models for distributional regression building on isotonic distributional regression (IDR) introduced in [12, 13]. A distributional regression model allows to estimate the full conditional distribution of the LoS given covariates. For the DIM used in this paper, we use a parametric model for a real-valued index function α, the DIM index, that depends on gender g, age a, SAPS II s, and NEMS m, that is

$$\alpha(g, a, s, m) = \beta_0 + \beta_1 \mathbf{1}\{g = \text{male}\} + \text{cr}_1(a) + \text{cr}_2(s) + \text{cr}_3(m),$$

where β_0 is the intercept, β_1 the coefficient for gender, and cr_1, cr_2, cr_3 are penalized cubic regression splines for the continuous variables age, SAPS II and NEMS; see the documentation of the `mgcv` package for details about the penalization. The model is fitted on the transformed LoS log(LoS+ 1). The log-transformation decreases the skewness of the data, while the addition of the constant 1 makes the resulting distribution more symmetric [6].

Furthermore, we assume that for the probability of the LoS Y of a randomly selected patient with covariates $(G, A, S, M) = (g, a, s, m)$ it holds that

$$\mathbb{P}(Y \leq y | (G, A, S, M) = (g, a, s, m)) = F_{\alpha(g,a,s,m)}(y), \quad \text{for all } y \in \mathbb{R} \tag{1}$$

with a family $(F_v)_{v \in \mathbb{R}}$ of stochastically ordered CDFs, that is $F_v(y) \leq F_w(y)$ for all $y \in \mathbb{R}$ if $v \geq w$.

We randomly split the training data in two and estimate α by $\hat{\alpha}$ on the first half. Given $\hat{\alpha}$, we use the second half of the training data to estimate F_v using IDR. In order to make the estimation procedure less dependent on the splitting of the training data, we use repeat this procedure 100 times and average the resulting estimated distributions to obtain our final estimate $\hat{F}\hat{\alpha}$.

There are dependencies between the covariates age, SAPS II and NEMS but we argue that it is still useful to include all of them in the model. The variable age is contained in SAPS II as a discretized effect with 6 levels. Age enters the model as a cubic regression spline with sufficiently high dimension, manually removing the age variable from SAPS II would essentially correspond to a basis transformation of the model and not affect the prediction results. The information provided by the NEMS is not redundant to SAPS II. NEMS is a crucial variable for COVID-19 patients since it contains information on the ventilation status, therapy with cardiovascular drugs and renal replacement treatment, which are not in the SAPS II. More precisely than the SAPS II, the NEMS reflects the actual therapeutic intensity a patients needs, and it is therefore likely to be one of the earliest markers for LoS.

Probabilistic predictions should be calibrated and sharp [14]. We assess probabilistic calibration by Probability Integral Transform (PIT) histograms, and use Pearson's chi-square test with 10 bins to test for uniformity. Marginal calibration is checked by comparing average predicted CDFs to empirical CDFs (ECDFs). Sharpness is assessed using the Continuous Ranked Probability Score (CRPS) and predictive power is compared with a Diebold-Mariano test based on the CRPS, see Section B of S1 Appendix.

The implementation is done in R 4.0 [15] using the packages mgcv [16] for the estimation of the index function, and isodistrreg for isotonic distributional regression [12]. Sample data and code are provided in the supplement S1 Code of this article.

3 Results

3.1 General

Summary statistics for the COVID-19 dataset and the training data are given in Table 1. The figures are correct for the June 15, 2020, snapshot. The proportion of men in the COVID-19 dataset is higher than in the training data set. The age structure of both datasets is similar with COVID-19 patients being slightly younger on average. COVID-19 patients generally have a higher NEMS in the first shift. The median and mean SAPS II is similar in both datasets.

Fig 2 provides a quantitative comparison of the LoS in the COVID-19 dataset and the training data. Panel (a) shows that the probability $\mathbb{P}(Y \geq y)$ of the LoS exceeding a fixed threshold y is larger for COVID-19 patients than in the training data up to about $y = 30$ days, and afterwards the relationship is reversed.

This observation does not exclude the possibility that given the covariates (G, A, S, M) for an individual patient, the conditional distribution of the LOS can be predicted well using the training data. Panel(b) of Fig 2 shows that the individual predictions are reasonable and are marginally calibrated up to about 25 days. The tail of the average forecast distribution is heavier than the tail of the empirical distribution of the COVID-19 dataset, meaning that very long LoS are less likely in the COVID-19 dataset.

The DIM predictions for the LoS of the COVID-19 patients have an average CRPS of 5.29 compared to 5.69, which is the average CRPS when predicting the LoS of the COVID-19 patients with the ECDF of the training data, that is, for all patients, independently of the covariates, the LoS is predicted by using the distribution of all the LoS values in the training data. This difference is highly significant with a p-value of less than $5 \cdot 10^{-4}$. This shows that the DIM predictions are significantly more informative than the ECDF forecast. The DIM predictions show better calibration than the ECDF predictions, see S3 Fig in S1 Appendix. Uniformity of the PIT is rejected for the ECDF forecasts (p-value $<10^{-4}$). For the DIM forecasts, uniformity of the PIT is not rejected (p-value: 0.384).

3.2 Age differences

Fig 3(a) gives the empirical CDFs of COVID-19 patients grouped by age. Young patients, less than 40 years, and very old patients, greater than 80 years have much shorter LoS than patients between 40 and 80. Patients between 40 and 65 tend to have a shorter LoS than patients between 65 and 80 except in cases of long LoS beyond 30 days. In Fig 3(c) the empirical CDFs

Table 1. Summary statistics of COVID-19 dataset and training data.

Variable	Data	Q25	Median	Mean	Q75	P-value
Age	training	55.0	67.0	63.8	75.0	$4.04 \cdot 10^{-3}$
	COVID-19	55.0	63.0	63.0	72.0	
LoS	training	4.5	9.1	12.4	15.8	$5.79 \cdot 10^{-5}$
	COVID-19	5.0	12.0	13.9	19.0	
NEMS	training	18.0	27.0	28.6	34.0	$<1.0 \cdot 10^{-16}$
	COVID-19	32.0	32.0	33.2	39.0	
SAPS II	training	35.0	46.0	48.5	59.0	$1.39 \cdot 10^{-1}$
	COVID-19	29.0	50.0	44.9	58.0	
Gender	training	Male: 61.9%		Female: 38.1%		$1.66 \cdot 10^{-8}$
	COVID-19	Male: 75.9%		Female: 24.1%		

P-values are for two-sided Wilcoxon's rank sum test for continuous variables and Fisher's exact test for gender.

https://doi.org/10.1371/journal.pone.0247265.t001

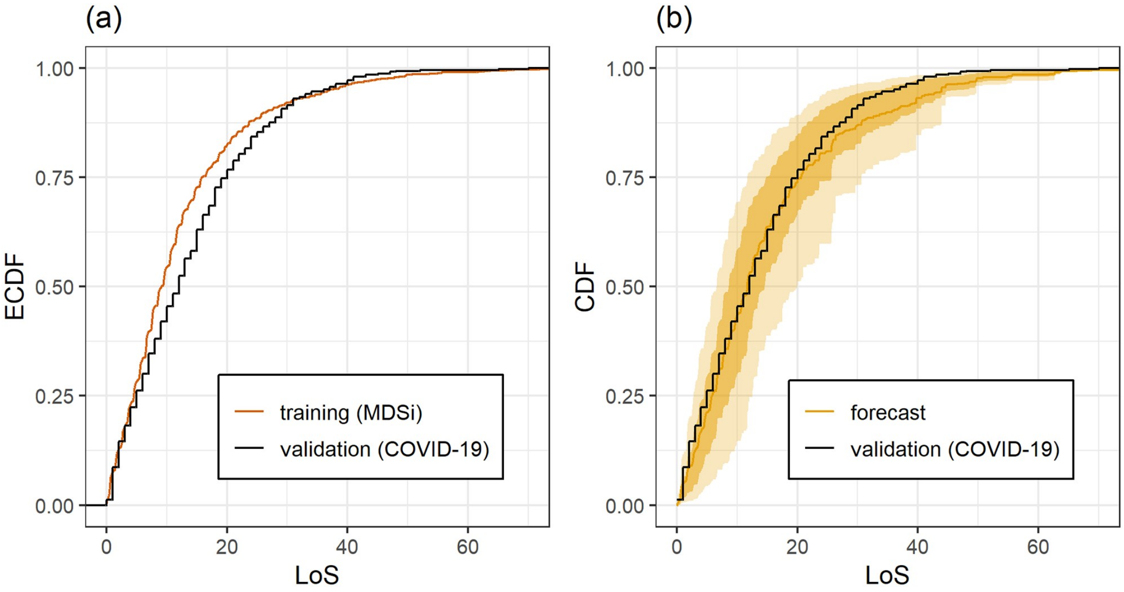


Fig 2. (a) EmpiricalCDF of the LoS in training and validation dataset. (b) Empirical CDF of the LoS in the validation dataset (black step function black, same as in panel (a)) and average LoS forecast for the COVID-19 patients (orange curve). Shaded areas show the pointwise 25% and 75% (10% and 90% for the outer bounds) quantiles of the predictive CDFs. For the average LoS forecast, the predictive CDFs of the COVID-19 patients are averaged pointwise, that is, the curves show the vertical average of the predictive CDFs for all patients in the COVID-19 dataset. The computation of the aggregated LoS CDFs is demonstrated in the sample code in S1 Code. The predictions take individual patient covariates into account and this allows to mitigate some differences between training and validation data observed in panel (a); for further discussion see text.

https://doi.org/10.1371/journal.pone.0247265.g002

are compared to the predictions based on the training data. The predictions for patients younger than 40 seem reasonable but their quality is hard to judge given the small sample size of this group in the COVID-19 dataset. For patients older than 80, the predicted LoS is longer than observed, but again, a definite statement should not be made due to small sample size. For patients between 40 and 65, marginal calibration is good until about 18 days. For higher thresholds, a longer LoS is predicted than observed. For patients between 65 and 80 years, the predictions give too much weight to LoS shorter than 25 days, and substantially overestimate the LoS beyond 25 days. Fig 3(b) shows that the training data leads to predictions of shorter LoS for patients younger than 40 and older than 80. In contrast to the COVID-19 data, the predicted LoS for patients between 65 and 80 is shorter than for patients between 40 and 65.

3.3 Gender differences

Fig 4(a) shows the empirical CDF of COVID-19 patients grouped by gender. Female patients show a slightly shorter LoS. The deviations of the predicted LoS from the observed LoS for male and female patients is displayed in Fig 4(c). Qualitatively the differences are similar with a slightly worse agreement of predictions and observations for female patients. The average predictive distributions for male and female patients are displayed in Fig 4(b). The predictions show a clear difference depending on gender with the same order as the COVID-19 data in that the LoS for women tends to be shorter than the one for men. However, the difference in average predicted LoS CDF is larger than the difference in ECDF based on the COVID-19 data.

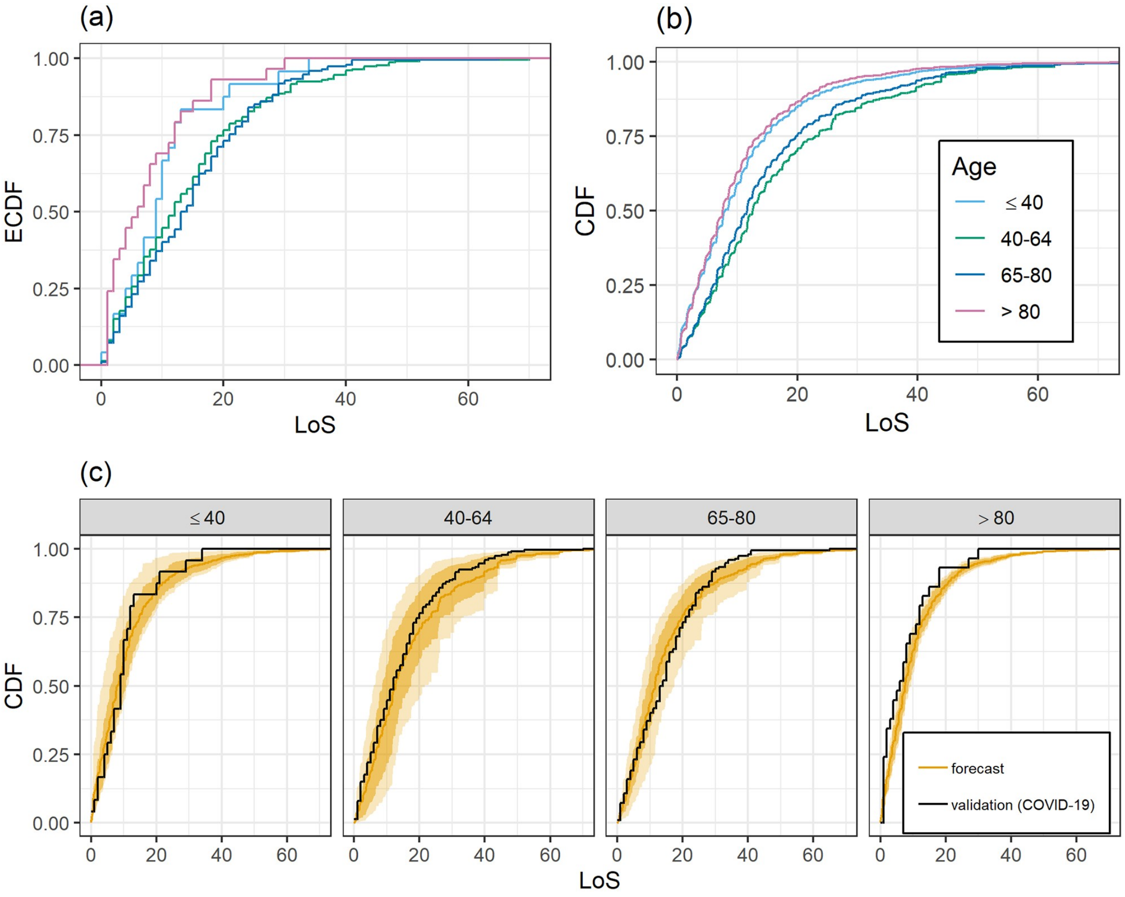


Fig 3. Depending on age: (a) Empirical LoS distributions of COVID-19 patients. (b) Average DIM forecasts for COVID-19 patients. (c) Empirical LoS distributions of COVID-19 patients and corresponding DIM forecasts. DIM forecasts are as in Fig 2.

https://doi.org/10.1371/journal.pone.0247265.g003

In order to gain some insight on the reasons for this effect, we checked if there is a significant difference in the LoS distribution of men and women in the training data. This is not the case. Furthermore, a comparison of the distribution of the DIM index computed for the men and women in the COVID-19 dataset shows that, indeed, the index values for women tend to be smaller than those for men, which explains the differences between the CDFs in Fig 4(b). In summary, it appears that a female patient with COVID-19 is likely to stay longer in the ICU than a similar female patient in the training data, whereas this effect is less pronounced for men.

3.4 Regional differences

We have split the COVID-19 dataset according to the location of the ICU within Switzerland. Region NE consisting of Northern and Eastern Switzerland and Region WT consisting of Western Switzerland and Ticino. Region WT was hit earlier and more severely by the COVID-19 crisis than Region NE. While ICU capacity limits were never reached in Region NE, ICU occupation was possibly critical in Region WT.

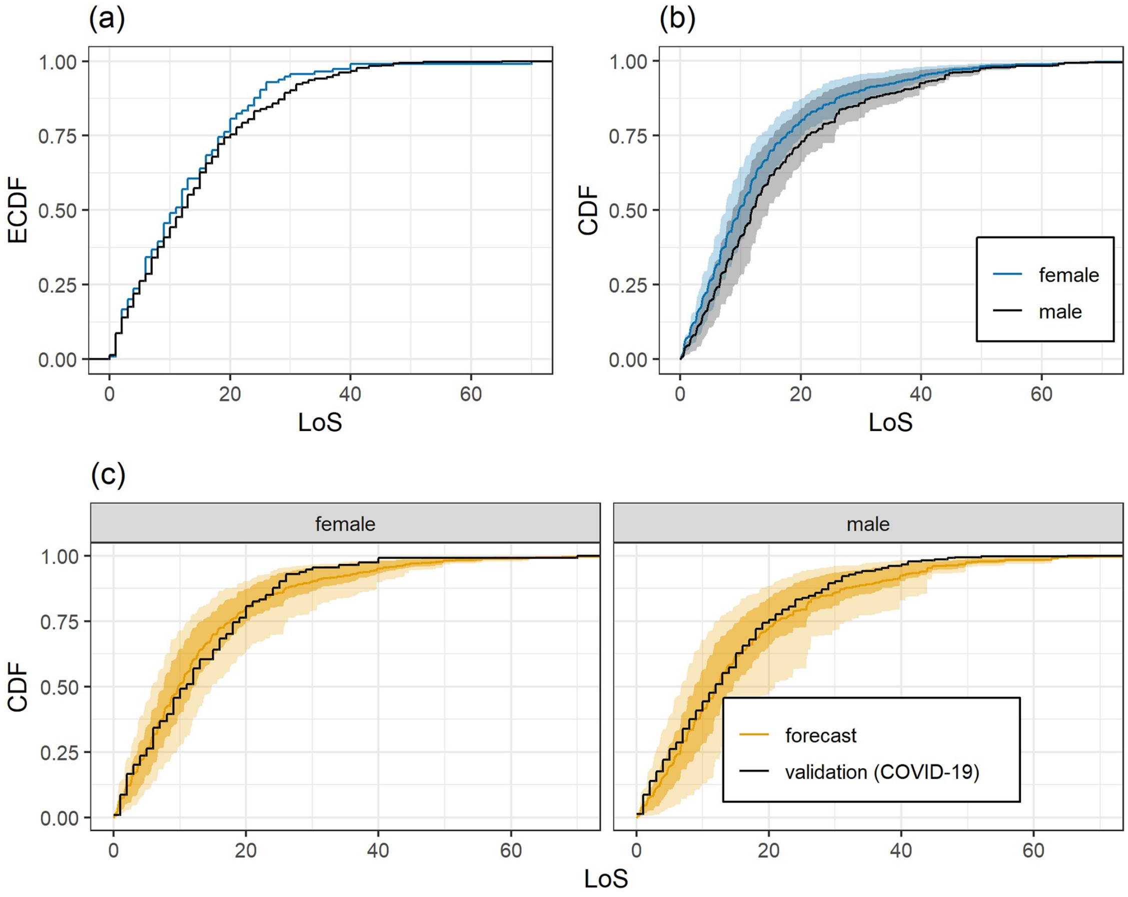


Fig 4. Depending on gender: (a) Empirical LoS distributions for COVID-19 patients. (b) Average DIM forecasts for COVID-19 patients. (c) Empirical LoS distributions of COVID-19 patients and corresponding DIM forecasts. DIM forecasts are as in Fig 2.

https://doi.org/10.1371/journal.pone.0247265.g004

The LOS distribution of COVID-19 patients is similar in both regions. The null hypothesis of equal LoS distribution in both regions cannot be rejected (two-sample Kolmogorov-Smirnov test p-value: 0.510, Wilcoxon rank sum test p-value: 0.607), see also S4 Fig in S1 Appendix. Comparing the regional LoS distributions to the DIM forecasts for the regions, we obtain that both regions show the same pattern: Good marginal calibration until about 25 days and then shorter LoS of the COVID-19 patients in comparison to the DIM predictions, see S5 Fig in S1 Appendix. The differences in the predictions for both regions are small, see S6 Fig in S1 Appendix.

4 Discussion

We have applied a new semi-parametric model, a DIM, for probabilistic predictions for the LoS of COVID-19 patients in Swiss ICUs. The model is trained with data from the MDSi, namely with data of patients with ARDS. Validation of the model using the COVID-19 dataset shows that the predictions are probabilistically calibrated, marginally calibrated (except for the tail of the distribution), and significantly more informative then an ECDF forecast based on the training data.

COVID-19 patients younger than 40 and older than 80 years tend to have a shorter stay in the ICU than the patient groups between 40–65 and 65–80 years. Predictions for patients older than 80 were longer than observed which could be an indicator of early treatment withdrawal in very old patients with severe COVID-19 disease. In the age groups 65-80 years, forecasts were shorter in the early phase than observations. This could be explained by prolonged recovery times compared with ARDS in elderly patients. The forecasts in both age groups (40–65 and 65–80 years) were longer after 25 to 30 days. In those patient groups, withdrawal of treatment is often executed after 20-30 days because of medical futility. The analysis of the LoS with respect to age suggests that there are differences between ARDS (training data) and COVID-19 in the sense that in terms of LoS COVID-19 patients might rather behave like slightly older ARDS patients keeping the other covariates fixed.

The difference between the LoS distribution of female and male COVID-19 patients is smaller than the difference between the predicted LoS distributions based on the training data, that is, non-COVID-19 patients with ARDS. For male patients the predictions agree better with the empirical distribution of observed LoS of the COVID-19 patients than for female patients. In terms of LoS, male COVID-19 patients behave more similarly to patients in the training data than female COVID-19 patients, making "longer than expected" LoS more likely for female than for male patients.

Despite the fact that the Western Switzerland and Ticino (Region WT) were hit earlier, and potentially less prepared for the COVID-19 crisis than Northern and Eastern Switzerland (Region NE), we do not see an impact on the LoS of COVID-19 patients.

There are somepossible shortcomings of our study. First, the training dataset is not on COVID-19 patients. Despite severe COVID-19 pneumonia behaving similar to ARDS, there are some important differences [17]. Furthermore, multiple organ involvement is frequent in severe COVID-19 disease [18, 19]. There have been discussions how and if classical ARDS and ARDS secondary to COVID-19 (C-ARDS) are different. Initially, substantial differences were postulated [20–22] but more recently consensus is growing that C-ARDS is most probably similar to classical ARDS in treatment intensity and therapeutic approach [23]. In view of this, the historical training data is as well chosen as historical data can be. Furthermore, the NEMS evaluates how severe or nursing intensive a patient is, independently of the diagnosis. Therefore, using is as a covariate in prediction is likely to mitigate confounders between training data and COVID-19 dataset. Second, a limitation is imposed by the use of MDSi as training dataset because the analysis is then constrained to the relatively few variables contained in MDSi. Clearly, there are further relevant predictors for COVID-19 patients. However, most of them concern mortality and not LoS, for example, coagulation status. These values are available in the RISC-19-ICU registry but not in the MDSi training data. Furthermore, we believe that a successful model for probabilistic predictions of LoS should rely on values that are routinely recorded and available early after hospitalization such as SAPS II and NEMS. Since they are compound variables, they are informative for the LoS. If training data sets with more covariates are available, the DIM model we propose in Section 2.2 could be adapted to variables specific to COVID-19 patients. This may lead to an increase in predictive skill. Third, there is possibly a bias towards a longer predicted LoS because of the data sampling process. We have assessed whether the patients with missing LoS value in the RISC-19-ICU registry have a substantially different distribution of covariate values than the patients with valid LoS value. This is not the case which is an indication that many of them, rather than having a censored LoS, have indeed not been updated. We have also repeated all of our analyses on the COVID-19 dataset restricted to patients with admission date before April 5, 2020. Here, the update and the censoring problem should be less. Qualitatively, we obtained the same results as the ones reported here. Nevertheless, it should be kept in mind that some of the very long LoS are likely to be censored in either case. Fourth, LoS is often not only dependent on epidemiological and physiologic variables but additionally on ICU resources, therapeutic restriction policies [24] and withdrawal strategies (https://www.samw.ch/de/Ethik/Themen-A-bis-Z/Intensivmedizin.html). Our forecasts predict a longer LoS compared with the observed LOS overall and in almost any patient subgroups after 25 days. This may be due to an earlier withdrawal of the intensive therapy compared to ARDS, especially in shortage of ICU resources. However we did not find any significant difference in LoS distribution between two regions of Switzerland with diverse ICU strain.

5 Conclusion

A new semiparametric model permits calibrated and informative probabilistic prediction of LoS of individual patients with severe COVID-19 in ICUs, given covariate information. These predictions would allow to simulate stochastic models for bed occupation in ICUs under different scenarios for the case mix. These scenarios could be different projections for the rate at which COVID-19 patients and other patients arrive in the ICUs.

Supporting information

S1 Appendix. Additional information about probabilistic forecasting, censoring of LoS in the COVID-19 dataset, and supplementary figures.
(PDF)

S1 Data. Minimal dataset to replicate the results of this study.
(ZIP)

S1 Code. Sample data and code to illustrate the computation and usage of probabilistic length of stay forecasts.
(ZIP)

Acknowledgments

The authors are grateful to the Swiss Society of Intensive Care Medicine for providing access to the MDSi.

RISC-19-ICU Investigators of Switzerland: RISC-19-ICU registry, University of Zurich (Hilty Matthias P, MD; Wendel Garcia Pedro D, MSc; Schüpbach Reto A, MD; Thierry Fumeaux, MD; Jonathan Montomoli, MD, PhD; Philippe Guerci, MD); Klinik für Operative Intensivmedizin, Kantonsspital Aarau, Aarau (Rolf Ensner, MD); Intensivstation, Kantonsspital Schaffhausen, Schaffhausen (Nadine Gehring, MD); Institut fuer Anesthaesie und Intensivmedizin, Zuger Kantonsspital AG, Baar (Peter Schott, MD; Severin Urech, MD); Department Intensivmedizin, Universitaetsspital Basel, Basel (Martin Siegemund, MD; Nuria Zellweger); Intensivmedizin, St. Claraspital, Basel (Adriana Lambert, MD; Lukas Merki, MD); Department of Intensive Care Medicine, University Hospital Bern, Inselspital, Bern (Marie-Madlen Jeitziner, RN, PhD; Beatrice Jenni-Moser, RN, MSc); Department Intensive Care Medicine, Spitalzentrum Biel, Biel (Marcus Laube, MD); Interdisziplinäre Intensivstation, Spital Bülach, Bülach (Bernd Yuen, MD; Thomas Hillermann, MD); Intensivstation, Regionalspital Emmental AG, Burgdorf (Petra Salomon, MD; Iris Drvaric, MD); Intensivmedizin, Kantonsspital Graubünden, Chur (Frank Hillgaertner, MD; Marianne Sieber); Institut fuer Anaesthesie und Intensivmedizin, Spital Thurgau, Frauenfeld (Alexander Dullenkopf, MD; Lina Petersen, MD);

Soins Intensifs, Hopital cantonal de Fribourg, Fribourg (Hatem Ksouri, MD, PhD; Govind Oliver Sridharan, MD); Division of Intensive Care, University Hospitals of Geneva, Geneva (Sara Cereghetti, MD; Filippo Boroli, MD; Jerome Pugin, MD, PhD); Division of Neonatal and Pediatric Intensive Care, University Hospitals of Geneva, Geneva (Serge Grazioli, MD; Peter C. Rimensberger, MD); Intensivstation, Spital Grabs, Grabs (Christian Bürkle, MD); Institut für Anaesthesiologie Intensivmedizin & Rettungsmedizin, See-Spital Horgen & Kilchberg, Horgen (Julien Marrel, MD; Mirko Brenni, MD); Soins Intensifs, Hirslanden Clinique Cecil, Lausanne (Isabelle Fleisch, MD; Jerome Lavanchy, MD); Anaesthesie und Intensivmedizin, Kantonsspital Baselland, Liestal (Anja Baltussen Weber, MD; Peter Gerecke, MD; Andreas Christ, MD); Dipartimento Area Critica, Clinica Luganese Moncucco, Lugano (Romano Mauri, MD; Samuele Ceruti, MD); Interdisziplinaere Intensivstation, Spital Maennedorf AG, Maennedorf (Katharina Marquardt, MD; Karim Shaikh, MD); Institut fuer Anaesthesie und Intensivmedizin, Spital Thurgau, Münsterlingen (Thomas Neff, MD; Tobias Hübner, MD); Intensivmedizin, Schweizer Paraplegikerzentrum Nottwil, Nottwil (Hermann Redecker, MD); Soins intensifs, Groupement Hospitalier de l'Ouest Lémanique, Hôpital de Nyon, Nyon (Thierry Fumeaux, MD; Mallory Moret-Bochatay, MD); Intensivmedizin & Intermediate Care, Kantonsspital Olten, Olten (Michael Studhalter, MD); Intensivmedizin, Spital Oberengadin, Samedan (Michael Stephan, MD; Jan Brem, MD); Anaesthesie Intensivmedizin Schmerzmedizin, Spital Schwyz, Schwyz (Daniela Selz, MD; Didier Naon, MD); Medizinische Intensivstation, Kantonsspital St. Gallen, St. Gallen (Gian-Reto Kleger, MD); Departement of Anesthesiology and Intensive Care Medicine, Kantonsspital St. Gallen, St. Gallen (Miodrag Filipovic, MD; Urs Pietsch, MD); Paediatric Intensive Care Unit, Children's Hospital of Eastern Switzerland, St. Gallen (Bjarte Rogdo, MD; Andre Birkenmaier, MD); Departement for intensive care medicine, Kantonsspital Nidwalden, Stans (Anette Ristic, MD; Michael Sepulcri, MD); Intensivstation, Spital Simmental-Thun-Saanenland AG, Thun (Antje Heise, MD); Klinik für Anaesthesie und Intensivmedizin, Spitalzentrum Oberwallis, Visp (Friederike Meyer zu Bentrup, MD, MBA); Service d'Anesthesiologie, EHNV, Yverdon- les-Bains (Marilene Franchitti Laurent, MD; Jean-Christophe Laurent, MD); Institute of Intensive Care Medicine, University Hospital Zurich, Zurich (Philipp Bühler, MD; Silvio Brugger, MD, PhD; Jan Bartussek, PhD; Martina Maibach, PhD; Annelies Zinkernagel, MD, PhD, Dorothea Heuberger, PhD; Srikanth Mairpady Shambat, PhD); Interdisziplinaere Intensivstation, Stadtspital Triemli, Zurich (Patricia Fodor, MD; Pascal Locher, MD; Giovanni Camen, MD); Abteilung für Anaesthesiologie und Intensivmedizin, Hirslanden Klinik Im Park, Zürich (Tomislav Gaspert, MD; Marija Jovic, MD); Institut für Anaesthesiologie und Intensivmedizin, Klinik Hirslanden, Zurich (Christoph Haberthuer, MD; Roger F. Lussman, MD).

Author Contributions

Conceptualization: Alexander Henzi, Gian-Reto Kleger, Johanna F. Ziegel.

Data curation: Matthias P. Hilty, Pedro D. Wendel Garcia.

Formal analysis: Alexander Henzi.

Investigation: Alexander Henzi, Gian-Reto Kleger, Johanna F. Ziegel.

Methodology: Alexander Henzi, Gian-Reto Kleger, Johanna F. Ziegel.

Project administration: Gian-Reto Kleger, Johanna F. Ziegel.

Software: Alexander Henzi, Matthias P. Hilty, Pedro D. Wendel Garcia.

Supervision: Johanna F. Ziegel.

Validation: Gian-Reto Kleger, Matthias P. Hilty, Pedro D. Wendel Garcia.

Visualization: Alexander Henzi.

Writing – original draft: Johanna F. Ziegel.

Writing – review & editing: Alexander Henzi, Gian-Reto Kleger, Matthias P. Hilty, Pedro D. Wendel Garcia, Johanna F. Ziegel.

References

1. Le Gall JR, Lemeshow S, Saulnier F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA. 1993; 270:2957–2963.
2. Miranda DR, Nap R, de Rijk A, Schaufeli W, Iapichino G, members of the TISS Working Group. Nursing activities score. Crit Care Med. 2003; 31:374–382. https://doi.org/10.1097/01.CCM.0000045567.78801.CC
3. Verburg IWM, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which models can I use to predict adult ICU length of stay? A systematic review. Crit Care Med. 2017; 45:e222–e231. https://doi.org/10.1097/CCM.0000000000002054 PMID: 27768612
4. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med. 2006; 34:2517–2529. https://doi.org/10.1097/01.CCM.0000240233.01711.D9
5. Vasilevskis EE, Kuzniewicz MW, Cason BA, Lane RK, Dean ML, Clay T, et al. Mortality probability model III and acute simplified physiology score II: Assessing their value in predicting length of stay and comparison to APACHE IV. Chest. 2009; 136:89–101. https://doi.org/10.1378/chest.08-2591 PMID: 19363210
6. Henzi A, Kleger GR, Ziegel JF. Distributional (Single) Index Models. Preprint. 2020;arXiv:2006.09219.
7. Wendel Garcia PD, Fumeaux T, Guerci P, Heuberger DM, Montomoli J, Roche-Campo F, et al. Prognostic factors associated with mortality risk and disease progression in 639 critically ill patients with COVID-19 in Europe: Initial report of the international RISC-19-ICU prospective observational cohort. EClinicalMedicine. 2020; p. 100449. https://doi.org/10.1016/j.eclinm.2020.100449 PMID: 32838231
8. Hilty MP, Wendel Garcia PD. hobbes8080/risc-19-icu: registry data transformation v1.0. Zenodo Data Repository. 2020. https://doi.org/10.5281/zenodo.3757064
9. Rothen HU, Stricker K, Einfalt J, Bauer P, Metnitz PG, Moreno RP, et al. Variability in outcome and resource use in intensive care units. Intensive Care Med. 2007; 33(8):1329–1336. https://doi.org/10.1007/s00134-007-0690-3 PMID: 17541552
10. Granholm A, Christiansen CF, Christensen S, Perner A, Mueller MH. Performance of SAPS II according to ICU length of stay: A Danish nationwide cohort study. Acta Anaesthesiol Scand. 2019; 63(9):1200–1209. https://doi.org/10.1111/aas.13415
11. Kleger GR. Die Aufenthaltsdauer kritisch kranker Patienten auf einer Intensivstation: Probabilistische Prädiktionsmodelle. Master's Thesis, University of Bern; 2018.
12. Henzi A, Ziegel JF, Gneiting T. Isotonic distributional regression. Preprint. 2019;arXiv:1909.03725.
13. Mösching A, Dümbgen L. Monotone least squares and isotonic quantiles. Electron J Stat. 2020; 14:24–49. https://doi.org/10.1214/19-EJS1659
14. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Series B Stat Methodol. 2007; 69:243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x
15. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: https://www.R-project.org/.
16. Wood SN. Generalized Additive Models: An Introduction with R. 2nd ed. Chapman and Hall/CRC; 2017.
17. Marini JJ, Gattinoni L. Management of COVID-19 Respiratory Distress. JAMA. 2020; 323:2329–2330. https://doi.org/10.1001/jama.2020.6825
18. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5 PMID: 31986264
19. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. 2020; 8(5):475–481. https://doi.org/10.1016/S2213-2600(20)30079-5 PMID: 32105632

20. Grasselli G, Tonetti T, Filippini C, Slutsky AS, Pesenti A, Ranieri VM. Pathophysiology of COVID-19-associated acute respiratory distress syndrome—Authors' reply. Lancet Respir Med. 2021; 9(1):e5–e6. https://doi.org/10.1016/S2213-2600(20)30525-7

21. Marini JJ, Gattinoni L. Time Course of Evolving Ventilator-Induced Lung Injury: The "Shrinking Baby Lung". Crit Care Med. 2020; 48(8):1203–1209. https://doi.org/10.1097/CCM.0000000000004416

22. Chiumello D, Busana M, Coppola S, Romitti F, Formenti P, Bonifazi M, et al. Physiological and quantitative CT-scan characterization of COVID-19 and typical ARDS: a matched cohort study. Intensive Care Med. 2020; 46:2187–2196. https://doi.org/10.1007/s00134-020-06281-2 PMID: 33089348

23. Trahtemberg U, Slutsky AS, Villar J. What have we learned ventilating COVID-19 patients? Intensive Care Med. 2020; 46:2458–2460.

24. Vincent JL, Creteur J. Ethical aspects of the COVID-19 crisis: How to deal with an overwhelming shortage of acute beds. Eur Heart J Acute Cardiovasc Care. 2020; 9(3):248–252. https://doi.org/10.1177/2048872620922788

Appendix S1:

Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units

Alexander Henzi[1], Gian-Reto Kleger[2], Matthias P. Hilty[3], Pedro D. Wendel Garcia[3], RISC-19-ICU Investigators for Switzerland[3], Johanna F. Ziegel[1]

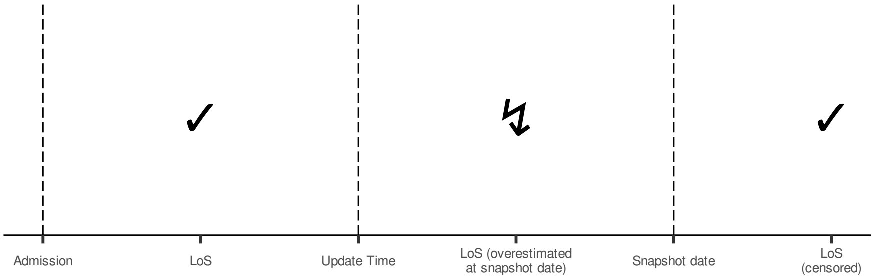
1 Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland

2 Division of Intensive Care Medicine, Cantonal Hospital, St.Gallen, Switzerland

3 The RISC-19-ICU registry board, University of Zurich, Switzerland and Institute of Intensive Care Medicine, University Hospital of Zürich, Switzerland

January 06, 2021

Fig I. Illustration of the relation between admission date, LoS, update time and snapshot date. For patients discharged before the update time, the (uncensored) LoS is available. The censored LoS is available only if a patient is discharged after the snapshot date, but not if the patient left the ICU between the update time and the snapshot date.



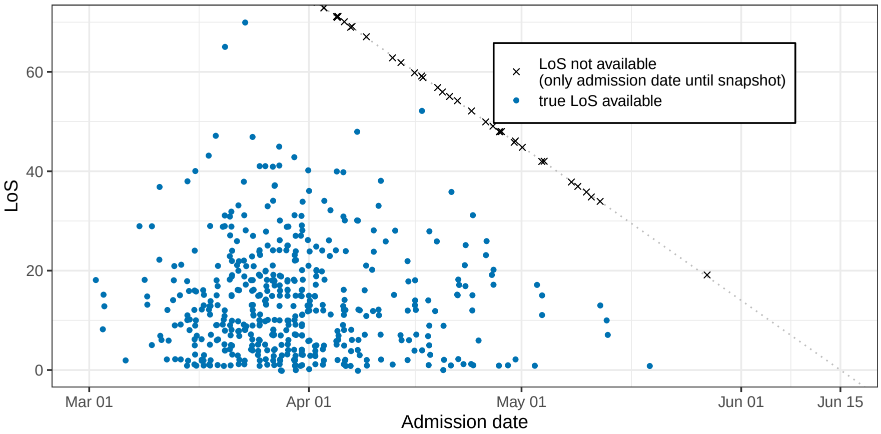
A. Censoring of LoS in the COVID-19 dataset

Since the COVID-19 pandemic is ongoing, it is likely that the set of patients in the RISC-19-ICU registry with available LoS has a selection bias towards shorter LoS. The natural approach to deal with this problem would be to treat the patients with missing LoS as censored observations with censoring time the number of days between admission and snapshot date. Unfortunately, this approach appears to be misleading and overestimates the LoS for the following reason. The data for each patient in the RISC-19-ICU registry is updated periodically by the corresponding ICU. We call the date of the last update for a given patient the update time. If the patient's LoS in the ICU has terminated before the update time, then we observe the LoS, if the patient is still in the ICU at the snapshot date, then the LoS is censored as above. However, there is the possibility that the patient has left the ICU between the update time and the shapshot data, and there is no possiblity to see this from the data. Fig I illustrates this problem, and Fig II shows how many patients are subject to this issue.

B. Evaluation of probabilistic predictions

Probabilistic predictions should be calibrated and sharp [1]. Calibration refers to the statistical compatibility of predictions and observations, and there are several tools available in the literature to assess calibration graphically and with statistical tests. The most prominent tool are so-called Probability Integral Transform (PIT) histograms, which are a histogram of $F_1(y_1)$, …, $F_n(y_n)$ [2,3]. Here, $(F_1, y_1), \dots, (F_n, y_n)$ are a generic notation for the available prediction-observation pairs. Predictions are called probabilistically calibrated if the PIT histogram is flat, and there are strong arguments that probabilistic calibration is an essential requirement for probabilistic forecasts [4]. The notion of probabilistic

Fig II. Patient admission dates and LoS. Dots show the LoS of patients who already left the ICU before the snapshot date (June 15). Black crosses show the time between the admission and the snapshot date for patients for which no discharge time is available in the database.



calibration has been reintroduced in under the name of D-calibrated in [5]. Probabilistic predictions are called marginally calibrated if $(1/n)\sum_{i=1}^{n}\mathbf{1}\{y_i \leq y\} = (1/n)\sum_{i=1}^{n} F_i(y)$ for all $y \in \mathbb{R}$, that is, the observed frequency of realizations of Y below any threshold y should be equal to the average prediction of this frequency [4].

Calibrated probabilistic predictions are not necessarily informative. Therefore, the authors of [1] postulated the principle that probabilistic predictions should *maximize sharpness subject to calibration.* Sharpness is a property of the forecasts only and it refers to how concentrated the predictive distribution is. A forecast is sharper if it yields shorter prediction intervals. Proper scoring rules allow to assess sharpness and calibration of a forecast simultaneously [6]. A widely used example is the Continuous Ranked Probability Score (CRPS) which is defined as

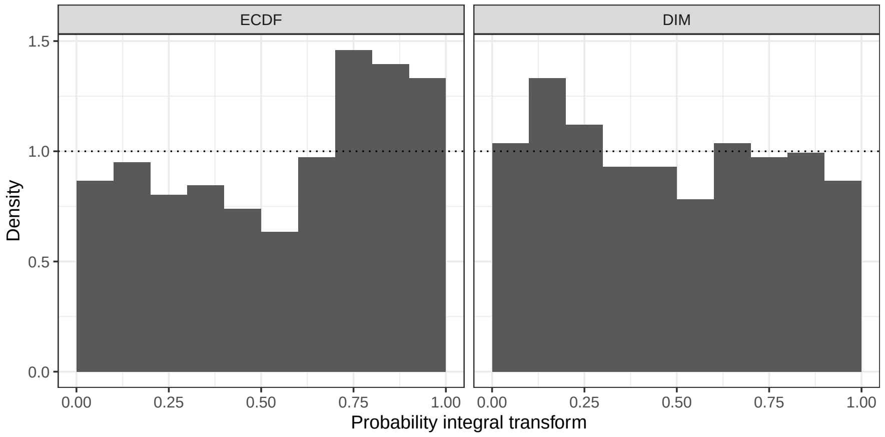
$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(t) - \mathbf{1}\{y \leq t\})^2 \mathrm{d}\, t.$$

for a CDF F and a real number y [7]. A forecast procedure is better the lower the average realized CRPS

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{CRPS}(F_k, y_k).$$

The significance of differences in forecast performance can be assessed by a Diebold-Mariano test [8].

Fig III. PIT histograms for the ECDF and the DIM predictions.



C. Diagnostic plots for calibration of DIM predictions

Fig III shows the PIT histograms for the ECDF predictions and the DIM predictions.

D. Figures on LoS by regions

Figs IV, V, VI summarise the COVID-19 dataset and the corresponding predictions split up by regions in Switzerland.

References

1. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Series B Stat Methodol. 2007;69:243–268.

2. Dawid AP. Statistical theory: The prequential approach. Journal of the Royal Statistical Society: Series A. 1984;147:278–290.

3. Diebold FX, Gunther TA, Tay AS. Evaluating density forecasts with applications to financial risk management. International Economic Review. 1998;39:863–883.

4. Gneiting T, Ranjan R. Combining predictive distributions. Electronic Journal of Statistics. 2013;7:1747–1782.

5. Andres A, Montano-Loza A, Greiner R, Uhlich M, Jin P, Hoehn B, et al. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. PLOS one. 2018;13(3):e0193523.

Fig IV. (a) Empirical distribution of LoS of COVID-19 patients in the regions NE and WT. (b) QQ-plot of the empirical distributions.

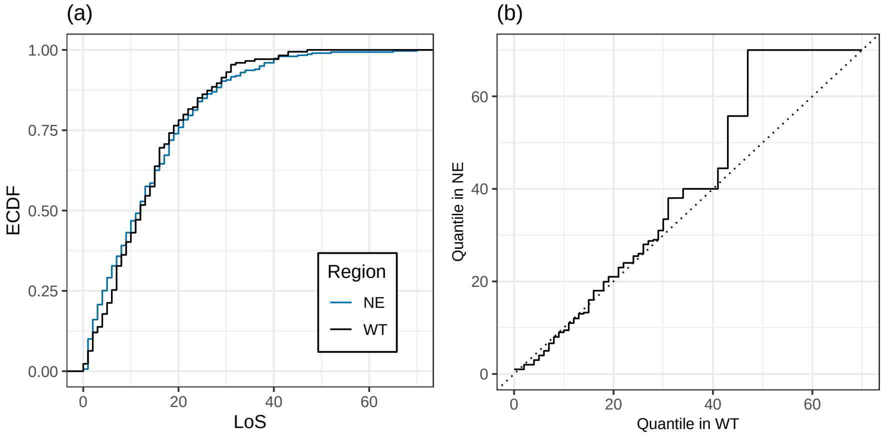


Fig V. Empirical LoS distributions of COVID-19 patients and corresponding DIM forecasts for the regions NE and WT. The DIM forecasts are as in Fig 2 in the article.

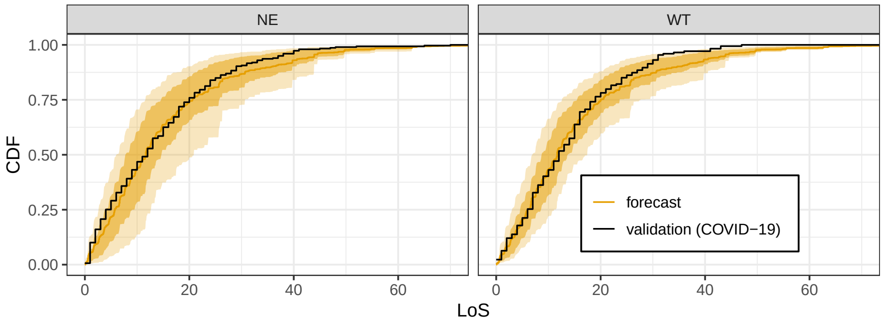
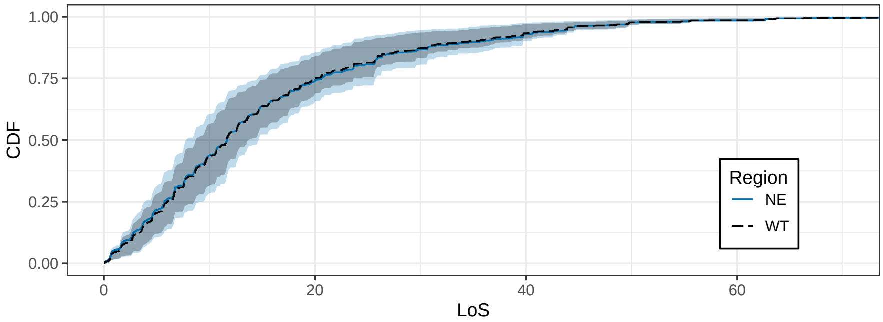


Fig VI. DIM forecasts for COVID-19 patients, depending on region.



6. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association. 2007;102:359–378.

7. Matheson JE, Winkler RL. Scoring rules for continuous probability distributions. Management Science. 1976;22:1087–1096.

8. Diebold FX, Mariano RS. Comparing predictive accuracy. Journal of Business & Economic Statistics. 1995;13:253–263.

Chapter 4

New methods for forecast evaluation

4.1 Valid sequential inference on probability forecast performance

The content of this section is published as

Henzi, A. and Ziegel, J.F. (2021+). Valid sequential inference on probability forecast performance. *Biometrika*, to appear.

The article is directly followed by its supplementary material, which is also available on `https://doi.org/10.1093/biomet/asab047`. The article is published under license Creative Commons CC BY.

Biometrika (2022), **00**, 0, *pp.* 1–17 doi: 10.1093/biomet/asab047
Printed in Great Britain

Valid sequential inference on probability forecast performance

BY ALEXANDER HENZI and JOHANNA F. ZIEGEL

Institute of Mathematical Statistics and Actuarial Science, University of Bern,
Alpeneggstrasse 22, 3012 Bern, Switzerland

alexander.henzi@stat.unibe.ch johanna.ziegel@stat.unibe.ch

SUMMARY

Probability forecasts for binary events play a central role in many applications. Their quality is commonly assessed with proper scoring rules, which assign forecasts numerical scores such that a correct forecast achieves a minimal expected score. In this paper, we construct e-values for testing the statistical significance of score differences of competing forecasts in sequential settings. E-values have been proposed as an alternative to p-values for hypothesis testing, and they can easily be transformed into conservative p-values by taking the multiplicative inverse. The e-values proposed in this article are valid in finite samples without any assumptions on the data-generating processes. They also allow optional stopping, so a forecast user may decide to interrupt evaluation, taking into account the available data at any time, and still draw statistically valid inference, which is generally not true for classical p-value-based tests. In a case study on post-processing of precipitation forecasts, state-of-the-art forecast dominance tests and e-values lead to the same conclusions.

Some key words: Consistent scoring function; E-value; Forecast dominance; Optional stopping; Probability forecast; Proper scoring rule; Sequential inference.

1. INTRODUCTION

Consider a forecast user who compares probability predictions $p_t, q_t \in [0,1]$, $t \in \mathbb{N}$, for a binary event $Y_{t+h} \in \{0,1\}$, where $h \geqslant 1$ is the time lag between the forecasts and the observations. At time t, the forecasts p_t and q_t, as well as any predictions and observations before t, are known. This setting encompasses many practical situations, such as probability-of-precipitation forecasts h days ahead or predictions of negative economic growth in the next quarter. The forecast user wants to draw conclusions about the relative performance of p_t and q_t, that is, to identify the better of the two forecasts.

Probability forecasts for binary events are arguably the simplest and best-understood type of probabilistic forecasts; see Winkler (1996) for an earlier overview and more recent reviews in Gneiting & Raftery (2007), Ranjan & Gneiting (2010) and Lai et al. (2011). The key requirements for probability forecasts are calibration, meaning that events with a predicted probability of p should occur at a frequency of p, and sharpness, which requires the forecast probabilities to be as informative as possible, i.e., close to 0 or 1. These properties are simultaneously assessed with proper scoring rules (Gneiting & Raftery, 2007), which coincide with consistent scoring functions for the mean (Gneiting, 2011) in the case of probability forecasts, and will be simply referred to as scoring functions in this article. A scoring function $S = S(p, y)$ maps a forecast probability p and an observation y to a numerical score, with smaller scores indicating a better forecast. More precisely, S satisfies

$$\mathbb{E}_\pi\{S(\pi, Y)\} \leqslant \mathbb{E}_\pi\{S(p, Y)\} \tag{1}$$

for all $p, \pi \in [0, 1]$, where $\mathbb{E}_\pi(\cdot)$ denotes the expected value under the assumption that $Y = 1$ with probability π. That is, the true event probability attains a minimal expected score, and S is strictly consistent if equality in (1) holds only for $p = \pi$. Well-known examples are the Brier score $(y - p)^2$ and the logarithmic score $-\log(|1 - y - p|)$.

To compare the predictions p_t and q_t, the forecast user would therefore collect a sample $y_{t+h}, p_t, q_t, t = 1, \ldots, T$, and compute the empirical score difference $(1/T)\sum_{t=1}^{T}\{S(p_t, y_{t+h}) - S(q_t, y_{t+h})\}$. To take into account the sampling uncertainty, such score differences are accompanied by p-values indicating whether the mean score differs significantly from zero. If the observations are not independent, as is usual in sequential settings, a number of asymptotic tests are available for computing p-values, prominent ones being the Diebold–Mariano test (Diebold & Mariano, 1995) and the test of conditional predictive ability proposed by Giacomini and White (Giacomini & White, 2006). Further examples are the martingale-based approaches of Seillier-Moiseiwitsch & Dawid (1993) and Lai et al. (2011), and more recent tests of forecast dominance (Ehm & Krüger, 2018; Yen & Yen, 2021).

In this article, we expand the tools for drawing inference on probability forecast performance by using e-values. E-values, where the ‘e’ refers to expectation, have been introduced as an alternative to p-values for testing. The term e-value was first used in the literature by Vovk & Wang (2021), but the concept also appears in Shafer (2021), under the name ‘betting score’, and in Grünwald et al. (2020); see also the series of working papers at `http://alrw.net/e/`. In brief, an e-value is a random variable $E \geqslant 0$, satisfying $\mathbb{E}(E) \leqslant 1$ under a given null hypothesis. By Markov’s inequality, this implies that $\mathrm{pr}(E > 1/\alpha) \leqslant \alpha$ for any $\alpha \in (0, 1)$, i.e., large realizations of an e-value can be taken as evidence against the null hypothesis, and the value $1/E$ is a conservative p-value. A main motivation for using e-values instead of p-values, explained in more detail in Shafer (2021), Grünwald et al. (2020) and Wang & Ramdas (2020), is their simple behaviour under combinations. The arithmetic average of e-values is again an e-value, and so is the product of independent or sequential e-values. E-values also have advantages over p-values with respect to false discovery rate control (Wang & Ramdas, 2020), which may be beneficial for the comparison of forecasts over many locations, such as over a fine latitude-longitude grid around the globe. The central property for this work is that e-values are valid under optional stopping and continuation; that is, the collection of data for computing an e-value may be stopped or continued based on seeing the past observations and e-values. It is well known that p-values in general do not have these properties.

Our main contribution is the result that for any scoring rule S and forecasts p and q for $Y \in \{0, 1\}$, there exists an e-value which satisfies $\mathbb{E}_\pi(E) \leqslant 1$ if and only if $\mathbb{E}_\pi\{S(p, Y) - S(q, Y)\} \leqslant 0$. This e-value allows one to draw inference on the relative performance of the forecasts p and q with respect to S based on only a single observation. In a sequential setting, e-values from different time-points can be merged by products into a nonnegative supermartingale or test-martingale, which are analysed in detail by Ramdas et al. (2020). This gives a statistical test of forecast dominance that is valid in finite samples without any further assumptions on the data-generating process. Moreover, the constructed e-values are valid under optional stopping, so a forecast user may decide to continue or stop forecast comparison based on only a part of the data. These advantages are inherent to any e-value, but we believe that they make e-values a particularly attractive tool in sequential forecast evaluation. The tests mentioned above for comparing probability forecasts are all only asymptotically valid, and the underlying assumptions are often difficult or impossible to verify. In the case of tests with asymptotic normality, the selection of the variance estimator for the test statistic may have a dramatic impact on the test validity; see, for example, Lazarus et al. (2018, Table 1). More serious is the problem of optional stopping. In a simple but realistic simulation example, we demonstrate that commonly used tests for forecast superiority at the level of 0.05 may yield rejection rates of up to 0.15 under optional stopping, grossly misleading and invalidating statistical inference. Although statisticians and practitioners should know that the sample size for classical tests must be determined in advance, we believe that optional stopping is quite common in forecast evaluation, where data arrive sequentially and it might be tempting to stop, or continue, an expensive or time-consuming experiment upon seeing enough, or just not enough, evidence against a hypothesis. Moreover, also in the analysis of past datasets, optional continuation may occur implicitly, in that methods are often first evaluated on a smaller, manageable part of the data and the analysis is continued if the results are promising. Last, but not least, even to a statistician fully aware of the problem of optional stopping, it may be desirable to have a tool that allows the stopping of an evaluation when enough evidence is collected, without having to bother about the implications for inference.

The advantages of e-values for forecast comparison relative to the currently available methods come at a price, namely lower power. This is well known, not only for e-values, and is a general phenomenon when tools for anytime-valid inference are compared with methods for inference with a fixed sample size; see, for example, Fig. 1 in Waudby-Smith & Ramdas (2021), which displays the widths of time-uniform and fixed-time confidence intervals for a mean. However, in the case study in this article, p-values from classical tests and e-values lead to qualitatively the same results.

2. Preliminaries

2.1. *Scoring functions for probabilities*

Throughout the article, $\mathbb{E}_{\mathbb{Q}}(\cdot)$ denotes the expected value of the quantity in parentheses under the probability distribution $\mathbb{Q}$. If the measure $\mathbb{Q}$ is the probability $\pi \in [0,1]$ of a binary event, we simply write $\mathbb{E}_{\pi}(\cdot)$.

When comparing probability forecasts with scoring functions, the choice of the scoring function plays a crucial role. While (1) guarantees that the true event probability always achieves a minimal expected score, different scoring functions may yield different rankings when misspecified forecasts are compared (Patton, 2020). This problem can be avoided by basing forecast comparisons on several or all scoring rules simultaneously. For probabilities of binary events, under mild regularity conditions stated in Gneiting et al. (2007, Theorem 2.3), all consistent scoring functions are of the form

$$S(p,y) = \int_{(0,1)} S_\theta(p,y)\,\mathrm{d}\nu(\theta), \tag{2}$$

where ν is a locally finite Borel measure on $(0,1)$ and

$$S_\theta(p,y) = (\theta - y)\{\mathbb{1}(p > \theta) - \mathbb{1}(y > \theta)\} = \begin{cases} \theta, & y = 0,\ p > \theta, \\ 1-\theta, & y = 1,\ p \leqslant \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In (3), $\mathbb{1}$ denotes the indicator function. This representation originally dates back to Schervish (1989); see also Ehm et al. (2016). The scoring function S is strictly consistent if and only if ν assigns positive mass to all nondegenerate intervals in $(0,1)$.

2.2. *Forecast dominance and hypotheses*

Let $(\Omega, \mathcal{F}, \mathbb{Q})$ be a probability space with a filtration $\mathcal{F}_t$, $t \in \mathbb{N}$. We assume that the competing forecasts p_t and q_t and the observation Y_t, constitute a random vector (Y_t, p_t, q_t) adapted to $\mathcal{F}_t$, and that (p_t, q_t) are forecasts for Y_{t+h} for some integer lag $h \geqslant 1$. The measure $\mathbb{Q}$ describes the joint dynamics of the forecasts and the observations.

When comparing forecasts using a given scoring function S, the quantity of interest is often not the unconditional expected score difference $\mathbb{E}_{\mathbb{Q}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h})\}$, which describes the average relative performance of p_t and q_t. More interesting is the question of whether, given the information $\mathcal{F}_t$ at the time of forecasting, the conditional event probability is closer to p_t than to q_t, i.e., $\mathbb{E}_{\mathbb{Q}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0$. This notion of forecast dominance is called conditional forecast dominance and was introduced by Giacomini & White (2006).

The definition of forecast dominance used here does not require knowledge about the processes generating (Y_t, p_t, q_t), which are often unknown or not well enough understood to formulate a suitable stochastic model. The relative performance of the forecasts p_t and q_t is governed by the underlying distribution $\mathbb{Q}$, and hypotheses about forecast dominance are hypotheses about the data-generating process. Denoting by $\mathcal{P}$ the set of probability measures on $(\Omega, \mathcal{F})$, we will construct tests for the following hypotheses:

$$\mathcal{H}_{S;c} = \left[\mathbb{P} \in \mathcal{P} : c_t\, \mathbb{E}_{\mathbb{P}}\{S(p_t, Y_{t+h}) - S(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0 \text{ a.s.}, t \in \mathbb{N}\right], \tag{4}$$

$$\mathcal{H}_c = \left[\mathbb{P} \in \mathcal{P} : \sup_{\theta \in [0,1]} c_t\, \mathbb{E}_{\mathbb{P}}\{S_\theta(p_t, Y_{t+h}) - S_\theta(q_t, Y_{t+h}) \mid \mathcal{F}_t\} \leqslant 0 \text{ a.s.}, t \in \mathbb{N}\right], \tag{5}$$

where a.s. stands for almost surely. Here, $(c_t)_{t\in\mathbb{N}}$ is a sequence of $\mathcal{F}_t$-measurable random variables $c_t \in \{0, 1\}$. If $c_t = 1$ for all t, we write $\mathcal{H}_{S;c} = \mathcal{H}_S$ and $\mathcal{H}_c = \mathcal{H}$. In this case, hypothesis (4) states that at all times t, forecast p_t is at least as good as forecast q_t under the scoring rule S, given the information available at the time of forecasting. Hypothesis (5) is stronger and states that p_t is preferred over q_t under all elementary scores (3), and it corresponds to what is denoted by H^S_- in Ehm & Krüger (2018, (2.5)). Recently, hypotheses of the type $\mathcal{H}$ or $\mathcal{H}_S$ have been called into question by Zhu & Timmermann (2020), who demonstrate that the null hypothesis of equal conditional predictive accuracy is basically never satisfied in realistic settings. Their criticism does not directly apply to one-sided hypotheses, but we emphasize that the null hypotheses $\mathcal{H}_S$ and $\mathcal{H}$ are rather strong in that they require conditional dominance at all time-points. Tests for these hypotheses are therefore most suitable for comparing a new method with an established benchmark or a state-of-the-art method, where rejecting the null means that the new method outperforms the benchmark at least in some situations, a minimal requirement.

The classical example for a situation with $\mathbb{P} \in \mathcal{H}$ is $p_t = \mathbb{P}(Y_{t+h} = 1 \mid \mathcal{F}_t)$, i.e., p_t is the ideal forecast in the sense of Gneiting & Ranjan (2013). For the hypotheses $\mathcal{H}_S$, one may easily construct situations with dominance relations also among noncalibrated forecasts; see the simulation examples in § 4.

In many practical situations, it cannot be expected that a certain forecast method will always outperform another one, and forecast users want to know under what conditions a particular forecast should be preferred. Choosing the sequence $(c_t)_{t\in\mathbb{N}}$ such that $c_t = 1$ if the condition holds and $c_t = 0$ otherwise, allows us to formalize this question. Here the variables c_t must be $\mathcal{F}_t$-measurable, i.e., known at the time of forecasting. In practice this is not a severe limitation, since the information that one forecast is more accurate than another under a given condition is useful only if this condition is known at the time of forecasting, and not ex post. But also from a theoretical point of view, forecast evaluation should only be conditioned on the forecasts themselves, and not on the observations or on information not available at the time of forecasting; see Lerch et al. (2017) for a detailed analysis of this issue in the case of extreme events.

3. E-values for testing forecast dominance

3.1. *One-period setting*

We first construct e-values for the comparison of probability forecasts in a one-period setting, where $Y = 1$ with probability π and the forecasts p and q are assumed to be fixed numbers in $(0, 1)$. These e-values give an absolute and valid interpretation of predictive performance with only a single observation, e.g., for a single time-point in the sequential setting of § 2.2 or in binary classification problems with independent forecast-observation pairs, where the competing forecasts are based on covariates and π is the probability of $Y = 1$ conditional on the covariate values. The null hypotheses that p is a better forecast than q with respect to a given score S or with respect to all scoring functions simultaneously correspond to

$$H_S = \left[\pi \in [0, 1] : \mathbb{E}_\pi\{S(p, Y) - S(q, Y)\} \leqslant 0\right],$$

$$H = \left[\pi \in [0, 1] : \sup_{\theta \in [0,1]} \mathbb{E}_\pi\{S_\theta(p, Y) - S_\theta(q, Y)\} \leqslant 0\right].$$

For $p < q$, a direct computation shows that H_S is the interval $[0, \kappa_\nu\{[p, q)\}]$ with

$$\kappa_\nu\{[a, b)\} = \frac{\int_{[a,b)} \theta \, \mathrm{d}\nu(\theta)}{\nu\{[a, b)\}} \quad (0 < a < b < 1).$$

The stronger null hypothesis H is the intersection of these intervals for all mixing measures ν, that is, $[0, p]$. In the case of $q > p$, the intervals H_S and H take the form $[\kappa_\nu\{[q, p)\}, 1]$ and $[p, 1]$, respectively. Table 1 gives the boundary $\kappa_\nu\{[p, q)\}$ for commonly used scoring functions.

For a set $\mathcal{P}$ of probability measures and disjoint $H, H' \subset \mathcal{P}$, we say that an e-value E has null hypothesis H and alternative H' if $\mathbb{E}_\mathbb{P}(E) \leqslant 1$ for all $\mathbb{P} \in H$ and $\mathbb{E}_\mathbb{Q}(E) > 1$ for all $\mathbb{Q} \in H'$. The following theorem characterizes e-values for testing H_S.

Theorem 1. *Let S be a consistent scoring function and let $p, q \in (0, 1)$ with $p \neq q$. Assume that the mixing measure ν of S satisfies $\nu\{[\min(p, q), \max(p, q))\} > 0$. Then a function $E = E(y)$ is an e-value with null hypothesis H_S and alternative* $[0, 1] \setminus H_S$ *if and only if for some* $\lambda \in (0, 1]$,

$$E(y) = E_{p,q;\lambda}(y) = 1 + \lambda \frac{S(p, y) - S(q, y)}{|S(p, \mathbb{1}\{p > q\}) - S(q, \mathbb{1}\{p > q\})|}. \tag{6}$$

Theorem 1 gives a family of e-values for testing forecast dominance with a given score S, and in a next step we tune the parameter λ in (6) such that the corresponding e-value has maximal power against a given alternative. The notion of power for e-values differs from the classical power for p-values, and it is motivated in detail by Shafer (2021) and Grünwald et al. (2020). An e-value can be interpreted as a bet against the null hypothesis, and a product $\prod_{t=1}^T E_t$ of e-values represents the accumulated capital at time T if the initial capital is 1 and all money is invested in the bet at each step. Maximizing the gains is equivalent to maximizing the growth rate $(1/T) \log \prod_{t=1}^T E_t = (1/T) \sum_{t=1}^T \log(E_t)$, a strategy sometimes called Kelly betting after Kelly Jr (1956). If an e-value maximizes $\mathbb{E}_\mathbb{P}\{\log(E)\}$ under a measure $\mathbb{P}$ representing an alternative

Table 1. *Commonly used scoring rules and the corresponding denominators in the GROW e-values under the assumption $p < q$. The case of $p > q$ is obtained by interchanging the roles of p and q. The mixing measure ν is given in the form of its Lebesgue density $h(\theta)$, $\theta \in (0, 1)$. For the spherical score, $\|p\| = (2p^2 - 2p + 1)^{1/2}$ denotes the Euclidean norm of the vector $(p, 1 - p)$*

Score	$S(p, y)$	Mixing density ν	$\kappa_\nu\{[p, q)\}$
Brier	$(p - y)^2$	2	$(p + q)/2$
Logarithmic	$-\log(\lvert 1 - y - p\rvert)$	$\theta^{-1}(1 - \theta)^{-1}$	$\log\left(\frac{1-p}{1-q}\right) / \log\left\{\frac{q(1-p)}{p(1-q)}\right\}$
Spherical	$1 - \lvert 1 - y - p\rvert / \|p\|$	$(2\theta^2 - 2\theta + 1)^{-3/2}$	$\frac{(q-1)\|p\| - (p-1)\|q\|}{(2q-1)\|p\| - (2p-1)\|q\|}$

hypothesis, it is said to be growth-rate-optimal, abbreviated GROW (Grünwald et al., 2020). One such alternative could be that $Y = 1$ with probability q, but one can maximize the power under any other alternative $\pi_1 \notin H_S$.

Theorem 2. *Under the assumptions of Theorem 1, for any $\pi_1 \notin H_S$, $\mathbb{E}_{\pi_1}\{\log(E_{p,q;\lambda})\}$ is maximal in λ if and only if*

$$\lambda = \begin{cases} (1 - \pi_1) + \pi_1 \dfrac{S(p, 1) - S(q, 1)}{S(p, 0) - S(q, 0)}, & p > q, \\ \pi_1 + (1 - \pi_1) \dfrac{S(p, 0) - S(q, 0)}{S(p, 1) - S(q, 1)}, & p < q. \end{cases}$$

The corresponding e-value equals

$$E_{p,q}^{\pi_1}(y) = \begin{cases} \dfrac{1 - \pi_1}{1 - \kappa_\nu\{[\min(p, q), \max(p, q))\}}, & y = 0, \\ \dfrac{\pi_1}{\kappa_\nu\{[\min(p, q), \max(p, q))\}}, & y = 1. \end{cases}$$

Theorem 2 shows that the GROW e-values for the comparison of probability forecasts take the form of likelihood ratios with the alternative probability in the numerator and the integral of the mixing measure ν over the interval $[\min(p, q), \max(p, q)$, suitably normalized, in the denominator. It is possible to obtain this result directly by applying Theorem 1 of Grünwald et al. (2020), since $\kappa_\nu\{[\min(p, q), \max(p, q))\}$ is the boundary of the null hypothesis H_S. We have chosen to take the indirect but more instructive approach via Theorem 1, because to the best of our knowledge this is the first application of e-values to forecast comparison, and similar approaches might be used to construct e-values for score differences in more general settings than the evaluation of binary event forecasts. In fact, Waudby-Smith & Ramdas (2021, Proposition 2) contains a similar representation of e-values to that in (6) for testing hypotheses about a constant mean.

For the test of the null hypothesis H, applying Theorem 1 of Grünwald et al. (2020) shows that the GROW e-value is the likelihood ratio.

Theorem 3. *Let $p, q \in (0, 1)$. Then the GROW e-value with null hypothesis H and alternative hypothesis that $Y = 1$ with probability $\pi_1 \notin H$ is*

$$E_{p,q}^{\pi_1 *}(y) = \begin{cases} (1 - \pi_1)/(1 - p), & y = 0, \\ \pi_1/p, & y = 1. \end{cases}$$

In testing with e-values, the GROW e-value for testing the point null hypothesis $\{p\}$ against the alternative π_1 is exactly the likelihood ratio, and Theorem 3 states that this is equivalent to testing forecast dominance with respect to all scoring functions. Dominance with respect to all scoring functions is a very strong requirement on p, since the null hypothesis is false as soon as the true probability π is on the same side of p as q, that is, in $(p, 1]$ for $p < q$ or in $[0, p)$ for $q < p$, and the choice of π_1 is restricted to these sets. Unlike the e-values $E_{p,q}^{\pi_1}$, $E_{p,q}^{\pi_1 *}$ does not depend directly on q, but rather indirectly via the admissible values for π_1.

3.2. *Sequential inference*

We now turn to the sequential model with observations Y_t and forecasts p_t and q_t defined on a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ with a filtration $\mathcal{F}_t$, $t \in \mathbb{N}$. In the $h = 1$ case, for any $\mathbb{Q} \in \mathcal{H}_{S;c}$ and any adapted sequence $\lambda_t \in [0, 1]$, $t \in \mathbb{N}$, with $E_{p_t,q_t;\lambda_t}$ as defined in (6),

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\} &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1}) \,\middle|\, \mathcal{F}_T\right\}\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\prod_{t=1}^{T-1} E_{p_t,q_t;\lambda_t}(Y_{t+1})\mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\}\right].\end{aligned}$$

If $c_t = 0$, then there is no hypothesis about p_t and q_t. For these cases, the definition in (6) may be extended to $\lambda = 0$, so that $E_{p_t,q_t;0} \equiv 1$ if $c_t = 0$. Then, if $\lambda_T = 0$ when $c_T = 0$,

$$\mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\} = (1 - c_T) + c_T \mathbb{E}_{\mathbb{Q}}\{E_{p_T,q_T;\lambda_T}(Y_{T+1}) \mid \mathcal{F}_T\} \leqslant 1$$

almost surely for $\mathbb{Q} \in \mathcal{H}_{S;c}$, so

$$\mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\} \leqslant \mathbb{E}_{\mathbb{Q}}\left\{\prod_{t=1}^{T-1} E_{p_t,q_t;\lambda_t}(Y_{t+1})\right\}.$$

Iterating this argument shows that the product $\prod_{t=1}^{T} E_{p_t,q_t;\lambda_t}(Y_{t+1})$ is an e-value for $\mathcal{H}_{S;c}$; more precisely, the process $\prod_{j=1}^{t} E_{p_j,q_j;\lambda_j}(Y_{j+1})$, $t = 1, 2, 3, \ldots$, is a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. For a general lag h, sequential conditioning at time steps of 1 is not possible, and one option is to average the products of all e-values with a time difference of h, in the spirit of the U-statistics merging functions suggested by Vovk & Wang (2021). We summarize this in the following proposition.

PROPOSITION 1. *Let* $(Y_t, c_t, p_t, q_t, \lambda_t) \in \{0, 1\}^2 \times (0, 1)^2 \times [0, 1]$ *be defined on a measurable space* $(\Omega, \mathcal{F})$ *and adapted to the filtration* $\mathcal{F}_t$ $(t \in \mathbb{N})$*, and assume that* $\lambda_t = 0$ *if* $c_t = 0$*. Further, let S be a strictly consistent scoring function. Then for all* $T \geqslant h + 1$,

$$e_T = \frac{1}{h}\sum_{k=1}^{h}\prod_{l\in I_k} E_{p_l,q_l;\lambda_l}(Y_{l+h})$$

with $I_k = \{k + hs : s = 0, \ldots, \lfloor (T - k)/h \rfloor - 1\}$ *are* $\mathcal{F}_T$*-measurable and are e-values under* $\mathcal{H}_{S;c}$.

Proposition 1 is an analogous result to Theorem 1 in the sense that it only characterizes possible e-values for testing forecast dominance, but the parameters λ_t could be any adapted sequence $(\lambda_t)_{t\in\mathbb{N}} \subset [0, 1]$. E-values for dominance testing under the conditions $(c_t)_{t\in\mathbb{N}}$ are obtained by setting all e-values for which the condition is not satisfied to 1. The forecast user may, and in fact has to, tune the $(\lambda_t)_{t\in\mathbb{N}}$ in order to attain good power against a given alternative. Recall that at any t, λ_t may be a function of all the forecasts and observations before time t. Instead of the parameters λ_t, it is usually more intuitive to think of an alternative probability η_t for the event $Y_{t+h} = 1$ and then directly use the GROW e-values $E^{\eta_t}_{p_t,q_t}$ constructed in Theorem 2. In that respect, testing forecast dominance with e-values differs from p-value-based tests of a zero score difference, which do not require the user to explicitly specify an alternative hypothesis. In the applications in § 4 and § 5, we will give guidance on the selection of alternative hypotheses and show that reasonable power can be attained with simple heuristic methods.

As a side remark, choosing an alternative hypothesis for e-values in sequential forecast dominance testing is similar to the conditional predictive ability tests of Giacomini & White (2006), where $\mathcal{F}_t$-measurable test functions are used to weight score differences and improve power. Whereas selection of the test functions in the Giacomini–White test is delicate, because they may have an impact on the variance estimates and the finite-sample validity of the tests, e-values remain valid under any choice of adapted weights $(\lambda_t)_{t\in\mathbb{N}}$.

Our final theoretical result states that the e-values e_T constructed above are also valid when T is replaced by a stopping time τ. This is a consequence of the fact that $(e_t)_{t\geqslant h+1}$ is a nonnegative supermartingale (see Ramdas et al., 2020, § 3).

Proposition 2. *Let $\tau \in \mathbb{N}$ be a stopping time. Then under the assumptions of Proposition 1,*

$$\mathbb{E}_{\mathbb{Q}}(e_\tau) \leqslant 1, \quad \mathbb{Q} \in \mathcal{H}_S.$$

To understand validity under optional stopping intuitively, recall that at time t the forecast user has to determine the parameter λ_t in the e-value $E_{p_t,q_t;\lambda_t}(Y_{t+h})$. Optional stopping at t_0 corresponds to setting $\lambda_t \equiv 0$, or equivalently $E_{p_t,q_t;\lambda_t}(Y_{t+h}) \equiv 1$, for $t \geqslant t_0$, i.e., ignoring all observations starting from time $t_0 + h$. In the case of forecast lag 1, this allows the forecast user to stop evaluation at any time, since λ_t in $E_{p_t,q_t;\lambda_t}(Y_{t+1})$ is defined at the same time as Y_t is observed. However, when $h > 1$, the coefficients λ_t in $E_{p_t,q_t;\lambda_t}(Y_{t+h})$ for $t = t_0 - h + 1, \ldots, t_0 - 1$ have already been determined in the past and may not be set to zero at t_0, since they must be $(\mathcal{F}_t)_{t\in\mathbb{N}}$-adapted. This implies that the stopped e-value depends on the unknown, future observations $Y_{t_0+1}, \ldots, Y_{t_0+h-1}$ and so is not deterministic at time t_0.

In the case $h = 1$, optional stopping is a powerful strategy when the goal is to assess forecast superiority at a significance level $\alpha \in (0, 1)$, because the stopping time

$$\tau_\alpha = \min\{T, \inf(t \geqslant 2 : e_t \geqslant 1/\alpha)\}$$

allows us to reject the null hypothesis as soon as the sequential e-value e_t exceeds $1/\alpha$. If $h > 1$, one may similarly define

$$\tau_{\alpha,h} = \min\Big(T, \inf\Big[t \geqslant h + 1 : e_t \geqslant \max_{j=t-h+1,\ldots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\}^{-1}/\alpha\Big]\Big),$$

which guarantees that when stopping at t_0, the level $1/\alpha$ is exceeded no matter what values $Y_{t_0+1}, \ldots, Y_{t_0+h-1}$ take; see the Supplementary Material. Instead of specifying a significance level α in advance, one may as well transform the sequence $(e_t)_{t\in\mathbb{N}}$ into so-called anytime-valid

p-values $p_{t_0} = \min\{1, \inf_{s=1,\ldots,t_0} 1/e_s\}$, which are valid simultaneously for all $t_0 \geqslant h+1$ (see Ramdas et al., 2020, § 3.1).

4. Simulation examples

4.1. *Basic properties*

For the simulation examples in this subsection and the next, we transform e-values E into p-values by taking their inverse $1/E$, so that direct comparisons with p-values are possible. Further variations of these simulation examples are presented in the Supplementary Material. An `R` package for the proposed methods and replication material for all results in this article are available at `https://github.com/AlexanderHenzi/eprob`.

In the first example, for varying $\mu \in (0,1)$, we simulate independent forecasts $p_t, q_t \sim \mathrm{Unif}(0,1)$, define $\pi_t = \mu q_t + (1-\mu)p_t$, and generate independent Bernoulli observations Y_{t+1} with mean π_t conditional on p_t and q_t. This represents a situation where forecasters only have access to partial information and both forecasts are not calibrated, i.e., $\mathbb{P}(Y_{t+1} = 1 \mid p_t) \neq p_t$ and $\mathbb{P}(Y_{t+1} = 1 \mid q_t) \neq q_t$. We choose S to be the Brier score, so that p_t outperforms q_t if and only if $\pi_t \in [0, (p_t+q_t)/2]$ if $p_t < q_t$ or $\pi_t \in [(p_t+q_t)/2, 1]$ if $p_t > q_t$, i.e., if and only if $\mu \leqslant 0.5$. When $\mu > 0.5$, the GROW e-value is obtained by choosing π_t as the alternative hypothesis probability, but in practice π_t is not known. The forecast user might assume that the true probability of $Y_{t+1} = 1$ lies somewhere between $(p_t+q_t)/2$ and q_t, and choose a convex mixture $\eta_t(\xi) = \xi(p_t+q_t)/2 + (1-\xi)q_t$ with some $\xi \in (0,1)$ as an alternative. Proposition 1 implies that for $k \in \mathbb{N}$ and $\xi_1, \ldots, \xi_k \in (0,1)$,

$$e_{t;\xi_j} = \prod_{i=1}^{t} E_{p_i,q_i}^{\eta_i(\xi_j)}(Y_{i+1}), \quad e_t = \frac{1}{k}\sum_{j=1}^{k} e_{t;\xi_j}$$

are e-values under $\mathcal{H}_S$. In Fig. 1, we compare the rejection rates at the 5% level, corresponding to e-values greater than or equal to 20, when the ξ_j are k equally spaced weights in $(0,1)$ for $k=1$ and $k=5$, i.e., $\xi_1 = 0.5$ if $k=1$ and $\xi_l = l/6$ for $l = 1, \ldots, 5$ in the case of $k=5$. We computed both the unstopped e-value e_T and the stopped variant $e_{\tau_{0.05}}$, and the e-values under alternatives $\eta_t = \pi_t$ and $\eta_t = q_t$. The rejection rates are compared with those of one-sided t-tests of the null hypothesis that the mean Brier score difference is nonpositive. Additionally, we report the rejection rates when the p-value is used for optional stopping at given time-points upon seeing a significant difference.

Our simulations illustrate the known fact that classical statistical tests are not valid under stopping. At the boundary of the null hypothesis, the rejection rate of the t-test amounts to 0.12 for $T = 600$ and optional stops at times 150, 300 and 450; given the number of optional stops, this phenomenon occurs independently of the sample size. As for the e-values, stopping, i.e., $e_{\tau_{0.05}}$, is always a more powerful but valid strategy compared to the e-value e_T. While the heuristic alternatives achieve a power close to that under the correct alternative hypothesis, the misspecified hypothesis $\eta_t = q_t$ is clearly weaker. Interestingly, the correct alternative $\eta_t = \pi_t$ has lower power than the heuristic alternatives close to the boundary of the null hypothesis. This is not an error: specifying $\eta_t = \pi_t$ yields the optimal growth rate for the e-value, but this does not necessarily mean that it gives optimal power for the stopped e-value at the threshold $1/\alpha = 20$ in finite samples. The t-test generally achieves higher power than the e-values, which is to be expected given the absence of assumptions on the data-generating process and the validity under optional stopping for the e-values. See also Waudby-Smith & Ramdas (2021).

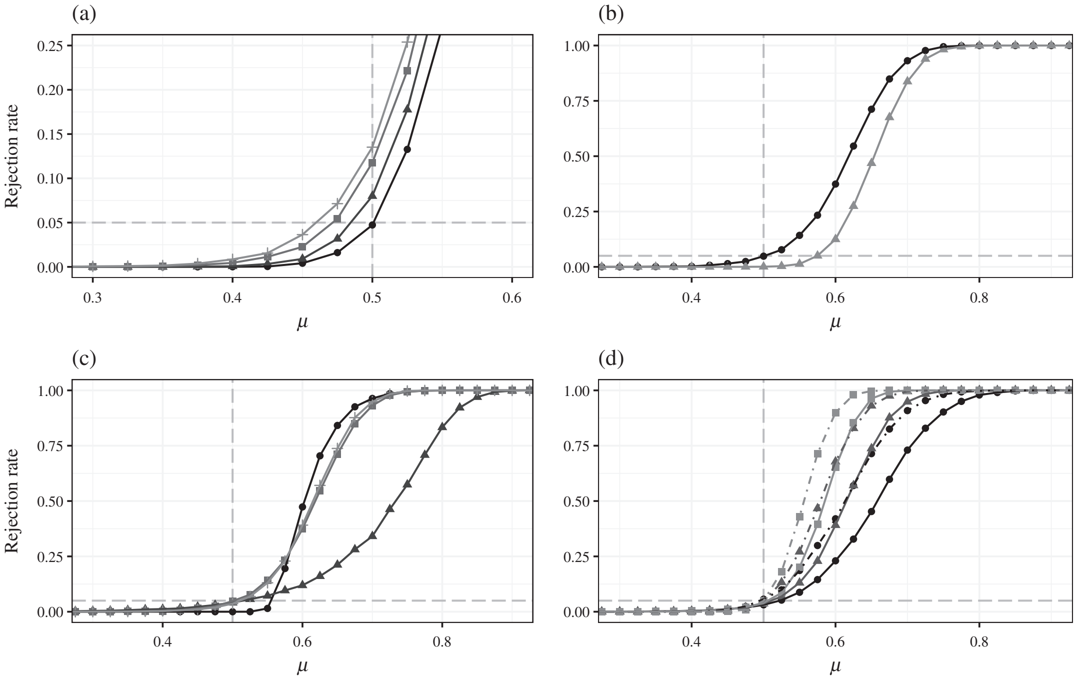


Fig. 1. Rejection rates of e-values and Student's t-test for the hypothesis that p_t dominates q_t with respect to the Brier score in the simulation of § 4.1. The sample size is $T = 600$ for panels (a)–(c) and the significance level is $\alpha = 0.05$ for all panels. (a) Rejection rate of t-test under optional stopping at one (triangles), three (squares) and five (crosses) equispaced time-points between 1 and $T = 600$, and without optional stopping (dots). (b) Rejection rates of stopped (dots) and unstopped (triangles) e-values with $k = 1$. (c) Rejection rates of e-values with different alternative hypotheses: q_t (triangles), π_t (dots), $k = 1$ (crosses) and $k = 5$ (squares). (d) Rejection rates of e-values with $k = 5$ (solid lines) and t-test without stopping (dot-dashed lines) for sample sizes $T = 300$ (dots), 600 (triangles) and 1200 (squares).

4.2. *Time series example*

We simulate Z_t from a moving-average process $Z_t = \epsilon_t + \theta \sum_{j=1}^{4} \epsilon_{t-j}$ and define

$$Y_t = \mathbb{1}\{Z_t > 0\}, \quad \pi_{t;h} = \mathbb{P}(Z_t > 0 \mid Z_{t-j}, j = h, \dots, 4) \quad (h = 1, \dots, 4). \tag{7}$$

The probability $\pi_{t;h}$ corresponds to the ideal forecast at lag h. We compare $q_{t;h} = \pi_{t;h}$ and $p_{t;h} = \pi_{t;h+1}$ for lags $h = 1, 2, 3$, so that $q_{t;h}$ always outperforms $p_{t;h}$. As the parameter θ decreases, serial dependence decreases and the forecasting skills of $p_{t;h}$ and $q_{t;h}$ become similar. The alternative hypothesis for the e-values is the correct alternative $\eta_{t;h} = q_{t;h}$, so that the effect of a higher lag can be analysed in isolation from the question of how to choose the alternative hypothesis. Rejection rates are compared with the Diebold–Mariano test at the 5% level.

Figure 2 shows the dependence of the rejection rates on the parameter θ for different sample sizes T. The e-values use the stopping time $\tau_{0.05}$ for lag 1 and the stopping time $\tau_{0.05;h}$ for lags $h = 2$ and $h = 3$. As in the previous simulations, the power of the e-values is below that of the p-values for the lag-1 forecasts, where the Diebold–Mariano test essentially coincides with the t-test. For lags 2 and 3 this difference increases, since the combination method for e-values becomes less powerful. With increasing lag, the rejection rates of both methods decrease, but the difference to lag 1 is smaller for the Diebold–Mariano test than for the e-value. In this example, the Diebold–Mariano test is valid because the forecasts are ideal and the data-generating process is stationary. For the e-values, validity is guaranteed without such assumptions, which may be a great advantage in applications.

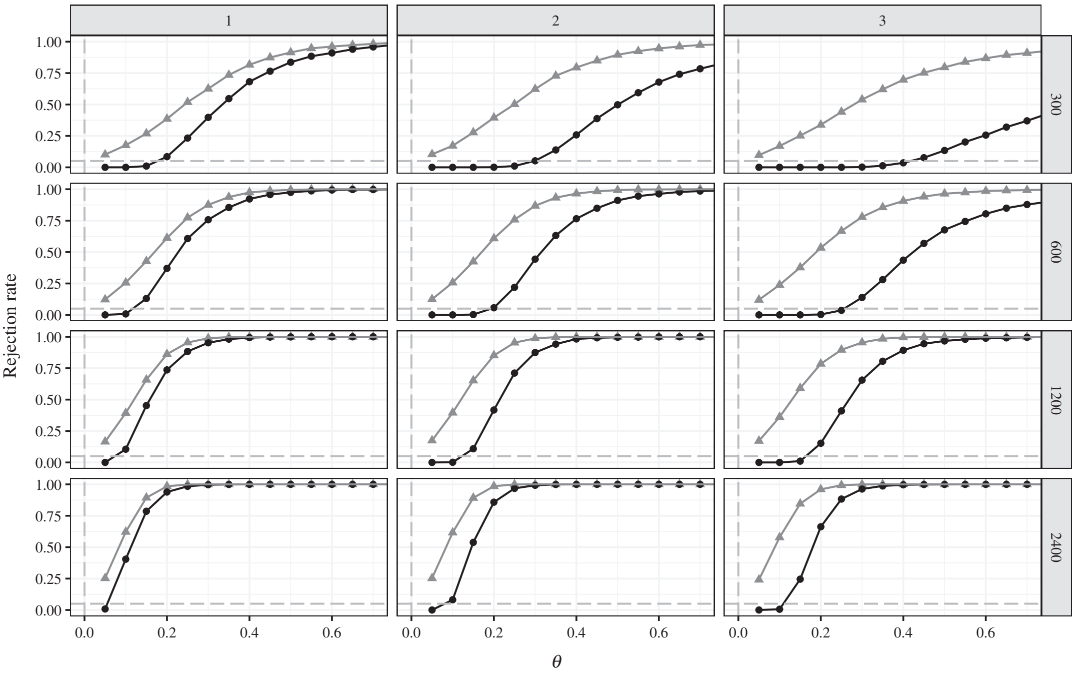


Fig. 2. Rejection rates of e-values (dots) and the Diebold–Mariano test (triangles) in the example (7) at the 5% level for different sample sizes T (rows) and lags h (columns).

5. Case study

5.1. *Data and methods*

Henzi et al. (2021) compared post-processing methods for precipitation forecasts with lags of one to five days at Brussels, Frankfurt, London Heathrow and Zurich airports. In their case study, probability-of-precipitation, or PoP, forecasts were evaluated with the Brier score, but no tests for significance of score differences were performed. We demonstrate here how to apply e-values to probability forecasts, and we will compare the results with state-of-the-art forecast dominance tests.

A detailed description of the dataset and methods is given in Henzi et al. (2021, § 5), so here we only summarize the key information. The dataset covers the period from 6 January 2007 to 1 January 2017, and upon accounting for missing values, the numbers of available observations are 3406 for Brussels, 3617 for Frankfurt, 2256 for London and 3241 for Zurich. Post-processing is applied to the ensemble forecasts of the European Centre for Medium-Range Weather Forecasts (Molteni et al., 1996; Buizza et al., 2005), which are issued on a latitude-longitude grid and consist of a high-resolution forecast, 50 perturbed ensemble forecasts at a lower resolution, and the control run for the perturbed forecasts. In simple terms, ensemble forecasts account for uncertainty by running a numerical weather prediction model several times, each time under slightly perturbed initial conditions; each run of the model yields a different forecast, and these forecasts together form a so-called ensemble (Leutbecher & Palmer, 2008). Ensemble forecasts are usually subject to biases and dispersion errors, which can be corrected by estimating the conditional distribution of the weather variable, given the numerical weather prediction ensemble. This statistical procedure is known as post-processing of ensemble forecasts (Vannitsem et al., 2018).

Henzi et al. (2021) proposed isotonic distributional regression, IDR, as a benchmark for such post-processing methods. IDR estimates conditional distributions nonparametrically and without any tuning parameters. The method is not specifically tailored to forecasting precipitation, and one would expect that a parametric model designed for this purpose will give more precise forecasts. One such method is heteroscedastic censored logistic regression, HCLR (Messner et al., 2014), which assumes that the square root of the precipitation follows a logistic distribution censored at zero. The implementation is as in Henzi et al. (2021). While the covariates in IDR are only the high-resolution forecast, the control forecast and the ensemble mean, the HCLR model additionally includes a scale parameter depending on the ensemble standard deviation.

In contrast to the study in Henzi et al. (2021), which uses an expanding window for the post-processing, we estimate both post-processed forecasts on half of the data for each airport for simplicity, and keep the remaining half for validation.

5.2. *Hypothesis tests*

We illustrate the usage of e-values in the following hypothesis tests. Firstly, we try to reject the null hypothesis that IDR PoP forecasts are better than HCLR PoP forecasts with respect to the Brier score. Secondly, we modify HCLR by dropping the scale parameter. It is expected that this variant, denoted by HCLR$_{-}$, will be outperformed by the original version of HCLR and also by IDR, since both IDR and HCLR$_{-}$ assume a monotone relationship between the covariates and the PoP, but the nonparametric IDR can estimate a broader class of functions. Finally, we further investigate the effect of the scale parameter on HCLR predictions for high precipitation. Suppose a weather forecaster issues a warning if the probability that the precipitation exceeds a high threshold is greater than 50%. As thresholds, we chose the empirical 90% quantile of precipitation in the training data for each airport. Intuitively, the HCLR model should yield more accurate warnings than HCLR$_{-}$, because it includes the ensemble standard deviation as an uncertainty measure.

The first and second sets of hypotheses are tested with the Brier score and the corresponding e-values. As an alternative probability, we take the convex mixtures $\eta_t = 0.25p_t + 0.75q_t$, which were explored in § 4, denoting by p_t the forecasting method that is expected to have a better performance than q_t under the null hypothesis. The hypothesis about the extreme precipitation warnings is a conditional comparison with the conditions $c_t = \mathbb{1}\{\max(p_t, q_t) \geqslant 0.5\}$. For this hypothesis, instead of dominance with respect to the Brier score, we test the stronger hypothesis of forecast dominance with respect to all scoring rules. The rationale is that the forecast dominance hypothesis should be easily rejected if the HCLR model truly issues the better tail forecasts; and, on the other hand, failing to reject may indicate either that, even with data of 10 years it is not possible to clearly discriminate the quality of such warnings, or that the ensemble standard deviation does not bring a benefit. For this hypothesis we define $\eta_t = q_t$, assuming that the conditional event probabilities should be much closer to those issued by HCLR than by HCLR$_{-}$. No optional stopping is applied in all e-values.

For comparison, we also compute p-values for the significance of score differences. The first two hypotheses are tested with one-sided Diebold-Mariano tests (Diebold & Mariano, 1995; see also Giacomini & White, 2006). To estimate the variance of the test statistics, we use the heteroscedasticity and autocorrelation consistent estimator with Bartlett weights; see Lerch et al. (2017, equation 2.18). For testing dominance of the tail probability forecasts, the test of Yen & Yen (2021) would allow arbitrary forecast lags, but it assumes strict stationarity. Since the sequence c_t selects only particular instances, with possibly strongly varying time gaps in between, stationarity is highly questionable. We therefore apply the dominance test of Ehm & Krüger (2018), which is valid under weaker assumptions, but is limited to lag 1. Strictly speaking, both the Diebold-Mariano test and the forecast dominance test are valid under larger null hypotheses than the

Table 2. *Brier scores for different PoP forecasting methods, along with e-values and p-values for testing significance of score differences. The columns under HCLR/IDR show e-values and p-values for tests of the null hypothesis that IDR PoP forecasts achieve a lower Brier score than HCLR forecasts, with analogous interpretations for the other forecast pairs*

		Average Brier score			HCLR/IDR		IDR/HCLR$_-$		HCLR/HCLR$_-$	
	Lag	IDR	HCLR	HCLR$_-$	*E*	*p*	*E*	*p*	*E*	*p*
BRU	1	0.107	0.117	0.118	0	0.9998	>100	$<10^{-4}$	>100	0.0702
	2	0.119	0.123	0.125	0.01	0.9471	>100	0.0101	13.602	0.0294
	3	0.134	0.133	0.136	0.425	0.4405	>100	0.1916	15.185	0.0019
	4	0.152	0.145	0.148	4.804	0.0138	1.943	0.9358	5.165	0.0074
	5	0.171	0.161	0.164	16.969	0.0002	0.415	0.9965	3.436	0.0003
FRA	1	0.109	0.111	0.114	0	0.7784	>100	0.0213	>100	$<10^{-4}$
	2	0.114	0.119	0.122	0.054	0.9643	>100	0.0002	>100	0.0004
	3	0.123	0.127	0.132	0.078	0.9352	>100	0.0001	26.569	$<10^{-4}$
	4	0.147	0.144	0.147	2.291	0.0966	9.618	0.5245	5.54	0.0001
	5	0.166	0.161	0.163	1.526	0.0305	2.362	0.8871	3.227	0.0051
LHR	1	0.135	0.138	0.139	0.029	0.8136	14.979	0.1314	2.845	0.3721
	2	0.138	0.143	0.143	0.188	0.9189	>100	0.0509	2.868	0.4369
	3	0.152	0.154	0.155	0.734	0.7549	40.905	0.1394	2.488	0.3400
	4	0.169	0.167	0.169	1.429	0.2455	1.7	0.5442	1.744	0.0785
	5	0.186	0.181	0.182	1.577	0.0753	0.379	0.9288	1.118	0.3216
ZRH	1	0.104	0.108	0.110	0.003	0.9306	>100	0.0055	61.747	0.0003
	2	0.110	0.112	0.114	0.116	0.7219	36.891	0.0304	10.276	0.0001
	3	0.121	0.118	0.121	1.516	0.0892	31.924	0.4410	5.098	0.0001
	4	0.138	0.132	0.134	4.069	0.0027	1.276	0.9588	2.771	0.0015
	5	0.165	0.156	0.159	15.151	$<10^{-4}$	0.842	0.9978	2.383	0.0002

IDR, isotonic distributional regression; HCLR, heteroscedastic censored logistic regression; HCLR$_-$, heteroscedastic censored logistic regression without the scale parameter; BRU, Brussels; FRA, Frankfurt; LHR, London Heathrow; ZRH, Zurich; E, e-values; *p*, *p*-values.

e-values, as they only require the average score difference between p_t and q_t to be nonpositive, whereas the null hypothesis for the e-values asks for conditional superiority at each time-point. A comparison is nevertheless interesting, since these two tests represent commonly used methods for testing the significance of score differences.

Tables 2 and 3 show the e-values and one-sided *p*-values for the hypotheses described above, computed separately for each airport and forecast lag. The e-values are not transformed to *p*-values here. For interpretation, Vovk & Wang (2020, § 3) suggested a discrete scale such that e-values in $(0, 1]$, $(1, 3.16]$, $(3.16, 10]$, $(10, 31.6]$, $(31.6, 100]$ and $(100, \infty)$ represent no, poor, substantial, strong, very strong and decisive evidence against the null hypothesis, respectively. E-values greater than 100 are not displayed to improve readability, but an untruncated version of Table 2 is included in the Supplementary Material so that it is possible to update the e-values with more recent data. For all hypotheses, the *p*-values and e-values largely lead to the same conclusions. HCLR does not outperform IDR for PoP forecasts at lags 1–3, but for the Brussels and Zurich airports there is substantial to strong evidence that it achieves lower Brier scores at lags 4 and 5. HCLR$_-$ is clearly outperformed by the more complex variant with the ensemble-dependent scale parameter at short lags; also, for the longer lead times there is some evidence that including the scale parameter improves the forecasts, except for London airport. As for the difference between IDR and HCLR$_-$, both the e-values and the *p*-values suggest that IDR yields the better forecasts at lags 1–3, but at lags 4 and 5 there are no rejections of the null hypothesis. Figure 3 shows how

Table 3. *Sample sizes, e-values and p-values for the comparison of tail probability forecasts; the sample size is the number of observations for which the condition* $\min(p_t, q_t) \geqslant 0.5$ *holds*

	Brussels		Frankfurt		London		Zurich	
Lag	*n*	*E* (*p*)	*n*	*E* (*p*)	*n*	*E* (*p*)	*n*	*E* (*p*)
1	116	>100 (0.050)	79	0.175 (0.814)	72	0.45 (0.724)	92	0.047 (0.892)
2	88	23.409	87	3.327	69	1.332	99	2.961
3	68	10.704	62	3.542	60	1.429	75	0.567
4	49	2.338	53	1.166	39	0.868	52	0.773
5	28	1.029	26	1.033	30	1.077	36	1.073

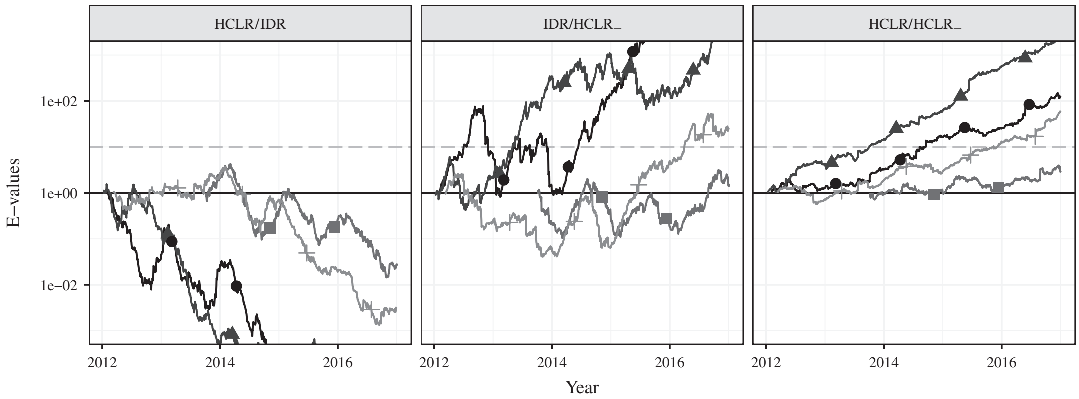


Fig. 3. E-values for the hypotheses tests at lag 1 for Brussels (dots), Frankfurt (triangles), London (squares) and Zurich (crosses). The abbreviations of the hypotheses are as in Table 2.

the cumulative products of the e-values for the hypothesis tests at lag 1 evolve over time. If the goal was to accumulate strong evidence against the hypotheses, say exceeding the level 10, then the hypothesis that IDR outperforms HCLR$_{-}$ could already be rejected with only 9% or 27% of the data for Brussels and Frankfurt airport, respectively, which is where the corresponding lines first cross the level 10. For Zurich airport, rejection happens at 85% of the total sample size.

Interestingly, in the comparison of HCLR and HCLR$_{-}$ for Brussels at lag 1, the p-value is nonsignificant, at 0.07, but the e-value gives decisive evidence, being greater than 100. We attribute this to the different null hypotheses of the tests. The mean difference in Brier score is only 0.001 with an estimated standard deviation of 0.03, giving little evidence against the null hypothesis of the Diebold-Mariano test. However, the null hypothesis for the e-value is smaller, requiring that HCLR$_{-}$ outperform HCLR at all time-points. Even if the score differences are only small, evidence eventually accumulates over the whole time period; see the rightmost panel of Fig. 3. The fact that the e-values in the HCLR/HCLR$_{-}$ comparison decrease with the forecast lag is an effect of the less powerful merging method for e-values with higher lag.

In the comparisons of extreme precipitation warnings, the p-value gives some evidence against the null hypothesis for Brussels airport, and the corresponding e-value is decisive, $E = 3703$. For the other lag-1 forecasts, both p-values and e-values do not indicate that including the ensemble standard deviation brings a benefit. As for the higher lags, for London and Zurich airports there is no evidence that HCLR outperforms HCLR$_{-}$, and for Brussels and Frankfurt airports there is evidence only at lags 2 and 3. Overall, the evidence in favour of the HCLR model for issuing extreme precipitation warnings as compared to HCLR$_{-}$ is surprisingly weak.

Acknowledgement

The authors acknowledge financial support from the Swiss National Science Foundation. They are grateful to Ruodu Wang for introducing them to e-values for hypothesis testing, and to Tobias Fissler and Aaditya Ramdas for helpful comments.

Supplementary material

Supplementary Material available at *Biometrika* online includes extensions of the simulation examples, a proof of the validity of the proposed stopping rule for lags $h > 1$, and a version of Table 2 without truncation of the e-values.

Appendix

Proof of Theorem 1. If $E(y)$ is of the stated form, then $E(y) \geqslant E\{\mathbb{1}(p > q)\} = 1 - \lambda \geqslant 0$, and one can easily verify that E has the given null hypothesis. Assume that $p < q$; the case of $p > q$ is analogous. Define $d_{p,q}(y) = S(p, y) - S(q, y)$ and, for $\pi \in [0, 1]$,

$$f(\pi) = \mathbb{E}_\pi\{d_{p,q}(Y)\} = (1 - \pi)d_{p,q}(0) + \pi d_{p,q}(1).$$

The elementary score representation (2) and $\nu\{[p, q)\} > 0$ imply that $d_{p,q}(0) < 0 < d_{p,q}(1)$, so $f(\pi)$ is strictly increasing in π and equal to zero for some $\pi_0 \in (0, 1)$. Let $E = E(y)$ be an e-value under H_S with alternative H_S^{c}, i.e., $E(y) \geqslant 0$ and

$$\mathbb{E}_\pi\{E(Y)\} = (1 - \pi)E(0) + \pi E(1) \leqslant 1 \iff f(\pi) \leqslant 0. \tag{A1}$$

Condition (A1) implies that $\mathbb{E}_\pi\{E(Y)\} = 1$ if and only if $f(\pi) = 0$, which yields

$$\frac{d_{p,q}(0)}{d_{p,q}(1) - d_{p,q}(0)} = \frac{E(0) - 1}{E(1) - E(0)}. \tag{A2}$$

Rearranging this equation gives $E(1) = 1 - \{1 - E(0)\}d_{p,q}(1)/d_{p,q}(0)$. It follows from (A1) and (A2) that $E(0) \in (0, 1)$, so with $\lambda = 1 - E(0)$ we obtain $E(y) = 1 + \lambda d_{p,q}(y)/|d_{p,q}(0)|$. Similar arguments for the case $p > q$ show that in general,

$$E(y) = 1 + \lambda \frac{d_{p,q}(y)}{|d_{p,q}\{\mathbb{1}(p > q)\}|}.$$

□

Proof of Theorem 2. All e-values for the given null hypothesis are of the form (6). To find the GROW e-value under the alternative that $Y = 1$ with probability π_1, we have to maximize

$$\mathbb{E}_{\pi_1}[\log\{E_{p,q;\lambda}(Y)\}] = (1 - \pi_1)\log\left[1 - \lambda \frac{d_{p,q}(0)}{d_{p,q}\{\mathbb{1}(p > q)\}}\right] + \pi_1 \log\left[1 - \lambda \frac{d_{p,q}(1)}{d_{p,q}\{\mathbb{1}(p > q)\}}\right],$$

where again $d_{p,q}(y) = S(p, y) - S(q, y)$. Let $p < q$; the $p > q$ case is analogous. Under this assumption $d_{p,q}(0) < 0 < d_{p,q}(1)$, and $g(\lambda) = \mathbb{E}_{\pi_1}[\log\{E_{p,q;\lambda}(Y)\}]$ is continuous in λ with $g(0) = 0$ and $\lim_{\lambda \to 1} g(\lambda) = -\infty$, so a maximum is attained at some $\lambda \in [0, 1)$. Define $h = d_{p,q}(1)/d_{p,q}(0) < 0$, so that

$$g(\lambda) = (1 - \pi_1)\log(1 - \lambda) + \pi_1 \log(1 - \lambda h), \quad g'(\lambda) = -\frac{1 - \pi_1}{1 - \lambda} - \pi_1 \frac{h}{1 - \lambda h}$$

and $g'(\lambda_0) = 0$ is equivalent to $\lambda_0 = \pi_1 + (1 - \pi_1)/h$. By the definition of H_S, $\pi_1 \notin H_S$ holds if and only if $\mathbb{E}_{\pi_1}\{d_{p,q}(Y)\} > 0$, which is equivalent to $\pi_1 + (1 - \pi_1)/h > 0$, so indeed $\lambda_0 > 0$ for all $\pi_1 \notin H_S$, and

$$E_{p,q;\lambda_0}(0) = 1 - \lambda_0 = (1 - \pi_1)\left(1 - \frac{1}{h}\right) = (1 - \pi_1)\frac{d_{p,q}(1) - d_{p,q}(0)}{d_{p,q}(1)},$$

$$E_{p,q;\lambda_0}(1) = 1 - \lambda_0 \frac{d_{p,q}(1)}{d_{p,q}(0)} = \pi_1 \frac{d_{p,q}(0) - d_{p,q}(1)}{d_{p,q}(0)}.$$

With $d_{p,q}(y) = \int \mathbb{1}\{p \leqslant \theta < q\}(\theta - y)\,\mathrm{d}\nu(\theta)$, it now follows that

$$\frac{d_{p,q}(1) - d_{p,q}(0)}{d_{p,q}(1)} = \frac{-\nu\{[p,q)\}}{-\nu\{[p,q)\} + \int_{[p,q)} \theta \,\mathrm{d}\nu(\theta)} = \frac{1}{1 - \kappa_\nu\{[p,q)\}}$$

and $1 - h = \{d_{p,q}(0) - d_{p,q}(1)\}/d_{p,q}(0) = \kappa_\nu\{[p,q)\}^{-1} > \pi_1^{-1}$, which gives the desired result. □

Proof of Theorem 3. A direct computation shows that $H = [0,p]$ if $p < q$ and $H = [p,1]$ if $p > q$, and that $\mathbb{E}_\pi\{E^{\pi_1 *}_{p,q}(Y)\} \leqslant 1$ for all $\pi \in H$ and $\mathbb{E}_\pi\{E^{\pi_1 *}_{p,q}(Y)\} > 1$ for $\pi \notin H$. The result then follows by Theorem 1 of Grünwald et al. (2020), with W_1 being the Dirac measure of the point $\{\pi_1\}$. □

Proof of Proposition 1. This follows as in the $h = 1$ case with sequential conditioning on $\mathcal{F}_{k+hl}$, $l = 1, \ldots, \lfloor (T-k)/h \rfloor$, for each of the h products $\prod_{l \in I_k} E_{p_l,q_l;\lambda_l}(Y_{l+h})$. □

References

Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. & Wei, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–97.

Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econ. Statist.* **13**, 253–63.

Ehm, W., Gneiting, T., Jordan, A. & Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Statist. Soc. B.* **78**, 505–62.

Ehm, W. & Krüger, F. (2018). Forecast dominance testing via sign randomization. *Electron. J. Statist.* **12**, 3758–93.

Giacomini, R. & White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545–78.

Gneiting, T. (2011). Making and evaluating point forecasts. *J. Am. Statist. Assoc.* **106**, 746–62.

Gneiting, T., Balabdaoui, F. & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. B.* **69**, 243–68.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.* **102**, 359–78.

Gneiting, T. & Ranjan, R. (2013). Combining predictive distributions. *Electron. J. Statist.* **7**, 1747–82.

Grünwald, P., de Heide, R. & Koolen, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*. IEEE.

Henzi, A., Ziegel, J. F. & Gneiting, T. (2021). Isotonic distributional regression. *J. R. Statist. Soc. B* **83**, 963–93.

Kelly Jr, J. L. (1956). A new interpretation of information rate. *Bell System Tech. J.* **35**, 917–26.

Lai, T. Z., Gross, S. T. & Shen, D. B. (2011). Evaluating probability forecasts. *Ann. Statist.* **39**, 2356–82.

Lazarus, E., Lewis, D. J., Stock, J. H. & Watson, M. W. (2018). HAR inference: Recommendations for practice. *J. Bus. Econ. Statist.* **36**, 541–59.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. & Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32**, 106–7.

Leutbecher, M. & Palmer, T. N. (2008). Ensemble forecasting. *J. Comp. Phys.* **227**, 3515–39.

Messner, J. W., Mayr, G. J., Wilks, D. S. & Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Weather Rev.* **142**, 3003–14.

Molteni, F., Buizza, R., Palmer, T. N. & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Patton, A. J. (2020). Comparing possibly misspecified forecasts. *J. Bus. Econ. Statist.* **38**, 796–809.

Ramdas, A., Ruf, J., Larsson, M. & Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167*.

Ranjan, R. & Gneiting, T. (2010). Combining probability forecasts. *J. R. Statist. Soc. B.* **72**, 71–91.

Schervish, M. J. (1989). A general method for comparing probability assessors. *Ann. Statist.* **17**, 1856–79.

Seillier-Moiseiwitsch, F. & Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Am. Statist. Assoc.* **88**, 355–59.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *J. R. Statist. Soc. A* **184**, 407–31.

Vannitsem, S., Wilks, D. S. & Messner, J., eds. (2018). *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam: Elsevier.

Vovk, V. & Wang, R. (2020). True and false discoveries with independent e-values. *arXiv:2003.00593*.

Vovk, V. & Wang, R. (2021). E-values: Calibration, combination, and applications. *Ann. Statist.* **49**, 1739–54.

Wang, R. & Ramdas, A. (2020). False discovery rate control with e-values. *arXiv:2009.02824v2*.

Waudby-Smith, I. & Ramdas, A. (2021). Estimating means of bounded random variables by betting. *arXiv:2010.09686v4*.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test* **5**, 1–60.

Yen, Y. & Yen, T. (2021). Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. *Int. J. Forecast.* **37**, 733–58.

Zhu, Y. & Timmermann, A. (2020). Can two forecasts have the same conditional expected accuracy? *arXiv:2006.03238*.

[*Received on* 15 *March* 2021. *Editorial decision on* 6 *September* 2021]

Biometrika (2020), **103**, 1, *p.* 1
Printed in Great Britain
Advance Access publication on 31 July 2018

Supplementary material for "Valid sequential inference on probability forecast performance"

BY ALEXANDER HENZI AND JOHANNA F. ZIEGEL
University of Bern, Institute of Mathematical Statistics and Actuarial Science,
Alpeneggstrasse 22, 3012 Bern, Switzerland.
alexander.henzi@stat.unibe.ch johanna.ziegel@stat.unibe.ch

1. OPTIONAL STOPPING FOR LAGS $h > 1$

In Section 3.2 of the article, the stopping rule

$$\tau_{\alpha,h} = \min\Big(T,\ \inf\big[t \geq h+1 : e_t \geq \max_{j=t-h+1,\dots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\}^{-1}/\alpha\big]\Big),$$

is defined for e-values of the form

$$e_T = \frac{1}{h}\sum_{k=1}^{h} \prod_{l \in I_k(T)} E_{p_l,q_l;\lambda_l}(Y_{l+h}),$$

where $I_k(T) = \{k + hs : s = 0, \dots, \lfloor (T-k)/h \rfloor - 1\}$. Assume that at time t, it is observed that $e_t \geq \max_{j=t-h+1,\dots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\}^{-1}/\alpha$, and that optional stopping is applied, i.e. $E_{p_s,q_s;\lambda_s}(Y_{t+s}) \equiv 1$ for $s \geq t$. The claim is that then $e_{t+h-1} \geq 1/\alpha$ no matter what values $Y_{t+1}, \dots, Y_{t+h-1}$ take. Because $E_{p_t,q_t;\lambda_t}(Y_{t+h}) \equiv 1$, we have $e_{t+h-1} = e_{t+h}$. For $k = 1, \dots, h$, let $s_k = k + h\lfloor (t-k)/h \rfloor$, so that $\{s_1, \dots, s_h\} = \{t-h+1, \dots, t\}$. Then, using that $I_k(t+h) \setminus \{s_k\} = I_k(t)$,

$$\begin{aligned}
e_{t+h-1} = e_{t+h} &= \frac{1}{h}\sum_{k=1}^{h} \left\{ E_{p_{s_k},q_{s_k};\lambda_{s_k}}(Y_{s_k+h}) \prod_{l \in I_k(t+h)\setminus\{s_k\}} E_{p_l,q_l;\lambda_l}(Y_{l+h}) \right\} \\
&\geq \frac{1}{h}\sum_{k=1}^{h} \left[E_{p_{s_k},q_{s_k};\lambda_{s_k}}\{\mathbb{1}(p_{s_k} > q_{s_k})\} \prod_{l \in I_k(t)} E_{p_l,q_l;\lambda_l}(Y_{l+h}) \right] \\
&\geq \min_{j=t-h+1,\dots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\} \cdot \frac{1}{h}\sum_{k=1}^{h} \prod_{l \in I_k(t)} E_{p_l,q_l;\lambda_l}(Y_{l+h}) \\
&= \left[\max_{j=t-h+1,\dots,t-1} E_{p_j,q_j;\lambda_j}\{\mathbb{1}(p_j > q_j)\}^{-1} \right]^{-1} e_t \ \geq\ 1/\alpha.
\end{aligned}$$

2. SIMULATION EXAMPLES: ADDITIONAL FIGURES

The simulation example in Section 4.1 in the article has been tested for robustness with respect to various parameters: significance levels ($\alpha = 0.001,\ 0.01,\ 0.05$), scoring functions (Brier score, spherical score, logarithmic score), sample sizes (150, 300, 600, 1200, 2400), tests for computing p-values (Student's t-test, Wilcoxon's signed rank test), and alternative hypotheses for constructing the e-values (parameter k as explained in Section 4.1 in the article).

For the spherical and the logarithmic score, the probability π_t was computed in such a way that $\mu = 0.5$ corresponds to a score difference of zero, namely, with $r_t = \mathbb{E}_\nu\big\{\theta \mid \theta \in [\min(p_t, q_t), \max(p_t, q_t))\big\}$, we set $\pi_t = p_t$ for $\mu = 0$, $\pi_t = r_t$ for $\mu = 0.5$, $\pi_t = q_t$ for $\mu = 1$, and interpolate linearly in between these three points for the other μ.

Figure S1 demonstrates that the rejection rates of the e-values are almost the same for all scoring functions.

Figure S2 shows how the rejection rates vary with the alternative hypothesis for the e-value. In particular, it can be seen that the alternative π_t is superior and q_t is inferior for all sample sizes and significance levels. As for the alternatives with the parameter k, smaller k give higher rejection rates for small sample sizes and lower rejection rates for larger samples.

Figure S3 shows that also the rejection rates of Student's t-test are essentially equal for the different scoring functions.

In Figure S4, it can be seen that the rejection rates of Student's t-test and Wilcoxon's signed rank test for this simulation are almost equal.

Figure S5 shows that close to $\mu = 0.05$, Student's t-test under optional stopping has too high rejection rates independent of the significance level and the sample size.

The simulation example in Section 4.2 was tested with different significance levels and scoring functions.

Figure S6 shows that the choice of the scoring function has a minor influence on the rejection rates for the sample sizes 300 and 600, and almost no effect for 1200 and 2400.

Figure S7 compares the rejection rates of the Diebold-Mariano test and the e-values for different significance levels.

3. CASE STUDY: ADDITIONAL MATERIAL

Table S1 contains the e-vales and p-values of Table 2 in scientific digit notation.

[*Received on 2 January 2017. Editorial decision on 1 April 2017*]

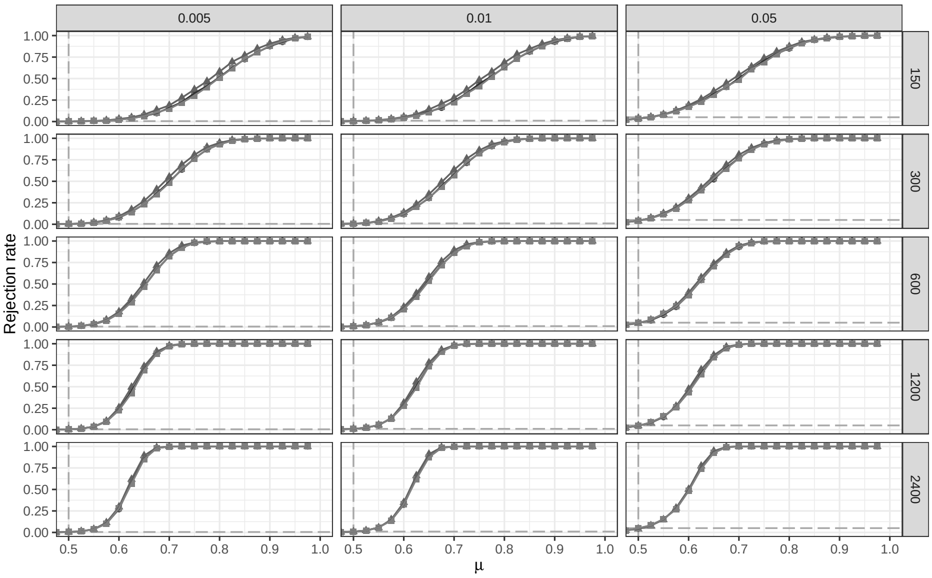


Fig. S1. Rejection rate of stopped e-value (alternative hypothesis with $k = 1$ as explained in the article) for Brier score (dots), spherical score (squares), logarithmic score (triangles), and different significance levels (columns) and sample sizes (rows).

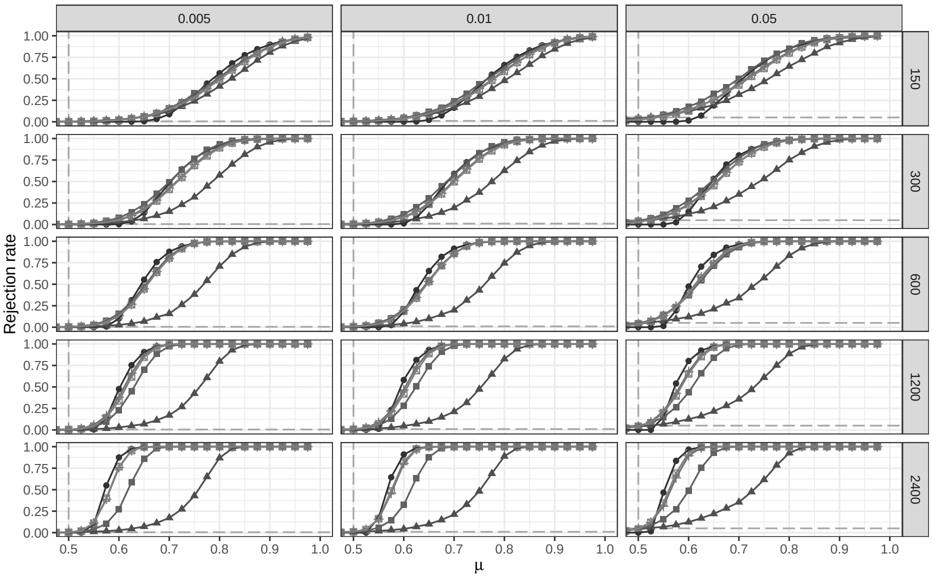


Fig. S2. Rejection of stopped e-values based on Brier score for different alternative hypotheses and different sample sizes and significance levels. The alternatives are π_t (dots), q_t (triangles), $k = 1$ (filled squares), $k = 3$ (crosses), $k = 5$ (squares with cross).

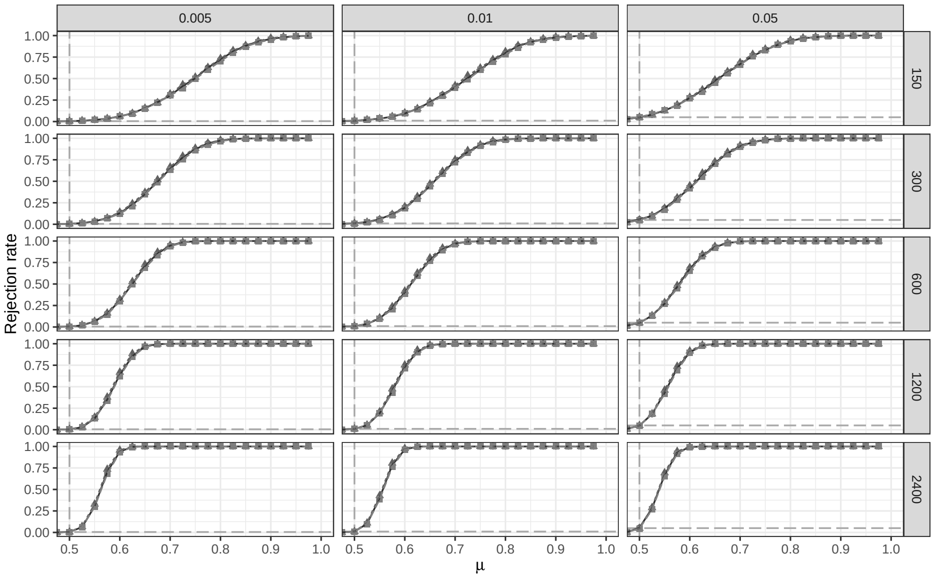


Fig. S3. Rejection rate of Student's t-test for Brier score (dots), spherical score (squares), and logarithmic score (triangles) differences, for different significance levels and sample sizes.

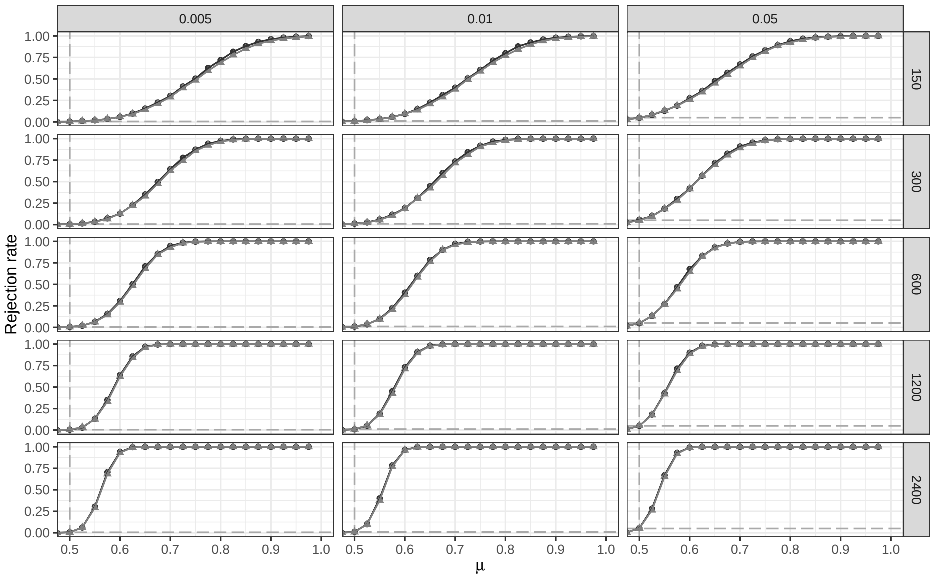


Fig. S4. Rejection rates of Student's t-test (circles) and Wilcoxon's signed rank test (triangles) for Brier score differences, for different significance levels and sample sizes.

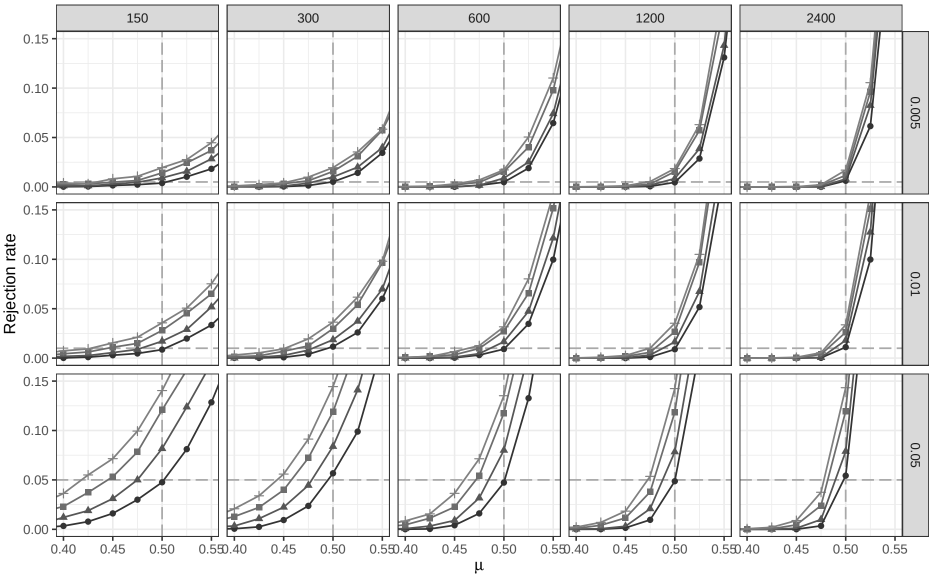


Fig. S5. Rejection rates of Student's t-test under optional stopping, for different significance levels and sample sizes. Optional stops are included at 1 (triangles), 3 (squares) and 5 equispaced time points in between 1 and the sample size T. Dots show the rejection rates without optional stopping.

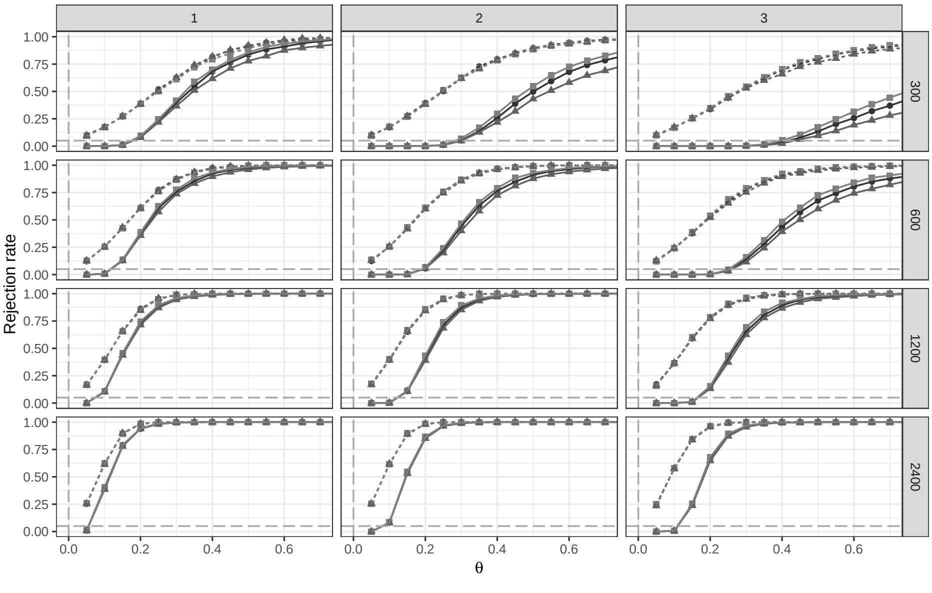


Fig. S6. Rejection rates of Diebold-Mariano test (dashed lines) and e-values (normal lines) for the Brier score (dots), spherical score (squares), and the logarithmic score (triangles), and for different lags (columns) and sample sizes (rows).

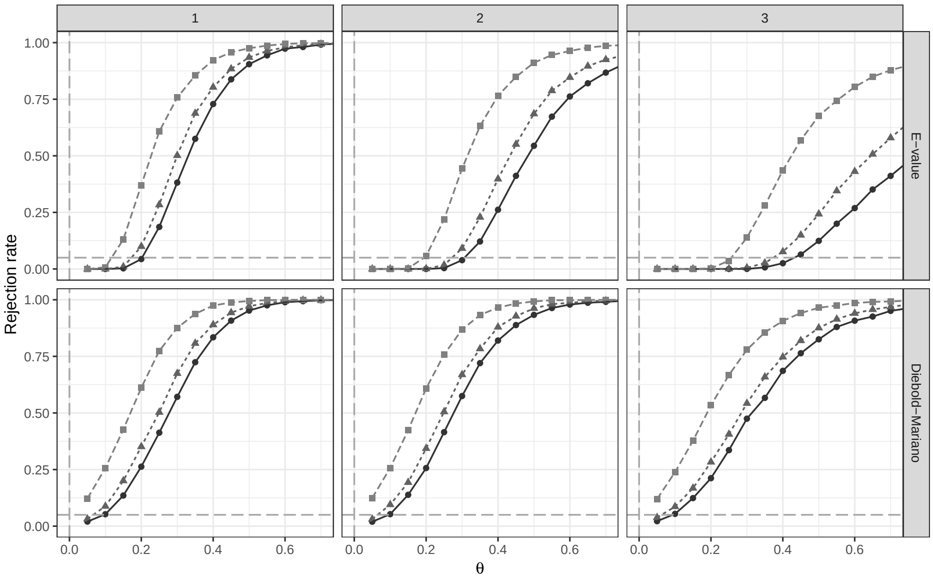


Fig. S7. Rejection rates of the Diebold-Mariano test and E-values for the significance levels 0.005 (dots), 0.01 (triangles), and 0.05 (squares), based on Brier score differences and a sample size of 600.

Table S1. *Brier scores for different probability of precipitation forecasting methods, and e-values (E) and p-values (p) for testing significance of score differences. The columns HCLR/IDR show e-values and p-values for tests tests of the null hypothesis that IDR probability of precipitation forecasts achieve a lower Brier score the HCLR forecasts; the interpretation is analogous for the other forecast pairs.*

		Average Brier score			HCLR/IDR		IDR/HCLR$_-$		HCLR/HCLR$_-$	
	Lag	IDR	HCLR	HCLR$_-$	E	p	E	p	E	p
BRU	1	0.107	0.117	0.118	5.6e−08	1.0e+00	5.0e+09	1.3e−05	1.3e+02	7.0e−02
	2	0.119	0.123	0.125	9.5e−03	9.5e−01	2.2e+02	1.0e−02	1.4e+01	2.9e−02
	3	0.134	0.133	0.136	4.3e−01	4.4e−01	5.4e+02	1.9e−01	1.5e+01	1.9e−03
	4	0.152	0.145	0.148	4.8e+00	1.4e−02	1.9e+00	9.4e−01	5.2e+00	7.4e−03
	5	0.171	0.161	0.164	1.7e+01	2.3e−04	4.1e−01	1.0e+00	3.4e+00	3.3e−04
FRA	1	0.109	0.111	0.114	1.4e−06	7.8e−01	1.6e+11	2.1e−02	2.4e+03	2.8e−06
	2	0.114	0.119	0.122	5.4e−02	9.6e−01	1.3e+06	2.3e−04	2.5e+02	4.2e−04
	3	0.123	0.127	0.132	7.8e−02	9.4e−01	3.8e+04	1.3e−04	2.7e+01	5.4e−06
	4	0.147	0.144	0.147	2.3e+00	9.7e−02	9.6e+00	5.2e−01	5.5e+00	5.9e−05
	5	0.166	0.161	0.163	1.5e+00	3.0e−02	2.4e+00	8.9e−01	3.2e+00	5.1e−03
LHR	1	0.135	0.138	0.139	2.9e−02	8.1e−01	1.5e+01	1.3e−01	2.8e+00	3.7e−01
	2	0.138	0.143	0.143	1.9e−01	9.2e−01	1.2e+02	5.1e−02	2.9e+00	4.4e−01
	3	0.152	0.154	0.155	7.3e−01	7.5e−01	4.1e+01	1.4e−01	2.5e+00	3.4e−01
	4	0.169	0.167	0.169	1.4e+00	2.5e−01	1.7e+00	5.4e−01	1.7e+00	7.8e−02
	5	0.186	0.181	0.182	1.6e+00	7.5e−02	3.8e−01	9.3e−01	1.1e+00	3.2e−01
ZRH	1	0.104	0.108	0.110	3.0e−03	9.3e−01	3.0e+04	5.5e−03	6.2e+01	3.2e−04
	2	0.110	0.112	0.114	1.2e−01	7.2e−01	3.7e+01	3.0e−02	1.0e+01	5.0e−05
	3	0.121	0.118	0.121	1.5e+00	8.9e−02	3.2e+01	4.4e−01	5.1e+00	1.0e−04
	4	0.138	0.132	0.134	4.1e+00	2.7e−03	1.3e+00	9.6e−01	2.8e+00	1.5e−03
	5	0.165	0.156	0.159	1.5e+01	2.3e−05	8.4e−01	1.0e+00	2.4e+00	1.7e−04

4.2 Sequentially valid tests for forecast calibration

The content of this section is published as an arXiv preprint,

Arnold, S., Henzi, A. and Ziegel, J. F. (2021). Sequentially valid tests for forecast calibration. *arXiv preprint arXiv:2109.11761.*

Sequentially valid tests for forecast calibration

Sebastian Arnold Alexander Henzi Johanna F. Ziegel

{sebastian.arnold,alexander.henzi,johanna.ziegel}@stat.unibe.ch

November 8, 2021

Abstract

Forecasting and forecast evaluation are inherently sequential tasks. Predictions are often issued on a regular basis, such as every hour, day, or month, and their quality is monitored continuously. However, the classical statistical tools for forecast evaluation are static, in the sense that statistical tests for forecast calibration are only valid if the evaluation period is fixed in advance. Recently, e-values have been introduced as a new, dynamic method for assessing statistical significance. An e-value is a non-negative random variable with expected value at most one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a conservative p-value. E-values are particularly suitable for sequential forecast evaluation, since they naturally lead to statistical tests which are valid under optional stopping. This article proposes e-values for testing probabilistic calibration of forecasts, which is one of the most important notions of calibration. The proposed methods are also more generally applicable for sequential goodness-of-fit testing. We demonstrate in a simulation study that the e-values are competitive in terms of power when compared to extant methods, which do not allow for sequential testing. In this context, we introduce test power heat matrices, a graphical tool to compactly visualize results of simulation studies on test power. In a case study, we show that the e-values provide important and new useful insights in the evaluation of probabilistic weather forecasts.

1 Introduction

Probabilistic forecasts incorporate the uncertainty about a future quantity Y comprehensively in the form of probability distributions. A minimal requirement for useful probabilistic forecasts is calibration, meaning that the predicted probabilities should conform with the actual observed event frequencies. This article develops novel statistical tools to validate *probabilistic calibration*, one of the most prominent and widely applied notions of calibration. Probabilistic calibration requires that Y should be below the α-quantile of the forecast distribution with a frequency of about $\alpha \cdot 100\%$, for all $\alpha \in (0, 1)$. More precisely, the predictive cumulative distribution function (CDF) F is evaluated at the outcome Y, and this quantity, suitably randomized in case of discontinuities of F, is called the probability integral transform (PIT) and should be uniformly distributed on $(0, 1)$ for a probabilistically calibrated forecast. Checks of the uniformity of the PIT, and of the closely related rank histogram, constitute a cornerstone of forecast evaluation (Diebold et al., 1998; Hamill, 2001; Gneiting et al., 2007).

From a statistical point of view, testing probabilistic calibration for forecasts with lag 1 is straightforward. For example, in the case of daily forecasts issued for the next day, it suffices to apply any goodness of fit test for the standard uniform distribution to a sample of the PIT from a series of forecasts and observations. However, it has been noted early on that statistical tests alone are not informative enough, because they do not indicate the type of misspecification (Diebold et al., 1998). Therefore, tests of calibration are commonly accompanied by a histogram plot of the PIT distribution, which allows to identify classical types of misspecification at a glance, namely, biased forecasts lead to PIT histograms skewed to the left or the right, and under- or overdispersed forecasts yield U-shaped or inverse U-shaped PIT histograms, respectively.

We argue that a drawback of the established tools for validating probabilistic calibration is that they do not fully account for the sequential nature of forecasting. The relationship between forecasts and observations is often complicated and forecast misspecification changes over time. However, the classical tools for validating calibration require the sample size to be fixed in advance and independently of the data. As an illustrative example, consider a weather forecaster who, after updating a prediction model, monitors the quality of daily forecasts and wants to check if the new forecasts are probabilistically calibrated. She aims at a sample size of one year, and plans to check uniformity of the PIT at the end of the observation period. If, by chance, the forecaster realizes after half of observation period that the forecast is strongly biased, then a p-value from a classical goodness of fit test with all data at this time point is not valid since the sample size depends on the data. On the other hand, if the forecaster does not look at the data until the end of the observation period, then the PIT distribution with the sample of the full year could again be close to uniform, for example if there is a change in the direction of the bias in the second half of the year, and the forecaster is unable to detect the misspecification. Such effects appear in practice as exemplified in the case study in Section 5. Any analysis of sub-periods has to be planned in advance, which is often difficult and cumbersome since it is usually not known in advance how forecast misspecification changes over time and what discretizations of the time domain are appropriate.

This article develops new methodology for checks of probabilistic and related notions of calibration in a sequential setting, based on the new concept of *e-values*. E-values have received an increasing interest in recent years, see Vovk and Wang (2021), Grünwald et al. (2019), Shafer (2021) (who uses the term *betting score*), Ramdas et al. (2020), and the references therein. Henzi and Ziegel (2021) gave a first application of e-values to forecast comparison; see also the more recent article by Choe and Ramdas (2021). An e-value is a non-negative random variable E such that for all distributions P in a set $\mathcal{P}$, the null hypothesis, the inequality $\mathbb{E}_P(E) \leq 1$ holds. E-values can be easily transformed into (conservative) p-values since $P(1/E \geq \alpha) \leq \alpha$ for $\alpha \in (0,1)$ by Markov's inequality, and large e-values give evidence against the null hypothesis. One main motivation for using e-values is their simple behaviour under combinations. Convex combinations of e-values are again e-values, and so is the product of independent e-values. In a a sequential setting, if $(E_t)_{t\in\mathbb{N}}$, is a sequence of e-values adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, then by the tower property of conditional expectations the process $e_t = \prod_{i=1}^{t} E_i$, $t \in \mathbb{N}$, is a non-negative supermartingale or test martingale and it satisfies Ville's inequality, that is $P(\sup_{t\in\mathbb{N}} e_t \geq 1/\alpha) \leq \alpha$; see Ramdas et al. (2020) for a comprehensive analysis of non-negative martingales for statistical testing. In the example of the weather forecaster from the previous paragraph, this implies that with e-values the forecaster may reject the hypothesis of calibration at the level α as soon as the process $(e_t)_{t\in\mathbb{N}}$ exceeds $1/\alpha$, without having to fix a sample size in advance. The forecaster is allowed to monitor the PIT and the process $(e_t)_{t\in\mathbb{N}}$ in real time. In the special case of a simple null hypothesis, $\mathcal{P} = \{P_0\}$, e-values take the form of likelihood ratios or Bayes factors (Grünwald et al., 2019). In particular, e-values for testing the null hypothesis that a quantity $Z \in (0,1)$ is uniformly distributed on the unit interval, short UNIF$(0,1)$, are Lebesgue densities on $[0,1]$. It is therefore simple to construct valid e-values for testing probabilistic calibration or, detached from the forecasting context, goodness-of-fit testing of the UNIF$(0,1)$ distribution, with "valid" referring to type one error guarantees. The non-trivial task in the construction of e-values is to achieve sufficient power to detect violations of the null hypothesis. Shafer (2021) calls strategies for constructing e-values "betting strategies", since an e-value can be interpreted as a bet against the null hypothesis and the process $(e_t)_{t\in\mathbb{N}}$ corresponds to the capital over time if all gains are reinvested into the new bet at each $t \in \mathbb{N}$.

The contributions of this article are as follows. In Section 3 we construct e-values for testing the null hypothesis that a quantity $Z \in [0,1]$ is distributed according to UNIF$(0,1)$, and for the analogous hypothesis that a discrete $R \in \{1, \dots, m\}$ follows a uniform distribution on the integers 1 to m, short $\mathrm{UNIF}(\{1, \dots, m\})$. These hypotheses appear naturally in calibration checks for probabilistic forecasts and ensemble forecasts, and precise definitions of forecast calibration are given in Section 2. Furthermore, we characterize and construct e-values for the weaker hypotheses that a random variable $Z \in [0, 1]$ with distribution P is stochastically smaller than $\mathrm{UNIF}(0, 1)$, short $P \leq_{\mathrm{st}} \mathrm{UNIF}(0, 1)$, which means $P(Z \leq z) \geq z$ for all $z \in (0, 1)$. This hypothesis appears in a new definition of calibration for quantile forecasts which is closely related to usual probabilistic calibration. Section 3 is of interest independent from the forecasting context, and the methods can also be applied to general goodness-of-fit or stochastic order testing problems in sequential settings. Proofs of theoretical results are deferred to Appendix A. In Section 4 we demonstrate that the e-values are competitive in terms of power when compared to established tests. Here, we suggest a new graphical tool to compactly display simulation results on test power, so-called test power heat matrices. Section 5 presents an application to testing calibration of postprocessed weather forecasts, and we show that the e-values give rise to novel and informative graphical tools for the sequential evaluation of forecast calibration.

2 Probabilistic calibration

Let Y be a real-valued outcome defined on a probability space $(\Omega, \mathcal{F}, P)$. We denote by F the CDF associated with a (random) probabilistic forecast for Y.

Definition 2.1. The *probability integral transform (PIT)* of a forecast F for an outcome Y is defined as $Z_F(Y) = F(Y-) + V(F(Y) - F(Y-))$, where $F(y-) = \lim_{z \to y, z < y} F(z)$ and V is a uniformly distributed random variable on $(0, 1)$ independent of the pair (F, Y). The forecast F is *probabilistically calibrated* if $Z_F(Y) \sim \mathrm{UNIF}(0, 1)$.

Of great importance in weather forecasting are ensemble forecasts (Bauer et al., 2015). An ensemble forecast is a collection of point forecasts generated by running a numerical weather prediction (NWP) model m times, typically $m = 20$ to 50, each time with different initial conditions, which allows to quantify the forecast uncertainty. We denote ensemble forecasts by vectors $\boldsymbol{X} = (X_1, \dots, X_m) \in \mathbb{R}^m$. To define calibration, let the (randomized) rank of Y equal

$$\mathrm{rank}_{\boldsymbol{X}}(Y) = 1 + \#\big\{i = 1, \dots, m \mid X_i < Y\big\} + W \ \in \{1, \dots, m + 1\}, \tag{1}$$

where W is a random variable that equals zero almost surely if $N = \#\big\{i = 1, \dots, m \mid X_i = Y\big\}$ is zero, and is uniformly distributed on $\big\{1, \dots, N\big\}$ otherwise.

Definition 2.2. An ensemble forecast $\boldsymbol{X}$ is *rank calibrated* if $\mathrm{rank}_{\boldsymbol{X}}(Y) \sim \mathrm{UNIF}\big(\{1, \dots, m{+}1\}\big)$.

Remark. Rank calibration is commonly assessed with the rank histogram, a plot of the empirical frequencies of the ranks over a sample (Anderson, 1996). We use a randomization of the rank in case of ties because with this convention the PIT and the rank are related via the equation $\mathrm{rank}_{\boldsymbol{X}}(Y) = 1 + \lfloor m Z_{F_{\boldsymbol{X}}}(Y) \rfloor$, where $F_{\boldsymbol{X}}$ is the empirical CDF (ECDF) of the ensemble $\boldsymbol{X}$. The definition of the rank given in (1) slightly generalizes the unified PIT introduced by Vogel et al. (2018, p. 374) for evaluating precipitation forecasts, which randomizes ranks in case of multiple occurrences of zero forecasts.

A closely related notion of calibration can be given for quantile forecasts. Let $\alpha_0 = 0 < \alpha_1 < \dots < \alpha_K < 1 = \alpha_{K+1}$ be K quantile levels. Instead of issuing a complete predictive CDF for the unknown quantity Y, we only aim to give point forecasts $q_1 \leq \dots \leq q_K$ for the quantiles of the distribution of Y at levels $\alpha_1, \dots, \alpha_K$. Recall, that q_i is an α_i-quantile of F if

$$F(q_i-) \leq \alpha_i \leq F(q_i), \quad i = 1, \dots, K.$$

Therefore the set of quantile forecasts can be interpreted as a partial disclosure of the predictive CDF F. With $q_0 = -\infty$ and $q_{K+1} = \infty$, define

$$F_u(y) := \sum_{i=1}^{K+1} (\alpha_i - \alpha_{i-1}) \mathbb{1}\{q_i \le y\}, \quad F_\ell(y) := \sum_{i=0}^{K} (\alpha_{i+1} - \alpha_i) \mathbb{1}\{q_i \le y\}, \quad y \in \mathbb{R}.$$

Proposition 2.1. *Let $0 < \alpha_1 < \cdots < \alpha_K < 1$ be K quantile levels. Any (deterministic) CDF F with corresponding quantiles $q_1 \le \cdots \le q_K$ satisfies*

$$F_u(y) \le F(y) \le F_\ell(y), \quad y \in \mathbb{R}. \tag{2}$$

Furthermore

$$F_u(q_i-) \le \alpha_i \le F_u(q_i) \quad \text{and} \quad F_\ell(q_i-) \le \alpha_i \le F_\ell(q_i), \quad i = 1, \ldots, K. \tag{3}$$

Similar to the classical PIT, we define

$$Z_{F_u}(Y) := V F_u(Y) + (1 - V) F_u(Y-) \text{ and } Z_{F_\ell}(Y) := V F_\ell(Y) + (1 - V) F_\ell(Y-),$$

where V is a uniformly distributed random variable on $(0, 1)$ independent of Y and the quantile predictions $q_1, \ldots, q_K$. In the sequel, $Z_{F_u}(Y)$ and $Z_{F_l}(Y)$ will be referred to as the *upper* and *lower quantile PIT*. By (2) these quantities satisfy

$$Z_{F_u}(Y) \le Z_F(Y) \le Z_{F_\ell}(Y)$$

almost surely, and $Z_{F_\ell}(Y) - Z_{F_u}(Y) \le \sup_{i=0,\ldots,K}(\alpha_{i+1} - \alpha_i)$ with equality if $\alpha_{i+1} - \alpha_i = c > 0$ for all i. In this case, which is the important special case of equispaced quantile levels, $Z_{F_\ell}(Y) = Z_{F_u}(Y) + c$.

Definition 2.3. A set of quantile forecasts $q_1 < \cdots < q_K$ is *probabilistically calibrated* if

$$Z_{F_u}(Y) \le_{\text{st}} \text{UNIF}(0, 1) \le_{\text{st}} Z_{F_\ell}(Y).$$

Remark. The functions F_u and F_ℓ are defective CDFs in the sense that $\lim_{y \to \infty} F_u(y) = \alpha_K < 1$ and $\lim_{y \to -\infty} F_\ell(y) = \alpha_1 > 0$, respectively, but they satisfy the remaining conditions for being a CDF. Note that the distribution of the upper (lower) quantile PIT is stochastically smaller (greater) than UNIF$(0, 1)$, which implies that its CDF is pointwise greater (smaller) than the uniform CDF for probabilistically calibrated quantile forecasts.

The definitions of calibration introduced so far do not include any notion of time. In practice, for example for the PIT, one observes a time series $(F_t, Y_t)_{t \in \mathbb{N}}$ of forecasts and observations, where F_t is the forecast for a lagged observation Y_{t+h}, with a fixed integer lag $h \ge 1$. The definition below formalizes calibration for such sequential settings.

Definition 2.4. Let $(\Omega, \mathcal{F}, P)$ be a probability space with a filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$, and h be a positive integer. Let further $(Y_t)_{t \in \mathbb{N}}$ be an adapted sequence of observations.

(i) A sequence of probability forecasts $(F_t)_{t \in \mathbb{N}}$ is *probabilistically calibrated at lag h* if

$$\mathcal{L}(Z_{F_t}(Y_{t+h}) \mid Z_{F_j}(Y_{j+h}), 0 \le j \le t - h) = \text{UNIF}(0, 1), \ t \in \mathbb{N}.$$

(ii) A sequence of ensemble forecasts $(\boldsymbol{X}_t)_{t \in \mathbb{N}}$ of size m is *rank calibrated at lag h* if

$$\mathcal{L}(\text{rank}_{\boldsymbol{X}_t}(Y_{t+h}) \mid \text{rank}_{\boldsymbol{X}_j}(Y_{j+h}), 0 \le j \le t - h) = \text{UNIF}(\{1, \ldots, m + 1\}), \ t \in \mathbb{N}.$$

(iii) A sequence of quantile forecasts $(q_{1;t},\dots,q_{K;t})_{t\in\mathbb{N}}$ is *probabilistically calibrated at lag h* if

$$Z_{F_{u;t}}(Y_{t+h}) \le_{\mathrm{st}} \mathrm{UNIF}(0,1) \le_{\mathrm{st}} Z_{F_{\ell;t}}(Y_{t+h})$$

conditional on $Z_{F_{u;j}}(Y_{j+h}), Z_{F_{\ell;j}}(Y_{j+h})$, $0 \le j \le t-h$, for $t\in\mathbb{N}$.

Note that for $t \le h$ there is no conditioning in all cases, and the requirements in (i)-(iii) are understood to hold unconditionally. Furthermore we silently assume existence of a sequence $(V_t)_{t\in\mathbb{N}}$ of adapted, independent UNIF$(0,1)$ variables defined on the probability space, independent of all other objects, in parts (i) and (iii) of Definition 2.4 to define the PIT and quantile PIT. Similarly, existence of an analogous sequence $(W_t)_{t\in\mathbb{N}}$ for the randomization of the ranks in part (ii) is assumed. For forecasts with lag $h>1$, the definition does not condition on $Z_{F_j}(Y_{j+h})$ with $t-h<j<t$, since the corresponding observations are not yet available at time t when the forecasts are issued, and in this case, the joint distribution of the PIT (or ranks, quantile PIT) which are less than t time units apart is not specified.

Remark. With lag $h=1$, the definition of calibration implies that the sequence of the PIT, $(F_t(Y_{t+1}))_{t\in\mathbb{N}}$, or of the ranks, $(\mathrm{rank}_{\boldsymbol{X}_t}(Y_{t+1}))_{t\in\mathbb{N}}$, are independent, since the conditional distributions in part (i) or (ii) do not depend on the past values in the sequence. Indeed, for a probabilistically calibrated sequence of forecasts $(F_t)_{t\in\mathbb{N}}$ and $v,w\in[0,1]$, it holds

$$P(Z_{F_1}(Y_2)\le v, Z_{F_2}(Y_3)\le w) = P(Z_{F_2}(Y_3)\le w \mid Z_{F_1}(Y_2)\le v)P(Z_{F_1}(Y_2)\le v) = vw,$$

and it follows inductively that $(Z_{F_t}(Y_{t+1}))_{t\in\mathbb{N}}$ are independent. Hence the definition of probabilistic calibration corresponds to the classic definition given in Diebold et al. (1998). However, for lag $h>1$ and for the quantile PIT, where no particular conditional distribution is assumed, there may be dependence in the sequence of PITs, ranks, or quantile PITs.

3 E-values

3.1 E-values in sequential settings

We proceed with formal definitions and properties of e-values in sequential settings. The notation largely follows Vovk and Wang (2021), but we formalize a new concept of lagged sequential e-values which is particularly relevant in forecast evaluation. Throughout this section, let $(\Omega,\mathcal{F})$ be the underlying measurable space and $\mathcal{P}$ be a suitable set of distributions.

Definition 3.1. Let $\mathcal{H},\mathcal{H}'\subset\mathcal{P}$. An *e-value for $\mathcal{H}$* is a non-negative random variable E such that $\mathbb{E}_P E\le 1$ for all $P\in\mathcal{H}$. An e-value for $\mathcal{H}$ is *testing $\mathcal{H}$ against $\mathcal{H}'$* if $\mathbb{E}_Q E>1$ for all $Q\in\mathcal{H}'$.

Definition 3.2. Let $(\mathcal{F}_t)_{t\in\mathbb{N}}$ be a filtration, h be a positive integer and $\mathcal{H},\mathcal{H}'\subset\mathcal{P}$. Adapted non-negative random variables $(E_t)_{t\in\mathbb{N}}$ are called *sequential e-values for $\mathcal{H}$ at lag h* if $\mathbb{E}_P(E_t\mid\mathcal{F}_{t-h})\le 1$ for all $P\in\mathcal{H}$ and for all $t\in\mathbb{N}$. Sequential e-values for $\mathcal{H}$ at lag h are *testing $\mathcal{H}$ against $\mathcal{H}'$* if $\mathbb{E}_Q(E_t\mid\mathcal{F}_{t-h})>1$ for all $Q\in\mathcal{H}'$ and for all $t\in\mathbb{N}$. For $t\le h$ expectations are understood unconditionally.

For lag $h=1$, sequential e-values can be combined by their cumulative product. For $h>1$ we can combine e-values with a U-statistics approach (see Vovk and Wang (2021) and Proposition 3.4 of Henzi and Ziegel (2021)).

Proposition 3.1. *Let $(E_t)_{t\in\mathbb{N}}$ be sequential e-values for $\mathcal{H}\subset\mathcal{P}$ at lag h adapted to the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Then for all $T\ge h+1$, with $I_k(T)=\{k+hs : s=0,\dots,\lfloor (T-k)/h\rfloor-1\}$,*

$$e_T=\frac{1}{h}\sum_{k=1}^{h}\prod_{l\in I_k(T)}E_l \tag{4}$$

is $\mathcal{F}_T$ measurable and an e-value for $\mathcal{H}$. In particular, e_τ is an e-value for $\mathcal{H}$ for any stopping time τ.

The combination formula (4) in Proposition 3.1 transforms sequential e-values into a supermartingale by averaging cumulative products of all e-values with lag h, which then allows to apply results of Ramdas et al. (2020) to obtain validity under optional stopping. For rejecting the null hypothesis at a fixed level $\alpha \in (0, 1)$, one may apply the aggressive stopping criterion

$$\tau = \inf\{t \in \mathbb{N} : e_t \geq 1/\alpha\}.$$

Example 3.1. Assume that at each time point $t \in \mathbb{N}$ we are assessing a probabilistic forecast F_t with prediction horizon $h \geq 1$ for a quantity of interest. At time t, we are given the current quantity Y_t and the forecast F_t for Y_{t+h}. We are interested in testing the null hypothesis that the forecasts are calibrated. A natural choice for the filtration is $\mathcal{F}_t = \sigma\big(Y_1, F_1, \ldots, Y_t, F_t\big)$. Forecast evaluation is normally based on the observation and on the information available at the time of forecasting. Therefore, an e-value E_t for testing calibration at time $t \geq h+1$ should satisfy $\mathbb{E}(E_t \mid \mathcal{F}_{t-h}) \leq 1$. But $\mathbb{E}(E_t \mid \mathcal{F}_j) \leq 1$ may be violated for $t - h < j < h$ even for calibrated forecasts, since the conditional expectation involves information not available at the time of forecasting. Therefore, e-values $(E_t)_{t\in\mathbb{N}}$ for testing calibration should be sequential at lag h. In this case, combination formula (4) can be applied.

In the following sections we construct sequential e-values for the continuous and for the discrete uniform distribution and for testing stochastic dominance relations with respect to the uniform distribution. General construction principles for e-values, and possible caveats, are explained in the special case of the continuous uniform distribution in Section 3.2, but also apply to the other situations in Sections 3.3 and 3.4. An R package implementing the methods is available on GitHub (`https://github.com/AlexanderHenzi/epit`), and technical details are given in Appendix B.

3.2 Continuous uniform distribution

Let Z be a random variable with values in $[0, 1]$. We are interested in testing whether Z is uniformly distributed, that is, constructing e-values for the hypothesis

$$\mathcal{H}_{\text{CUF}} := \{\text{UNIF}(0, 1)\}. \tag{5}$$

The underlying set of distributions, $\mathcal{P}$, simply consists of all distributions on the interval $[0, 1]$. As a first strategy, we suggest to test $\mathcal{H}_{\text{CUF}}$ against the family of beta distributions which we denote by $\mathcal{H}'$. Any $P \in \mathcal{H}'$ can be parametrized by a vector in the set

$$\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 \mid \alpha > 0,\ \beta > 0\}.$$

Let $P_{(\alpha,\beta)}$ denote the beta distribution with parameters (α, β), so that $\mathcal{H}_{\text{CUF}} = \{P_{(1,1)}\}$. As mentioned in the introduction, the hypothesis $\mathcal{H}_{\text{CUF}}$ is simple, and for any $(\alpha, \beta) \neq (1, 1)$ the density, or likelihood ratio, with respect to UNIF$(0, 1)$,

$$E^{\alpha,\beta}(Z) := \frac{1}{B(\alpha, \beta)} Z^{\alpha-1}(1 - Z)^{\beta-1}$$

is an e-value testing $\mathcal{H}_{\text{CUF}}$ against $\{P_{(\alpha,\beta)}\}$, where $B(\cdot, \cdot)$ denotes the beta function. Grünwald et al. (2019) suggest to determine e-values in such a way that the expected logarithm of the e-value is maximal in the worst case scenario, and refer to e-values with this property as growth rate optimal in worst case (GROW). Following this criterion, parameters $(\alpha^*, \beta^*) \in \Theta$ would have to be found such that

$$\inf_{(\alpha,\beta)\in\Theta} \mathbb{E}_{Z\sim P_{(\alpha,\beta)}}[\log(E^{\alpha^*,\beta^*}(Z))]$$

is maximal. However, this approach is only feasible if either α or β (or their ratio or difference) is fixed, which yields a one-parameter exponential family for which results of Grünwald et al.

(2019) are applicable. In many situations this is a prohibitive limitation, since no sufficient prior knowledge is available to restrict the parameters. On the other hand, if both α and β can take any positive values, the GROW e-value is constant 1, because the infimum in the equation above is negative unless $\alpha^* = \beta^* = 1$.

As a different strategy in sequential settings, we propose to estimate (α, β) by maximum likelihood estimation (MLE) to optimize power for the next e-value, in the spirit of the betting strategies suggested by Waudby-Smith and Ramdas (2020) for estimating a bounded mean. Given a sequence of observations $(z_t)_{t\in\mathbb{N}} \subseteq [0, 1]$, one can successively calculate e-values for $\mathcal{H}_{\mathrm{CUF}}$ as follows: For $t \geq 2$ estimate parameters $(\hat{\alpha}_t, \hat{\beta}_t)$ by MLE, that is

$$(\hat{\alpha}_t, \hat{\beta}_t) = \underset{(\alpha,\beta)\in\Theta}{\arg\max} \sum_{i=1}^{t} \log\left(p_{(\alpha,\beta)}(z_i)\right), \tag{6}$$

where $p_{(\alpha,\beta)}$ denotes the Lebesgue density of $P_{(\alpha,\beta)}$. Set $E_1 = E_2 = 1$ and calculate

$$E_{t+1} = E^{\hat{\alpha}_t,\hat{\beta}_t}(z_{t+1}) \tag{7}$$

to obtain a sequence $(E_t)_{t\in\mathbb{N}}$ of e-values for testing the null hypothesis that the sequence $(z_t)_{t\in\mathbb{N}}$ is i.i.d. $\mathrm{UNIF}(0, 1)$.

To construct e-values at lag h, parameter estimation can be performed separately on all all subsamples with indices $\{k + hs \mid s = 0, 1, \ldots\}$, $k = 1, \ldots, h$. That is, for $t \geq 2h$ calculate

$$(\hat{\alpha}_t^k, \hat{\beta}_t^k) = \underset{(\alpha,\beta)\in\Theta}{\arg\max} \sum_{s:k+hs\leq t} \log\left(p_{(\alpha,\beta)}(z_{k+hs})\right), \quad k = 1, \ldots, h. \tag{8}$$

Set $E_1 = \cdots = E_{2h} = 1$ and, for $t = h, h + 1, \ldots,$

$$E_{k+th} = E^{\hat{\alpha}_t^k,\hat{\beta}_t^k}(z_{k+th}), \quad k = 1, \ldots, h.$$

Then $(E_t)_{t\in\mathbb{N}}$ are sequential e-values at lag h for the null hypothesis that $z_t \sim \mathrm{UNIF}(0, 1)$ conditional on $z_1, \ldots, z_{t-h}$ for all t, and these e-values can be combined with the formula (4).

Remark. Estimating the parameters by maximum likelihood can be considered as a sample version of the GROW criterion, with Z distributed according to the empirical distribution of $z_1, \ldots, z_n$ instead of taking the infimum over all parameters in Θ. The strategy to maximize the expected logarithm of a product is also referred to as Kelly betting, in reference to Kelly Jr (1956).

The beta family of distributions is flexible enough to adapt the most common violations of uniformity which occur in practice, namely increasing, decreasing, unimodal and U-shaped densities. This also covers the typical shapes of the PIT distribution for biased and over- or underdispersed probabilistic forecasts. However, in certain applications or data-rich situations, it may be desirable not to restrict the shape of the e-values to a parametric family. A powerful tool for such cases is kernel density estimation, which allows with a sample $\boldsymbol{\zeta}^k = (\zeta_1, \ldots, \zeta_k) \in [0, 1]^k$ to approximate any density on the unit interval by a mixture

$$E^{K,b,\boldsymbol{\zeta}^k}(Z) = \sum_{i=1}^{k} \frac{1}{b} K\left(\frac{Z - \zeta_i}{b}\right),$$

where K is a suitable kernel density and $b > 0$ the bandwidth. The selection of the bandwidth, and even of the kernel K, can be done in a sequential fashion like the parameter estimation for the beta e-values. For the e-value at time t, the sample $\boldsymbol{\zeta}^{t-1}$ can be taken as $(z_1, \ldots, z_{t-1})$ for lag 1 forecasts, and the e-value $E^{K,b,\boldsymbol{\zeta}^{t-1}}$ is evaluated at the observation z_t. For higher lags, the procedure is separated by subsamples with lag h like for MLE in (8).

Compared to the e-values based on beta distributions, the kernel density approach offers more flexibility, which on the other hand also implies more implementation decisions, especially due to the complicating fact that the domain $[0, 1]$ is a bounded interval. We describe our implementation in Appendix B. Furthermore, while MLE for parameter estimation in the beta e-values is theoretically motivated by maximizing the growth rate in sequential settings, estimation methods for kernel densities are often based on different criteria, such as integrated mean squared error, which do not have a natural interpretation in the context of e-values. However, this does not mean that the beta e-values necessarily have a higher power even if the underlying distribution can be approximated well by a beta distribution. For example, if the goal is to reject the null hypothesis at a level α, then the growth-rate optimal e-value does not always have the maximal power with respect to this particular criterion (see for example the simulations in Henzi and Ziegel, 2021, Section 4.1).

In the practical implementation of e-values, some details should be taken into account. Under the null hypothesis (5), the boundary points 0 and 1 occur with probability zero, but in applications, observations of exactly 0 or 1 appear in most datasets, for example due to rounding. This may be problematic for the construction of the e-values (for example, the estimator (6) diverges if $z_i \in \{0, 1\}$ for some i), and it may lead to e-values equal to zero or infinity. In our implementation, we decided to ignore observations in $\{0, 1\}$ both in the parameter estimation and when computing the e-values; the latter corresponds to setting $E^{\alpha,\beta}(z) = E^{K,b,\boldsymbol{\zeta}}(z) = 1$, for $z \in \{0, 1\}$, which is a valid strategy since it does not change the expectation of the e-value under the null hypothesis. The rationale is that if zeros or ones occur only rarely, then omitting them should not influence the results. On the other hand, if they occur frequently then it is questionable whether a test of the UNIF$(0, 1)$ hypothesis is really necessary in the given problem since the null hypothesis is obviously false.

A second practical issue is that e-values of exactly zero should be prevented since the e-values lose their power once a level of zero is reached. For the beta distributions, zeros can only occur when $Z \in \{0, 1\}$, but the kernel e-values may be zero also inside $(0, 1)$ when there are no data points in some region. A simple correction is to replace the e-values $(E_t)_{t\in\mathbb{N}}$ by convex combinations $(\lambda_t + (1 - \lambda_t)E_t)_{t\in\mathbb{N}}$ for some $\lambda_t > 0$. We set $\lambda_t = 1/t$ in our implementation, since the danger of zero e-values is typically larger for smaller sample sizes, where the sequential parameter estimation is less stable or observations may be sparse in some subsets of $(0, 1)$. When constructing e-values sequentially, one may also set the first n_0 e-values to 1 and start the sequential parameter estimations with a slightly larger sample size, which increases stability. We set $n_0 = 10$ for both the beta and kernel e-values; the minimum n_0 to perform MLE for the beta e-values is $n_0 = 2$.

3.3 Discrete uniform distribution

For $m \geq 1$ the null hypothesis in the discrete case is

$$\mathcal{H}_{\mathrm{DUF}} := \big\{\mathrm{UNIF}(\{1, \dots, m\})\big\},$$

and the underlying set $\mathcal{P}$ consists of all probability distributions on $\{1, \dots, m\}$. Any $P \in \mathcal{P}$ can be parametrized by m weights in the set $\{\boldsymbol{w} \in [0, 1]^m \mid \sum_{i=1}^m w_i = 1\}$, and $\mathcal{H}_{\mathrm{DUF}} = \{P_{\boldsymbol{w}_0}\}$ for $\boldsymbol{w}_0 = (1/m)_{i=1}^m$. Let R be a random variable with values in $\{1, \dots, m\}$. Since $\mathcal{H}_{\mathrm{DUF}}$ is a simple null hypothesis, the likelihood ratio

$$E(R) = \frac{p_1(R)}{p_0(R)} = m\, p_1(R)$$

is an e-value testing $\mathcal{H}_{\mathrm{DUF}}$ against the simple alternative hypothesis $\{P_1\}$, where $P_1 \in \mathcal{P}$ has probability mass function p_1. Like in the continuous case, we suggest a parametric and a nonparametric method for constructing p_1 sequentially.

For parametric e-values, we propose to use the beta-binomial probability mass function

$$p^{\alpha,\beta}(r) = \binom{m-1}{r-1} \frac{B(\alpha - r + 1, \beta + m - r)}{B(\alpha, \beta)}, \quad \alpha, \beta > 0$$

with support in $\{1, \dots, m\}$. This yields e-values with properties similar to the beta e-values, and estimation can again be performed sequentially with the maximum likelihood method. Like the beta distribution on $[0, 1]$, the beta-binomial distribution can approximate increasing, decreasing, unimodal and U-shaped probability mass functions on $\{1, \dots, m\}$.

The most natural nonparametric method for obtaining p_1 is the empirical distribution. That is, for a given sample $r_1, \dots, r_t \in \{1, \dots, m\}$, $p_1(R) = p_{1;t}(R)$ can be set as the empirical frequency of R in the sample up to time t, and at time $t+1$, the frequencies are updated accordingly with the value of r_{t+1}. These empirical frequencies are in fact the maximum likelihood estimator given the available sample and therefore also fit into the GROW approach. A drawback of this procedure is that the e-values may attain zero if one of the frequencies $p_{1;t}(j)$, $j = 1, \dots, m$, is zero. To prevent this, one may start with a particular $\mathcal{P}_1 \in \mathcal{P}$, which serves as a first guess for what the actual frequencies will look like. For example, a neutral first guess is $P_1 = P_{\boldsymbol{w}_0}$, and at time t, the weights could be updated with the formula

$$\boldsymbol{w}_t = \left(\frac{k_1^t + 1}{t + m}, \dots, \frac{k_m^t + 1}{t + m} \right),$$

where $k_j^t = \#\{i = 1, \dots, t \mid r_i = j\}$. Here we successively update with the empirical distribution and each component of the weight vector contains one artificial observation. In comparison with the beta-binomial weights, it has to be expected that for even moderate m (say, 20 or 50, as common in ensemble forecasting), much larger sample sizes are required to recover the actual underlying distribution.

3.4 Stochastic ordering with respect to the uniform distribution

Instead of testing whether $Z \in [0, 1]$ is distributed according to $\mathrm{UNIF}(0, 1)$, one is sometimes only interested in whether it attains systematically lower or higher values than expected under $\mathrm{UNIF}(0, 1)$. This is formalized by the hypotheses

$$\mathcal{H}_{\mathrm{ST}} = \{P \in \mathcal{P}([0, 1]) \mid P \leq_{\mathrm{st}} \mathrm{UNIF}(0, 1)\}, \tag{9}$$

$$\overline{\mathcal{H}}_{\mathrm{ST}} = \{P \in \mathcal{P}([0, 1]) \mid P \geq_{\mathrm{st}} \mathrm{UNIF}(0, 1)\}, \tag{10}$$

where $\mathcal{P}([0, 1])$ is the set of all distributions on $[0, 1]$. The quantile forecasts described in Section 2 give one motivation to test these hypotheses. More generally, for a random variable Y and a strictly increasing CDF G, tests for $\mathcal{H}_{\mathrm{ST}}$ or $\overline{\mathcal{H}}_{\mathrm{ST}}$ applied to $G(Y)$ allow to evaluate if the distribution of Y is stochastically smaller or greater than G. Note that testing whether a random variable $Z \in [0, 1]$ has distribution in $\mathcal{H}_{\mathrm{ST}}$ is equivalent to testing whether the distribution of $1 - Z$ lies in $\overline{\mathcal{H}}_{\mathrm{ST}}$.

The null hypotheses $\mathcal{H}_{\mathrm{ST}}$ and $\overline{\mathcal{H}}_{\mathrm{ST}}$ are composite hypotheses, so the construction of e-values is more involved than for the continuous and discrete uniform distribution. The following result characterizes non-conservative e-values for $\mathcal{H}_{\mathrm{ST}}$ and $\overline{\mathcal{H}}_{\mathrm{ST}}$, with non-conservative meaning that they have expectation 1 under the $\mathrm{UNIF}(0, 1)$ distribution.

Proposition 3.2. *Let f be a Lebesgue density on $[0, 1]$. Then $\mathbb{E}_P(f(Z)) \leq 1$ for all $P \in \mathcal{H}_{\mathrm{ST}}$ ($P \in \overline{\mathcal{H}}_{\mathrm{ST}}$) if and only if there exists an increasing (decreasing) density $\tilde{f}$ and a Lebesgue null set A such that $f(x) = \tilde{f}(x)$ for all $x \notin A$ and $f(x) < \tilde{f}(x)$ for all $x \in A$.*

Remark. Vovk and Wang (2021, Section 2) call random variables $p \in [0, 1]$ which satisfy $P(p \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ *p-variables*, and a decreasing function $f : [0, 1] \mapsto [0, \infty)$ a *p-to-e* *calibrator* if $f(p)$ is an e-value for all p-variables p. In simple words, a p-to-e calibrator is a function which transforms p-values into e-values. This is closely related to the stochastic dominance hypotheses in this section. The set $\overline{\mathcal{H}}_{\mathrm{ST}}$ contains the distributions of all p-variables, and Proposition 3.2 states that decreasing functions are indeed (essentially) the only p-to-e-calibrators. Furthermore, a p-to-e calibrator f is called *admissible* if there exists no other calibrator g such that $g \geq f$ and $g \neq f$, which is equivalent to $f(0) = \infty$ and f being upper semicontinuous and integrating to one (Vovk and Wang, 2021, Proposition 2.1). For testing stochastic dominance, this result in conjunction with Proposition 3.2 implies that all reasonable e-values for $\overline{\mathcal{H}}_{\mathrm{ST}}$ ($\mathcal{H}_{\mathrm{ST}}$) are obtained by choosing left-continuous decreasing (right-continuous increasing) Lebesgue densities f, so the null set A in Proposition 3.2 should be the empty set. While it is valid to set $f(0) = \infty$ for $\overline{\mathcal{H}}_{\mathrm{ST}}$, this may be not desirable in practical applications when values of 0 may occur but should not immediately lead to a rejection of the null hypothesis.

For constructing e-values in a sequential setting, a suitable estimator for decreasing (or increasing) density functions is the Grenander estimator (Grenander, 1956), which is the maximum likelihood estimator among all decreasing density functions and therefore fits into the GROW approach. The Grenander estimator produces piecewise constant density functions, and as a smooth alternative, we propose the estimator by Turnbull and Ghosh (2014) based on mixtures of Bernstein polynomials, that is, beta densities. This estimator was originally proposed for the estimation of unimodal densities, but monotone densities can be easily accommodated by setting the mode to zero or one. Estimation is based on minimizing a squared distance between the ECDF of a sample $z_1, \dots, z_n$ under constraints on the mixture weights to ensure monotonicity. Different from the Grenander estimator, there is a tuning parameter, namely the maximum degree in the Bernstein polynomials, for which Turnbull and Ghosh (2014) propose several selection criteria. Sequential updating of the estimator, the construction of lag h e-values, and potential corrections to avoid e-values of zero can be done as described for the case of the $\mathcal{H}_{\mathrm{CUF}}$ hypothesis in Section 3.2.

The Grenander estimator has the additional advantage that it automatically adapts in the case when it is known (or cannot be excluded) that the distributions of interest have discrete support. To see this for the hypothesis $\mathcal{H}_{\mathrm{ST}}$, assume that the support is $0 = s_1 < \dots < s_k < s_{k+1} = 1$; here $s_1 = 0$ and $s_k < 1$ are necessary conditions for $P \in \mathcal{H}_{\mathrm{ST}}$. If f is an increasing density, then $\mathbb{E}_P(f(Z)) \leq 1$ by Proposition 3.2, but the piecewise constant density $g = g(z; f)$ defined by

$$g(z; f) = \frac{\int_{s_i}^{s_{i+1}} f(z)\,\mathrm{d}z}{s_{i+1} - s_i},\ z \in [s_i, s_{i+1}),\ i = 1, \dots, k-1,\ \ g(z; f) = \frac{\int_{s_i}^{s_{i+1}} f(z)\,\mathrm{d}z}{1 - s_k},\ z \geq s_k, \tag{11}$$

is also increasing and satisfies $g(s_i; f) \geq f(s_i)$, $i = 1, \dots, k$, so $g(\cdot; f)$ yields a more powerful e-value than f. If f is computed with the Grenander estimator and all observations are in $\{s_1, \dots, s_k\}$, then f is already piecewise constant on the intervals $[s_i, s_{i+1})$, and therefore $g(z; f) = f(z)$. The density (11) can also be interpreted as the likelihood ratio between the probabilities $g_i = \int_{s_i}^{s_{i+1}} f(z)\,\mathrm{d}z$ and the discretization of the uniform distribution which puts mass $p_i = s_{i+1} - s_i$ on the points s_i.

Remark. Assume that we are interested in the hypothesis

$$\mathcal{H} = \big\{ \boldsymbol{P} \in \mathcal{P}\big([0,1] \times [0,1]\big) \mid P_1 \leq_{\mathrm{st}} \mathrm{UNIF}(0,1) \leq_{\mathrm{st}} P_2 \big\}, \tag{12}$$

where $\mathcal{P} = \mathcal{P}([0,1] \times [0,1])$ denotes the set of all bivariate distributions on $[0,1] \times [0,1]$ and P_1, P_2 denote the marginal distributions of some $\boldsymbol{P} \in \mathcal{P}$. Then

$$\mathcal{H} = \big\{ \boldsymbol{P} \in \mathcal{P} \mid P_1 \leq_{\mathrm{st}} \mathrm{UNIF}(0,1) \big\} \cap \big\{ \boldsymbol{P} \in \mathcal{P} \mid \mathrm{UNIF}(0,1) \leq_{\mathrm{st}} P_2 \big\} = \mathcal{H}_{\mathrm{ST};1} \cap \overline{\mathcal{H}}_{\mathrm{ST};2}.$$

Since we can write $\mathcal{H}$ as an intersection of two hypotheses it follows immediately that $(E_1 + E_2)/2$ is an e-value for $\mathcal{H}$ if E_1, E_2 are e-values for $\mathcal{H}_{\mathrm{ST};1}, \overline{\mathcal{H}}_{\mathrm{ST};2}$ respectively. E-values for $\mathcal{H}_{\mathrm{ST};1}, \overline{\mathcal{H}}_{\mathrm{ST};2}$

can be constructed with the methods proposed in this section, since the hypotheses only impose restrictions on one of the marginals.

Example 3.2. In this example we show how to use the e-values of Proposition 3.2 and the above remark to check probabilistic calibration of quantile forecasts as defined in Section 2. Assume that for given quantile levels $0 < \alpha_1 < \cdots < \alpha_K < 1$ we sequentially predict quantiles $(q_{1;t}, \ldots, q_{K;t})_{t \in \mathbb{N}}$ at lag 1 and observe the quantities $(y_t)_{t \in \mathbb{N}}$. We calculate the sequence of upper quantile PITs $(z_t)_{t \in \mathbb{N}}$ and lower quantile PITs $(\overline{z}_t)_{t \in \mathbb{N}} \subseteq [0, 1]$, where

$$z_t = Z_{F_{u;t}}(y_{t+1}) \quad \text{and} \quad \overline{z}_t = Z_{F_{\ell;t}}(y_{t+1}).$$

For $t \geq 1$ and upper quantile PIT values $z_1, \ldots, z_t$ we estimate an increasing density f_t. Analogously, we estimate a decreasing density $\overline{f}_t$ with the lower quantile PIT $\overline{z}_1, \ldots, \overline{z}_t$. By Proposition 3.2, $E_{t+1} = f_t(z_{t+1})$ is an e-value for $\mathcal{H}_{\mathrm{ST};1}$ and $\overline{E}_{t+1} = \overline{f}_t(z_{t+1})$ is an e-value for $\overline{\mathcal{H}}_{\mathrm{ST};2}$. Sequential e-values for probabilistic calibration of the quantile forecasts are obtained by $\bar{E}_t = (E_t + \overline{E}_t)/2$, as explained in the above remark. For for $h > 1$, we refer to the usual procedure where we have to estimate densities separately on subsamples with indices $\{k + hs \mid s = 0, 1, \ldots\}$ for $k = 1, \ldots, h$.

4 Simulation study

To evaluate the power of the e-values, we generate independent observations $Y \sim \mathcal{N}(0, 1)$ and define forecasts $F = \mathcal{N}(\varepsilon, 1 + \delta)$, where $\epsilon, \delta \in \{-0.5, -0.4, \ldots, 0.5\}$ are the bias and dispersion error, respectively. Figure 1 illustrates the distribution of the PIT $Z_F(Y) = F(Y)$ for different combinations of bias and dispersion error. For $\delta = \varepsilon = 0$ the PIT is uniformly distributed. To obtain comparable simulations for testing the discrete uniform distribution, we generate 20 independent ensemble forecasts $\boldsymbol{X} = (X_1, \ldots, X_m)$ according to F, and test for uniformity of $\mathrm{rank}_{\boldsymbol{X}}(Y) \in \{1, \ldots, 21\}$. The tests for stochastic order are applied to the PIT $Z_F(Y)$, and we only test if the distribution of $Z_F(Y)$ is stochastically greater than $\mathrm{UNIF}(0, 1)$. For testing calibration of quantile forecasts, we take $K = 19$ equispaced quantiles (levels $0.05, 0.1, \ldots, 0.95$) of the distribution F and compute the e-values as described in Example 3.2. Since both F and the distribution of Y are absolutely continuous, the lower and upper quantile PITs are discrete in this case with values in $\{0.05, 0.1, \ldots, 1\}$ and $\{0, 0.05, \ldots, 0.95\}$, respectively.

We display the result of our simulation experiments with test power hear matrices; see Figure 2 and the additional figures in the Supplementary Material. While this graphical display is self-explanatory, we emphasize that it allows to compare test power across several tests with respect to two directions of alternatives at a single glance. Figure 2 shows the rejection rates of different tests in the simulation examples at a level of $\alpha = 0.05$ with a sample size of $n = 360$. All e-values apply the stopping criterion $\tau = \min(360, \inf\{t \geq 1 : e_t \geq 1/\alpha\})$, and we refer to Appendix B for implementation details. The results for different values of α and n are qualitatively similar and presented in the Supplementary Material. For the continuous uniform distribution, we compare the beta e-values and the kernel e-values to the Kolmogorov-Smirnov test (abbreviated `ks.test` in the following).[1] While the `ks.test` has a higher power against biased forecasts, it is less sensitive to dispersion errors than both e-values. The beta e-values generally achieve a higher power than the e-values based on kernel density estimation, but this difference becomes smaller for larger sample sizes; see the Supplementary Material. For the discrete uniform distribution, we take the chisquare test for comparison. The e-values based on the betabinomial distribution are most sensitive to violations of uniformity, whereas constructing e-values with the empirical

[1] The quantile PIT has a discrete distribution in this simulation study, but the `ks.test` as implemented in R is still applicable since it applies an asymptotic distribution for the test statistic which is sufficiently precise for the sample sizes considered here. We refer the reader to the detailed description and references in the R documentation of `ks.test`.

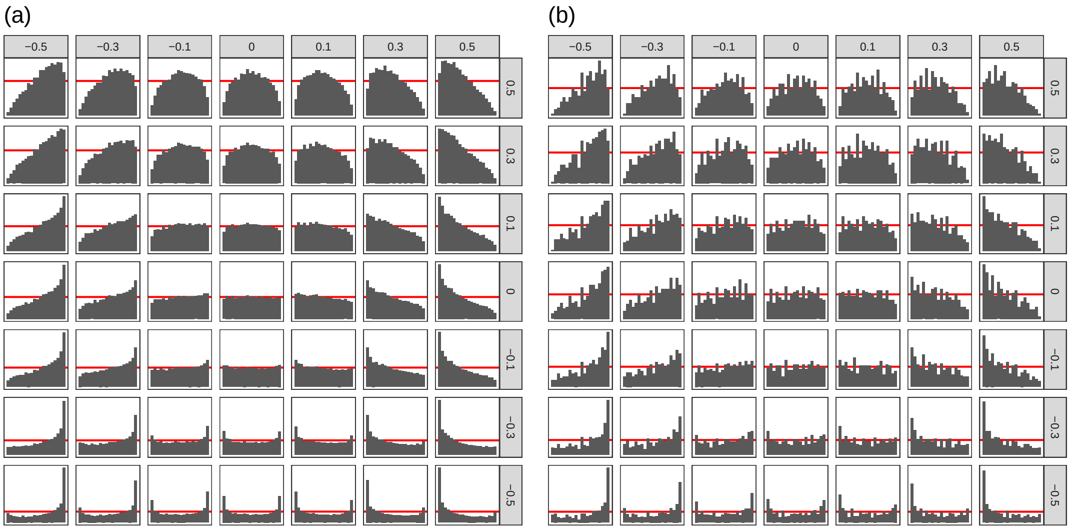


Figure 1: Histograms of the PIT in the simulation study, with 20 equispaced bins and a sample size of (a) $n = 10'000$ (theoretical appearance of the underlying distribution), and (b) $n = 360$ (PIT histogram in a typical simulation). The rows in the figure panels give the dispersion error δ, and the columns give the bias ε. The horizontal red line shows the uniform density. Note the different scaling of the y-axis in the panel rows.

frequencies of the ranks is not powerful for the given simulation, since the empirical distribution only recovers the shape of the underlying distribution very slowly. For testing the null hypothesis that the PIT is stochastically greater than UNIF(0, 1), we apply a one sided version of the `ks.test`, which turns out to be more powerful than the e-values. Nevertheless, the e-values with Bernstein polynomials achieve a similar power when the forecast is underdispersed. For testing calibration of the quantile forecasts, one-sided `ks.tests` are applied to the upper and lower quantile PIT and corrected with the Bonferroni method, so that probabilistic calibration can be rejected if at least one of the corrected p-values is below 0.05. The e-values based on the Grenander estimator are more sensitive to forecast dispersion errors than the `ks.test`, but less sensitive to the bias. The Bernstein e-values achieve a lower power, which is due to the fact that they do not automatically adapt to the discreteness of the quantile PIT.

To summarize, in all simulations the e-values are able to achieve similar power as established methods which do not possess the advantages of e-values, such as validity under optional stopping. For the discrete uniform distribution, we suggest to use the betabinomial e-values unless the sample size is large or the number of distinct values m is small. In stochastic dominance testing with smooth distributions, it is generally better to apply the Bernstein e-values. The Grenander estimator should be preferred for testing calibration of quantile forecasts when both the underlying forecast distribution and the distribution of the outcome are continuous.

5 Case study

5.1 Data and methods

Ensemble prediction systems have tremendously improved the precision of weather forecasts in the past decades (Bauer et al., 2015). However, it is well known that ensemble forecasts remain subject to biases and dispersion errors, which require statistical correction, so called postprocessing, and a variety of methods is available for this task and applied by weather

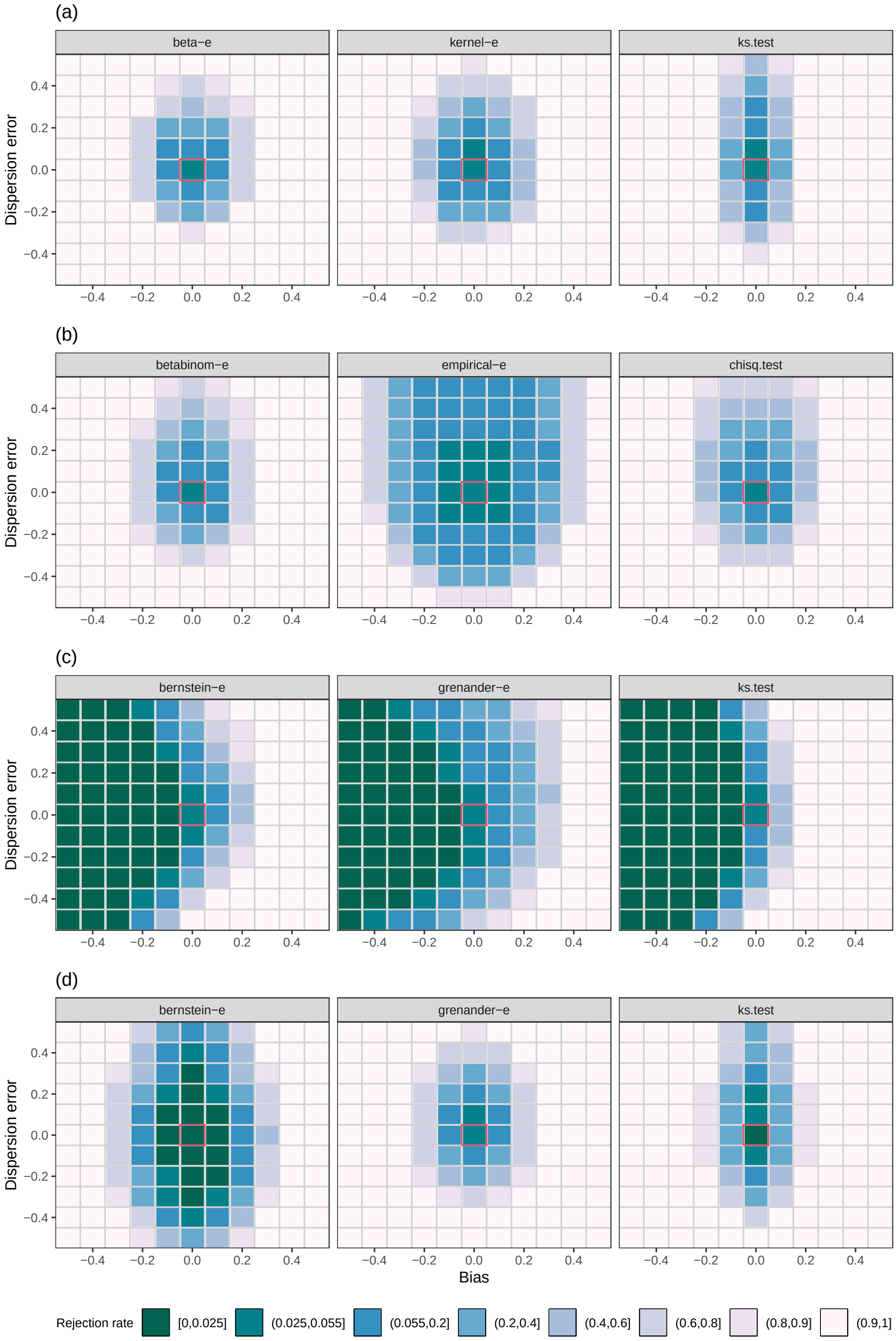


Figure 2: Rejection rates of different tests for (a) the continuous uniform distribution (b) the discrete uniform distribution (c) stochastic dominance (d) calibration of quantile forecasts, at the level $\alpha = 0.05$ with a sample size of $n = 360$, depending on the bias and dispersion error. The red box highlights the rejection rates for bias and dispersion error equal to zero. Rejection rates are computed over 5000 simulations.

Table 1: Meteorological station information (latitude, longitude, World Meteorological Organization (WMO) station identifier, station name).

Latitude	Longitude	WMO ID	Name	Latitude	Longitude	WMO ID	Name
54.18	7.90	10015	Helgoland	51.13	13.75	10488	Dresden-Klotzsche
53.63	9.98	10147	Hamburg-Fuhlsbüttel	50.87	7.17	10513	Köln-Bonn
53.65	11.38	10162	Schwerin	50.98	10.97	10554	Erfurt-Weimar
53.05	8.80	10224	Bremen	49.75	6.67	10609	Trier-Petrisberg
52.47	9.68	10338	Hannover	50.05	8.60	10637	Frankfurt/Main
52.13	11.60	10361	Magdeburg	49.77	9.97	10655	Würzburg
52.38	13.07	10379	Potsdam	49.52	8.55	10729	Mannheim
52.57	13.32	10382	Berlin-Tegel	48.68	9.23	10738	Stuttgart-Echterdingen
51.30	6.77	10400	Düsseldorf	49.50	11.05	10763	Nürnberg
51.50	9.95	10444	Göttingen	49.05	12.10	10776	Regensburg
51.42	12.23	10469	Leipzig/Halle	48.43	10.93	10852	Augsburg

forecasters (Vannitsem et al., 2018). Ensemble postprocessing methods try to estimate the conditional distribution of the variable of interest given the ensemble forecasts. Postprocessed forecasts usually achieve a better calibration than the raw ensemble forecasts, but they may still be miscalibrated if the relationship between forecasts and observations changes over time or if the postprocessing method (say, a parametric model), is not appropriate for the variable at hand. The PIT is one important tool for identifying misspecification of postprocessed forecasts.

In this case study we apply the e-values to test calibration of postprocessed weather forecasts for 22 SYNOP weather stations in Germany. The dataset is part of the data analysed by Hemri et al. (2014) and was kindly provided by Sebastian Lerch. Forecast data are available through the European Centre for Medium-Range Weather Forecasts (ECMWF) Meteorological Archival and Retrieval System (`https://www.ecmwf.int/en/forecasts`) and via TIGGE (Bougeault et al., 2010; Swinbank et al., 2016). Station observations can be downloaded from NOAA's Integrated Surface Database (`https://www.ncdc.noaa.gov/isd`). Station information is given in Table 1. We postprocess the ensemble predictions from the ECMWF, which consists of 50 perturbed forecasts (Molteni et al., 1996; Buizza et al., 2005). The variables considered are 2 meter temperature, wind gust speed, and accumulated precipitation, for lead times of 24, 48, and 72 hours. Data is available from January 1, 2002, to March 20, 2014, and all data until and including the year 2008 is used for training the postprocessing models and the remaining part for validation. The validation dataset consists of 1855 to 1896 days per station, slightly varying due to different numbers of missing values.

Postprocessing is performed separately for each forecast lag and for seasons, namely, the model parameters are estimated on data from the calendar months April to September and October to March for forecasts within the respective periods. The postprocessing for all variables is based on the Ensemble Model Output Statistics (EMOS) approach with heteroscedastic regression: The conditional distribution of the variable of interest is approximated by a parametric location-scale family, with the location parameter being an affine function of the ensemble mean and the scale parameter beging the exponential of an affine transformation of the ensemble standard deviation. For temperature forecasts, the parametric family are Gaussian distributions. Wind gust speed is modelled with the density of a logistic distribution truncated at zero and rescaled so that it integrates to one. Forecasts for accumulated precipitation are based on the censored logistic distribution, where the probability mass on the non-positive numbers gives the probability of zero precipitation. Parameters are estimated by maximum likelihood for the temperature and wind speed forecasts. For precipitation forecasts, parameters are estimated by minimizing the continuous ranked probability score (CRPS) for precipitation, or by maximum likelihood in case the minimization of the CRPS criterion did not converge. The implementation is in R with the `crch` package (Messner et al., 2016).

To evaluate probabilistic calibration we apply the e-values based on kernel density estimation. To make full use of the large sample size, we use the data of the first year in the validation (more precisely, the first 366 days) only for the computation of a reliable first guess of the density of the PIT, and set all e-values for this period to 1. For lag 2 and lag 3 forecasts, this gives sample sizes of 183 or 122, respectively, for each of the lagged sequences of e-values. Apart from this modification, the implementation is as described in Appendix B.2. The e-values based on beta distributions work less well than the kernel densities because the shape of the PIT distribution is often more complicated than just unimodal or U-shaped. We also applied the e-values for the discrete uniform distribution on the raw ensembles, which lead to very fast rejection of the null hypothesis and extremely high e-values (see Table 1 in the Supplementary Material).

5.2 Results

Panels (a) and (b) of Figures 3, 4 and 5 display the PIT histograms and e-values for selected stations, with the common choice of 20 bins for plotting the histograms. For many stations, the PIT histograms indicate severe deviations from uniformity, and the e-values give decisive evidence against the null hypothesis of probabilistic calibration. For higher lags, where e-values cannot be merged by product, the power is generally lower than for lag 1. If the goal is purely to check whether the violation of calibration is significant, then Figure 6 demonstrates that the e-values indeed correlate well with the distance of the PIT from the uniform density.

As argued in the introduction, evaluating probabilistic calibration only at the end of an observation period is often not informative since forecast misspecification changes over time, and this change of forecast misspecification can indeed be seen in the e-values. Consider first the 24 hour temperature forecasts for station 10015, Helgoland (Figure 3). The forecasts are biased, with temperatures often being higher than expected under the forecast distribution. Interestingly, the cumulative product of the e-values displayed in panel (b) of Figure 3 exhibits a clear seasonal pattern: Evidence against calibration is usually gained in the first half of each calender year, but not in the second half. To further investigate this effect, we plot the kernel density estimates of the PIT (with the same method as used for constructing the e-values) separated by time periods. Panel (c) of Figure 3 shows for each half year the density of the PIT based on data until (but not including) the given period. For lag 1 forecasts, this is the e-value $E^{K,b,\boldsymbol{\zeta}^t}$, where $\boldsymbol{\zeta}^t$ are all PIT values before the period and b is the bandwidth estimated with data $\boldsymbol{\zeta}^t$. The second density function is estimated based on data within the given time period. For example, the solid line in the second plot in Figure 3 (c) uses PIT values from 2009 until the end of June 2010, and the dashed density is based the PIT from July until December 2010. If the two densities exhibit similar deviations from uniformity, then evidence against the null hypothesis of calibration is gained, since the observed PIT lies in regions where the e-value is greater than one. It can be seen that the bias of the forecast indeed only occurs in the months January to June, where the e-value increases, but the forecasts are relatively well calibrated from July to December. Improving the postprocessing method should therefore take into account that there is a different seasonal behaviour of the forecasts and observations, which is not captured by performing separate parameter estimation for the months April to September and October to March, and this seasonal behaviour is directly visible in the e-values in panel (b).

A similar observation can be made for the 24 hour wind speed forecasts for station 10162, Schwerin, in Figure 4. The PIT histogram looks close to uniform, and the e-value at the end of the observation period is close to zero and therefore suggests that the forecast is calibrated. However, when looking at the full time domain, there is in fact strong evidence against probabilistic calibration: At the end of the year 2011 the e-value reaches a level of more than 10^7, which corresponds to a highly significant p-value of 10^{-7}. Rejecting calibration based on this observation is statistically valid, because the probability that the process exceeds this level at any time is less or equal to 10^{-7}. The density estimates, in the same spirit as for the previous station, show that the forecasts are in fact biased over the whole observation period, but the

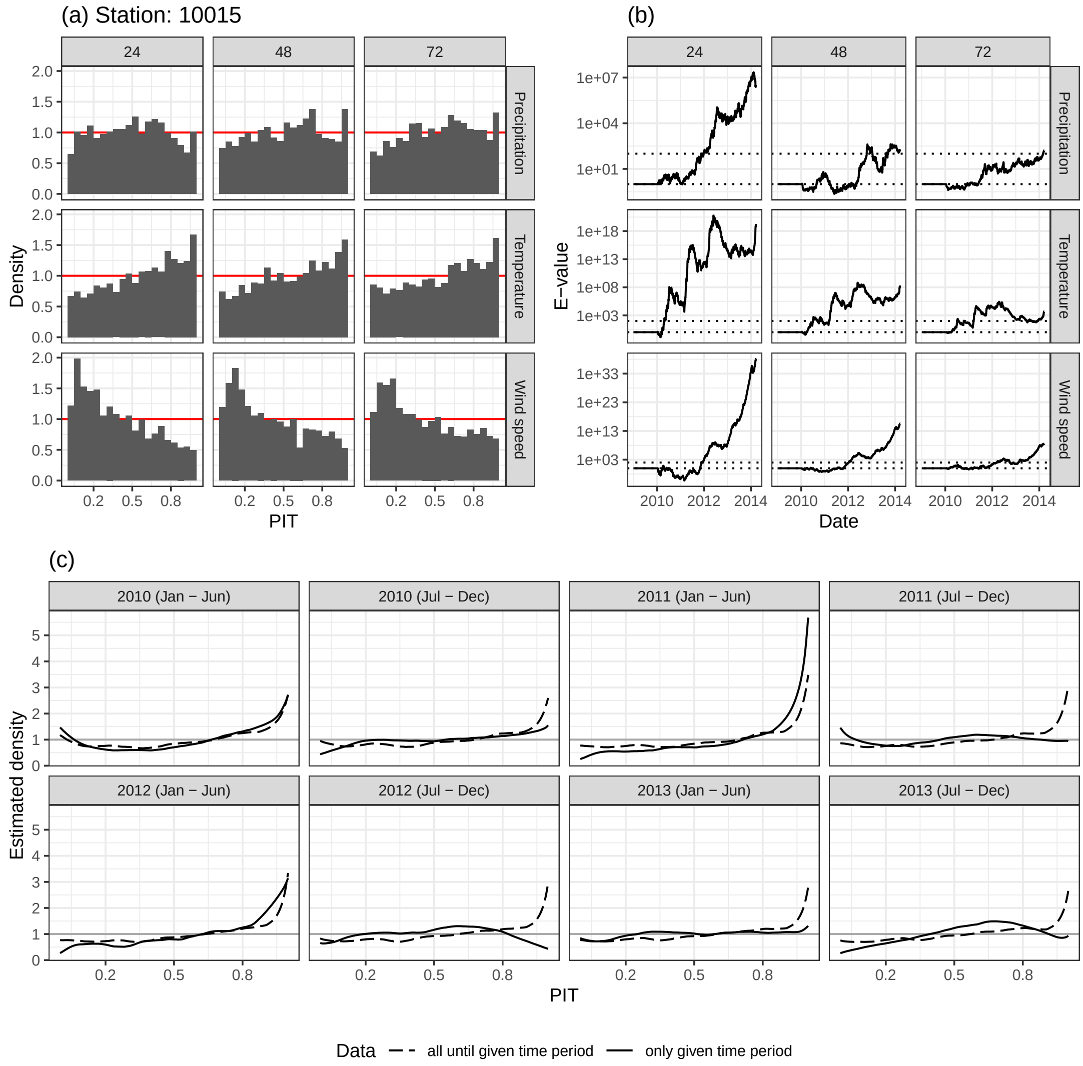


Figure 3: (a) PIT histograms of forecasts for station with ID 10015, for all variables and lead times. (b) E-values (e_t) testing uniformity of the PIT of the given forecasts, where the dotted horizontal show the levels 1 and 100. (c) Density estimates of the PIT for given time periods. The same density estimation method is used as for the computation of the e-values. The dashed density is based on all data until (but not including) the period indicated in the caption, and the solid lines represent the density of the PIT only within the given period.

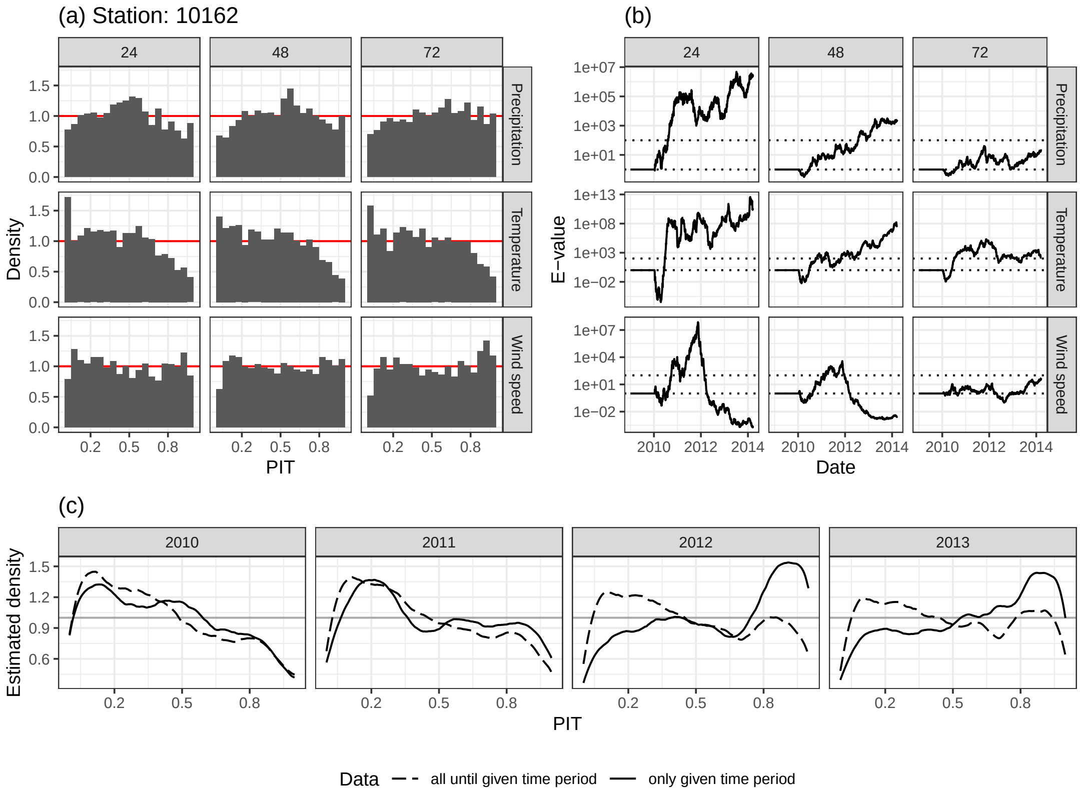


Figure 4: Calibration checks for station 10162. The plots are as described in Figure 3.

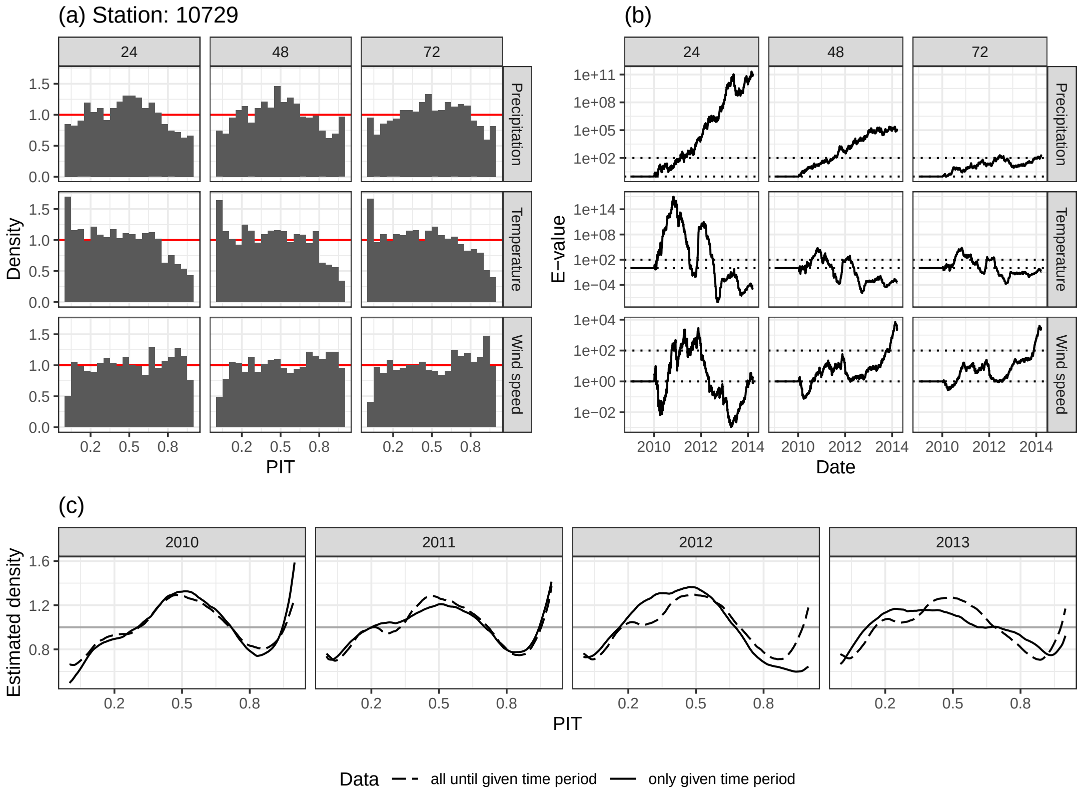


Figure 5: Calibration checks for station 10729. The plots are as described in Figure 3.

direction of the bias changes at the end of 2011. This change in forecast misspecification is clearly visible in the plot of the e-value over time. Hence a decreasing e-value does not necessarily indicate that the forecast is calibrated, but that there is a change in the miscalibration (either to calibration or to a different type of miscalibration).

Finally we consider the 48 hour precipitation forecasts for station 10729, Mannheim (Figure 5). The e-value grows steadily over time and reaches a level of 10^5, indicating that the underdispersion visible in the PIT is indeed significant. The kernel density estimates in panel (c) of Figure 5 confirm that this underdispersion is consistent over the whole time period and not varying, as one could expect from the plot of the e-values.

To summarize, by examining how e-values develop over time, changes in forecast calibration or miscalibration become visible at a glance. Furthermore, e-values make it is possible to detect forecast miscalibration which cannot be seen directly in a PIT histogram based on the complete data, and yield valid p-values for rejecting calibration at any time point without having to stratify the data in advance. A stratified analysis by season or year, as in the panel (c) of the figures in this section, does of course not necessarily require e-values. However, it has been demonstrated that e-values may simplify this process by indicating whether or at what time points forecast misspecification changes.

6 Discussion

Forecasting is an inherently sequential task. Most forecasts exhibit non-stationary errors, for example due to seasonal effects, and forecasters adapt and improve their methods and models over time, which results in systematic changes of forecast performance. For this reason forecast

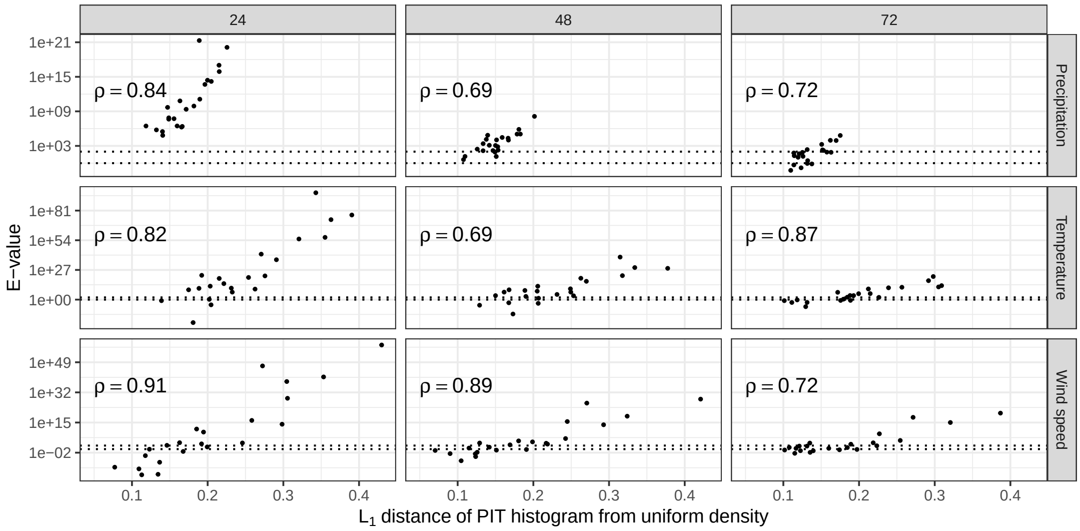


Figure 6: E-values for each station compared to the integrated absolute difference (L_1 distance) between the PIT histogram and the uniform density. The e-values are the ones obtained at the end of the observation period, and ρ gives the Spearman rank correlation between the e-values and the L_1 distance for the given station and lead time. The dotted horizontal lines show the nominal levels of 1 and 100.

evaluation should be sequential as well. Indeed, most practitioners and institutions continuously evaluate the quality of their forecasts; for example, the EMCWF analyses their forecast methods annually in their reports available on `https://www.ecmwf.int/en/publications/annual-reports`. From the theoretical side, there is a lack of methods tailored for sequential forecast evaluation, which do not simply rely on a discretization of the time domain and applying static methods for fixed sample sizes.

E-values, which are arguably the suitable tool for sequential forecast evaluation, have received increasing interest in recent years, but research is still mainly of theoretical nature and has not yet systematically focused on the evaluation of probabilistic forecasts. We have shown how e-values can be applied to obtain sequentially valid tests for probabilistic calibration, which is one of the most important notions of forecast calibration. The e-values which are provided in this paper are also of stand-alone interest and can be applied in other areas of statistics.

Simulation studies are an important tool to understand rejection rates (power) of newly proposed tests across a range of relevant alternatives. Often, if several parameters are varied in the study, readers are overwhelmed by too many numbers in large tables or too many lines in graphs. We suggest to display summaries of rejection rates as test power heat matrices as given in Figure 2. These diagrams allow to see a power comparison of several tests against many alternatives at one glance.

Our paper focuses on probabilistic calibration. A topic for future work is to derive valid tests for sequential forecast evaluation for other notions of calibration like auto-calibration. In contrast to probabilistic calibration, the notion of auto-calibration extends readily also to multivariate forecasts.

Acknowledgements

The authors are grateful to Sebastian Lerch for providing data for the case study and thank Timo Dimitriadis, Tilmann Gneiting and the members of his group for valuable discussions and inputs. This work was supported by the Swiss National Science Foundation.

References

J. L. Anderson. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9:1518–1530, 1996.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, and S. Worley. The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91:1059–1072, 2010.

R. Buizza, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133:1076–1097, 2005.

Y. J. Choe and A. Ramdas. Comparing sequential forecasters. *Preprint, arXiv* arXiv:2110.00115, 2021.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39:863–883, 1998.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268, 2007.

U. Grenander. On the theory of mortality measurement: Part II. *Scandinavian Actuarial Journal*, 2:125–153, 1956.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Preprint,* arXiv:1906.07801, 2019.

T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129:550–560, 2001.

S. Hemri, M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden. Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41(24):9197–9205, 2014.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika, to appear*, 2021.

A. Henzi, A. Moesching, and L. Duembgen. Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Preprint,* arXiv:2006.05527, 2020.

M. C. Jones and P. J. Foster. A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, 6:1005–1013, 1996.

J. L. Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

J. W. Messner, G. J. Mayr, and A. Zeileis. Heteroscedastic censored and truncated regression with crch. *The R Journal*, 8:173–181, 2016.

F. Molteni, R. Buizza, T. N. Palmer, and T. Petroliagis. The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119, 1996.

H. Muller and J. Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50:61–76, 1994.

M. Papadakis, M. Tsagris, M. Dimitriadis, S. Fafalios, I. Tsamardinos, M. Fasiolo, G. Borboudakis, J. Burkardt, C. Zou, K. Lakiotaki, and C. Chatzipantsiou. *Rfast: A Collection of Efficient and Extremely Fast R Functions*, 2020. URL https://CRAN.R-project.org/package=Rfast. R package version 2.0.1.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *Preprint, arXiv arXiv: 2009. 03167*, 2020.

G. Santafe, B. Calvo, A. Perez, and J. A. Lozano. *bde: Bounded Density Estimation*, 2015. URL https://CRAN.R-project.org/package=bde. R package version 1.0.1.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184:407–431, 2021.

B. Stellato, G. Banjac, P. Goulart, and S. Boyd. *osqp: Quadratic Programming Solver using the 'OSQP' Library*, 2019. URL https://CRAN.R-project.org/package=osqp. R package version 0.6.0.3.

R. Swinbank, M. Kyouda, P. Buchanan, L. Froude, T. M. Hamill, T. D. Hewson, J. H. Keller, M. Matsueda, J. Methven, F. Pappenberger, M. Scheuerer, H. A. Titley, L. Wilson, and M. Yamaguchi. The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97:49–67, 2016.

B. C. Turnbull and S. K. Ghosh. Unimodal density estimation using bernstein polynomials. *Computational Statistics & Data Analysis*, 72:13–29, 2014.

S. Vannitsem, D. S. Wilks, and J. Messner, editors. *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, Amsterdam, 2018.

P. Vogel, P. Knippertz, A. H Fink, A. Schlueter, and T. Gneiting. Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, 33:369–388, 2018.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49:1736–1754, 2021.

M. P. Wand and M. C. Jones. *Kernel smoothing*. Chapman & Hall, London, 1995.

M. P. Wand and M. C. Jones. *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2021. URL https://CRAN.R-project.org/package=KernSmooth. R package version 2.23-20.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Preprint, arXiv: 2010. 09686*, 2020.

A Proofs of theoretical results

Proof of Proposition 2.1. For $y < q_1$ and $z > q_K$ the conditions $F_u(y) = 0 \leq F(y)$ and $F(z) \leq 1 = F_\ell(z)$ are always satisfied. If $y \in (q_i, q_{i+1})$ for some $i = 1, \ldots, K-1$, then

$$F_u(y) = \alpha_i \leq F(q_i) \leq F(y) \leq F(q_{i+1}-) \leq \alpha_{i+1} = F_\ell(y).$$

For the second claim consider $j \in \{1, \ldots, K\}$ and define $\mathcal{I}_j = \{i \in \{1, \ldots, K\} \mid q_i = q_j\}$. Then,

$$\begin{aligned} F_u(q_j-) &= \min_{i \in \mathcal{I}_j} \alpha_{i-1} \leq \max_{i \in \mathcal{I}_j} \alpha_i = F_u(q_j), \\ F_\ell(q_j-) &= \min_{i \in \mathcal{I}_j} \alpha_i \leq \max_{i \in \mathcal{I}_j} \alpha_{i+1} = F_\ell(q_j), \end{aligned}$$

which shows equation (3). □

Proof of Proposition 3.2. We show the claim for $\mathcal{H}_{\mathrm{ST}}$. The arguments for $\overline{\mathcal{H}}_{\mathrm{ST}}$ are analogous.
Sufficiency: Assume that there exists an increasing function $\tilde{f}$ and a Lebesgue null set A such that $f(x) = \tilde{f}(x)$ for all $x \in [0,1] \setminus A$ and $\tilde{f}(x) > f(x)$ for $x \in A$. Let $P \in \mathcal{H}_{\mathrm{ST}}$, then $P \leq_{\mathrm{st}} \mathrm{UNIF}(0,1)$ and

$$\mathbb{E}_P(f(Z)) \leq \mathbb{E}_P(\tilde{f}(Z)) \leq \mathbb{E}_{\mathrm{UNIF}(0,1)}(\tilde{f}(Z)) = 1, \tag{13}$$

using that $f(x) \leq \tilde{f}(x)$ for all $x \in [0,1]$, isotonicity of $\tilde{f}$, and the fact that $\mathbb{E}_{F_1}(g(X)) \leq \mathbb{E}_{F_2}(g(X))$ for all increasing functions g if $F_1 \leq_{\mathrm{st}} F_2$.
Necessity: Let f be a density on $[0,1]$ such that there exist no Lebesgue null set A and increasing Lebesgue density $\tilde{f}$ such that $f(x) = \tilde{f}(x)$ for $x \notin A$ and $f(x) < \tilde{f}(x)$ for $x \in A$. We show that then $\mathbb{E}_P(f(Z)) > 1$ for some $P \in \mathcal{H}_{\mathrm{ST}}$.

Case 1: There is an increasing Lebesgue density $\tilde{f}$ such that $f(x) = \tilde{f}(x)$ for all $x \notin A$, where A is a Lebesgue null set. Then there must exist $a \in [0,1]$ such that

$$f(a) > \tilde{f}(a). \tag{14}$$

If (14) only holds for $a = 1$, then $f(x) \leq \tilde{f}(x)$ for $x \neq 1$ and $f(1) > \tilde{f}(1)$. This yields a contradiction, because with the isotonic function $\check{f}$ defined as $\check{f}(x) = \tilde{f}(x)$ for $x < 1$ and $\check{f}(1) = f(1)$, we have that $f(x) = \check{f}(x)$ for all $x \in ([0,1] \setminus A) \cup \{1\}$, and $f(x) < \check{f}(x)$ for $x \in A \setminus \{1\}$. Hence we can assume that (14) holds for some $a < 1$. If $\tilde{f}(x) \geq f(a)$ for all $x > a$ and for all a such that (14) holds, then similar to before, define $\check{f}(x) = \tilde{f}(x)$ for $x \neq a$ and $\check{f}(a) = f(a)$ for all a for which (14) is true. Then $\check{f}$ is again an increasing function almost surely equal to f and satisfies $\check{f} \geq f$, a contradiction (a figure illustrating this special case can be found in the Supplementary Material). Therefore there must exist $a \in [0,1)$ such that $f(a) > \tilde{f}(b)$ for some $b \in (a,1]$. This implies $f(a) > \tilde{f}(y)$ for all $y \in [a,b]$, by monotonicity of $\tilde{f}$. Choose a, b such that this condition holds, and define the CDF G by

$$G(x) = \begin{cases} x, & x \in [0,a), \\ b, & x \in [a,b), \\ x, & x \in [b,1]. \end{cases}$$

Then $G(x) \geq x$ for $x \in [0,1]$, so $G \in \mathcal{H}_{\mathrm{ST}}$, and

$$\begin{aligned} \mathbb{E}_G(f(Z)) &= \int_{[0,a)} f(z)\,\mathrm{d}z + (b-a)f(a) + \int_{[b,1]} f(z)\,\mathrm{d}z \\ &= \int_{[0,a)} \tilde{f}(z)\,\mathrm{d}z + (b-a)f(a) + \int_{[b,1]} \tilde{f}(z)\,\mathrm{d}z \\ &> \int_{[0,a)} \tilde{f}(z)\,\mathrm{d}z + \int_{[a,b)} \tilde{f}(z)\,\mathrm{d}z + \int_{[b,1]} \tilde{f}(z)\,\mathrm{d}z = 1, \end{aligned}$$

using the fact that $f = \tilde{f}$ Lebesgue almost surely.

Case 2: There exists no monotone increasing Lebesgue density $\tilde{f}$ such that $f(x) = \tilde{f}(x)$ for all $x \in [0,1] \setminus A$, where A is a Lebesgue null set. For $x \in [0,1]$, define $F(x) = \int_0^x f(z)\,\mathrm{d}z$. Then F is not convex because otherwise, F would be differentiable almost everywhere and its derivative would be increasing and equal to f for all x not contained in some set of Lebesgue measure zero. This implies that there are points $0 \leq x_1 < x_2 < x_3 \leq 1$ such that

$$\frac{\int_{[x_1,x_2]} f(z)\,\mathrm{d}z}{x_2 - x_1} = \frac{F(x_2) - F(x_1)}{x_2 - x_1} > \frac{F(x_3) - F(x_2)}{x_3 - x_2} = \frac{\int_{[x_2,x_3]} f(z)\,\mathrm{d}z}{x_3 - x_2}. \tag{15}$$

Let $c := (x_3 - x_1)/(x_2 - x_1) = 1 + (x_3 - x_2)/(x_2 - x_1) > 1$ and define

$$G(x) = \begin{cases} x, & x < x_1, \\ x_1 + c(x - x_1), & x \in [x_1, x_2), \\ x_3, & x \in [x_2, x_3), \\ x, & x \in [x_3, 1]. \end{cases}$$

Then $G(x) \geq x$ for $x \in [0,1]$, and by (15),

$$\begin{aligned} \mathbb{E}_G(f(Z)) &= \int_{[0,x_1)} f(z)\,\mathrm{d}z + c \int_{[x_1,x_2)} f(z)\,\mathrm{d}z + \int_{[x_3,1]} f(z)\,\mathrm{d}z \\ &= \int_{[0,x_1)} f(z)\,\mathrm{d}z + \int_{[x_1,x_2)} f(z)\,\mathrm{d}z + \frac{x_3 - x_2}{x_2 - x_1} \int_{[x_1,x_2)} f(z)\,\mathrm{d}z + \int_{[x_3,1]} f(z)\,\mathrm{d}z \\ &> \int_{[0,1]} f(z)\,\mathrm{d}z = 1. \end{aligned}$$

□

B Implementation details

B.1 Beta e-values

The parameters (α, β) in the beta e-values are estimated by maximum likelihood with Newton's method for maximization. The moment matching estimator is taken as a starting point, and the Newton iterations are continued until the likelihood between subsequent iterations does not differ by more than 10^{-6} or until a maximum number of 20 iterations is reached. For stability, the values of (α, β) are truncated to lie in $[0.001, 100]$, and parameter estimation is only started after 10 observations are available (the first 10 e-values are set to 1). The implementation of Newton's method for maximizing the likelihood uses code adapted from the `Rfast` package (Papadakis et al., 2020).

B.2 Kernel e-values

The kernel e-values use the boundary kernel densities as suggested by Muller and Wang (1994). In their original form, these kernel functions may attain negative values, so the non-negativity correction by Jones and Foster (1996) is applied. This estimation method is implemented in the `bde` package in R (Santafe et al., 2015, function `jonesCorrectionMuller94BoundaryKernel`). The resulting density may sometimes not integrate to one. Therefore, it is evaluated on the discrete grid $0, 0.01, \dots, 0.99, 1$ and rescaled so that this discretized version has integral one. To estimate the bandwidth, the direct plug-in approaches as described in Section 3.6 of Wand and Jones (1995) and implemented in the `KernSmooth` package (Wand and Jones, 2021) are applied, with 2 levels of functional estimation for the plug-in rule. In this article, all results with the kernel e-value are based on the boundary corrected Epanechnikov kernel, and only the bandwidth is updated sequentially.

B.3 Betabinomial e-values

For the estimation of the parameters (α, β) in the betabinomial e-values, Newton's method is applied to maximize the likelihood. The moment matching estimators are taken as starting point, and iterations are continued until the sum of the absolute differences between the parameter estimates, $|\alpha_k - \alpha_{k-1}| + |\beta_k - \beta_{k-1}|$, is smaller than 10^{-7} or until a maximum number of 20 iterations is reached. The stopping criterion is different from the estimation of the beta-distribution, since the evaluation of the log-likelihood function for the betabinomial distribution is more costly. The values of α and β are truncated to lie in $[0.001, 100]$. Parameter estimation starts with 20 observations (the first 20 e-values are set to 1), because the smaller number of 10 observations, which is applied in the beta e-values, led to diverging parameter estimates in some simulation examples. The implementation of Newton's method for maximizing the likelihood uses code adapted from the `Rfast` package (Papadakis et al., 2020).

B.4 E-values based on empirical frequencies

The e-values for the discrete uniform distribution based on the empirical frequencies start with a minimum number of 10 observations, all previous e-values are set to 1. For each element of the discrete set, one artificial observation is included at the beginning, so that the frequencies in the t-th step equal $(k_j^t + 1)/(m + t)$, $j = 1, \dots, m$, where $k_j^t = \#\{i = 1, \dots, t \mid r_i = j\}$.

B.5 Grenander e-values

The e-values based on the Grenander estimator start with a minimum number of 10 observations. The Grenander estimator is recomputed with each new observation, applying the abridged pool-adjacent violaters algorithm by Henzi et al. (2020). To avoid e-values of exactly zero, the correction $\tilde{E}_t = 1/t + (1 - 1/t)E_t$ is applied.

B.6 Bernstein e-values

The estimation of monotone densities with mixtures of Bernstein polynomials is based on adapted `R` code by Turnbull and Ghosh (2014). The mixture weights are computed by minimizing the error defined in Equation (5) in Turnbull and Ghosh (2014), subject to constraints on the weights to ensure monotonicity. This leads to a quadratic programming problem, which is solved with `osqp` from the identically named `R` package (Stellato et al., 2019). The `osqp` algorithm is faster and more stable than the quadratic programming solver applied in the original version of the code. The relative and absolute convergence tolerance parameters are set to 10^{-5} and the maximum number of iterations to 4000. A minimum number of 10 observations is required to compute the e-values, and the first 10 e-values are set to 1. The maximal degree of the Bernstein polynomials is fixed at 20 and not estimated. To avoid zero e-values, the correction $\tilde{E}_t = 1/t + (1 - 1/t)E_t$ is applied.

C Supplementary material

C.1 Simulation example

The rejection rates in all figures for the simulation study are computed over 5000 simulations.

Figures S1 and S2 are like Figure 2 in the article but with sample sizes $n = 180$ and $n = 720$. Figures S3 and S4 have sample size $n = 360$ but $\alpha = 0.01$ and $\alpha = 0.1$ instead of $\alpha = 0.05$.

Figure S5 shows the rejection rates of the tests for the discrete uniform distribution for ensembles of size $m = 10$, 20, 50, with $n = 360$ and $\alpha = 0.05$.

Table S1: Number of observations (after the first $n_0 = 366$ days where e-values are not computed) until the e-value with the given method (empirical distribution and betabinomial distribution) exceeds the level 10^8 for the first time, for each weather station, variable, and leadtime. Note that due to the fact that both methods have very similar power, the values often coincide for the two different methods.

Station ID	Variable	Empirical distr.			Betabinomial			Station ID	Variable	Empirical distr.			Betabinomial		
	Leadtime	24	48	72	24	48	72		Leadtime	24	48	72	24	48	72
10015	Precipitation	19	39	86	19	39	86	10488	Precipitation	17	73	340	17	73	340
	Temperature	37	99	141	37	99	141		Temperature	43	137	688	27	137	688
	Wind speed	24	416	1428	27	416	1428		Wind speed	34	153	309	34	153	309
10147	Precipitation	21	100	328	21	100	328	10513	Precipitation	19	71	152	17	71	152
	Temperature	22	205	550	22	205	550		Temperature	13	109	382	13	109	382
	Wind speed	71	307	1096	69	307	1096		Wind speed	17	57	96	17	57	96
10162	Precipitation	19	128	141	19	128	141	10554	Precipitation	21	37	140	21	37	140
	Temperature	29	157	442	31	157	442		Temperature	21	45	151	21	45	151
	Wind speed	21	100	271	21	100	271		Wind speed	18	137	597	18	137	597
10224	Precipitation	31	123	233	31	123	233	10609	Precipitation	19	49	317	19	49	317
	Temperature	36	165	538	36	165	538		Temperature	30	109	243	30	109	243
	Wind speed	47	442	1258	47	442	1258		Wind speed	31	96	150	29	96	150
10338	Precipitation	23	66	144	23	66	144	10637	Precipitation	18	92	172	19	92	172
	Temperature	18	137	460	18	137	460		Temperature	17	76	388	17	76	388
	Wind speed	27	98	621	26	98	621		Wind speed	69	196	485	60	196	485
10361	Precipitation	14	55	244	13	55	244	10655	Precipitation	11	37	320	11	37	320
	Temperature	15	169	679	15	169	679		Temperature	15	46	420	14	46	420
	Wind speed	27	76	99	26	76	99		Wind speed	44	220	629	48	220	629
10379	Precipitation	14	105	333	14	105	333	10729	Precipitation	25	43	316	21	43	316
	Temperature	20	111	418	19	111	418		Temperature	17	73	174	17	73	174
	Wind speed	24	286	687	22	286	687		Wind speed	21	140	409	19	140	409
10382	Precipitation	10	42	111	10	42	111	10738	Precipitation	26	77	129	24	77	129
	Temperature	16	165	693	16	165	693		Temperature	34	75	675	33	75	675
	Wind speed	22	219	366	21	219	366		Wind speed	26	139	143	18	139	143
10400	Precipitation	19	72	76	18	72	76	10763	Precipitation	12	59	105	12	59	105
	Temperature	14	160	498	16	160	498		Temperature	22	86	398	22	86	398
	Wind speed	25	140	690	25	140	690		Wind speed	37	104	264	27	104	264
10444	Precipitation	33	79	185	33	79	185	10776	Precipitation	21	93	472	21	93	472
	Temperature	38	302	476	38	302	476		Temperature	21	48	393	20	48	393
	Wind speed	42	149	429	42	149	429		Wind speed	43	153	287	44	153	287
10469	Precipitation	20	71	340	19	71	340	10852	Precipitation	21	33	105	21	33	105
	Temperature	18	44	73	17	44	73		Temperature	19	45	67	19	45	67
	Wind speed	36	190	707	32	190	707		Wind speed	47	392	1531	45	392	1531

Figure S6 show the rejection rates of the tests for calibration of quantile forecasts for $K = 9$, 19 equispaced quantiles, with $n = 360$ and $\alpha = 0.05$.

C.2 Case study

The e-values for testing the discrete uniform distributions have been applied to the rank histograms of the raw ECMWF ensemble forecasts. Both methods (e-values based on the empirical distribution and on the betabinomial distribution) clearly reject uniformity with only a few observations, since the raw ensemble forecasts have strong biases and dispersion errors when evaluated against station observations. Like for the PIT, the first $n_0 = 366$ e-values are set to 1 and the corresponding ranks are used to estimate the rank histogram with the empirical distribution or the betabinomial distribution. Table S1 shows how many observations after $n_0 = 366$ are required until the sequential e-values first cross the level 10^8.

C.3 Illustration for the proof of Proposition 3.2

Figure S7 illustrates the special case in the proof of Proposition 3.2 where f is almost surely equal to a Lebesgue density $\tilde{f}$ and $f(a) > \tilde{f}(a)$ for some $a \in [0, 1)$ but $f(x) \leq \tilde{f}(x)$ for all $x > a$. In this case, the function $\check{f}$ defined by $\check{f}(a) = f(a)$ for all such a and $\check{f}(x) = \tilde{f}(x)$ for all other x is increasing, almost surely equal to f, and $\check{f}(x) \geq f(x)$ for all $x \in [0, 1]$.

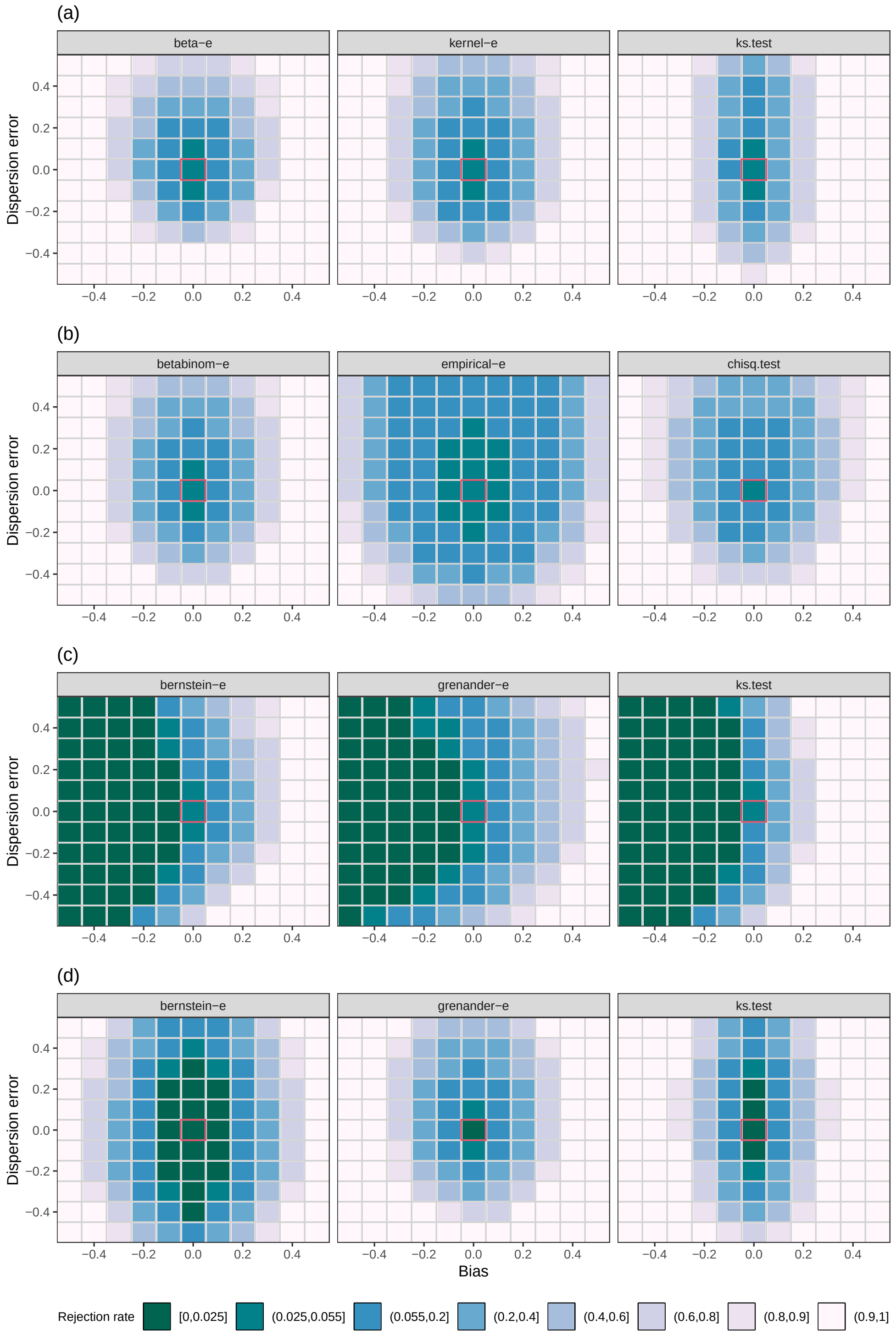


Figure S1: Rejection rates of different tests for (a) the continuous uniform distribution (b) the discrete uniform distribution (c) stochastic dominance (d) calibration of quantile forecasts, at the level $\alpha = 0.05$ with a sample size of $n = 180$, depending on the bias and dispersion error. The red box highlights the rejection rates for bias and dispersion error equal to zero.

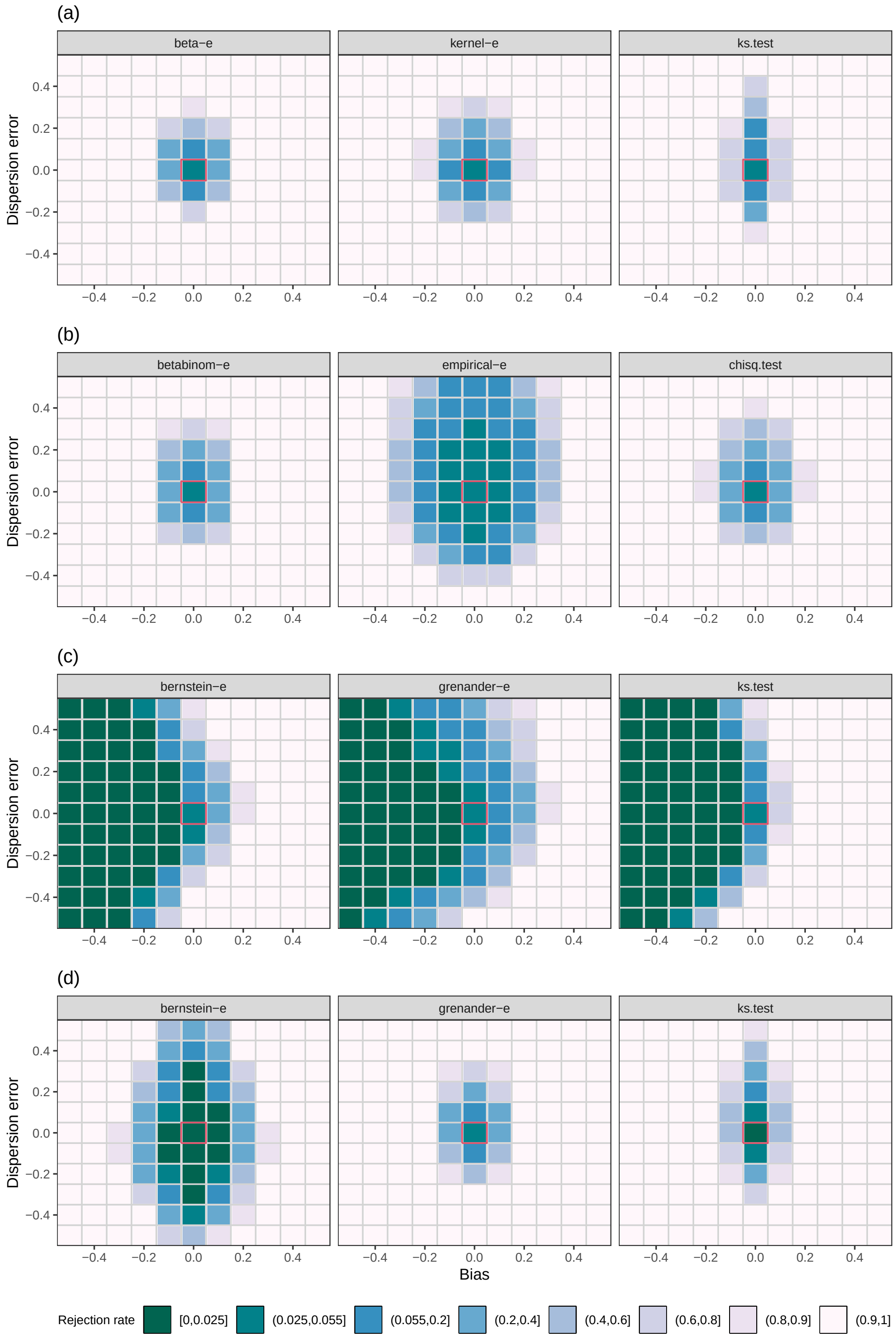


Figure S2: Rejection rates of different tests for (a) the continuous uniform distribution (b) the discrete uniform distribution (c) stochastic dominance (d) calibration of quantile forecasts, at the level $\alpha = 0.05$ with a sample size of $n = 720$, depending on the bias and dispersion error. The red box highlights the rejection rates for bias and dispersion error equal to zero.

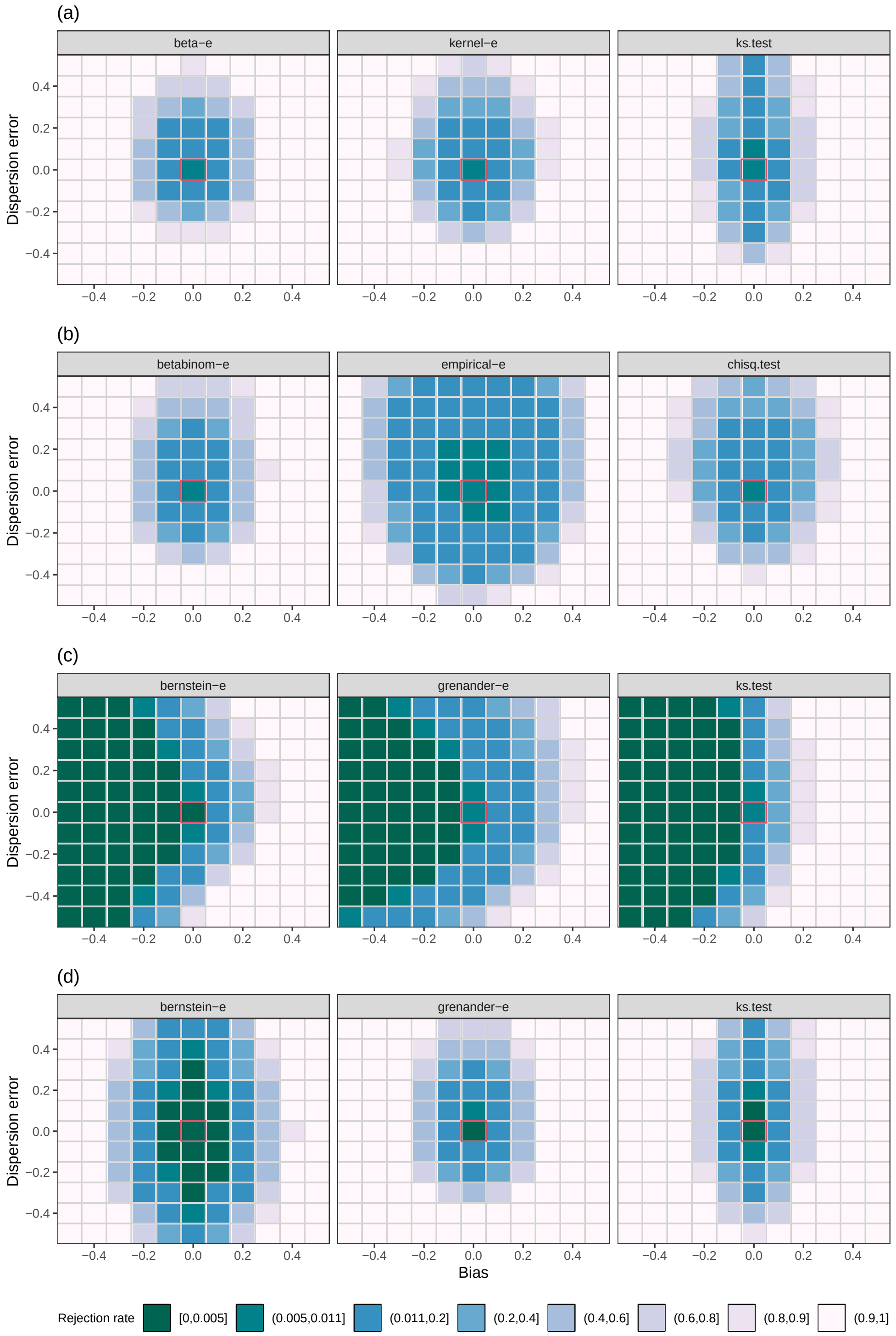


Figure S3: Rejection rates of different tests for (a) the continuous uniform distribution (b) the discrete uniform distribution (c) stochastic dominance (d) calibration of quantile forecasts, at the level $\alpha = 0.01$ with a sample size of $n = 360$, depending on the bias and dispersion error. The red box highlights the rejection rates for bias and dispersion error equal to zero.

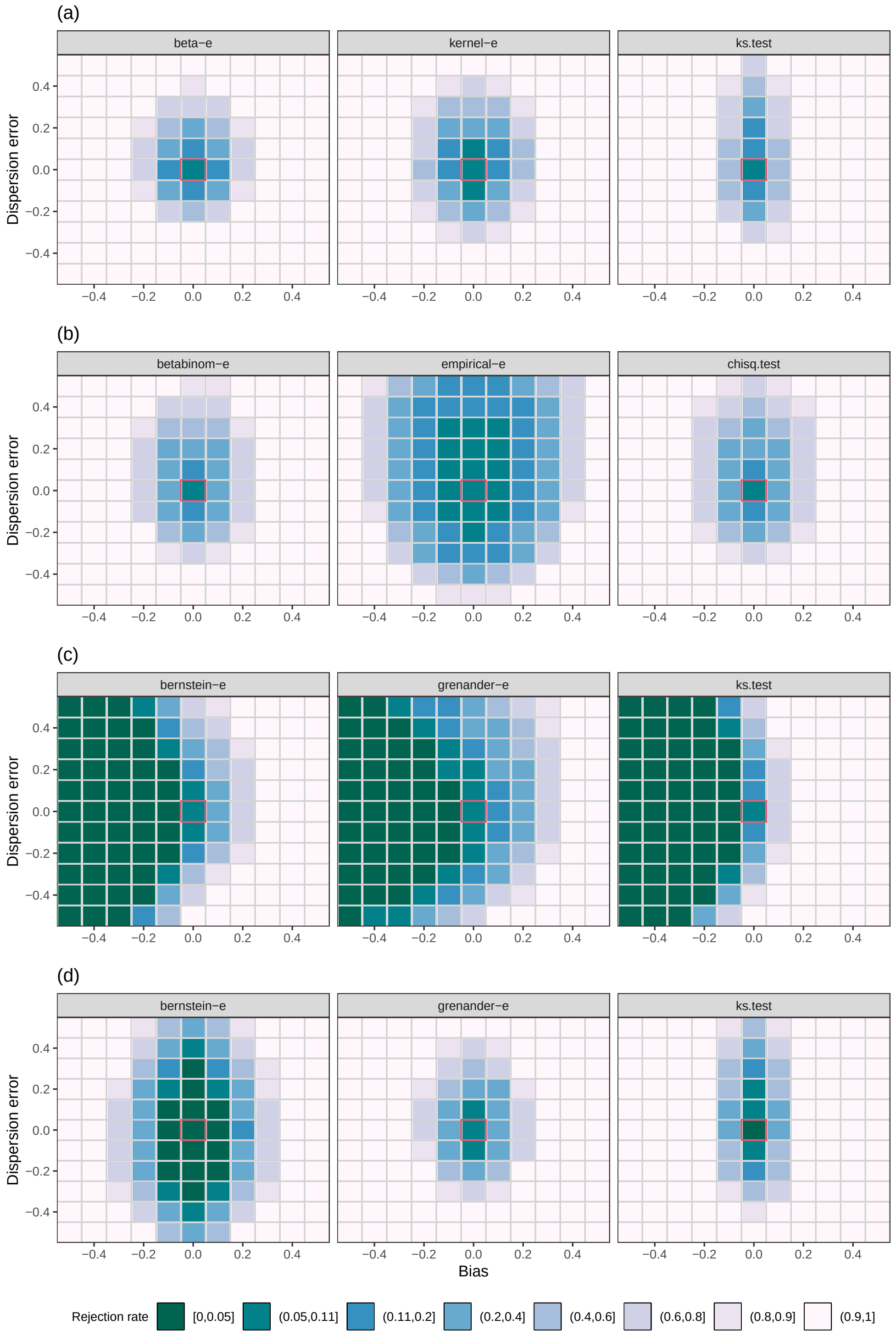


Figure S4: Rejection rates of different tests for (a) the continuous uniform distribution (b) the discrete uniform distribution (c) stochastic dominance (d) calibration of quantile forecasts, at the level $\alpha = 0.1$ with a sample size of $n = 360$, depending on the bias and dispersion error. The red box highlights the rejection rates for bias and dispersion error equal to zero.

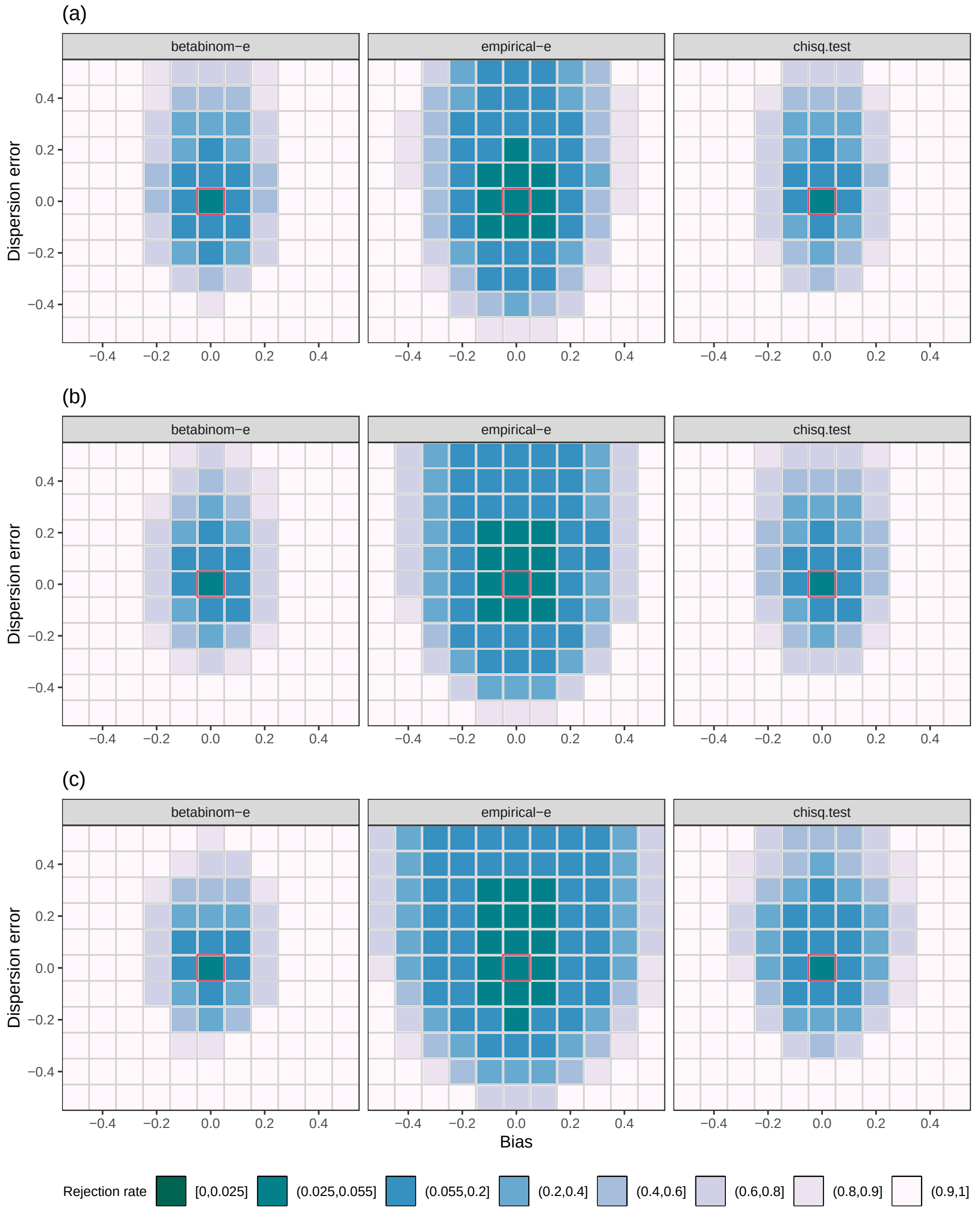


Figure S5: Rejection rates of different tests for uniformity of the rank histogram with ensemble sizes of (a) $m = 10$ (b) $m = 20$ (c) $m = 50$, with $n = 360$ and at the level $\alpha = 0.05$.

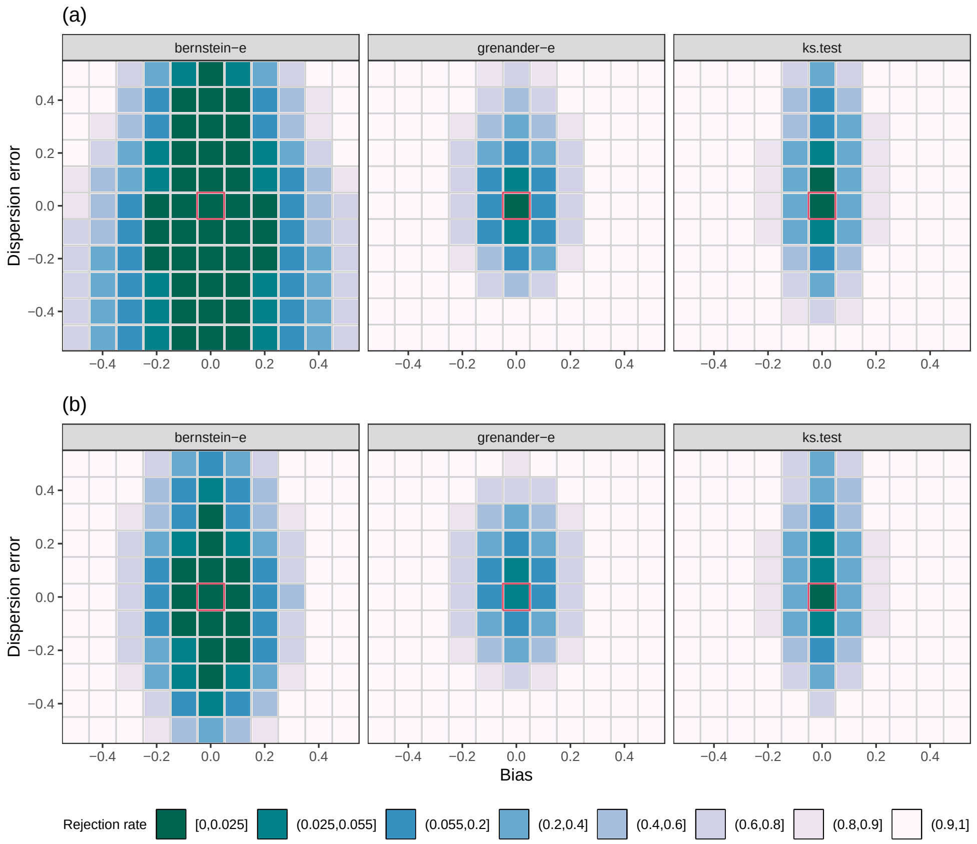


Figure S6: Rejection rates of different tests for calibration of quantile forecasts, with (a) $K = 9$ and (b) $K = 19$ equispaced quantiles, $n = 360$, and $\alpha = 0.05$.

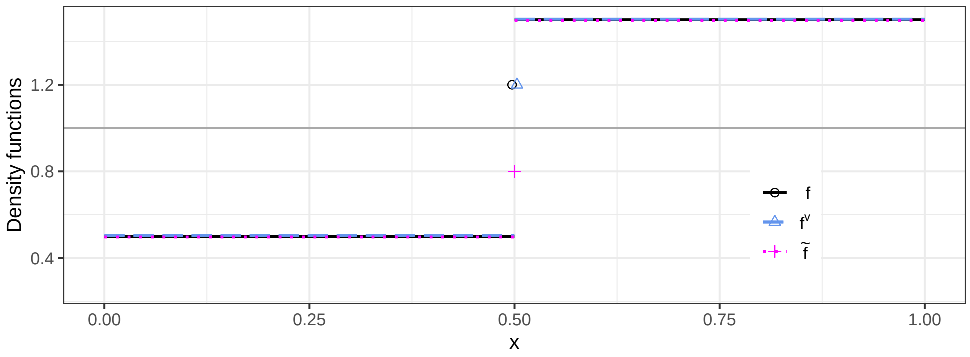


Figure S7: Illustration of the functions f, $\tilde{f}$, and $\check{f}$ in the special case in the proof of Proposition 3.2. Here $f(x) = \tilde{f}(x) = \check{f}(x) = 0.5$ for $x < 0.5$ and $f(x) = \tilde{f}(x) = \check{f}(x) = 1.5$ for $x > 0.5$. At $a = 0.5$ we have $\tilde{f}(a) = 0.8 < 1.2 = f(a)$, and we set $\check{f}(a) = f(a) = 1.2$.

4.3 A safe Hosmer-Lemeshow test

The content of this section is published as an arXiv preprint,

Dimitriadis, T., Henzi, A., Puke, M. and Ziegel, J. (2022). A safe Hosmer-Lemeshow test. *arXiv preprint arXiv:2203.00426.*

A safe Hosmer-Lemeshow test

Timo Dimitriadis[1], Alexander Henzi[2], Marius Puke[3], and Johanna Ziegel[2]

[1]Alfred Weber Institute of Economics, Heidelberg University, Germany,
timo.dimitriadis@awi.uni-heidelberg.de

[2]Institute of Economics, University of Hohenheim, Germany,
marius.puke@uni-hohenheim.de

[2]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland,
alexander.henzi@stat.unibe.ch, johanna.ziegel@stat.unibe.ch

March 4, 2022

Abstract

This technical report proposes an alternative to the Hosmer-Lemeshow (HL) test for evaluating the calibration of probability forecasts for binary events. The approach is based on e-values, a new tool for hypothesis testing. An e-value is a random variable with expected value less or equal to 1 under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a p-value. In a simulation study, the proposed e-values detect even slight miscalibration for larger sample sizes, but with a reduced power compared to the original HL test.

1 Introduction

Suppose that we have a sample of observations $(p_i, y_i)_{i=1}^n$ from $(P_i, Y_i)_{i=1}^n$ such that (P_i, Y_i) has the same distribution as $(P, Y) \in [0,1] \times \{0,1\}$, $i = 1, \dots, n$. The interpretation is that P_i is a prediction for the probability that $Y_i = 1$. The random variables are defined on some underlying probability space $(\Omega, \mathcal{F})$ and denotes $\mathcal{P}$ all probability measures on $(\Omega, \mathcal{F})$. Hosmer and Lemeshow (1980) propose a test for the null of perfect calibration

$$\mathcal{H}_{\mathrm{HL},n} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y_i|P_i) = P_i \ \mathbb{P}\text{-almost surely}, \ i = 1, \dots, n\}.$$

The Hosmer-Lemeshow (henceforth HL) test is based on partitioning the interval $[0,1]$ in $g \in \mathbb{N}$ bins and counting the observed numbers of events, o_{1g}, and no event occurrences, o_{0g}, in each bin. Based on that binning and counting procedure, the HL test statistic to test for perfect calibration of the probability predictions is

$$\widehat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - \widehat{e}_{1k})^2}{\widehat{e}_{1k}} + \frac{(o_{0k} - \widehat{e}_{0k})^2}{\widehat{e}_{0k}} \right],$$

where $\hat{e}_{1k}$ and $\hat{e}_{0k}$ are the expected event and no event occurrences in bin k, respectively (Hosmer et al., 2013). Asymptotically, $\widehat{C} \sim \chi^2_{g-2}$ for $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}$. Technically, the choice of the binning procedure is up the user of the HL test and conventionally implemented via quantile based binning strategies with $g = 10$ which results in equally populated bins (decile-of-risk). Less commonly, the test is based on equally spaced bins, where the interval $[0,1]$ is divided into g equidistant bins.

While the choice of g obviously influences the test statistic, there are other known issues with the HL test based on quantile binning. Using data on birth weight and maternal behaviors, Hosmer et al. (1997) show that six major statistical software packages resulted in six different p-values ranging from 0.02 to 0.16. Bertolini et al. (2000) find that in mortality data from 1393 intensive care patients in Italy, the standard implementation of the HL test is extremely unstable upon sheer reordering of the same data set (that has ties in the values p_i). The authors observe p-values between 0.01 and 0.95 across all possible rearrangements. Those crude examples imply that researchers can tailor any desired test decision to their will and casts doubt on the test's trustworthiness; see Kuss (2002) and the references therein for a summary of disconcerting and paradoxical results regarding the HL test. In light of the reproducibility crises and also under the consideration of the disadvantages outlined above, it seems surprising that the HL test remains the literature's favorite for checking the calibration of binary prediction models and is still commonly used in current medical and epidemiological studies; see amongst many others Neblett Fanfair et al. (2012); Ostrosky-Zeichner et al. (2017); Lee et al. (2020).

We suggest a new Hosmer-Lemeshow test using e-values, henceforth called eHL test. E-values, where 'e' abbreviates the word 'expectation', were proposed recently as an alternative to p-values in testing problems. In a nutshell, an e-value is a realization of a non-negative random variable whose expected value is at most one under a given null hypothesis. This already signals that an e-value itself allows for meaningful interpretations since an e-value greater than one provides evidence against the null hypothesis. Additionally, an e-value can be transformed to a conservative p-value by Markov's inequality. From a game-theoretic perspective, the e-value has a simple financial meaning in the sense that the e-value can be seen as the factor by which a skeptic multiplies her money when betting against the null hypothesis; see Shafer and Vovk (2019); Shafer (2021). An important advantage of e-values over p-values emerges in sequential testing exercises, where e-values convince by their uncomplicated behaviors in combinations: the arithmetic average of e-values also is an e-value, likewise the product of independent or successive e-values; see Shafer (2021); Grünwald et al. (2020); Wang and Ramdas (2020). In practice, this appeals because more evidence can be added later, i.e. evidence across studies can easily be combined. As a result, e-values are valid under optional stopping and continuation, which is not generally true for p-values. That is, the process of collecting data for obtaining an e-value might be stopped or continued based on examining past realizations and e-values; see Henzi and Ziegel (2021+) who exploit that property in forecast dominance tests. However, optional stopping cannot be implemented sensibly for goodness-of-fit tests under the null of perfect calibration since researchers are usually not interested in rejecting it. Hence, the proposed eHL test offers a safe alternative to a fragile state-of-the-art approach by avoiding ad-hoc choices and software instabilities.

The remainder of the report is structured as follows. Section 2 introduces the proposed e-test. Section 3 assess the empirical performance of the test in a simulation study. Section 4 concludes. Replication material for all results is available on GitHub (`https://github.com/marius-cp/eHL`)

2 Construction of HL e-values

2.1 Preliminaries

An e-variable for $\mathcal{H}_{\mathrm{HL},n}$ is a non-negative random variable E (that is allowed to take the value $+\infty$) such that $\mathbb{E}_{\mathbb{P}}(E) \leq 1$ for all $\mathbb{P} \in \mathcal{P}$. An e-value is a realization of an e-variable. An e-variable E always yields a valid p-variable $1/E$ (a p-value is a realized p-variable) by Markov's inequality, since

$$\mathbb{P}\Big(\frac{1}{E} \leq \alpha\Big) = \mathbb{P}\Big(E \geq \frac{1}{\alpha}\Big) \leq \alpha \mathbb{E}_{\mathbb{P}}(E) \leq \alpha, \quad \text{for all } \mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}. \tag{1}$$

We will reject the null hypothesis $\mathcal{H}_{\mathrm{HL},n}$ if we observe a large value of E. If we want to ensure a classical p-guarantee then we have to determine the rejection region for a given α by (1). Vovk and Wang (2021) show that this is essentially the only way to transform an e-variable into a p-variable.

We say that an e-variable has the alternative hypothesis $\mathcal{H}' \subset \mathcal{P}$ if $\mathbb{E}_{\mathbb{Q}}(E) > 1$ for all $\mathbb{Q} \in \mathcal{H}'$.

2.2 Sample size one

We will first construct e-values for the sample size one Hosmer-Lemeshow null hypothesis

$$\mathcal{H}_{\mathrm{HL},1} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y|P) = P\}.$$

Grünwald et al. (2020); Shafer (2021) show that e-variables are (essentially) likelihood ratios. To see this in the special case here, assume that $q \in [0,1]$ and $\mathbb{P}(P \in \{0,1\}) = 0$. Then, an e-variable for $\mathcal{H}_{\mathrm{HL},1}$ is given by

$$E_q(P,Y) = \frac{q^Y(1-q)^{1-Y}}{P^Y(1-P)^{1-Y}} = \begin{cases} q/P, & \text{if } Y = 1, \\ (1-q)/(1-P), & \text{if } Y = 0. \end{cases}$$

The variable $E_q(P,Y)$ is clearly non-negative, and for $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},1}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}(E_q(P,Y)) &= \mathbb{E}_{\mathbb{P}}\left(\mathbb{E}_{\mathbb{P}}(Y \mid P)\frac{q}{P} + \mathbb{E}_{\mathbb{P}}(1-Y \mid P)\frac{1-q}{1-P}\right) \\ &= \mathbb{E}_{\mathbb{P}}\left(P\frac{q}{P} + (1-P)\frac{1-q}{1-P}\right) = 1. \end{aligned}$$

To find alternative hypotheses for the e-variable E_q, let $\pi = \mathbb{E}_{\mathbb{Q}}(Y \mid P)$. Then,

$$\mathbb{E}_{\mathbb{Q}}(E_q(P,Y) \mid P) = \pi\frac{q}{P} + (1-\pi)\frac{1-q}{1-P} > 1$$

holds if and only if, $\pi > p$ and $q > P$, or, $\pi < P$ and $q < P$. This shows that if $q < P$, E_q has the alternative

$$\mathcal{H}' = \{\mathbb{Q} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{Q}}(Y \mid P) < P\}, \tag{2}$$

and if $q > P$, E_q has the alternative

$$\mathcal{H}' = \{\mathbb{Q} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{Q}}(Y \mid P) > P\}. \tag{3}$$

It is possible to show that basically any e-variable for $\mathcal{H}_{\mathrm{HL},1}$ is of the form $E = E_q(P,Y)$ for some q (depending P) but this requires some more arguments; it follows by the construction in Henzi and Ziegel (2021+), see also Waudby-Smith and Ramdas (2021). The connection of $E_q(P,Y)$ to the e-variables in Henzi and Ziegel (2021+) of type $E = 1 + \lambda D$ with $D \geq -1$ such that $\mathbb{E}_{\mathbb{P}}(D) = 0$ for $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},1}$, follows from the fact that λ in this representation can be bijectively mapped to q. In this context,

$$E = 1 + \lambda(P - Y) \tag{4}$$

is an e-variable for $\mathcal{H}_{\mathrm{HL},1}$ for any λ that is $\sigma(P)$-measurable with $-(1/P) \leq \lambda \leq 1/(1-P)$. If $P = 1$, there is no restriction on λ from above, and analogously if $P = 0$, there is no restriction from below. By choosing $\lambda = (P-q)/(P(1-P))$, we obtain that $E = E_q(P,Y)$.

2.3 Combining e-values in the iid case

We assume now that $(P_i, Y_i)_{i=1}^n$ are independent and identically distributed (iid). For testing $\mathcal{H}_{\mathrm{HL},n}$, we suggest the e-variable

$$E^{\mathrm{id}}_{\mathrm{HL},n} = \prod_{i=1}^{n} E_{q_i}(P_i, Y_i), \tag{5}$$

where q_i is $\sigma(P_1, \dots, P_i, Y_1, \dots, Y_{i-1})$-measurable. For $\mathbb{P} \in \mathcal{H}_{\mathrm{HL},n}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}} E^{\mathrm{id}}_{\mathrm{HL},n} &= \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n} E_{q_i}(P_i, Y_i) | P_1, \dots, P_n, Y_1, \dots, Y_{n-1}\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i) \mathbb{E}_{\mathbb{P}}\Big(E_{q_n}(P_n, Y_n) | P_1, \dots, P_n, Y_1, \dots, Y_{n-1}\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big(1 + \frac{P_n - q_n}{P_n(1-P_n)} \mathbb{E}_{\mathbb{P}}(P_n - Y_n | P_1, \dots, P_n, Y_1, \dots, Y_{n-1})\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big(1 + \frac{P_n - q_n}{P_n(1-P_n)} \mathbb{E}_{\mathbb{P}}(P_n - Y_n | P_n)\Big)\Big) \\
&= \mathbb{E}_{\mathbb{P}}\Big(\prod_{i=1}^{n-1} E_{q_i}(P_i, Y_i)\Big) = \mathbb{E}_{\mathbb{P}} E^{\mathrm{id}}_{\mathrm{HL},n-1} = \dots = 1,
\end{aligned}$$

where we used the equivalent representation of $E_q(P, Y)$ in (4). In particular, from the above derivation it is easy to see that $(E^{\mathrm{id}}_{\mathrm{HL},n})_{n\in\mathbb{N}}$ is a test martingale.

The e-variable $E^{\mathrm{id}}_{\mathrm{HL},n}$ depends on the ordering of $(P_i, Y_i)_{i=1}^n$ through the choice of q_i. Let S_n denote all permutations of $\{1, \dots, n\}$, and for $\sigma \in S_n$ define $E^{\sigma}_{\mathrm{HL},n}$ as $E^{\mathrm{id}}_{\mathrm{HL},n}$ for the random variables $(P_{\sigma(i)}, Y_{\sigma(i)})_{i=1}^n$ instead of $(P_i, Y_i)_{i=1}^n$. Generally,

$$\sup_{\sigma \in S_n} E^{\sigma}_{\mathrm{HL},n}$$

is not an e-variable for $\mathcal{H}_{\mathrm{HL},n}$, so one would guess that there are opportunities to fish for (spurious) significance by choosing some specific ordering of a sample of observations $(p_i, y_i)_{i=1}^n$. If there is a natural ordering of the observations such as a time stamp then the problem usually does not occur in applications since a different ordering of the observations hard to justify. Indeed, when the observations are sequential (and possibly dependent), the e-variable defined at (5) is also an e-variable for the hypothesis

$$\mathcal{H}_{\mathrm{HL},n,seq} = \{\mathbb{P} \in \mathcal{P} \mid \mathbb{E}_{\mathbb{P}}(Y_i | P_1, \dots, P_i, Y_1, \dots, Y_{i-1}) = P_i \ \mathbb{P}\text{-almost surely}, \ i = 1, \dots, n\}.$$

Contrary to classical theory, the sequential case is easier to treat than the iid case and has been the focus of many works employing e-values including for example (Waudby-Smith and Ramdas, 2021; Henzi and Ziegel, 2021+).

Coming back to our situation with iid data, an alternative to (5) could be

$$E_{\mathrm{HL},n,\mathrm{sym}} = \frac{1}{n!} \sum_{\sigma \in S_n} E^{\sigma}_{\mathrm{HL},n}.$$

This strategy is essentially the merging technique for independent e-values in Section 4 of Vovk and Wang (2021). Computationally, this seems rather intractable and we also might lose a lot of power by averaging.

A second open problem is the choice of the quantities q_i. If the goal is to maximize the growth rate of the e-value, the true conditional probabilities $\pi = \mathbb{E}_{\mathbb{Q}}(Y \mid P)$ are the optimal choice, since

$$\mathbb{E}_{\mathbb{Q}}(\log(E_q(Y, P)) \mid P) = \pi(\log(q) - \log(P)) + (1 - \pi)(\log(1 - q) - \log(1 - P)),$$

and the derivative of this quantity with respect to q equals

$$\frac{d}{dq} \mathbb{E}_{\mathbb{Q}}(\log(E_q(Y, P)) \mid P) = \frac{\pi}{q} - \frac{1 - \pi}{1 - q} = \frac{\pi - q}{q(1 - q)} \leq 0 \iff q \leq \pi.$$

We suggest to address both problems above by the following strategy: Split the dataset into two parts, say $\{1, \ldots, n/2\}$ and $\{n/2+1, \ldots, n\}$ for simplicity. Estimate the conditional expectations $p \mapsto \mathbb{E}(Y \mid P = p)$ on the first part of the data with some statistical model, and generate predictions q_i for $\mathbb{E}(Y \mid P_i)$, $i \geq n/2+1$. These e-values can then combined by products like in $E^{\text{id}}_{\text{HL},n}$ thanks to independence. To avoid that the e-value depends on the split of the dataset, we repeat this procedure several times and average the e-values from all splits. This is essentially a bootstrapping procedure in which we try to estimate the true conditional probabilities $\mathbb{E}(Y \mid P_i)$ and verify if they are far away from the predictions P_i. In the end we obtain a valid e-value for the hypothesis $\mathcal{H}_{\text{HL},n}$, which is different from bootstrap p-values or confidence intervals, which may be difficult to interpret.

While the above procedure is valid no matter what method for the estimation of $\mathbb{E}(Y \mid P)$ is applied, we suggest to use isotonic regression, which is solved by the pool-adjacent-violators (PAV) algorithm (Ayer et al., 1955). Similar procedures are used for recalibration of binary classifiers in machine learning application; see e.g. Zadrozny and Elkan (2002) or Flach (2012). Recently, Dimitriadis et al. (2021) related the isotonic regression approach to reliability plots, which are a key diagnostic tool in meteorological forecasting. In connection to the aforementioned literature isotonic regression would be an attractive choice because it maximizes

$$(q_1, \ldots, q_{n/2}) \mapsto \sum_{i=1}^{n/2} \log(1-q_i)(1-Y_i) + \log(q_i)Y_i$$

over all $q_1 \leq \cdots \leq q_{n/2}$ (assuming that $P_1 \leq \cdots \leq P_{n/2}$), that is, it maximizes the growth rate among all monotone estimators. In particular we proceed as follows to estimate q_i:

1. Split the data set into two parts: For $s \in (0,1)$
   a) randomly select $\lfloor ns \rfloor$ observations without replacement
   b) extract the non-selected observations.
2. Estimate the conditional expectations on the first part of the data with isotonic mean regression and generate predictions q_i for $\mathbb{E}(Y_i|P_i)$, with $i = \lceil ns \rceil, \ldots, n$. Compute e-values like in (5).
3. Repeat the procedure B times and average the e-values from all those splits.

The assumption of monotonicity is a restriction, but it is reasonable given that a model with $\mathbb{E}(Y \mid P = p)$ not monotone in p is not particularly useful anyway and one would probably even not need to perform a eHL test to discard it. Note, if one nonetheless does not like the monotonicity assumption of the procedure described above, one might replace step two by sufficiently general nonparametric methods, for example some nearest neighbor approach. Another option is the use of flexible parametric models to estimate the curve $p \mapsto \mathbb{E}(Y \mid P = p)$, for instance parametrized by the CDFs of Beta-distributions.

3 Simulations

This section evaluates the empirical performance of the proposed HL goodness-of-fit test based on e-values. Therefore, we simulate a sample (P_i, Y_i) with iid observations $i = 1, \ldots, n$ according to the classical setup of Hosmer et al. (1997) which is, if at all, just slightly modified in more recent contributions; see e.g. Hosmer and Hjort (2002), Xie et al. (2008) and Allison (2014) or Canary et al. (2017) and Nattino et al. (2020). While those contributions investigate several different simulation setups to examine the empirical performance of the tests when a misspecified logistic model is used to issue predictions, we here focus on the case of quadratic misspecification.

3.1 Data generation: quadratic misspecification

We simulate data from a logistic model with two covariates using the logit link function by following the tradition which was established in the above mentioned work. In doing so, we define the true conditional event probability as

$$\pi_i = \pi(X_i) = \mathbb{P}(Y_i = 1 | X_i, \beta_0, \beta_1, \beta_2) = \frac{\exp(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)}{1 + \exp(\beta_0 + \beta_1 X_i + \beta_2 X_i^2)}, \tag{6}$$

where $X_i \sim \mathrm{U}(-3, 3)$. Based on that we simulate the outcome variable $Y_i \sim \mathrm{Bernoulli}(\pi_i)$. We use the generated data to estimate a misspecified model without considering the quadratic term. The probability of a positive outcome is predicted by

$$P_i = P(X_i) = \mathbb{P}(Y_i = 1 | X_i, \widehat{\beta}_0, \widehat{\beta}_1) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)}. \tag{7}$$

Depending on the severity of the misspecification, expressed through the magnitude of β_2, a goodness-of-fit test should detect the model fit to be poor. To consider that, we vary the magnitude of β_2 by parameterizing the lack of linearity in the true model. For that, we follow the literature and specify the parameters such that the true regression curve crosses the points $\pi(-1.5) = 0.05$, $\pi(3) = 0.95$ and $\pi(-3) = j$, where j is a positive value. This results in a system of nonlinear equations which we can solve for the true model parameters β_0, β_1, and β_2; see Appendix A for details on the computation. For instance, setting $j = 0.007$ results in the parameter vector $\beta \approx (-0.9578, 1.3165, 0)$. Thus, under this condition the specification of the models in Equation (6) and (7) coincide (ignoring the difference between β and sample estimate $\widehat{\beta}$) and we expect to obtain calibrated predictions. For a j that deviates from 0.007, the model that generally omits the squared effect is expected not to predict the realizations perfectly because the lack of linearity in the logit function becomes increasingly more pronounced. Note that some simulation studies do not specify exactly which j was chosen. Instead, such studies qualitatively describe the strength of the quadratic effect in the population model by using three categories (e.g., Allison (2014) or Xie et al. (2008) refer only to true models that exhibit a strong, moderate, or a weak lack of linearity).

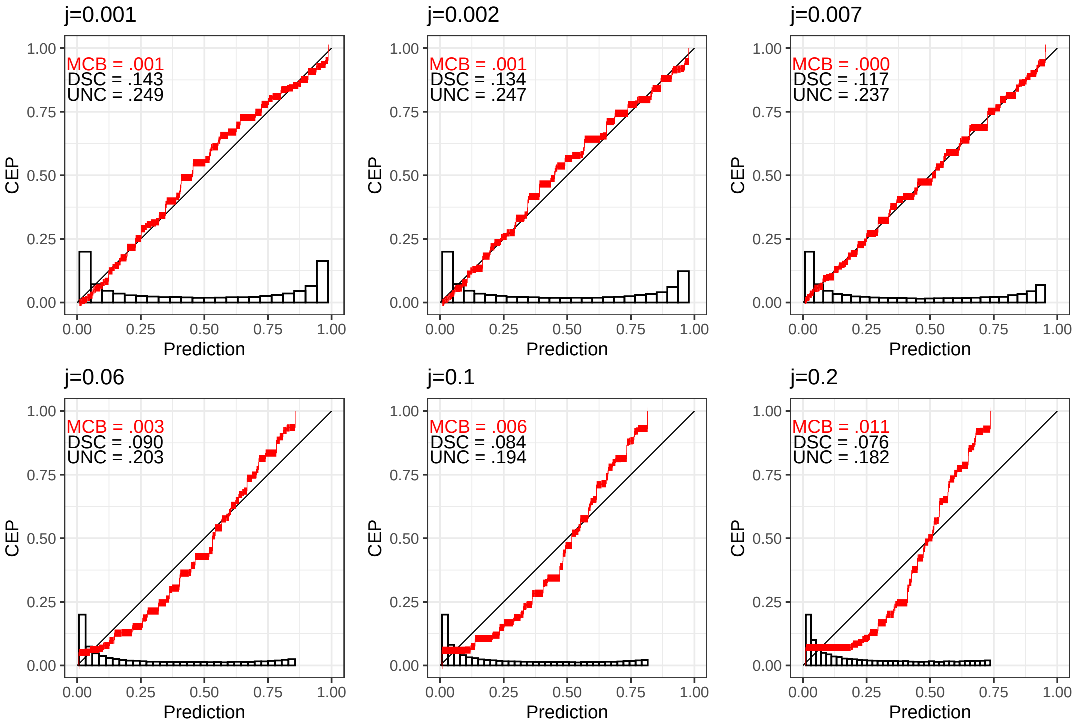
3.2 Isotonic regression to estimate the conditional event probability

As shown, the e-variable for the data instance i depends on q_i. When setting $q_i = \pi_i$ the growth rate of the e-value is maximized. Since π_i is unknown in practice, we need to employ a suitable estimation procedure, which ensures that the estimate of q_i is independent of Y_i. Following the procedure described in Section 2.3, we split the data into two parts, where the estimation sample holds $\lfloor ns \rfloor$ randomly chosen observations. The remaining $\lceil ns \rceil$ are assigned to the test sample. Using the estimation sample, we employ the estimation procedure based on the PAV-recalibrated probabilities (see Section 2.3), such as for the CORP-estimates of Dimitriadis et al. (2021). This effectively results in a stepwise regression curve which we use to obtain estimates for q_i in the test sample. Based on that we construct $E^{\mathrm{id}}_{\mathrm{HL},n}$.

Using this estimation strategy, we may encounter two problems. Firstly, within the test sample, we may find a prediction P_i that lies outside the available values of the estimation sample. In that case, we interpolate constantly. In other words, if the range of the estimation sample is $[P_l, P_m]$ and we for example find $P_i > P_m$, then we use the q_i at P_m. Secondly, whenever isotonic regression estimates $\pi_i \in \{0, 1\}$ occur, an e-value of zero might be obtained, which should be avoided. In these cases we set $q_i = P_i$. Thus, we ignore each of those sample points by forcing $E_{q_i} = 1$, which is not affecting the validity of the e-value since q_i can be chosen arbitrarily under the null hypothesis.

Figure 1 presents CORP reliability diagrams for selected choices of j; for a detailed description of such plots see Dimitriadis et al. (2021). The plots use simulated data according to the

Figure 1: CORP reliability diagrams based on Dimitriadis and Jordan (2020) for the predictions from the correctly ($j = 0.007$) specified and some misspecified models. All plots are based on 50,000 observations. The data is generated as described in Section 2.



procedure above, where we set $n = 50,000$. Thus we can consider the red curve as a highly accurate estimate of the true conditional event probability (CEP) under the imposed lack of linearity. Based on that, the motivation for presenting Figure 1 is threefold. Firstly, the plots reflect the severity of the misspecification. For example, consider the subfigure for the correctly specified model, i.e. when $j = 0.007$. In that case, the predictions seem to be well-calibrated, which is also indicated by the miscalibration (MCB) component of the CORP Brier score decomposition. In contrast, the remaining graphics of Figure 1 show the reliability of the misspecified models, where we set $j \neq 0.007$ for the underlying DGP. That results in poorly calibrated predictions for $j = 0.2$ since the corresponding model omits the quadratic term, which obviously is of importance. Secondly, Figure 1 is suitable for visualizing the estimation procedure we introduced above. Let us assume for now that the red curve is the correctly estimated CEP in the estimation sample and suppose that we observe a prediction $P_i = 0.4$ in the test sample. Given that prediction, we would obtain a GROW e-value if isotonic regression assigns q_i equal to the intersection with the red curve. Further, as follows from the alternative hypothesis, q_i and P_i are required the be on the same side of the diagonal (black solid line) to gain evidence against the null. Thirdly, the CORP reliability diagrams reveal a critique of the simulation setup, which is traditionally used to evaluate the performance of goodness-of-fit tests. While this classical simulation setup allows us to control X_i and the parameter vector β, the distribution of P_i is changing; e.g. compare the histograms. As an alternative, we should also consider simulating P_i directly such that the distribution of the prediction remains unchanged over j.

Table 1: Rejections of the tests under the null hypothesis of calibrated predictions in *percent*. For the eHL test we present the sample splits $s = \{0.2, 0.3, \cdots, 0.6\}$. For the oracle eHL we implement an full sample ($s = 0$) and a half sample version ($s = 0.5$). The nominal significance level of the cHL is 5%. The eHL tests are rejected whenever $E^{\text{id}}_{\text{HL},n} \geq 20$.

		oracle eHL		eHL				
obs	cHL	0	0.5	0.2	0.3	0.4	0.5	0.6
512	5.08	0	0	0.14	0.14	0.18	0.14	0.14
1024	4.58	0	0	0.08	0.24	0.24	0.28	0.28
2048	4.68	0	0	0.04	0.10	0.16	0.12	0.12
4096	4.42	0	0	0.02	0.06	0.12	0.10	0.14

3.3 Results

We apply several versions of the eHL goodness-of-fit tests for $n \in \{1024, 2048, 4096\}$ and j on a grid of values between 0.001 and 0.2. For the eHL, we reject the H_0 if $E^{\text{id}}_{\text{HL},n} \geq 20$ which follows from Markov's inequality in Equation (1) for $\alpha = 0.05$. Concerning the sample splits we chose $s \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. Concerning the bootstrap replications, we always set $B = 10$, which should be sufficient since the aim is to eliminate dependencies from the sample split. We compare the rejection rates to a classical HL test (cHL) with ten equally populated bins. For the cHL we use a nominal significance level of 5%. Note, for the cHL we do not split the data set. While the two tests above are feasible and thus ready to implement in practical applications, it appeals from a theoretical perspective to simulate an eHL test where q_i is set equal to the true CEP. Since such a test can only be conducted in a simulation setting, we refer to it as oracle eHL. For the oracle eHL we implement a full and a half sample version, i.e. we use $s = 0$ or $s = 0.5$. We simulate all tests 5000 times.

The empirical sizes of the tests are reported in Table 1. From that table follows that the cHL test is well-sized. In contrast, the eHL tests seem to be considerably undersized. Even though the rejections under the null hypothesis are drastically close to zero for all eHL tests, this should not be very surprising. An explanation might be that the so-called p-guarantee is obtained by Markov's inequality, which yields a conservative test by construction. That is, we only know that the rejection rate for $E^{\text{id}}_{\text{HL},n} \geq 20$ is not greater than 5%. Remarkably, when assigning the true CEP to q_i (oracle eHL), the respective tests never reject the null hypothesis when it is true.

Figure 2 presents the rejection rates for the tests. The solid colored lines illustrate the eHL tests. The shapes indicate the proportion s of the data that was assigned to the estimation sample. In comparison to other splits, we find higher rejection rates when choosing s closely around 0.5. Intuitively, we would anticipate obtaining more powerful e-values when in large samples $s < 0.5$ is set. Also, this should not affect the accuracy of the estimate for q_i negatively. However, this cannot be seen for the sample sizes chosen here. Therefore, it might be interesting to consider larger sample sizes as well. Irrespectively of the sample splits, we find the eHL tests performing similar to the cHL when $n \geq 2048$ and $j \geq 0.05$, which Hosmer et al. (1997) still consider as slight degrees of misspecification. See also the CORP reliability diagrams in Figure 1 in that context. In contrast to the eHL, the cHL seems to detect such degrees of misspecification even in smaller samples. Also be aware that the first plot in Figure 2 ($n = 512$) illustrates the rejection rates for a wider range of j, i.e. we use $0.001 \leq j \leq 0.2$. The other plots only visualize rejection rates for $0.001 \leq j \leq 0.1$.

The performance of the full sample oracle eHL approaches that of the cHL with the number of observations. For $n = 4096$ the rejection rates of both tests are almost indistinguishable. When comparing the half sample oracle eHL to the feasible eHL tests for $n \in \{512, 1024\}$, we

Figure 2: Rejection rates of eHL tests are represented by the colored solid lines. The shapes indicate the value of s used to estimate q_i via isotonic regression. Further, the full and half sample oracle eHL is visualized (black dotted and blue dashed curve). The H_0 of an e-value test is rejected if $E^{\mathrm{id}}_{\mathrm{HL},n} \geq 20$. The classical HL is performed using ten equally populated bins (gray dot-dashed). The H_0 of the cHL is rejected if the p-value is smaller than a nominal significance level of 5% (horizontal gray dashed line). For $j = 0.007$, the DGP samples under the H_0 (vertical gray dashed line). Note that the x-axis for the plot with $n = 512$ also includes values of $j > 0.1$.

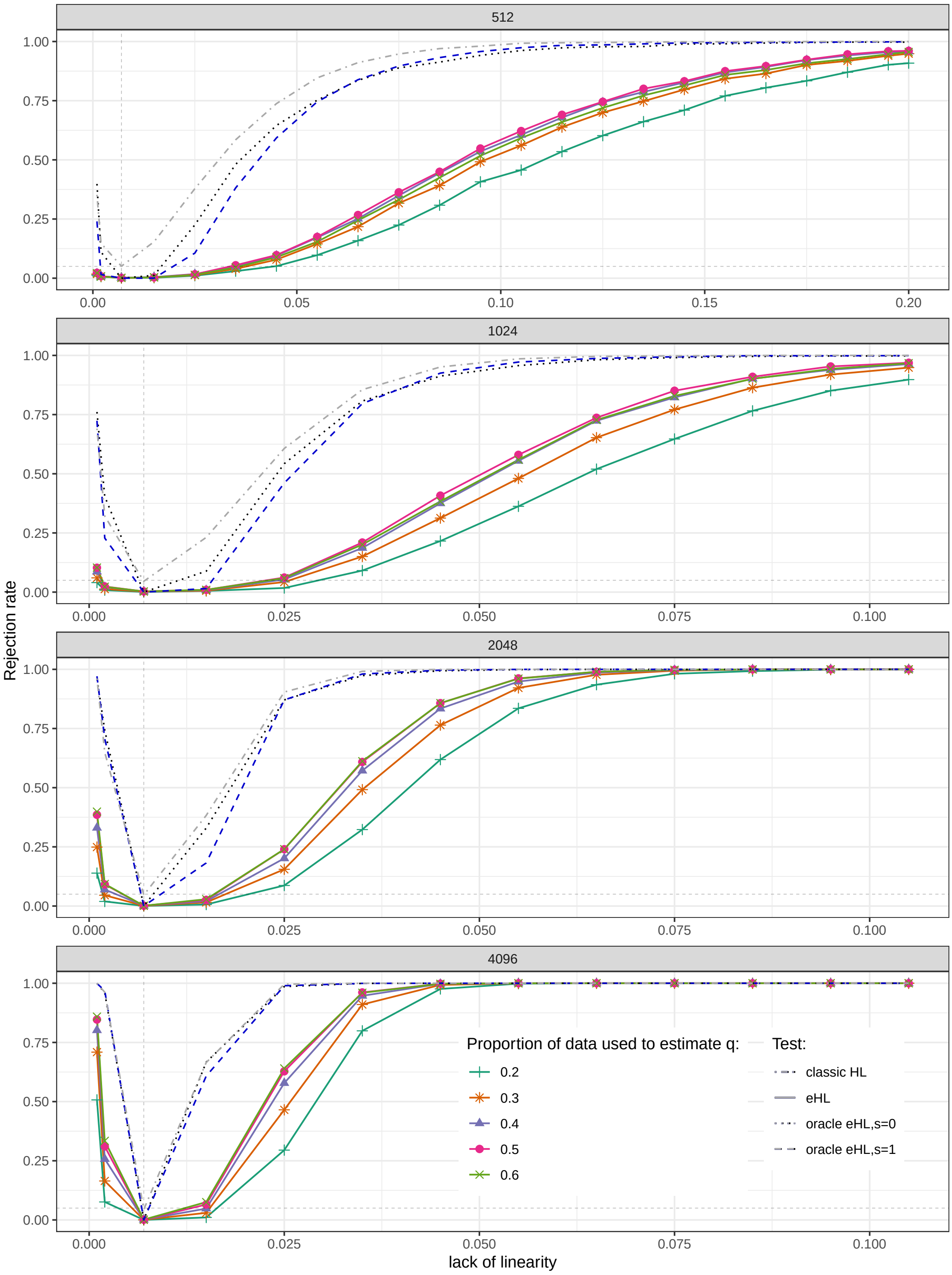
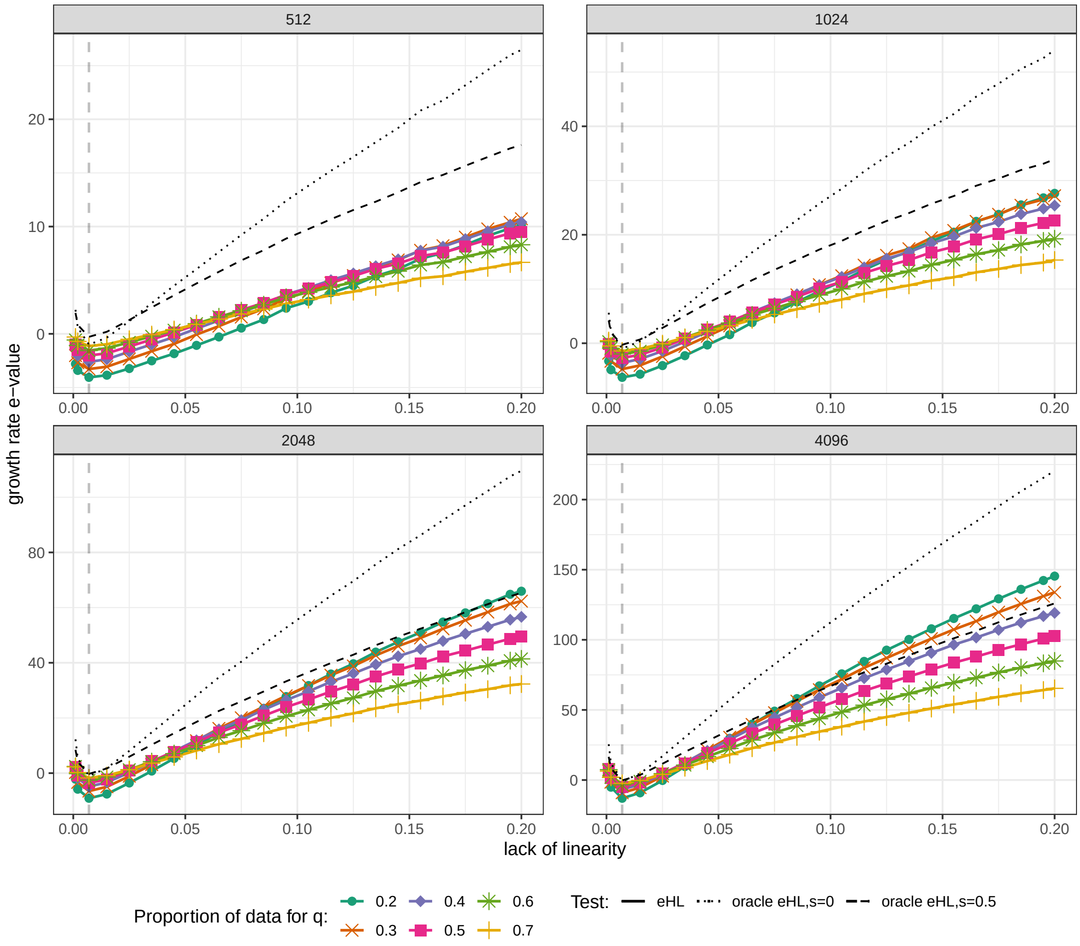


Figure 3: Growth rate of the e-value for a given lack of linearity, j. All solid lines are the eHL tests based on isotonic regression estimations for q_i, where the shapes indicate the proportion of data used to estimate q_i. The full sample oracle eHL is the dotted black line. The black dashed line is the half sample oracle eHL.



might conclude that the estimates of q_i are not always precise enough to obtain comparable power properties. However, with increasing n, the differences between both tests are reducing gradually; e.g. compare the rejection rates for $j = 0.025$. Thus, it seems plausible that the difference in the power comes primarily from losing $s \cdot 100\%$ observations for the feasible eHL tests, and perhaps not from the fact that we use e-value tests or that we estimate the CEP poorly.

The full sample oracle eHL can be shown to exhibit the maximal growth rate, which is defined as $(1/n)\log\prod_{i=1}^n E_i = (1/n)\sum_{i=1}^n \log(E^{\text{id}}_{\text{HL},n})$ and illustrated for given values of j in Figure 3. As expected, the full sample oracle eHL yields steeper growth curves than the other tests. Interestingly, the half sample oracle eHL yields a growth rate that is comparable to the feasible eHL tests for $n \geq 2048$. Thus, in contrast to the rejection rates, the growth rate plots match the intuition raised before: smaller choices of s are reasonable in large samples. This becomes especially evident when examining the subplot with $n = 4096$. The inconsistent results between growth rates and rejection rates might be explained by the fact that an optimal growth rate is not necessarily implying optimal power at our given threshold of 20. We believe that the same argument may also explain the diverging performance of the half sample oracle eHL test, which can be observed when comparing Figure 2 and 3.

4 Discussion

This technical report proposes an e-test for perfect calibration, which is a safe testing counterpart to the widely used Hosmer-Lemeshow test. The proposed eHL test follows a simple betting interpretation (see Shafer (2021)) where the e-value can be seen as the factor by which we multiply the bet against the hypothesis of perfect calibration. Intuitively, when accumulating money by the bet, we gain evidence against the null. Here, the e-value depends on the probability prediction, its corresponding realization, and an arbitrary value, which we suggest estimating in a two-step approach by isotonic regression. Further, we assess the empirical performance of the test to detect quadratic model misspecifications. The simulations show that in samples of more than 2000 observations, the eHL test allows to reliably detect levels of quadratic misspecification, which Hosmer et al. (1997, p. 973) denote to be slight. The intrinsic flexibility of the e-values allows the application of stable data-driven methods (here isotonic regression) instead of the typical binning and counting technique in the HL test. However, this flexibility comes at the cost of lower power small samples of less than 2000 observations.

Finally, we want to stress that a major advantage of e-values over p-values is their validity in sequential testing. Theoretically the proposed testing framework allows to combine evidence from multiple equally specified binary regression models.

References

P. J. Allison. Measures of fit for logistic regression. *Paper 1485-2014, SAS Global Forum 2014*, pages 1–12, 2014.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26: 641–647, 1955.

G. Bertolini, R. D'Amico, D. Nardi, A. Tinazzi, and G. Apolone. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics*, 5:251–253, 2000.

J. D. Canary, L. Blizzard, R. P. Barry, D. W. Hosmer, and S. J. Quinn. A comparison of the Hosmer–Lemeshow, Pigeon–Heyse, and Tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods. *Communications in Statistics - Simulation and Computation*, 46:1871–1894, 2017.

T. Dimitriadis and A. I. Jordan. `reliabilitydiag`: Reliability diagrams using isotonic regression. R package version 0.1.3, 2020. URL https://cran.r-project.org/package=reliabilitydiag.

T. Dimitriadis, T. Gneiting, and A. I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118, 2021.

P. Flach. *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge University Press, 2012.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Preprint, arXiv: 1906.07801*, 2020.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, to appear, 2021+.

D. W. Hosmer and N. L. Hjort. Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 21:2723–2738, 2002.

D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9:1043–1069, 1980.

D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 1997.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Ltd, 2013.

O. Kuss. Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine*, 21:3789–3801, 2002.

L. Y. Lee, J.-B. Cazier, V. Angelis, R. Arnold, V. Bisht, N. A. Campton, J. Chackathayil, V. W. Cheng, H. M. Curley, M. W. Fittall, L. Freeman-Mills, S. Gennatas, A. Goel, S. Hartley, D. J. Hughes, D. Kerr, A. J. Lee, R. J. Lee, S. E. McGrath, C. P. Middleton, N. Murugaesu, T. Newsom-Davis, A. F. Okines, A. C. Olsson-Brown, C. Palles, Y. Pan, R. Pettengell, T. Powles, E. A. Protheroe, K. Purshouse, A. Sharma-Oates, S. Sivakumar, A. J. Smith, T. Starkey, C. D. Turnbull, C. Várnai, N. Yousaf, U. C. M. P. Team, R. Kerr, and G. Middleton. Covid-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a prospective cohort study. *Lancet (London, England)*, 395:1919–1926, 2020.

G. Nattino, M. L. Pennell, and S. Lemeshow. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics*, 76:549–560, 2020.

R. Neblett Fanfair, K. Benedict, J. Bos, S. D. Bennett, Y.-C. Lo, T. Adebanjo, K. Etienne, E. Deak, G. Derado, W.-J. Shieh, C. Drew, S. Zaki, D. Sugerman, L. Gade, E. H. Thompson, D. A. Sutton, D. M. Engelthaler, J. M. Schupp, M. E. Brandt, J. R. Harris, S. R. Lockhart, G. Turabelidze, and B. J. Park. Necrotizing cutaneous mucormycosis after a tornado in joplin, missouri, in 2011. *New England Journal of Medicine*, 367:2214–2225, 2012.

L. Ostrosky-Zeichner, R. Harrington, N. Azie, H. Yang, N. Li, H. Zhao, V. Koo, and E. Q. Wu. A risk score for fluconazole failure among patients with candidemia. *Antimicrobial Agents and Chemotherapy*, 61:e02091–16, 2017.

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184:407–431, 2021.

G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, Ltd, 2019.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49:1736 – 1754, 2021.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Preprint, arXiv:2009.02824*, 2020.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Preprint, arXiv:2010.09686*, 2021.

X.-J. Xie, J. Pendergast, and W. Clarke. Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors. *Computational Statistics & Data Analysis*, 52:2703–2713, 2008.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 694–699, New York, NY, USA, 2002. Association for Computing Machinery.

A Details on the data generation

The literature mentioned in the introduction obtains the parameter values β_0, β_1, and β_2 by solving the system of nonlinear equations

$$\begin{aligned}\frac{\exp(\beta_0 + (-1.5)\beta_1 + 1.5^2\beta_2)}{1 + \exp(\beta_0 + (-1.5)\beta_1 + (-1.5)^2\beta_2)} &= 0.05,\\ \frac{\exp(\beta_0 + 3\beta_1 + 3^2\beta_2)}{1 + \exp(\beta_0 + 3\beta_1 + 3^2\beta_2)} &= 0.95,\\ \frac{\exp(\beta_0 + (-3)\beta_1 + (-3)^2\beta_2)}{1 + \exp(\beta_0 + (-3)\beta_1 + (-3)^2\beta_2)} &= j,\end{aligned}$$

where j is a positive value. While one can solve the system for any positive value of j, the literature just uses $j \geq 0.007$. Within our simulation we also use $j = 0.001, j = 0.002$. For example, we obtain a model with a parameter vector $\beta = (-2.3366, 0.8569, 0.3013)$ if $j = 0.1$.

4.4 Honest calibration assessment for binary outcome predictions

The content of this section is published as an arXiv preprint,

Dimitriadis, T., Duembgen, L., Henzi, A., Puke, M. and Ziegel, J. (2022). Honest calibration assessment for binary outcome predictions. *arXiv preprint arXiv:2203.04065.*

Honest calibration assessment for binary outcome predictions

Timo Dimitriadis[1,2], Lutz Dümbgen[3], Alexander Henzi[3], Marius Puke[4], and Johanna Ziegel[3]

[1]Heidelberg University, Germany
[2]Heidelberg Institute for Theoretical Studies (HITS), Germany
[3]University of Bern, Switzerland
[4]University of Hohenheim, Germany

timo.dimitriadis@awi.uni-heidelberg.de, lutz.duembgen@stat.unibe.ch, alexander.henzi@stat.unibe.chh, marius.puke@uni-hohenheim.de, johanna.ziegel@stat.unibe.ch

March 8, 2022

Abstract

Probability predictions from binary regressions or machine learning methods ought to be calibrated: If an event is predicted to occur with probability x, it should materialize with approximately that frequency, which means that the so-called calibration curve $p(x)$ should equal the bisector for all x in the unit interval. We propose honest calibration assessment based on novel confidence bands for the calibration curve, which are valid only subject to the natural assumption of isotonicity. Besides testing the classical goodness-of-fit null hypothesis of perfect calibration, our bands facilitate inverted goodness-of-fit tests whose rejection allows for the sought-after conclusion of a sufficiently well specified model. We show that our bands have a finite sample coverage guarantee, are narrower than existing approaches, and adapt to the local smoothness and variance of the calibration curve p. In an application to model predictions of an infant having a low birth weight, the bounds give informative insights on model calibration.

Keywords: Binary regression, calibration validation, isotonic regression, confidence band, goodness-of-fit, universally valid inference

1 Introduction

Let $x_1 \leq \cdots \leq x_n$ be given covariates in $[0, 1]$ and $Y_1, \ldots, Y_n \in \{0, 1\}$ independent binary observations such that $\mathbb{P}(Y_i = 1) = p(x_i)$ for some unknown function $p\colon [0, 1] \to [0, 1]$. In practice, the covariates can be probability predictions for the components of $\mathcal{Y} := (Y_i)_{i=1}^n$, e.g., stemming from a test sample of binary regressions, machine learning methods, or any other statistical model for binary data. A reliable interpretation of these predictions relies on the property of calibration, meaning that the so-called calibration curve p is sufficiently close to the identity. For instance, if a fetus is predicted to have a low birth weight with probability $x = 5\%$, decisions on a potential medical treatment rely on this probability prediction being accurate enough, $|p(x) - x| < \varepsilon$ for some small $\varepsilon > 0$.

Testing calibration, closely related to goodness-of-fit testing, is crucial in applications (Tutz, 2011; Hosmer et al., 2013) and is still regularly carried out by the classical test of Hosmer and Lemeshow (1980), which groups the predictions x_i into bins and applies a χ^2-test. It is however subject to multiple criticisms: First, its ad hoc choice of bins can result in untenable

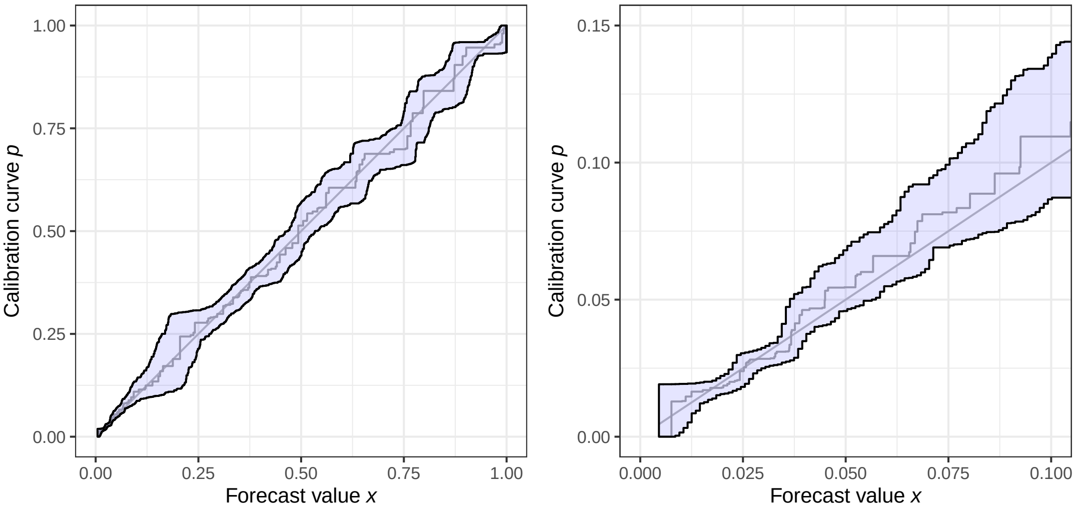


Figure 1: Left: Calibration bands for the first model specification of the low birth weight application in Section 6. The blue band denotes the calibration band and the grey step function the isotonic regression estimate. Right: Magnified version focusing on predicted probabilities below 10%.

instabilities (Bertolini et al., 2000; Allison, 2014). Second, placing the hypothesis of calibration in the null only allows for rejecting calibration rather than showing that a model is sufficiently well calibrated, where the latter would be highly desirable for applied researchers. Third, the test rejects essentially all, even acceptably well-specified models in large samples (Nattino et al., 2020a; Paul et al., 2013), resulting in calls for a goodness-of-fit tests with inverted hypotheses (Nattino et al., 2020b).

We propose a statistically sound solution to these criticisms by constructing honest, simultaneous confidence bands (L^α, U^α) for the function p. That is, for a given small number $\alpha \in (0,1)$, we compute data-dependent functions $L^\alpha = L^\alpha(\cdot, \mathcal{Y})$ and $U^\alpha = U^\alpha(\cdot, \mathcal{Y})$ on $[0,1]$ such that

$$\mathbb{P}\{L^\alpha \leq p \leq U^\alpha \text{ on } [0,1]\} \geq 1 - \alpha, \tag{1}$$

which we call calibration band. It allows for the desirable conclusion that with confidence $1-\alpha$, the true calibration curve p lies inside the band, simultaneously for all values of the predicted probabilities.

Hence, it resolves the above mentioned criticisms of classical goodness-of-fit tests. Figure 1 shows the bands in a large data example for probit model predictions for the binary outcome of a fetus having a low birth weight. See Section 6 for additional details. The test of Hosmer and Lemeshow clearly rejects calibration even though our bands indicate a well-calibrated model, especially for the, in this application, most important region of small probability predictions shown in the magnified right panel of the figure. Our bands also nest a goodness-of-fit test with classical hypotheses by checking whether the band contains the bisector $b(x) = x$ for all relevant values $x \in [0,1]$. It is important to notice that even though we build our bands on the model predictions, the methodology applies equally to both, causal and predictive regressions. An open-source implementation in the statistical software `R` is available under https://github.com/marius-cp/calibrationband.

Our confidence bands are valid in finite samples subject only to the mild assumption that the function p is increasing,

$$p(x) \;\leq\; p(x'), \quad 0 \leq x \leq x' \leq 1, \tag{2}$$

which is natural in the context of assessing calibration as decreasing parts of the calibration curve p can be dismissed as nonsensical predictions resulting from severely misspecified models (Dimitriadis et al., 2021; Roelofs et al., 2020). As can be expected for a non-parametric, pathwise and almost universally valid confidence band, we require large data sets of at least above 5 000 observations to obtain sensibly narrow bands. These are exactly the sample sizes where the classical goodness-of-fit tests become uninformative by rejecting all models in applications. For example, in a simulation study on assessing a logistic regression model with minor misspecification, Kramer and Zimmerman (2007) find that the Hosmer-Lemeshow test at level $\alpha = 0.1$ achieves a power of 18.5% for $n = 5\,000$ and essentially 100% for $n = 50\,000$.

A theoretical analysis shows that the proposed confidence band adapts locally to the smoothness of the function p and to the variance of the observations. Adaptivity to the smoothness means that the width of the bands decreases faster with the sample size n in regions where p is constant, and at a slower rate where p is steeper. This property is known for more general confidence bands for a monotone mean function developed by Yang and Barber (2019), which are proved to be more conservative than our bands in the case of binary outcomes. Adaptivity to the variance means that the band is substantially narrower at x if $p(x)$ is close to zero or one, compared to $p(x)$ near 0.5. In many practical applications, including the low birth weight predictions analyzed in this article, predicted probabilities close to zero or one are of most relevance and a sharp assessment of calibration is particularly important.

Existing methods for the construction of confidence bands in this setting are rare with the following two exceptions: First, Nattino et al. (2014) propose the use of confidence bands based on a parametric assumption on the function p, which we show to have incorrect coverage in almost all of our simulation settings. Second, the nonparametric bands of Yang and Barber (2019) are valid but shown to be wider than our bands as we show in theory and simulations.

We explain the absence of competing methods by their theoretical difficulties. Using asymptotic theory of the isotonic regression estimator is complicated as it requires the estimation of nuisance quantities such as the derivative of the unknown function p, the convergence rate depends on the functional form of p, it is subject to more restrictive assumptions and only results in bands with a pointwise interpretation (Wright, 1981). Resampling schemes are theoretically found to be inconsistent for the isotonic regression (Sen et al., 2010; Guntuboyina and Sen, 2018). Other non-parametric approaches in the literature for constructing confidence bands for functions, many of them presented in the review by Hall and Horowitz (2013), are often pointwise, not simultaneous, and require the selection of tuning parameters that may lead to instabilities, similar to the choice of the bins in the Hosmer-Lemeshow test. In contrast, the confidence bands proposed here are simple to compute and do not involve any implementation decisions resulting in a stable and reproducible method as recently called for by Stodden et al. (2016); Yu and Kumbier (2020).

2 Construction of the confidence bands

In what follows we focus on confidence bounds $L_i^\alpha = L_i^\alpha(\mathcal{Y})$ and $U_i^\alpha = U_i^\alpha(\mathcal{Y})$ for $p_i = p(x_i)$, where $1 \le i \le n$. Indeed, if

$$\mathbb{P}(L_i^\alpha \le p_i \le U_i^\alpha \text{ for } 1 \le i \le n) \ge 1 - \alpha,$$

then

$$\begin{aligned} U^\alpha(x) &= U_i^\alpha, \quad x \in (x_{i-1}, x_i],\ 1 \le i \le n+1, \\ L^\alpha(x) &= L_i^\alpha, \quad x \in [x_i, x_{i+1}),\ 0 \le i \le n, \end{aligned}$$

defines a confidence band (L^α, U^α) satisfying (1) with the auxiliary values $x_0 := -\infty$, $L_0^\alpha := 0$ and $x_{n+1} := \infty$, $U_{n+1}^\alpha := 1$.

Our confidence bands are based on the classical confidence bounds of Clopper and Pearson (1934) for a binomial parameter. Suppose that Z is a binomial random variable with parameters m and $q \in [0,1]$. For $\delta \in (0,1)$ let

$$\begin{aligned} u^{\delta}(Z,m) &= \max\{\xi \in [0,1] \colon \mathrm{pbin}(Z,m,\xi) \geq \delta\} \\ &= \begin{cases} \mathrm{qbeta}(1-\delta, Z+1, m-Z), & Z < m, \\ 1, & Z = m, \end{cases} \\ \ell^{\delta}(Z,m) &= \min\{\xi \in [0,1] \colon \mathrm{pbin}(Z-1,m,\xi) \leq 1-\delta\} \\ &= \begin{cases} \mathrm{qbeta}(\delta, Z, m+1-Z), & Z > 0, \\ 0, & Z = 0. \end{cases} \end{aligned}$$

Here $\mathrm{pbin}(\cdot, m, \xi)$ denotes the distribution function of the binomial distribution with parameters m and ξ, while $\mathrm{qbeta}(\cdot, a, b)$ stands for the quantile function of the beta distribution with parameters $a, b > 0$. Then

$$\mathbb{P}\{q \leq u^{\delta}(Z,m)\} \geq 1-\delta \quad \text{and} \quad \mathbb{P}\{q \geq \ell^{\delta}(Z,m)\} \geq 1-\delta.$$

For the representation of $\ell^{\delta}(Z,m)$ and $u^{\delta}(Z,m)$ in terms of beta quantiles we refer to Johnson et al. (2005).

Assumption (2) allows to construct confidence bands for p as follows. For arbitrary indices $1 \leq j \leq k \leq n$, the random sum

$$Z_{jk} = \sum_{i=j}^{k} Y_i$$

is stochastically larger than a binomial random variable with parameters $n_{jk} = k - j + 1$ and p_j, and it is stochastically smaller than a binomial variable with parameters n_{jk} and p_k. Thus, as explained in Lemma B.1,

$$\mathbb{P}\{p_j \leq u^{\delta}(Z_{jk}, n_{jk})\} \geq 1-\delta, \qquad \mathbb{P}\{p_k \geq \ell^{\delta}(Z_{jk}, n_{jk})\} \geq 1-\delta. \tag{3}$$

If we combine these bounds for all pairs (j,k) in a given set $\mathcal{J}$, then we may claim with confidence $1 - 2|\mathcal{J}|\delta$ that simultaneously for all $(j,k) \in \mathcal{J}$,

$$p_i \leq u^{\delta}(Z_{jk}, n_{jk}) \ \ \forall\, i \leq j, \qquad p_i \geq \ell^{\delta}(Z_{jk}, n_{jk}) \ \ \forall\, i \geq k.$$

Specifically, let $\mathcal{J}$ be the set of all index pairs (j,k) such that $j \leq k$ and $x_{j-1} < x_j$ and $x_k < x_{k+1}$. If there are tied covariate values, $\mathcal{J}$ selects the outermost indices of the tied values. Hence, if $\{x_1, \ldots, x_n\}$ contains $N \leq n$ different points, then $|\mathcal{J}| = (N^2+N)/2$. Consequently, for a given confidence level $1 - \alpha \in (0,1)$, we may combine the bounds $u^{\delta}(Z_{jk}, n_{jk})$ and $\ell^{\delta}(Z_{jk}, n_{jk})$ with $\delta = \alpha/(N^2+N)$ to obtain a first confidence band.

Theorem 2.1. *For $1 \leq i \leq n$ let*

$$U_i^{\alpha,\mathrm{raw}} = \min_{(j,k) \in \mathcal{J} \colon x_j \geq x_i} u^{\alpha/(N^2+N)}(Z_{jk}, n_{jk}), \tag{4}$$

$$L_i^{\alpha,\mathrm{raw}} = \max_{(j,k) \in \mathcal{J} \colon x_k \leq x_i} \ell^{\alpha/(N^2+N)}(Z_{jk}, n_{jk}). \tag{5}$$

If p satisfies the isotonicity assumption (2), then the resulting confidence band $(L^{\alpha,\mathrm{raw}}, U^{\alpha,\mathrm{raw}})$ satisfies requirement (1).

In the definition (4), taking the minimum over index pairs (j,k) with $x_j \geq x_i$ is equivalent to the minimum over $j \geq i$ if $x_1 < \cdots < x_n$. When there are ties in the covariate, it is possible that the minimum is attained with an index $j < i$ but $x_j = x_i$. Analogously, it is possible that the maximum in (5) is attained with an index $k > i$ but $x_k = x_i$.

The confidence band proposed in Theorem 2.1 has two potential drawbacks. First, a natural nonparametric estimator for the function p under the assmption (2) is given by a minimizer $\hat{p}$ of $\sum_{i=1}^n \{h(x_i) - Y_i\}^2$ over all isotonic functions $h\colon [0,1] \to [0,1]$ (Dimitriadis et al., 2021). This minimizer is unique on the set $\{x_1, \ldots, x_n\}$. But there is no guarantee that $L^{\alpha,\mathrm{raw}} \le \hat{p} \le U^{\alpha,\mathrm{raw}}$. Second, the upper and lower bounds in (4) and (5) may even cross, resulting in an empty, and hence, nonsensical confidence band. These problems can be dealt with by using the non-crossing confidence band $(L^{\alpha,\mathrm{nc}}, U^{\alpha,\mathrm{nc}})$ with

$$L_i^{\alpha,\mathrm{nc}} = \min\{L_i^{\alpha,\mathrm{raw}}, \hat{p}(x_i)\}, \qquad U_i^{\alpha,\mathrm{nc}} = \max\{L_i^{\alpha,\mathrm{raw}}, \hat{p}(x_i)\}. \tag{6}$$

This band satisfies $L^{\alpha,\mathrm{nc}} \le \hat{p} \le U^{\alpha,\mathrm{nc}}$ on $[0,1]$, no matter how $\hat{p}(x)$ is defined for $x \notin \{x_1, \ldots, x_n\}$. Our simulations experiments indicate that $(L^{\alpha,\mathrm{raw}}, U^{\alpha,\mathrm{raw}}) = (L^{\alpha,\mathrm{nc}}, U^{\alpha,\mathrm{nc}})$ with probability $\gg 1 - \alpha$; see Section 5 for details.

A potential obstacle in the practical application of the confidence bands proposed in this section is that their computation requires $\mathcal{O}(N^2)$ steps. This can be relieved by reducing the number of distinct values in the covariate by rounding, which often has almost no visible effect on the appearance of the bands. If differences in the covariate smaller than K^{-1} for some $K \in \mathbb{N}$ are regarded as negligible, one can round up $x_1, \ldots, x_n$ to the next multiple of K^{-1} for the computation of the upper bound, and round off $x_1, \ldots, x_n$ to the next lower multiple of K^{-1} to compute the lower bound. This still yields a valid confidence band for the function p but guarantees that $N \le K + 1$, which also implies a less conservative correction of the confidence level. The number $K \in \mathbb{N}$ should not be too small since the resulting confidence bands are constant on intervals of length K^{-1} thereby limiting their adaptivity.

3 Relation to Yang and Barber (2019)

The methods of Yang and Barber (2019) may be adapted to the present regression setting with covariates $x_1 \le \cdots \le x_n$ as follows: With the isotonic estimator $\hat{p}$ introduced before, let

$$Z_{jk}^{\mathrm{iso}} = \sum_{i=j}^{k} \hat{p}(x_i).$$

Set

$$U_i^{\alpha,\mathrm{YB}} = \min_{(j,k)\in\mathcal{J}\colon x_j \ge x_i} \Bigl[\frac{Z_{jk}^{\mathrm{iso}}}{n_{jk}} + \sqrt{\frac{\log\{(N^2+N)/\alpha\}}{2n_{jk}}}\Bigr], \tag{7}$$

$$L_i^{\alpha,\mathrm{YB}} = \max_{(j,k)\in\mathcal{J}\colon x_k \le x_i} \Bigl[\frac{Z_{jk}^{\mathrm{iso}}}{n_{jk}} - \sqrt{\frac{\log\{(N^2+N)/\alpha\}}{2n_{jk}}}\Bigr]. \tag{8}$$

This defines a confidence band $(L^{\alpha,\mathrm{YB}}, U^{\alpha,\mathrm{YB}})$ with the following property:

$$\mathbb{P}\{L^{\alpha,\mathrm{YB}} \le \tilde{p} \le U^{\alpha,\mathrm{YB}} \quad \text{on } [0,1]\} \ge 1 - \alpha, \tag{9}$$

where $\tilde{p}\colon [0,1] \to [0,1]$ is any fixed isotonic function minimizing $\sum_{i=1}^n \{\tilde{p}(x_i) - p_i\}^2$. Thus one obtains a confidence band with guaranteed covergage probability $1 - \alpha$ for an isotonic approximation of p, even if (2) is violated. The proof of (9) follows from the arguments of Yang and Barber (2019), noting that the random variables Y_i are subgaussian with scale parameter $\sigma = 1/2$. That means, $\mathbb{E}\exp(t(Y_i - p_i)) \le \exp(\sigma^2 t^2/2)$ for all $t \in \mathbb{R}$, which implies that for arbitrary $\eta \ge 0$,

$$\mathbb{P}\{\pm(Z_{jk} - \mathbb{E}Z_{jk}) \ge \eta\} \le \exp(-2n_{jk}\eta^2),$$

see Hoeffding (1963). The following result shows that the confidence bands $(L^{\alpha,\mathrm{raw}}, U^{\alpha,\mathrm{raw}})$ and $(L^{\alpha,\mathrm{nc}}, U^{\alpha,\mathrm{nc}})$ are always contained in the band $(L^{\alpha,\mathrm{YB}}, U^{\alpha,\mathrm{YB}})$.

Theorem 3.1. *For all $\alpha \in (0,1)$ and any data vector $\mathcal{Y} \in \{0,1\}^n$,*

$$L^{\alpha,\mathrm{YB}} \le L^{\alpha,\mathrm{nc}} \le L^{\alpha,\mathrm{raw}}, \quad U^{\alpha,\mathrm{raw}} \le U^{\alpha,\mathrm{nc}} \le U^{\alpha,\mathrm{YB}} \quad \text{on } [0,1].$$

For the applications considered in the present paper, the validity of a confidence band in case of p violating (2) is not essential. It should be mentioned, however, that the band $(L^{\alpha,\mathrm{YB}}, U^{\alpha,\mathrm{YB}})$ has a computational advantage. For the computation of $U_i^{\alpha,\mathrm{YB}}$ in (7), it suffices to take the minimum over endpoints of constancy regions of $\hat{p}$, that is, all $(j,k) \in \mathcal{J}$ such that $j = \min(s \colon x_s \ge x_i)$ and $\hat{p}(x_k) < \hat{p}(x_{k+1})$ or $k = n$, see Proposition B.3 in the appendix. Likewise, for the computation of $L_i^{\alpha,\mathrm{YB}}$ in (8), it suffices to take the maximum over all $(j,k) \in \mathcal{J}$ such that $\hat{p}(x_{j-1}) < \hat{p}(x_j)$ or $j = 1$ and $k = \max(s \colon x_s \le x_i)$. While the computation of $(L^{\alpha,\mathrm{raw}}, U^{\alpha,\mathrm{raw}})$ or $(L^{\alpha,\mathrm{nc}}, U^{\alpha,\mathrm{nc}})$ requires $\mathcal{O}(N^2)$ steps, the following lemma implies that the computation of $(L^{\alpha,\mathrm{YB}}, U^{\alpha,\mathrm{YB}})$ requires only $\mathcal{O}(N \min\{n^{2/3}, N\})$ steps.

Lemma 3.2. *The cardinality of $\{\hat{p}(x_i) \colon 1 \le i \le n\}$ is smaller than $3n^{2/3}$.*

4 Theoretical properties of the confidence bands

This section illustrates consistency and adaptivity properties of the confidence band $(L_n^{\alpha,\mathrm{raw}}, U_n^{\alpha,\mathrm{raw}})$, where the subscript n indicates the sample size, and we consider a triangular scheme of observations $(x_i, Y_i) = (x_{ni}, Y_{ni})$, $1 \le i \le n$. We are interested in situations in which the observed covariates x_{ni} could be the realizations of the order statistics of a random sample. Thus we have to extend the framework of Yang and Barber (2019) and consider the following assumption.

(A) Let $\mathrm{Leb}(\cdot)$ denote Lebesgue measure, and let $W_n(B) = \sum_{i=1}^n \mathbb{1}(x_{ni} \in B)$ for $B \subset [0,1]$. There exist constants $C_1, C_2 > 0$ such that for sufficiently large n,

$$W_n(B) \ge C_1 n \mathrm{Leb}(B)$$

for arbitrary intervals $B \subset [0,1]$ such that $\mathrm{Leb}(B) \ge C_2 \log(n)/n$.

This assumption comprises the setting of Yang and Barber (2019). Let G be a differentiable distribution function on $[0,1]$ such that G' is bounded away from 0. If $x_{ni} = G^{-1}(i/n)$ for $1 \le i \le n$, then (A) is satisfied for any $C_1 < \inf_{[0,1]} G'$ and arbitrary $C_2 > 0$. The arguments in Mösching and Dümbgen (2020, Section 4.3) can be modified to show that if $x_{n1}, \ldots, x_{nn}$ are the order statistics of n independent random variables with distribution function G, then Condition (A) is satisfied almost surely, provided that $C_1, C_2 > 0$ are chosen appropriately.

Theorem 4.1. *Suppose that condition (A) is satisfied. Let $\rho_n = \log(n)/n$.*
(i) Suppose that p is constant on some non-degenerate interval $[a,b] \subset [0,1]$. Then

$$\sup_{x \in [a,b']} \bigl\{U_n^{\alpha,\mathrm{raw}}(x) - p(x)\bigr\}^+ + \sup_{x \in [a',b]} \bigl\{p(x) - L_n^{\alpha,\mathrm{raw}}(x)\bigr\}^+ = \mathcal{O}_p(\rho_n^{1/2})$$

for any fixed interval $[a',b'] \subset (a,b)$.
(ii) Suppose that p is Lipschitz-continuous on a non-degenerate interval $[a,b] \subset [0,1]$. Then

$$\sup_{x \in [a, b-\rho_n^{1/3}]} \bigl\{U_n^{\alpha,\mathrm{raw}}(x) - p(x)\bigr\}^+ + \sup_{x \in [a+\rho_n^{1/3}, b]} \bigl\{p(x) - L_n^{\alpha,\mathrm{raw}}(x)\bigr\}^+ = \mathcal{O}_p(\rho_n^{1/3}).$$

(iii) Suppose that for some constant $\gamma \ge 1$, $p(x) = \mathcal{O}(x^\gamma)$ as $x \to 0$. Then,

$$\sup_{x \in [0,1]} \frac{\mathbb{E}\{U_n^{\alpha,\mathrm{raw}}(x)\}}{x^\gamma + \rho_n^{1/2}} = \mathcal{O}(1).$$

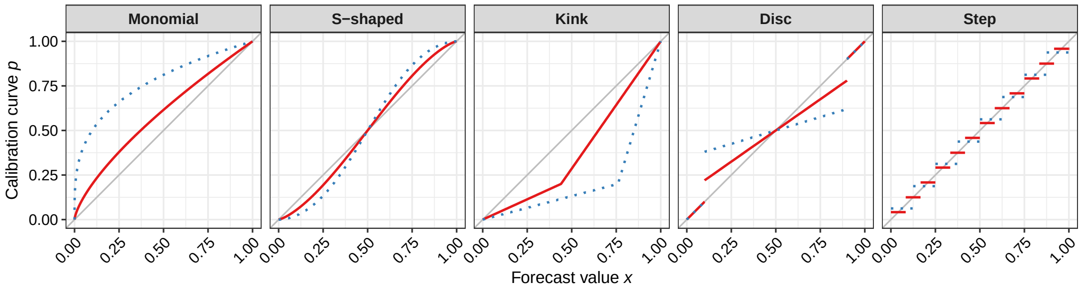


Figure 2: Illustration of the five simulated calibration curves $p_s(\cdot)$, where the solid red line corresponds to the shape parameter value $s = 0.3$ and the dashed blue line to $s = 0.7$.

An analogous statement holds for the lower bound $L_n^{\alpha,\mathrm{raw}}$.
(iv) Suppose that p *is discontinuous at some point* $x_o \in (0,1)$*. Then for any number* q *strictly between* $p(x_o-)$ *and* $p(x_o+)$ *and a suitable constant* $C > 0$,

$$U_n^{\alpha,\mathrm{raw}}(x_o - C\rho_n) < q < L_n^{\alpha,\mathrm{raw}}(x_o + C\rho_n)$$

with asymptotic probability one as $n \to \infty$.

Parts (i-ii) of this theorem are analogous to the results of Yang and Barber (2019, Sections 4.4 and 4.6). Part (iii) demonstrates that our bounds are particularly accurate in regions where $p(x)$ is close to 0 or 1. Presumably, the conclusions in parts (iii-iv) are not satisfied for the confidence band $(L_n^{\alpha,\mathrm{YB}}, U_n^{\alpha,\mathrm{YB}})$.

5 Simulations

Here, we illustrate that our calibration bands have correct coverage in the sense of (1) and are narrower than existing techniques. We consider both, the raw method in (4) and (5) and the non-crossing variant in (6). Both methods are combined with the rounding technique to three digits after the comma as described in the end of Section 2 in order to facilitate faster computation at a minimal cost in accuracy. For comparison, we use the isotonic bands of Yang and Barber (2019) given in (7) and (8) with a minimal variance factor of $\sigma^2 = 1/4$ and the parametric bands of Nattino et al. (2014), implemented in the `GivitiR` package in the statistical software `R` (R Core Team, 2022). Replication material for the simulations and applications is available under https://github.com/marius-cp/replication_DDHPZ22.

We use 1000 replications, a significance level of $\alpha = 0.05$ and simulate the predictions $X \sim \mathrm{U}[0,1]$. The binary outcomes are generated by $Y \sim \mathrm{Bern}\{p_s(X)\}$ based on five distinct functional forms of the calibration curve $p_s(x)$ for $x \in [0,1]$ depending on a shape parameter $s \in \mathcal{S} := \{0, 0.1, \ldots, 1\}$. All specifications of $p_s(x)$ satisfy the isotonicity assumption in (2) and they cover smooth, non-smooth as well as discontinuous setups. The choice $s = 0$ results in perfectly calibrated forecasts with $p_0(x) = x$ whereas the misscalibration increases with s. In particular, we consider the following specifications, which are illustrated in Figure 2 for two exemplary shape values $s \in \{0.3, 0.7\}$.

1. Monomial: First, we use a calibration curve defined by $p_s(x) = x^{1-s}$, where $s \in \mathcal{S} \setminus \{1\}$. This is already used in the simulations assessing the CORP reliability diagram in Dimitriadis et al. (2021, Appendix A).

2. S-shaped: Second, the calibration curve follows an S-shaped form $p_s(x) = \big(1 + ((1-x)/x)^{1+s}\big)^{-1}$, where $s \in \mathcal{S}$ pronounces the curves for larger values of s.

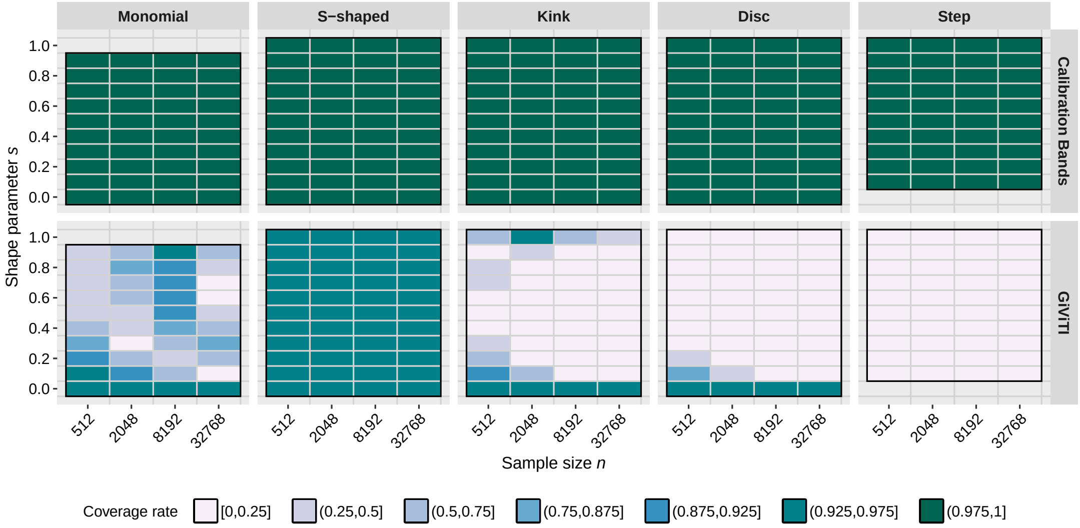


Figure 3: Empirical coverage rates of our calibration bands and the GiVitI bands for $1 - \alpha = 0.95$ averaged over all forecast values $x \in [0, 1]$, for the five specifications of the calibration curve $p_s(\cdot)$, different shape values s and a range of sample sizes n. Notice that the choices $s = 1$ in the monomial, and $s = 0$ in step specification are not defined.

3. Kink: Third, $p_s(x)$ linearly interpolates the points $(0, 0), (0.2 + 0.8s, 0.2)$ and $(1, 1)$ for $s \in \mathcal{S}$, resulting in a kink at the point $(0.2 + 0.8s, 0.2)$ for all $s > 0$.

4. Disc: Fourth, we have a perfect calibration $p_s(x) = x$ close to the borders $x \notin (0.1, 0.9)$, and a rotating, miscalibrated disc, $p_s(x) = (1 - s)x + s/2$ within $x \in [0.1, 0.9]$, where the rotation increases with $s \in \mathcal{S}$.

5. Step: Fifth, we use a step function with $s^\star \in \{5, 6, \ldots, 14\}$ equidistant steps in the unit interval. Formally, it is given by $p_s(x) = \big\{\lfloor s^\star x\rfloor + \mathbb{1}(x \neq 1)\big\}/s^\star$, where $s^\star = 15 - 10s$ and $s \in \mathcal{S} \setminus \{0\}$. Notice that this choice does not nest a correctly specified model, but its misspecification still increases with s.

Figure 3 presents the average coverage rates for a range of sample sizes between 512 and 32 768. We find that, as predicted by our theory, our calibration bands have conservative coverage throughout all simulation setups and sample sizes. We observe coverage rates between 0.998 and 1, with the majority of 179 out of the 212 displayed coverage values being exactly one. We dispense with a presentation of the coverage rates of the non-crossing and Yang and Barber (2019) bands, as both are guaranteed to be larger by Theorem 3.1. The non-crossing bands differ from the raw ones in less than one out of a hundred thousand simulated forecast values. These deviations occur exclusively for large values of s in the Step and Disc specifications within constancy regions of the calibration curve p.

The parametric bands of Nattino et al. (2014) rarely achieve correct coverage rates unless in the cases $s = 0$ and for the S-shaped calibration curve. This can be explained as these bands are based on the assumption of a certain parametric form of $p_s(x)$, which is rarely satisfied. The results get worse for the non-smooth and the two discontinuous specifications.

Figure 4 displays the average widths of the non-crossing and Yang and Barber (2019) bands. We present the theoretically wider non-crossing bands instead of the raw versions thereof. Their average widths is however non-distinguishable in these displays. We fix a medium degree of miscalibration $s = 0.5$. The upper plot panel displays the widths averaged over all simulation runs and forecast values depending on the sample size n. We find that the size of both bands

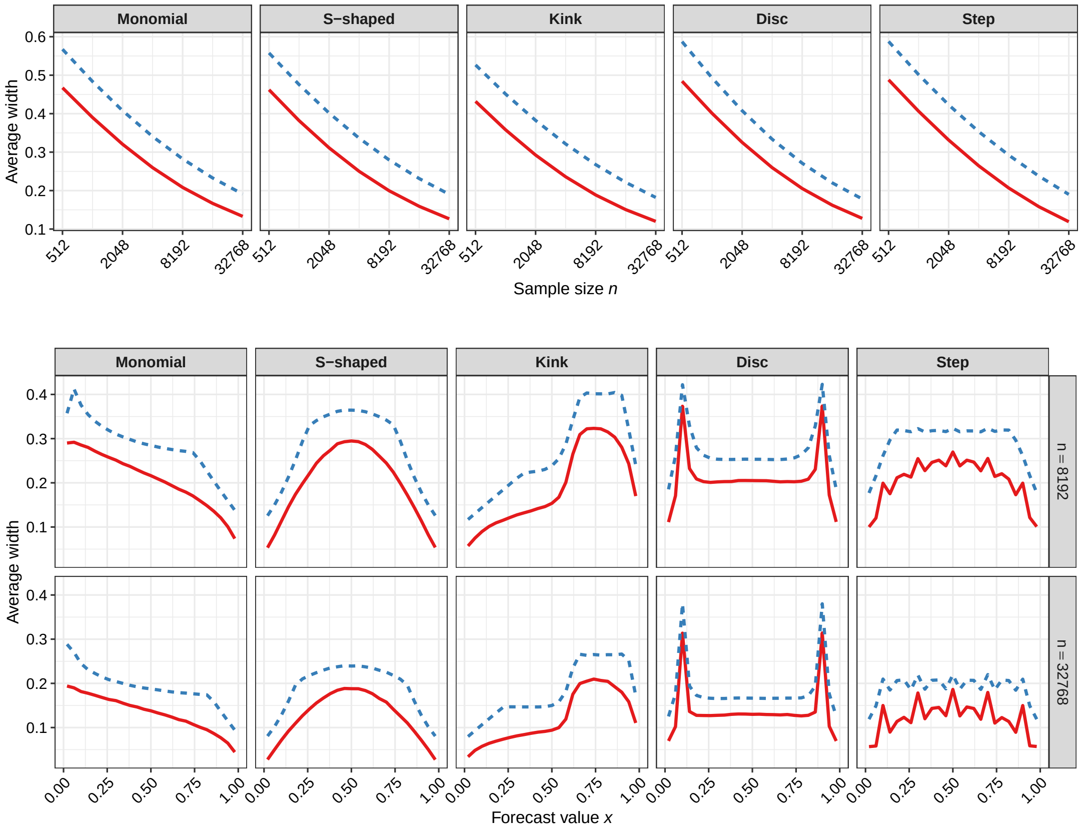


Figure 4: Top: Average empirical widths of the 95% confidence bands by sample size for the non-crossing and Yang and Barber (2019) bands for each of the five specifications of $p_s(x)$ given in the main text for a fixed degree of misspecification $s = 0.5$. Bottom: Average empirical widths by forecast value x for two sample sizes. In both panels, the solid red line corresponds to the non-crossing bands and the dashed blue line to the Yang and Barber (2019) bands.

shrinks with n and that we can reconfirm the ordering established in Theorem 3.1. We further see that our bands are only narrow enough for practical use in large samples. The relative gain in width of our bands is the highest for large sample sizes, exactly for which we propose the application of our method for calibration validation. It is worth noting that the bands of Yang and Barber (2019) are more generally valid than for the special case of binary observations.

The lower plot panel shows the widths averaged over the simulation replications, but depending on the forecast value x for two selected sample sizes. It shows that the relative gains in width upon the bands of Yang and Barber (2019) are particularly pronounced close to the edges of the unit interval. These regions of predicted probabilities close to zero or one are often of the highest interest in assessing calibration as for example in the subsequent application to low birth weight probability predictions.

6 Application: Predicting low birth weight probabilities

We apply our calibration bands to binary regressions predicting the probability of a fetus having a low birth weight, defined as weighting less than 2500 grams at birth (World Health Organization, 2015). We use U.S. Natality Data from the National Center for Health Statistics (2017), which provides demographic and health data for 3 864 754 births in the year 2017. For the data set at

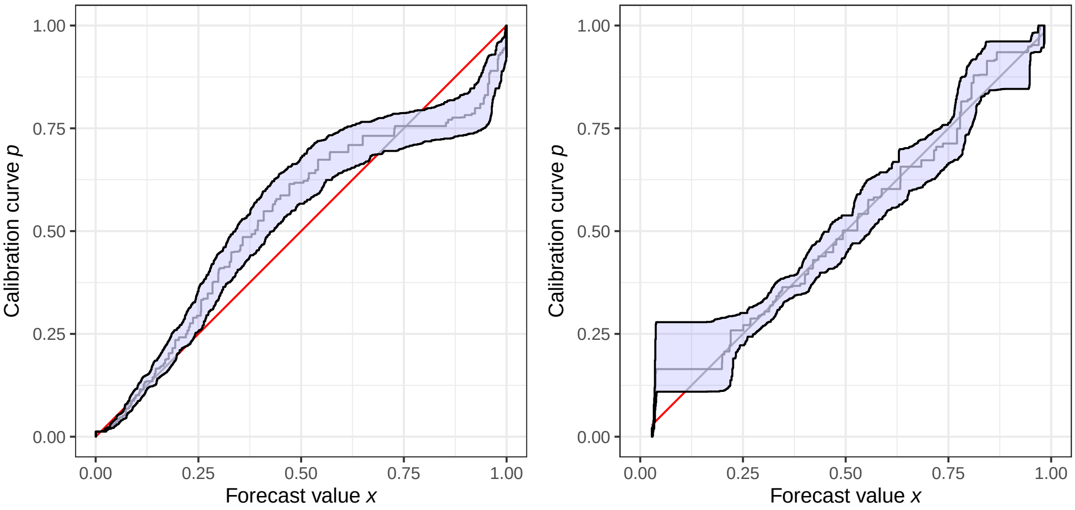


Figure 5: Calibration bands for the second model specification on the left and for the third specification on the right for the low birth weight application. The blue band denotes the calibration band and the grey step function the isotonic regression estimate. The bisector is given in red color whenever it is not contained in the calibration band.

hand, a low birth weight is observed in 8.1% of the cases.

We estimate three binary regression models by maximum likelihood on the same randomly drawn subset that contains all but 1 000 000 observations that we leave for external model validation. All three models contain standard risk factors such as the mother's age, body mass index and smoking behavior but they differ as follows. The first model uses a probit link function, and the explanatory variable week of gestation is categorized into four left-closed and right-open intervals with lower interval limits of 0, 28, 32 and 37 weeks, pertaining to the standard definitions of the World Health Organization of extremely, very, moderate and non preterm (Quinn et al., 2016). Through this categorization, the model specification can capture the week of gestation in a non-linear fashion. In contrast, the second model uses the week of gestation as a continuous explanatory variable and the third specification employs the cauchit instead of the probit link function, which is known to produce less confident predictions close to zero and one (Koenker and Yoon, 2009). Additional details of the model specifications are given in Appendix 1.

The classical Hosmer-Lemeshow test rejects perfect calibration of all three models with p-values of essentially zero for both, internal and external model validation, which leaves an applied researcher without any useful conclusions on model calibration. We show our calibration bands for the first model in Figure 1 and for the other two model specifications in Figure 5. We use the non-crossing method with rounding to three digits with a confidence level of $1 - \alpha = 95\%$.

For the first model, the calibration bands encompass the bisector for all forecast values, meaning that we cannot reject the null hypothesis of perfect calibration at the 5% level. More importantly, we are 95% certain that the true calibration curve lies within the the band at any point $x \in [0, 1]$, implying that we are confident that the model is at least as well calibrated as specified by the band. This is especially notable in the important region of predictions below 10% in the magnified right panel of Figure 1, where the confidence bands are remarkably close to the bisector implying a particularly well calibrated model. E.g., we are confident to conclude that a prediction of $x = 5\%$ occurs on average with a probability between 4.6% and 6.7%.

In contrast, we reject calibration for both, the second and third model specifications as shown in Figure 5. However, these bands are much more informative than a simple test rejection as they directly show the exact form of model miscalibration. E.g., for the second model specification, we can conclude that the predicted probabilities are particularly miscalibrated for values larger than 20% whereas the third specification entails miscalibrated probabilities for predictions below 10% that are presumably of the highest importance for medical decision making. While small predicted values from the second specification might still be treated as relatively reliable, they should be interpreted with great caution when stemming from the third model. The wider bands for the third model specification between predicted probabilities of 5% and 20% are caused by relatively little predictions in this interval.

Acknowledgement

A. Henzi and J. Ziegel gratefully acknowledge financial support from the Swiss National Science Foundation.

A Model specifications in the low birth weight application

We give some additional details on the model specifications of the application here. The first two models are based on the probit link function whereas the third one uses the cauchit link function (Koenker and Yoon, 2009). The second model uses the week of gestation as a continuous variable whereas the first and third models use the week of gestation as a categorical variable with left-closed and right-open intervals with lower interval limits of 0, 28, 32 and 37 weeks, which corresponds to the standard categorization of the World Health Organization (Quinn et al., 2016).

Additionally, all three models contain the following common explanatory variables: the mother's age and its squared term, her body mass index prior to pregnancy, her smoking behavior as a categorical variable with left-closed and right-open intervals with lower limits of 0, 1, 9, and 20 cigarettes per day averaged over all three trimesters, individual binary variables for mother's diabetes, any form of hypertension, mother's education below or equal to eight years, employed infertility treatments, a cesarean in a previous pregnancy, a preterm birth in a previous pregnancy, current multiple pregnancy, the sex of the unborn child, and an infection of one of the following: gonorrhea, syphilis, chlamydia, hepatitis b, hepatitis c. Additional details on the data are given in the user guide under https://data.nber.org/natality/2017/natl2017.pdf.

B Proofs and Technical Lemmas

Lemma B.1. *Let $Y_1, \dots, Y_m$ be independent Bernoulli variables with expectations $p_1 \leq \cdots \leq p_m$, and let $Z = Y_1 + \cdots + Y_m$. Then for any $\delta \in (0,1)$,*

$$\mathbb{P}\{p_1 \leq u^\delta(Z,m)\} \geq 1-\delta \quad \textit{and} \quad \mathbb{P}\{p_m \geq \ell^\delta(Z,m)\} \geq 1-\delta.$$

Proof of Lemma B.1. For the upper bound, note that $u^\delta(z,m)$ is increasing in z. If $b = \min\{z \in \{0,\dots,m\} : u^\delta(z,m) \geq p_1\}$, then $\mathbb{P}\{p_1 \leq u^\delta(Z,m)\} = \mathbb{P}(Z \geq b)$. By Shaked and Shanthikumar (2007, Example 1.A.25), Z is stochastically larger than $\tilde{Z}$ with binomial distribution with parameters m and p_1, so $\mathbb{P}(Z \geq b) \geq \mathbb{P}(\tilde{Z} \geq b) \geq 1-\delta$, where the last inequality follows from the validity of the Clopper-Pearson confidence bounds. The proof for the lower bound is similar. □

The proof of Theorem 3.1 uses standard results for isotonic least squares regression and the following inequalities of Hoeffding (1963, Theorem 1).

Lemma B.2. *Let $Y_1, Y_2, \ldots, Y_m$ be independent random variables with values in $[0,1]$ and expectations $p_1, p_2, \ldots, p_m$. Suppose that $q = m^{-1}\sum_{i=1}^m p_i \in (0,1)$, and set $\hat{q} = m^{-1}\sum_{i=1}^m Y_i$. Then for arbitrary $r \in [0,1]$,*

$$\begin{aligned} \mathbb{P}(\hat{q} \le r) &\le \exp\{-mK(r,q)\} \le \exp\{-2m(r-q)^2\} \quad \textit{if } r \le q, \\ \mathbb{P}(\hat{q} \ge r) &\le \exp\{-mK(r,q)) \le \exp\{-2m(r-q)^2\} \quad \textit{if } r \ge q, \end{aligned}$$

where $K(r,q) := r\log(r/q) + (1-r)\log[(1-r)/(1-q)]$.

Corollary 1. *For integers $m \ge 1$, $z \in \{0,1,\ldots,m\}$ and any number $\delta \in (0,1)$,*

$$\begin{aligned} u^\delta(z,m) &\le \max\bigl\{\xi \in [\hat{q},1] : K(\hat{q},\xi) \le \log(1/\delta)/m\bigr\} \le \hat{q} + \sqrt{\log(1/\delta)/(2m)}, \\ \ell^\delta(z,m) &\ge \min\bigl\{\xi \in [0,\hat{q}] : K(\hat{q},\xi) \le \log(1/\delta)/m\bigr\} \ge \hat{q} - \sqrt{\log(1/\delta)/(2m)}, \end{aligned}$$

where $\hat{q} = z/m$.

In addition, the proof of Theorem 3.1 makes use of the following proposition which is of independent interest, since it implies a more efficient method for computing the bounds of Yang and Barber (2019).

Proposition B.3. *For an arbitrary observation vector $\mathcal{Y} \in \mathbb{R}^n$, let $\hat{p}\colon [0,1] \to \mathbb{R}$ be an increasing function minimizing $\sum_{i=1}^n \{Y_i - \hat{p}(x_i)\}^2$. For some $\tau > 0$ and any index $1 \le i \le n$, let*

$$U_i = \min_{(j,k)\in\mathcal{J}\colon x_j \ge x_i} \Bigl(\frac{Z_{jk}^{\mathrm{iso}}}{n_{jk}} + \frac{\tau}{\sqrt{n_{jk}}}\Bigr), \qquad L_i = \max_{(j,k)\in\mathcal{J}\colon x_k \le x_i} \Bigl(\frac{Z_{jk}^{\mathrm{iso}}}{n_{jk}} - \frac{\tau}{\sqrt{n_{jk}}}\Bigr).$$

Then, the minimum for U_i is attained at some $(j,k) \in \mathcal{J}$ such that $j = \min(s\colon x_s \ge x_i)$ and $\hat{p}(x_k) < \hat{p}(x_{k+1})$ or $k = n$. The maximum for L_i is attained at some $(j,k) \in \mathcal{J}$ such that $\hat{p}(x_{j-1}) < \hat{p}(x_j)$ or $j = 1$ and $k = \max(s\colon x_s \le x_i)$.

Proof of Proposition B.3. Consider the statement about U_i. The claim about j follows from the fact that for fixed k, $Z_{jk}^{\mathrm{iso}}/n_{jk}$ is increasing and $n_{jk} = n - j + k$ is decreasing in $j \le k$. As to the upper index k, note that U_i is the minimum of $u_{jk} = Z_{jk}^{\mathrm{iso}} n_{jk}^{-1} + \tau n_{jk}^{-1/2}$ over all $k \ge j = \min(s : x_s \ge x_i)$ such that $(j,k) \in \mathcal{J}$. Let $j \le k_1 < k_2$ be indices such that $\hat{p}(x_k) = \hat{q}$ for $k_1 < k \le k_2$. Then, for $k_1 \le k \le k_2$,

$$Z_{jk}^{\mathrm{iso}} = Z_{jk_1}^{\mathrm{iso}} + (k - k_1)\hat{q} = B + n_{jk}\hat{q}$$

with

$$B = Z_{jk_1}^{\mathrm{iso}} - n_{jk_1}\hat{q} \begin{cases} \le 0, & \\ = 0 & \text{if } \hat{p}(x_j) = \hat{q}. \end{cases}$$

Consequently, for $k_1 \le k \le k_2$,

$$u_{jk} = \hat{q} + Bn_{jk}^{-1} + \tau n_{jk}^{-1/2}$$

is a concave function of $n_{jk}^{-1} \in [n_{jk_2}^{-1}, n_{jk_1}^{-1}]$, and it is increasing in n_{jk}^{-1} if $\hat{q} = \hat{p}(x_j)$. This implies that

$$u_{jk} \ge \begin{cases} \min(u_{jk_1}, u_{jk_2}), & \\ u_{jk_2} & \text{if } \hat{q} = \hat{p}(x_j). \end{cases}$$

Consequently, the minimum of u_{jk} over all $k \ge j$ is attained at some $k \ge j$ such that $\hat{p}(x_k) < \hat{p}(x_{k+1})$ or $k = n$, and this entails that $(j,k) \in \mathcal{J}$. The statement about L_i follows from the one about U_i when $x_1, \ldots, x_n$ are replaced by $1 - x_n, \ldots, 1 - x_1$ and $Y_1, \ldots, Y_n$ by $-Y_n, \ldots, -Y_1$. □

Proof of Theorem 3.1. The inequalities $L_i^{\alpha,\mathrm{nc}} \le L_i^{\alpha,\mathrm{raw}}$ and $U_i^{\alpha,\mathrm{raw}} \le U_i^{\alpha,\mathrm{nc}}$, as well as $L_i^{\alpha,\mathrm{YB}} \le \hat{p}(x_i) \le U_i^{\alpha,\mathrm{YB}}$ hold by construction. It is therefore sufficient to show that $L_i^{\alpha,\mathrm{YB}} \le L_i^{\alpha,\mathrm{raw}}$ and $U_i^{\alpha,\mathrm{raw}} \le U_i^{\alpha,\mathrm{YB}}$. As to the inequality $U_i^{\alpha,\mathrm{raw}} \le U_i^{\alpha,\mathrm{YB}}$, we know that $U_i^{\alpha,\mathrm{YB}}$ equals

$$u_{jk}^{\mathrm{YB}} = Z_{jk}^{\mathrm{iso}} n_{jk}^{-1} + \tau n_{jk}^{-1/2}$$

for some $(j,k) \in \mathcal{J}$ with $j = \min\{s : x_s \ge x_i\}$ and $\hat{p}(x_k) < \hat{p}(x_{k+1})$ or $k = n$, where $\tau = \sqrt{\log\{(N^2+N)/\alpha\}/2}$. As explained later, this implies that

$$Z_{jk} \le Z_{jk}^{\mathrm{iso}} \quad \text{if } \hat{p}(x_k) < \hat{p}(x_{k+1}) \text{ or } k = n. \tag{10}$$

But then it follows from Corollary 1 that $U_i^{\alpha,\mathrm{YB}} = u_{jk}^{\mathrm{YB}}$ is greater than or equal to

$$Z_{jk} n_{jk}^{-1} + \tau n_{jk}^{-1/2} \ge u^{\alpha/(N^2+N)}(Z_{jk}, n_{jk}) \ge U_i^{\alpha,\mathrm{raw}}.$$

Inequality (10) follows from a standard result about isotonic regression (see for example Henzi et al., 2022, Characterization II). The index interval $\{j,\ldots,k\}$ may be partitioned into index intervals $\{\ell,\ldots,m\} = \{j,\ldots,n\} \cap \{s : \hat{p}(x_s) = \hat{q}\}$, where $\hat{q}$ is any value in $\{\hat{p}(x_j),\ldots,\hat{p}(x_k)\}$. For such an index interval, $Z_{\ell m} \le Z_{\ell m}^{\mathrm{iso}}$, with equality if $\hat{q} > \hat{p}(x_j)$.

The inequality for the lower bound follows from the one for the upper bound when $x_1,\ldots,x_n$ are replaced by $1-x_n,\ldots,1-x_1$ and $Y_1,\ldots,Y_n$ by $1-Y_n,\ldots,1-Y_1$. □

Proof of Lemma 3.2. Let $\hat{q}_1 < \cdots < \hat{q}_b$ be the different elements of $\{\hat{p}(x_i) \colon 1 \le i \le n\}$, where we assume that $b \ge 2$. There exists a partition of $\{1,\ldots,n\}$ into index intervals $I_1,\ldots,I_b$ such that $\hat{q}_\ell = |I_\ell|^{-1} \sum_{i \in I_\ell} Y_i$. For any integer $d \ge 1$, let M_d be the number of indices ℓ such that $|I_\ell| = d$. Since $\sum_{i\in I_\ell} Y_i \in \{0,1,\ldots,d\}$, the numbers M_d satisfy the following constraints: $M_d \in [0, d+1]$, and $\sum_{d=1}^n M_d d = n$. The question is, how large the number $b = \sum_{d=1}^n M_d$ can be under these constraints, where we drop the restriction that the M_d are integers. Suppose that $M_c < c+1$ and $M_{c'} > 0$ for integers $1 \le c < c'$. Then we may replace $(M_c, M_{c'})$ with $(M_c + \gamma/c, M_{c'} - \gamma/c')$, where γ is the minimum of $(c+1-M_c)c$ and $M_{c'}c'$. This does not affect the constraints, but the sum $\sum_{d=1}^n M_d$ increases strictly, while $M_c = c+1$ or $M_{c'} = 0$. Eventually, we obtain an integer $d_o \ge 1$ such that $M_d = d+1$ if $1 \le d \le d_o$ and $M_d = 0$ for $d \ge d_o + 2$. In particular,

$$n \ge \sum_{d=1}^{d_o} (d+1)d = \frac{(d_o+2)(d_o+1)d_o}{3} > \frac{d_o^3}{3},$$

whence $d_o < (3n)^{1/3}$, while

$$b \le \sum_{d=1}^{d_o+1} (d+1) = \frac{d_o(d_o+3)}{2} \le C n^{2/3},$$

where $C = 3^{2/3}(1 + 3/6^{1/3})/2 < 3$. □

For the proof of Theorem 4.1, we need an inequality for the auxiliary function $K(\cdot,\cdot)$ in Lemma B.2 which follows from Dümbgen (1998, Proposition 2.1).

Lemma B.4. *For arbitrary $q \in [0,1]$, $\xi \in (0,1)$ and $\gamma > 0$, the inequality $K(q,\xi) \le \gamma$ implies that*

$$|\xi - q| \le \sqrt{2\gamma q(1-q)} + |1-2q|\gamma.$$

Proof of Theorem 4.1. For notational convenience, we often drop the additional subscript n, e.g. we write x_i instead of x_{ni}. For symmetry reasons, it suffices to verify the assertions about $U^{\alpha,\mathrm{raw}}$.

In what follows, let C be a generic (large) constant which does not depend on n, but the value of which may change in each instance. It follows from Corollary 1 and Lemma B.4 that simultaneously for all $(j,k) \in \mathcal{J}$,

$$u^{\alpha/(N^2+N)}(Z_{jk}, n_{jk}) \le \hat{p}_{jk} + C \min\biggl\{ \sqrt{\frac{\log(n)\hat{p}_{jk}(1-\hat{p}_{jk})}{n_{jk}}} + \frac{\log(n)}{n_{jk}}, \sqrt{\frac{\log(n)}{n_{jk}}} \biggr\}, \tag{11}$$

where $\hat{p}_{jk} = Z_{jk}/n_{jk}$. Moreover, one can deduce from Lemma B.2 that simultaneously for all $(j,k) \in \mathcal{J}$,

$$\hat{p}_{jk} \le p_{jk} + C\sqrt{\frac{\log(n)}{n_{jk}}} \tag{12}$$

with asymptotic probability one, where $p_{jk} = \mathbb{E}(\hat{p}_{jk}) = n_{jk}^{-1}\sum_{i=j}^{k} p_i \in [p_j, p_k]$.

As to part (ii), let $B(x) = [x, x+\rho_n^{1/3}]$ for $x \in [a, b-\rho_n^{1/3}]$. For sufficiently large n, the length $\rho_n^{1/3}$ of these intervals is greater than $C_2\rho_n$, so it follows from assumption (A) that $B(x) \cap \{x_1, \ldots, x_n\} = \{x_{j(x)}, \ldots, x_{k(x)}\}$ with $(j(x), k(x)) \in \mathcal{J}$ satisfying

$$n_{j(x)k(x)} = W_n\{B(x)\} \ge C_1 n \rho_n^{1/3}.$$

Consequently, $\log(n)/n_{j(x)k(x)} \le C_1^{-1}\rho_n^{2/3}$, so we may deduce from inequalities (11), (12) and Lipschitz-continuity of p on $[a,b]$ that with asymptotic probability one, simultaneously for all $x \in [a, b-\rho_n^{1/3}]$,

$$\begin{aligned} U_n^{\alpha,\mathrm{raw}}(x) \le u^{\alpha/(N^2+N)}(Z_{j(x)k(x)}, n_{j(x)k(x)}) &\le \hat{p}_{j(x)k(x)} + C\rho_n^{1/3}, \\ \hat{p}_{j(x)k(x)} &\le p_{j(x),k(x)} + C\rho_n^{1/3}, \\ p_{j(x)k(x)} &\le p(x) + C\rho_n^{1/3}. \end{aligned}$$

Clearly, this implies the assertion about $U^{\alpha,\mathrm{raw}}$ in part (ii).

Part (i) can be verified similarly. With $\delta = b - b' > 0$, let $B(x) = [x, x+\delta]$ for $x \in [a, b']$. Now it follows from assumption (A) that for sufficiently large n, $B(x) \cap \{x_1, \ldots, x_n\} = \{x_{j(x)}, \ldots, x_{k(x)}\}$ with $(j(x), k(x)) \in \mathcal{J}$ satisfying $n_{j(x)k(x)} \ge C_1 n\delta$, uniformly for all $x \in [a, b']$. Now it follows from inequalities (11), (12) and p being constant on $[a,b]$ that with asymptotic probability one, simultaneously for all $x \in [a, b']$,

$$\begin{aligned} U_n^{\alpha,\mathrm{raw}}(x) \le u^{\alpha/(N^2+N)}(Z_{j(x)k(x)}, n_{j(x)k(x)}) &\le \hat{p}_{j(x)k(x)} + C\rho_n^{1/2}, \\ \hat{p}_{j(x)k(x)} &\le p_{j(x),k(x)} + C\rho_n^{1/2}, \\ p_{j(x)k(x)} &= p(x). \end{aligned}$$

This implies the assertion about $U^{\alpha,\mathrm{raw}}$ in part (i).

As to part (iii), it suffices to show that $\mathbb{E}\{U_n^{\alpha,\mathrm{raw}}(x_n)\} \le C(x_n^\gamma + \rho_n^{1/2})$ for any sequence of numbers $x_n \in [0,1]$ converging to 0. Let $t_n = \max(x_n, \rho_n^{1/2})$ and $B_n = [t_n, 2t_n]$. For sufficiently large n, $\mathrm{Leb}(B_n) \ge \rho_n^{1/2} \ge C_1\rho_n$, so $B_n \cap \{x_1, \ldots, x_n\} = \{x_{j_n}, \ldots, x_{k_n}\}$ with $(j_n, k_n) \in \mathcal{J}$ satisfying

$$n_{j_n,k_n} = W_n(B_n) \ge C_1 n t_n.$$

In particular, $\log(n)/n_{j_n k_n} \le C_1^{-1}\rho_n/t_n$, and the assumption that $p(x) = \mathcal{O}(x^\gamma)$ as $x \to 0$ implies that $p_{j_n k_n} = \mathcal{O}(t_n^\gamma)$. Hence, it follows from (11) that

$$\begin{aligned}
\mathbb{E}\{U_n^{\alpha,\mathrm{raw}}(x_n)\} &\le \mathbb{E}\{u^{\alpha/(N^2+N)}(Z_{j_n k_n}, n_{j_n k_n})\} \\
&\le \mathbb{E}\Big\{\hat{p}_{j_n k_n} + C\big(\sqrt{t_n^{-1}\rho_n \hat{p}_{j_n k_n}} + t_n^{-1}\rho_n\big)\Big\} \\
&\le p_{j_n k_n} + C\big(\sqrt{t_n^{-1}\rho_n p_{j_n k_n}} + t_n^{-1}\rho_n\big) \\
&= \mathcal{O}\big(t_n^\gamma + \rho_n^{1/2} t_n^{(\gamma-1)/2} + t_n^{-1}\rho_n\big) = \mathcal{O}_p(x_n^\gamma + \rho_n^{1/2}),
\end{aligned}$$

where the last inequality follows from Jensen's inequality.

To verify part (iv), let $B_n = [x_o - C_3\rho_n, x_o)$ for some $C_3 \ge C_2$. For sufficiently large n, $B_n \cap \{x_1, \ldots, x_n\} = \{x_{j_n}, \ldots, x_{k_n}\}$ with $(j_n, k_n) \in \mathcal{J}$ satisfying

$$n_{j_n k_n} \ge C_1 C_3 n\rho_n \ \ge \ C_1 C_3 \log(n) \quad \text{and} \quad p_{j_n k_n} \le p(x_o-) < q.$$

Consequently, $\log(n)/n_{j_n k_n} \le (C_1 C_3)^{-1}$ and thus with asymptotic probability one,

$$\begin{aligned}
U_n^{\alpha,\mathrm{raw}}(x_o - C_3\rho_n) \le u^{\alpha/(N^2+N)}(Z_{j_n k_n}, n_{j_n k_n}) &\le \hat{p}_{j_n k_n} + C C_3^{-1/2}, \\
\hat{p}_{j_n k_n} &\le p(x_o-) + C C_3^{-1/2}.
\end{aligned}$$

Consequently, $U_n^{\alpha,\mathrm{raw}}(x_o - C_3\rho_n) < q$ with asymptotic probability one, provided that C_3 is sufficiently large. □

References

Allison, P. J. (2014). Measures of fit for logistic regression. *Paper 1485-2014, SAS Global Forum 2014*, pages 1–12.

Bertolini, G., D'Amico, R., Nardi, D., Tinazzi, A., and Apolone, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of epidemiology and biostatistics*, 5:251–253.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118:e2016191118.

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics*, 26:288–314.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41:1892–1921.

Henzi, A., Moesching, A., and Dümbgen, L. (2022+). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodology and Computing in Applied Probability*. to appear.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9:1043–1069.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. Wiley Series in Probability and Statistics. Wiley, Hoboken, N.J, third edition.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, third edition.

Koenker, R. and Yoon, J. (2009). Parametric links for binary choice models: A Fisherian–Bayesian colloquy. *Journal of Econometrics*, 152:120–130.

Kramer, A. A. and Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The hosmer-lemeshow test revisited. *Critical care medicine*, 35:2052–2056.

Mösching, A. and Dümbgen, L. (2020). Monotone least squares and isotonic quantiles. *Electronic Journal of Statistics*, 14:24–49.

National Center for Health Statistics (2017). NCHS' Vital Statistics Natality Birth Data. https://data.nber.org/data/natality.html. Online; accessed 13 January 2021.

Nattino, G., Finazzi, S., and Bertolini, G. (2014). A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Statistics in Medicine*, 33:2390–2407.

Nattino, G., Pennell, M. L., and Lemeshow, S. (2020a). Assessing the goodness of fit of logistic regression models in large samples: A modification of the hosmer-lemeshow test. *Biometrics*, 76:549–560.

Nattino, G., Pennell, M. L., and Lemeshow, S. (2020b). Rejoinder to "assessing the goodness of fit of logistic regression models in large samples: A modification of the hosmer-lemeshow test". *Biometrics*, 76:575–577.

Paul, P., Pennell, M. L., and Lemeshow, S. (2013). Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, 32:67–80.

Quinn, J.-A., Munoz, F. M., Gonik, B., Frau, L., Cutland, C., Mallett-Moore, T., Kissou, A., Wittke, F., Das, M., Nunes, T., Pye, S., Watson, W., Ramos, A.-M. A., Cordero, J. F., Huang, W.-T., Kochhar, S., Buttery, J., and Brighton Collaboration Preterm Birth Working Group (2016). Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine*, 34(49):6047–6056.

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. (2020). Mitigating bias in calibration error estimation. *Preprint.* https://arxiv.org/abs/2012.08668.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *The Annals of Statistics*, 38(4):1953–1977.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders.* Springer Series in Statistics. Springer, New York.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P., and Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241.

Tutz, G. (2011). *Regression for Categorical Data.* Cambridge University Press, Cambridge.

World Health Organization (2015). *International statistical classification of diseases and related health problems.* World Health Organization. 10th revision, fifth edition. https://apps.who.int/iris/handle/10665/246208. Online; accessed 13 January 2021.

Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Annals of Statistics*, 9:443–448.

Yang, F. and Barber, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, 13:646–677.

Yu, B. and Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929.

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Henzi Alexander

Registration Number: 13-131-560

Study program: PhD in Statistics

Bachelor □ Master □ Dissertation ☑

Title of the thesis: Isotonic distributional regression

Supervisor: Prof. Dr. Johanna Ziegel

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of September 5th, 1996 and Article 69 of the University Statute of June 7th, 2011 is authorized to revoke the doctoral degree awarded on the basis of this thesis.

For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with theses submitted by others.

Bern, 15.03.2022

Place/Date

Signature AHenzi