Set-valued Data: Regression, Design and Outliers

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by
Qiyu Li
from China

Supervisor of the doctoral thesis:
Prof. Dr. Ilya Molchanov
Institute of Mathematical Statistics and Actuarial Science

Accepted by the Faculty of Science.

Bern, 5. March 2021

The Dean
Prof. Dr. Zoltan Balogh

Acknowledgement

First, I would like to thank heartily my supervisor Ilya Molchanov, whose encouragement, guidance and support made this thesis possible. This thesis consists of three research papers, written in collaboraruon with Francesca Molinari (Cornell University), Sida Peng (Microsoft) and Ignacio Cascos (Universidad Carlos III de Madrid). It has been my pleasure to work with and learn from them. Special thanks to Ana Colubi (University of Oviedo) for being the external examiner of this doctoral thesis.

This thesis was written while working as a teaching assistant at the Institute of Mathematical Statistics and Actuarial Science (IMSV) in University of Bern and as a senior biostatistician at the Swiss Group for Clinical Cancer Research (SAKK). I am thankful to my colleagues at both organizations for the pleasant working environments and intriguing conversations during the breaks. Special thanks to Lutz Dümbgen for his valuable advice, as well as to Dirk Klingbiel (my previous line manager at SAKK) and Stefanie Hayoz (my current line manager at SAKK) for their support and allowing me to use the clinical data of SAKK.

Finally, I would like to convey my deepest gratitude to my family for their encouragement and mental support.

Contents

1 Introduction **1**
- 1.1 Two examples of interval-valued data . . . 1
- 1.2 Partial identification and random set . . . 2
- 1.3 Goal and structure of this thesis . . . 3

2 Preliminaries **6**

3 Main results and discussions **8**
- 3.1 Local regression smoothers with set-valued outcome data . . . 8
- 3.2 Optimal design for multivariate multiple linear regression with set-identified response . . . 14
- 3.3 Random sets: depth and outliers . . . 18
  - 3.3.1 Band depth [Type A] . . . 19
  - 3.3.2 Simplicial depth [Type B] . . . 20
  - 3.3.3 Depth based on nonlinear expectations [Type C] . . . 20
  - 3.3.4 Half-space depth [Type D] . . . 23
  - 3.3.5 Other set-valued data . . . 23

Bibliography **28**

Appendix **29**
- **Paper A** . . . 29
- **Paper B** . . . 52
- **Paper C** . . . 62

Chapter 1

Introduction

1.1 Two examples of interval-valued data

Data in the form of intervals naturally appear in many contexts. Let us consider the snapshots of two data sets. Data set (a) presents the temperature on April 1, 2020 in different cities and (b) includes patients registered in a clinical study. The variables of interest are daily temperature in Data (a) and age at registration which was derived from the year of birth and registration date in Data (b). Comparing with conventional numerical data where one observes the exact value, in these examples temperature and age are presented as intervals. The reason of having such data is because the exact value could not be identified or on purpose converted to intervals. For the data set (a), the lowest and the highest temperatures form an interval which describes the variability of the temperature on a particular day. In order to protect the patients' privacy, the date of birth in (b) is not allowed to be collected in a clinical study, and so only the year of birth is available. Therefore, the age could not be exactly identified and we only know that the true age of each patient lies in a certain interval of one year length.

In these examples, the observed variables are said to be *partially identified*, while observations are said to be *point-identified* if their exact values are available. Comparing with the point-identified data where we can observe the realization of singleton-valued variable of interest, in the partially identified data one can only observe sets to which a realization of a variable of interest belongs.

Table 1.1: Data examples

(a)

City	Temperature (℃)
Paris	1 - 13
Berlin	2 - 9
Beijing	2 - 16
New York	4 - 13
Tokyo	9 - 13
⋮	⋮

(b)

ID	Year of birth	Registration date	Age at registration
1	1950	JAN 01, 2019	68-69
2	1967	MAR 31, 2019	51-52
3	1962	SEP 02, 2019	56-57
4	1983	DEC 18, 2019	35-36
5	1970	JAN 31, 2020	49-50
⋮	⋮	⋮	⋮

The setting is very close but still different from the case where the data are sets and one deals with a sample of sets. In case of partially identified models the underlying quantity of interest is the distribution of a point sampled from sets, while for set-valued data the distribution of a set is of primary importance. In the thesis we address statistical inference problems in both settings, which are closely related: the distribution of a set yields the distributions of points sampled from it, while the distribution of a point sampled from a set provides information about the distribution of the set itself. The framework goes beyond the interval-valued data and deals with general set-valued observations that are interpreted either as samples of sets or as partially identified observations of a multivariate parameter.

1.2 Partial identification and random set

The simplest way to handle partially identified data is to replace the observed interval with a real number, typically by its minimum, maximum or middle point, so that the conventional statistical methods can be used to estimate the parameter of interest and to perform statistical inference. However, this approach neglects the nature of the data-generation process that results in interval-identified or set-identified observations. With the onset of research on *partial identification* in the area of econometrics in the early 1990s (see, e.g., Manski [2003]), the nature of this data-generation process has been incorporated into the statistical analysis.

Partial identification is an approach to handle

1) data which is not an all-or-nothing (point-identified or missing) concept and

2) models which is not necessary point-identified (not all parameters could be point-identified) but can still provide valuable information.

In this case, the term "partial identification" is quite self-explaining. The variables age and daily temperature in the previous examples can be called *partially identified*. Note that the partial identification does not only deal with *incomplete data* (partially identified data as in the examples), but also concerns *incomplete models*, see Ciliberto and Tamer [2009], Tamer [2003]. The latter setting is out of scope of this thesis.

The main philosophy behind the partial identification is *the law of decreasing credibility*:

> "The credibility of inference decreases with the strength of the assumptions maintained", see Manski [2003].

The conventional method based on replacing observed sets with a single value is equivalent to adding an assumption that the set can be represented by a single value. This doesn't always make sense. The partial identification analysis suggests first focusing on what can be learned from the observed data without any assumption (except the basic restriction on the sampling process) and then combining the empirical evidence with plausible assumptions to study the effect of the assumptions on what one learns. With this procedure, one obtains first all possible values (a set of values) for the parameter of interest. This set may be reduced by strengthening the assumptions. In the ultimate

case, this set turns into a singleton. However, this should not be the ultimate goal, but we should rather focus on

- obtaining a useful characterization of the parameter of interest with available data and a set of plausible assumptions;
- estimating the set of values;
- conducting hypothesis testing and drawing conclusions.

For example, Manski [2003, 2007] considered learning the distribution function from the interval data and proposed to estimate it based on the worst case bounds. It is simple to understand and easy to compute. In another example in the context of linear models with interval data, sets of parameters that are consistent (in the Hausdorff metric) with the argmin of a particular objective function could be constructed, see Manski and Tamer [2002]. However, these papers did not discuss the inference issues.

In order to develop statistical inference techniques, Beresteanu and Molinari [2008] relied on the concept of a *random set* as a mathematical technique to handle the partial identification analysis in the linear regression with interval outcomes. In this paper, Beresteanu and Molinari considered the partially identified data as a set-valued sample drawn from the distribution of a random set and developed a whole statistical procedure involving the structure of the model, estimation of the linear regression parameter, the central limit theorem, hypothesis testing and construction of confidence regions.

The studies of random sets can be traced back to early 1930s where this concept was first mentioned, see Kolmogorov [1950]. Later, its mathematical theory including the exact definition of a random closed set and the relevant techniques was introduced by Matheron [1975] and later discussed in depth by Molchanov [2017]. Random set theory provides a relevant mathematical basis to partial identification analysis, such as distributions of random sets and their selections, operations with sets (Minkowski sum and union), limit theorems for Minkowski sums, involving important techniques from probabilities in Banach spaces.

One of the key ideas lies in interpreting a random set as a family of random singletons (or random vectors), and these random vectors are called *selections* of the random set. In many cases, it is possible to find a countable family of integrable selections to fill a random closed set, and then the expectation of the random closed set is defined as the set of expectations of all integrable selections; we call it the *selection expectation* or *Aumann expectation*. The practical calculation of the selection expectation is performed in terms of the *support function* of a random set, which measures the (signed) distance between its supporting hyperplanes and the origin. The Minkowski sum of random sets is defined as set of sums of all their selections. The support function uniquely identifies a convex closed set and therefore the Minkowski sum of convex closed sets can be equivalently obtained as the arithmetic sum of their support function. These are very powerful tools to make inference for set-valued data.

1.3 Goal and structure of this thesis

Taking these concepts as starting point, we are interested in

- How do numeric explanatory variables affect the set-valued outcome?
- How should we plan an experiment with set-valued outcome?
- How to identify the outlier in set-valued data?

To answer these questions, we studied three aspects of set-valued data, namely regression, optimal design and outlier identification. The first two aspects were motivated by Beresteanu and Molinari [2008] where linear regression with interval outcome have been thoroughly studied. In this thesis we want to investigate an extension of results on nonparametric regression to general set-valued data in $\mathbb{R}^d$. Beside the modelling, we also interested in using such models to design an experiment. Taking the linear model in Beresteanu and Molinari [2008], we focus on identifying the location of design points which ensure the best properties of the unbiased estimator of the parameter.

Finally, we deal directly with set-valued samples. Generally speaking, there are two reasons of having set-valued data. One reason is the partial identification problem where only one selection of the observed set from each observation is the true value but it is impossible to identify which selection (as age in Data (b)). The other is that due to the nature of the observation only set can be taken as value, such as daily temperature (in Data (a)). Further details regarding various interpretation of set-valued data could be found in Couso and Dubois [2014].

Particles, like stones or sand grains (see Stoyan and Stoyan [1994]) provide another source of set-valued data. Statistics of particles is different from the inference for interval-valued data like daily temperature, since statistical inference for particles should be invariant with respect to their positions and rotations, and, possibly, scaling. Regardless of the reason of having set-valued data, we aims to explore possible way to identify outliers in samples of sets.

This thesis consists of three papers (see Appendix), each of them addressing one question above.

Paper A Q. Li, I. Molchanov, F. Molinari, S. Peng: Local regression smoothers with set-valued outcome data. *International Journal of Approximate Reasoning*, **128**: 129-150, 2021.

Paper B Q. Li and I. Molchanov: Optimal design for multivariate multiple linear regression with set-identified response. *Statistical Planning and Inference*, **203**: 215-223, 2019.

Paper C I. Cascos, Q. Li, I. Molchanov: Depth and outliers for samples of sets and random sets distributions. *Australian and New Zealand Journal of Statistics*, `https://doi.org/10.1111/anzs.12326`, 2021

In the next chapter we introduce the notation and recall some important concepts from the theory of random sets. In Chapter 3 we provide an overview of results from these three papers with unified notation from Chapter 2, so that it may differ a bit from the papers. In Section 3.1, we propose a way of fitting local constant and local linear regression to set-valued outcome observations in $\mathbb{R}^d$. The proposed estimator is shown to

be consistent and its mean squared error and asymptotic distribution are derived. Additionally, we show how to find the best bandwidth by leave-one-out cross-validation and provide a method to build error tubes around the estimator, considered as confidence intervals for the set-valued prediction. In the framework of Beresteanu and Molinari [2008], Section 3.2 identifies optimal experimental designs under several the objective functions corresponding to the classical A, G, E and MV optimal designs. By adding some mild conditions, we are able to show that these objective functions can be simplified and they coincide with their classic objective functions of point-identified data. Finally, in Section 3.3 we provide different approaches to identify outliers in relation to the distribution of a random convex set or a sample of convex sets. Beside a generalization of the classic depth concept to set-valued data, we propose a new concept relaying on sub- and superlinear expectations. These nonlinear expectations have been studied Peng [2004, 2019] for random variables and later elaborated by Molchanov and Mühlemann [2021] for distributions of random convex sets.

Chapter 2

Preliminaries

Throughout the thesis, we work in Euclidean space $\mathbb{R}^d$ equipped with the Euclidean norm $\|\cdot\|$. The inner product is denoted by $\langle\cdot,\cdot\rangle$. The bold letters are used for random vectors in $\mathbb{R}^d$, while capital bold letters stand for random sets in $\mathbb{R}^d$. The normal font lower case and capital letters denote deterministic vectors in $\mathbb{R}^d$ and deterministic subsets of $\mathbb{R}^d$. For $x \in \mathbb{R}$, we denote the positive and negative parts of x respectively by $x^+ = \max(0, x)$ and $x^- = -\min(0, x)$. In the following we recall some important definitions and concepts for sets.

Operation. Let $\mathcal{F}(\mathbb{R}^d)$ be the family of closed sets in $\mathbb{R}^d$, and let $\mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$ denote its subfamily consisting convex and closed sets. Furthermore, we denote the collection of compact subsets of $\mathbb{R}^d$ by $\mathcal{K}(\mathbb{R}^d)$ and the family of non-empty compact convex sets by $\mathcal{K}_{\mathcal{C}}(\mathbb{R}^d)$. Note that $\mathcal{K}(\mathbb{R}^d) \subset \mathcal{F}(\mathbb{R}^d)$ and $\mathcal{K}_{\mathcal{C}}(\mathbb{R}^d) \subset \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$.

The *(Minkowski) sum* of two subsets A, B of $\mathbb{R}^d$ is defined as

$$A + B = \{a + b : a \in A, b \in B\}.$$

The *scaling* of set A is denoted by $cA = \{ca : a \in A\}$ with $c \in \mathbb{R}$. In particular,

$$-A = \{-x : x \in A\}.$$

Norm and distance. Let A and B be subsets of $\mathbb{R}^d$. The *directed Hausdorff distance* from A to B is defined by

$$d_H(A, B) = \inf\{\varepsilon \in \mathbb{R} : A \subseteq B^\varepsilon\},$$

where $B^\varepsilon = \{x \in \mathbb{R}^d : \inf_{b\in B} \|x - b\| \leq \varepsilon\}$ is the ε-neighborhood of B. The *Hausdorff distance* is defined as

$$H(A, B) = \max\{d_H(A, B), d_H(B, A)\}.$$

The *norm* is defined as

$$\|A\|_H = H(A, \{0\}) = \sup\{\|a\| : a \in A\}.$$

Support function. Let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$. The *support function* of a subset A of $\mathbb{R}^d$ in a direction $u \in \mathbb{S}^{d-1}$ is given by

$$s(A, u) = \sup_{a \in A} \langle a, u \rangle.$$

If $A \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$, then A is uniquely identified by its support function. The support function has the following properties:

- $s(tA, v) = ts(A, v)$ for $t \geq 0$,
- $s(A_1 + A_2, v) = s(A_1, v) + s(A_2, v)$ for $A_1, A_2 \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$.

We define the *width function* of A in direction $v \in \mathbb{S}^{d-1}$ as

$$w(A, v) = s(A, v) + s(A, -v). \tag{2.1}$$

The Hausdorff distance between $A, B \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$ can be written as

$$H(A, B) = \sup_{v \in \mathbb{S}^{d-1}} |s(A, v) - s(B, v)|,$$

meaning that the Hausdorff distance between closed convex sets is the uniform (L_∞) distance between their support functions, see Lemma 1.8.14 in Schneider [2014]. Other distances based on the L_p-norm are also available:

$$L_p(A, B) = \left(\int_{v \in \mathbb{S}^{d-1}} |s(A, v) - s(B, v)|^p \, dv \right)^{1/p},$$

for any $p \in [1, \infty)$.

Random set. Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be a probability space, where Ω is the space of elementary events equipped with σ-algebra $\mathfrak{F}$ and probability measure $\mathbf{P}$. A map $\mathbf{Y} : \Omega \to \mathcal{F}(\mathbb{R}^d)$ is called *random closed set* if

$$\mathbf{Y}^{-1}(K) = \{\omega \in \Omega : \mathbf{Y}(\omega) \cap K \neq \emptyset\} \in \mathfrak{F} \tag{2.2}$$

for each compact set K in $\mathbb{R}^d$.

A *random convex closed set* $\mathbf{Y}$ is a map from Ω to $\mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$ satisfying the same measurability condition (2.2). A similar definition applies for a compact random set as well as a compact convex random set.

Expectation. A random variable $\boldsymbol{y}$ with value in $\mathbb{R}^d$ is called a *(measurable) selection* of a random closed set $\mathbf{Y}$ if $\boldsymbol{y}(\omega) \in \mathbf{Y}(\omega)$ for almost all $\omega \in \Omega$. We denote this by $\boldsymbol{y} \in \mathbf{Y}$ a.s. Assuming that $\mathbf{Y}$ admits at least one integrable selection (then $\mathbf{Y}$ is said to be integrable), the *Aumann expectation* of $\mathbf{Y}$ is defined as

$$\mathbf{E}(\mathbf{Y}) = \operatorname{cl}\left\{\mathbf{E}\boldsymbol{y} : \boldsymbol{y} \in \mathbf{Y} \text{a.s. and } \mathbf{E}\|\boldsymbol{y}\| < \infty\right\},$$

where cl denotes the topological closure.

If $\mathbf{Y}$ is integrably bounded, that is, $\|\mathbf{Y}\|$ is integrable, then the closure on the right-hand side can be omitted, all selections of $\mathbf{Y}$ are integrable, $\mathbf{E}(\mathbf{Y})$ is convex and

$$\mathbf{E}(s(\mathbf{Y}, u)) = s(\mathbf{E}(\mathbf{Y}), u), \quad u \in \mathbb{S}^{d-1}.$$

Chapter 3

Main results and discussions

3.1 Local regression smoothers with set-valued outcome data

We are interested in estimating $\mathbf{E}(\mathbf{Y}|\boldsymbol{x} = x_0)$ using a local regression smoother based on an i.i.d. sample $(\boldsymbol{x}_i, \mathbf{Y}_i)_{i=1}^n$ drawn from $(\boldsymbol{x}, \mathbf{Y})$, where x_0 is a given value from the support of a random vector $\boldsymbol{x} \in \mathbb{R}^r$ and $\mathbf{Y}$ is an integrably bounded random compact convex set in $\mathbb{R}^d$. This was motivated by fitting linear regression for such data with $d = 1$ (so that $\mathbf{Y}$ is a random interval) proposed by Beresteanu and Molinari [2008]. They interpreted $(\boldsymbol{x}, \mathbf{Y})$ as a family of the pairs $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{y}$ belongs to $\mathbf{Y}$ almost surely, that is, $\boldsymbol{y}$ is a selection of $\mathbf{Y}$. Then

$$\mathbf{E}(\mathbf{Y}|\boldsymbol{x} = x_0) = \{\mathbf{E}(\boldsymbol{y}|\boldsymbol{x} = x_0) : \boldsymbol{y} \in \mathbf{Y} \text{ a.s.}\} \tag{3.1}$$

is the conditional selection expectation of $\mathbf{Y}$.

Let $\boldsymbol{\theta}$ be a vector taking values in $\mathbb{R}^r$. Each choice $(\boldsymbol{x}, \boldsymbol{y})$ yields a value of $\boldsymbol{\theta}$ such that $\mathbf{E}(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\theta}$, and this value $\boldsymbol{\theta}$ can be easily estimated by the classical ordinary least squares. The collection of all these $\boldsymbol{\theta}$'s based on different choices of $(\boldsymbol{x}, \boldsymbol{y})$ from $(\boldsymbol{x}, \mathbf{Y})$ forms a compact and convex set, denoted by $\boldsymbol{\Theta}$. It can be considered as the set of the best linear prediction parameters for $(\boldsymbol{x}, \mathbf{Y})$. Consequently, its estimator is the family of the least square estimator of $\boldsymbol{\theta}$'s. Equation (3.1) can be expressed as

$$\mathbf{E}(\mathbf{Y}|\boldsymbol{x} = x_0) = \{x_0^\top\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

Based on this model setting, Bontemps et al. [2012] extended the familiar Sargan test for over-identifying restrictions and Chandrasekhar et al. [2012] provided inference methods for the best linear approximation of any function $f(x)$ that is known to lie within two identified bounding functions. Fisher [2010] extended the model of Beresteanu and Molinari [2008] for the case where the regressor is set-valued and provided the inference. By simulations, he observed that the estimator for $\boldsymbol{\Theta}$ is convex in most of time but not in general. Kaido [2016] proposed an estimator for weighted average derivatives of conditional mean and conditional quantile functionals when either the outcome variable or a regressor is interval-valued. Adusumilli and Otsu [2017] proposed empirical likelihood methods for random sets to conduct inference in the class of problems analyzed by Beresteanu and Molinari [2008].

The approach of Beresteanu and Molinari [2008] differs from other approaches in the literature relying on set-valued arithmetics; see Schollmeyer and Augustin [2015] for a discussion bridging this literature to other papers on set-valued data. For example, the model based on the interval arithmetics proposed by Blanco-Fernández et al. [2013b,a] assumes that $\mathbf{E}(\mathbf{Y}|\boldsymbol{x}) = A\boldsymbol{x} + B$, where $\mathbf{Y}$, A and B are intervals and $\boldsymbol{x} \in \mathbb{R}$. This interval model can be transformed into two linear relationships

$$\begin{aligned} \text{mid}\mathbf{Y} &= \text{mid}A\boldsymbol{x} + \text{mid}B, \\ \text{spr}\mathbf{Y} &= \text{spr}A\boldsymbol{x} + \text{spr}B, \end{aligned}$$

where *mid* and *spr* denote the mid-point and half of the length of a interval, respectively. If the regressor is an interval, denoted by $\mathbf{X}$, the linear model was initially assumed to be $\mathbf{E}(\mathbf{Y}|\mathbf{X}) = a\mathbf{X} + B$ with $a \in \mathbb{R}$ (see Sinova et al. [2012], González-Rodríguez et al. [2007]) and generalized afterwards to a more flexible model

$$\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \alpha\ (\text{mid}\mathbf{X})[1,1] + \beta\ \text{spr}\mathbf{X}[-1,1] + \gamma[1,1] + B,$$

where $\alpha, \beta, \gamma \in \mathbb{R}$ (see Blanco-Fernández et al. [2012], Blanco-Fernández et al. [2011]). Further discussions of the linear regression problems related to such models can be found in the papers by Diamond [1990] and Gil et al. [2001]. Maatouk [2003] proposed using weighted least-squares to estimate the parameters in such models.

Comparing to the literature above, our approach works for both interval-valued ($d = 1$) and general set-valued outcome ($d > 1$). Furthermore, working with local regression smoother we give fewer specifications on the conditional expectation than in the literature related to interval arithmetics. Finally, our proposal is distinct from the literature on data coarsening, see Gill et al. [1997], Heitjan [1994], Heitjan and Rubin [1991], where the key assumption "coarsening at random" restricts directly the conditional distribution of the random set $\mathbf{Y}$ with $\mathbf{P}(\mathbf{Y} = A|\boldsymbol{x} = x_0)$ being constant for all $x_0 \in A$.

Before introducing our approach, we shortly recall the standard construction of the local polynomial estimator for the observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^n$, where $\boldsymbol{y}_i$ is point-identified in $\mathbb{R}$, see Fan and Gijbels [1996]. The aim is to estimate $m(x_0) = \mathbf{E}(\boldsymbol{y}|\boldsymbol{x} = x_0)$, assuming that the data have been generated from the model

$$\boldsymbol{y} = m(\boldsymbol{x}) + \varepsilon, \tag{3.2}$$

where $\mathbf{E}(\varepsilon) = 0$ and $\boldsymbol{x}$ and ε are independent. Assume that the $(p+1)$th derivative of $m(x)$ at x_0 exists. Then we use the Taylor expansion to approximate $m(x)$ for x in a neighbourhood of x_0 as

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p.$$

Fix a kernel function $K(\cdot)$ and a tuning parameter h_n called the bandwidth. We assume that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Generally speaking, kernel function can also be negative, see Condition 1 (iv) imposed by Fan [1993]. However, in this thesis we put more restrictions (see Assumption A in Paper A) on the kernel function due to technical reason.

By minimizing the weighted mean squared error

$$\sum_{i=1}^{n}\left[\boldsymbol{y}_i-\sum_{j=0}^{p}\theta_j(\boldsymbol{x}_i-x_0)^p\right]^2 K\left(\frac{\boldsymbol{x}_i-x_0}{h_n}\right)$$

with respect to $\theta_0,\dots,\theta_p$, we obtain the estimator

$$\hat{m}(x_0)=\hat{\theta}_0=\sum_{i=1}^{n}\ell_i(x_0)\boldsymbol{y}_i,$$

which is the sum of weighted outcomes. The weights $\ell_i(x_0)$ depend on $i=1,\dots,n$, the reference point x_0, the choice of K and h_n, but not on the observed outcomes $\boldsymbol{y}_i$. While $\ell_i(\cdot)$ can be negative, they sum up to one. To simplify the notation, denote

$$\boldsymbol{\kappa}_{in}=K((\boldsymbol{x}_i-x_0)/h_n),$$

so that

$$\boldsymbol{s}_j=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\kappa}_{in}(\boldsymbol{x}_i-x_0)^j,\qquad j=0,1,2.$$

In case of $p=0$ (local constant regression), $\hat{m}(x_0)$ is the Nadaraya–Watson estimator with $\ell_i(x_0)=\boldsymbol{\kappa}_{in}/n\boldsymbol{s}_0$. If $p=1$ (local linear regression), then

$$\ell_i(x_0)=\frac{\boldsymbol{\kappa}_{in}}{n}\frac{\boldsymbol{s}_2-(\boldsymbol{x}_i-x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2+n^{-4}}. \tag{3.3}$$

Our goal is to provide a local linear regression estimator for the expectation of each random variable $\boldsymbol{y}\in\mathbf{Y}$ conditional on $\boldsymbol{x}$ by assuming that the tuple $(\boldsymbol{x},\boldsymbol{y})$ almost surely belongs to a random set $\{\boldsymbol{x}\}\times\mathbf{Y}$. Each choice of $(\boldsymbol{x},\boldsymbol{y})\in\{\boldsymbol{x}\}\times\mathbf{Y}$ gives rise to a function m as in (3.2), and we denote by $\mathcal{M}$ the family of all regression functions generated in this way, so that $M(x)=\{m(x):m\in\mathcal{M}\}$ is the set of values of all possible regression functions at x. Assume that $(\boldsymbol{x}_i,\mathbf{Y}_i)_{i=1}^n$ are generated from the model

$$s(\mathbf{Y}_i,v)-s(M(\boldsymbol{x}_i),v)=\varepsilon_i(v),\qquad v\in\mathbb{S}^{d-1}, \tag{3.4}$$

where $M(\boldsymbol{x}_i)=\mathbf{E}(\mathbf{Y}|\boldsymbol{x}_i)$ and $\varepsilon_1(\cdot),\dots,\varepsilon_n(\cdot)$ are i.i.d. copies of a square integrable random function $\varepsilon(v),v\in\mathbb{S}^{d-1}$, such that $\mathbf{E}[\varepsilon_i(v)|\boldsymbol{x}_i]=0$ for all $v\in\mathbb{S}^{d-1}$. More details about this model and the discussion of the existence of the function $\varepsilon(\cdot)$ can be found in the discussion concerning Assumption B in Paper A. The proposed estimator of $M(x_0)$ is the weighted Minkowski average of $\mathbf{Y}_i$ defined as follows

$$\hat{M}(x_0)=\sum_{i=1}^{n}\ell_i(x_0)\mathbf{Y}_i, \tag{3.5}$$

where $\ell_i(x_0)$ is defined by (3.3). By representing $\ell_i=\ell_i^+-\ell_i^-$ and using the fact that $s(-A,v)=s(A,-v)$ for a convex compact set A, the support function of the estimator is

$$s(\hat{M}(x_0),v)=\sum_{i=1}^{n}\ell_i(x_0)s(\mathbf{Y}_i,v)+\sum_{i=1}^{n}\ell_i^-w(\mathbf{Y}_i,v),$$

where $w(\cdot, v)$ is the width function defined in (2.1).

To quantify the properties of our estimator, the following L_2 distance is used to define the mean square error

$$\begin{aligned}\mathrm{MSE}(\hat{M}(x_0, v), M(x_0, v)) &= \mathbf{E}\left(\int_{\mathbb{S}^{d-1}} (s(\hat{M}(x_0), v) - s(M(x_0), v))^2 \, dv\right) \\ &= \int_{\mathbb{S}^{d-1}} b^2_{x_0}(v) \, dv + \int_{\mathbb{S}^{d-1}} \sigma^2_{x_0}(v) \, dv,\end{aligned}$$

where $b^2_{x_0}(v)$ and $\sigma^2_{x_0}(v)$ are squared bias and variance in the direction v. Using the properties of support function (see Chapter 2) and (3.4), these two terms can be expressed as

$$b^2_{x_0}(v) = \mathbf{E}\left(\sum_{i=1}^n \ell_i(s(M_i(\boldsymbol{x}_i), v) - s(M(x_0), v)) + \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i), v)\right)^2 \tag{3.6}$$

and

$$\sigma^2_{x_0}(v) = \mathbf{E}\left(\sum_{i=1}^n \ell_i \varepsilon_i(v) + \sum_{i=1}^n \ell_i^- (\varepsilon_i(v) - \varepsilon_i(-v))\right)^2.$$

Comparing to the classical case where $\boldsymbol{y}_i$ is point-identified, the negative part (with ℓ_i^-) in (3.6) plays essential role for the set-valued outcome. Moving $\boldsymbol{x}_i$ closer to x_0, the width $w(M(\boldsymbol{x}_i), v)$ does not vanish. Therefore, the bias may not tend to zero if some weights are negative and not close to zero. Much of our asymptotic analysis is concerned with establishing the asymptotic behavior of these negative weights. This is also the reason why the assumption on the kernel function (Assumption A of Paper A) is stricter than those imposed by Fan [1993] and Fan and Gijbels [1992]. Still, many kernel functions satisfy our assumption. Imposing additionally a general assumption on the density $f(\cdot)$ of $\boldsymbol{x}$ (Assumption D of Paper A), we show that the second moment of the sum of all negative weights converges to 0.

Proposition 3.1.1 (Proposition 4.2 in Paper A)**.** *Let $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Under Assumptions A and D, for sufficiently large r,*

$$\mathbf{E}\left(\sum_{i=1}^n \ell_i^-\right)^2 = \frac{1}{h_n} \mathcal{O}\left(\left(1/\sqrt{nh_n}\right)^r\right).$$

Note that we use o and $\mathcal{O}$ to denote the deterministic order of magnitude uniformly in the Hölder class $\mathcal{H}(1, \gamma)$ (see Assumption D in Paper A).

Proposition 3.1.1 shows that the negative part converges to 0 much faster than $h_n^4 + 1/nh_n$ (see Theorem 3.1.2 below) by choosing sufficiently large r. This can be explained by the construction of $\boldsymbol{s}_j$ and $\ell_i(x_0)$. Write $Z_n = \mathcal{O}_r(a_n)$ if

$$\sup_{f \in \mathcal{H}(1,\gamma)} \mathbf{E}|Z_n|^r = \mathcal{O}(a_n^r).$$

If the rth moment of Z_n exists, we write

$$Z_n = \mathbf{E}Z_n + \mathcal{O}_r[(\mathbf{E}|Z_n - \mathbf{E}Z_n|^r)^{1/r}].$$

Since in Assumption A we require that the kernel function has compact support and is bounded, we have

$$\int z^r K(z)\,dz < \infty, \quad \text{for any integer } r > 1.$$

Therefore, the rth moment of $\boldsymbol{s}_j$ always exists and $\boldsymbol{s}_j$ can be expressed as

$$\begin{aligned}
\boldsymbol{s}_j &= \mathbf{E}\boldsymbol{s}_j + h_n^{j+1}\mathcal{O}_r\big(1/\sqrt{nh_n}\big) \\
&= h_n^{j+1}\left(f(x_0)\int z^j K(z)\,dz + \mathcal{O}(h_n) + \mathcal{O}_r(\frac{1}{\sqrt{nh_n}})\right) \\
&= h_n^{j+1}\left(f(x_0)\int z^j K(z)\,dz + \mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}})\right) \\
&= \begin{cases} h_n\left(f(x_0) + \mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}})\right), & j = 0, \\ h_n^2\mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}}), & j = 1, \\ h_n^3\left(f(x_0)\int z^2K(z)\,dz + \mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}})\right), & j = 2. \end{cases}
\end{aligned}$$

The dominant terms of $\boldsymbol{s}_0$ and $\boldsymbol{s}_2$ are $h_n f(x_0)$ and $h_n^3 f(x_0)\int z^2K(z)\,dz$, respectively, while for $\boldsymbol{s}_1$ it is 0. Thus, the dominant part of $\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2$ is positive. Considering (3.3), $\ell_i(x_0)$ is negative only if $\boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)$ is negative. By Assumption A, the kernel function $\boldsymbol{\kappa}$ is bounded and $|\boldsymbol{x}_i - x_0| \le c_K h_n$. This leads to the fact that the dominant part of $\boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)$ is not negative. This explains reasons behind Proposition 3.1.1.

With this result in hand, we establish the asymptotic behavior of MSE and a limit theorem for the support function of the estimator by adding assumption on the model structure (Assumption B in Paper A) and differentiability of the support function of the theoretical response function (Assumption C in Paper A).

Theorem 3.1.2 (Theorem 4.3 in Paper A)**.** *Under Assumptions A, B, C and D, if $h_n = cn^{-\beta}$ with $0 < \beta < 1$ and a constant $c > 0$, the mean squared error of the local linear estimator (3.5) is*

$$\begin{aligned}
\mathrm{MSE}(x_0) = \frac{h_n^4\left(\int z^2K(z)\,dz\right)^2}{4}\int_{\mathbb{S}^{d-1}} s''(M(x_0),v)^2\,dv + \frac{\int_{\mathbb{S}^{d-1}}\mathbf{E}(\varepsilon(v)^2)\,dv}{nh_n f(x_0)}\int K^2(z)\,dz \\
+ o\left(h_n^4 + \frac{1}{nh_n}\right).
\end{aligned}$$

Let $\zeta(v)$, $v \in \mathbb{S}^{d-1}$, be a centered Gaussian process on the unit sphere with the covariance

$$\mathbf{E}[\zeta(v)\zeta(u))] = \frac{\mathbf{E}(\varepsilon(v),\varepsilon(u))}{f(x_0)}\int K(z)^2\,dz. \tag{3.7}$$

Theorem 3.1.3 (Theorem 4.4 in Paper A)**.** *Assume that $h_n = cn^{-\beta}$ with $0 < \beta < 1$, and fix x_0 from support of $\boldsymbol{x}$. Under Assumptions A, B, C and D, the stochastic process*

$$\sqrt{nh_n}\left(s(\hat{M}(x_0),v) - s(M(x_0),v) - h_n^2\frac{1}{2}s''(M(x_0),v)\int z^2K(z)\,dz\right)$$

constructed using the local linear estimator in (3.5) converges in distribution in the space of continuous functions on $\mathbb{S}^{d-1}$ with the uniform metric to the Gaussian process ζ.

The optimal bandwidth which minimize the MSE in Theorem 3.1.2 is $h_{n,\text{MSE}} = Cn^{-1/5}$ with constant C that does not depend on n but on some unknown quantities such as $s''(M(x_0), v)$, $\mathbf{E}(\varepsilon(v)^2)$ and $f(x_0)$. This problem exist also in the classical case where the outcome is singleton-valued. One possible way to determine the optimal bandwidth is to substitute the unknown term by its estimators (plug-in type bandwidth). Other practical approaches were summarised in Chapter 4 of Fan and Gijbels [1996]. We propose to use leave-one-out cross-validation with the cross-validation score defined as

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{S}^{d-1}} (s(\mathbf{Y}_i, v) - s(\hat{M}_{(-i)}(\boldsymbol{x}_i), v))^2\, dv, \tag{3.8}$$

where $\hat{M}_{(-i)}(x) = \sum_{j=1}^{n} \mathbf{Y}_j \ell_{j,(-i)}(x)$ and

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if } j = i, \\ \frac{\ell_j(x)}{\sum_{k\neq i}\ell_k(x)} & \text{if } j \neq i. \end{cases}$$

This procedure assigns zero weight to $\boldsymbol{x}_i$ and renormalizes other weights to sum to one.

Beside the selection of bandwidth, we also interested in assessing statistical uncertainty of the estimator using pointwise error tubes, which are similar to the pointwise confidence interval for the singleton-valued outcome. Given x_0, the error tube is constructed as

$$\hat{\mathcal{C}}(x_0) = \hat{M}(x_0) + \frac{c_\alpha}{\sqrt{nh_n}} B, \tag{3.9}$$

where $B = \{b : \|b\| \leq 1\}$ is the unit ball. According to Theorem 3.1.3, c_α is chosen so that

$$\mathbf{P}\left(\max_{v:\, \|v\|=1} \{\zeta(v)\}^+ > c_\alpha\right) = \alpha, \tag{3.10}$$

where ζ is the centered Gaussian process with covariance kernel (3.7). The critical value c_α can be obtained by simulation or bootstrap. The validity of the bootstrap can be established as in Theorem 4.13 in Molchanov and Molinari [2018].

We have performed some simulation studies for the cases when $d = 1$ (outcome is interval-valued) and $d = 2$ (see Section 6 and Appendix E in Paper A) in order to investigate the coverage probability of the proposed error tube by varying the optimal bandwidth $h_{n,CV}$ from that obtained using the cross-validation by multiplying $1, 1/2, 1/3$. All these simulation results lead to the same conclusion. The coverage probability based on $h_{n,CV}$ is lower than the nominal level of 95%, while using $1/3h_{n,CV}$ (undersmoothing) we observed the opposite result. Choosing $1/2h_{n,CV}$ (less undersmoothing than using factor $1/3$) the coverage probability is very close to the nominal level. This can be explained by the fact that choosing optimal $h_n = Cn^{-1/5}$ the bias term in Theorem 3.1.3 does not vanish because $nh_n^5 \not\to 0$. It is also the case in the classical local polynomial regression with singleton-valued outcomes. To deal with this problem, one can use undersmoothing as an approach to reduce the bias. However, over undersmoothing increases the variance of the estimator, which leads to conservative error tubes.

Remark 3.1.4 (Appendix C in Paper A). In the local constant case, the weights $i = \boldsymbol{\kappa}_{in}/(n\boldsymbol{s}_0)$ are always nonnegative. Then the estimator $M(x_0)$ can be constructed as

the convex set whose support functions is obtained by calculating the Nadaraya–Watson estimator for the sample $s(\mathbf{Y}_i, v), i = 1, \dots, n$, in each particular direction v. In other words, $M(x_0)$ is the sum of the observed sets $\mathbf{Y}_i$ multiplied by nonnegative coefficients i. Therefore, the bias and variance of the set-valued local constant estimator can be obtained similarly to the singleton-valued data case. For this, it suffices to assume that the function $s(M(x), v)$ is Lipschitz in x with the same constant for all v, which is equivalent to requiring that $M(x), x \in \mathbb{R}$, is Lipschitz in the Hausdorff metric.

3.2 Optimal design for multivariate multiple linear regression with set-identified response

The basic problem in the theory of optimal design for regression models aims to identify the locations of design points which ensure the best properties of the unbiased estimator of the parameter of interest, see Atkinson et al. [2007], Silvey [1980]. Taking the basic linear regression as an example, we believe that the random tuple $(\boldsymbol{x}, \boldsymbol{y})$ taking value from $(\{1\} \times \mathbb{R}^r, \mathbb{R})$ satisfies

$$\boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{\theta} + \varepsilon,$$

where $\boldsymbol{\theta}$ is a vector of $(r+1)$ numerical unknown parameters and ε is a centered random variable with $\mathrm{Var}(\varepsilon) = \sigma^2$. In the setting of experimental design, $\boldsymbol{x}$ contains factors affecting the outcome $\boldsymbol{y}$ and all these factors can be chosen arbitrarily from some given domain. Let us denote this domain by $\mathcal{I}$, which is a compact subset of $\{1\} \times \mathbb{R}^r$. Our goal is to choose n points $x_1, \dots, x_n$ from $\mathcal{I}$, so that the estimator of $\boldsymbol{\theta}$ based on the sample $(x_i, y_i)_{i=1}^n$ is optimal with respect to a pre-specified criteria.

Let

$$\mathscr{X} = (x_1^\top, \dots, x_n^\top)^\top = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nr} \end{pmatrix}$$

be the *design matrix* with n rows and $r+1$ columns. If the linear model has full rank, the least square estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \Sigma^{-1} \mathscr{X}^\top \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

where $\Sigma = \mathscr{X}^\top \mathscr{X}$. Then $\mathbf{E}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2 \Sigma^{-1}.$$

The most common optimality criteria are the following ones.

- D-optimal design aims to minimize the determinant of Σ^{-1} which represents the volume of the confidence ellipsoid

$$\left\{ \theta \in \mathbb{R}^{r+1} : (\theta - \hat{\boldsymbol{\theta}})^\top \Sigma (\theta - \hat{\boldsymbol{\theta}}) \le c \right\}.$$

- A-optimal design minimizes the trace of Σ^{-1} and therefore minimizes the sum of the variance of all estimated parameters.

- E-optimal design minimizes the largest eigenvalue of Σ^{-1}, and thus minimizes the least well estimated contrast $a^\top \hat{\boldsymbol{\theta}}$ under the constraint $\|a\| = a^\top a = 1$.

- MV-optimal design: as in the E-optimal design, if the L_1 norm $\|a\|_1 = \sum_{i=1}^{r+1} |a_i| = 1$ instead of the L_2 norm is used, it gives another criterion called minimum variance (MV) optimality, which was first introduced for block designs by Jacroux [1983] and later has been generalized for simple as well as weighted linear regression for all choices in a compact design space by Torsney and López-Fidalgo [1995], López-Fidalgo et al. [1998]. The MV-optimal design minimizes the maximum of the variance of individual parameters, and therefore minimizes the maximum diagonal element of Σ^{-1}.

- G-optimal design minimizes $\max_{x \in \mathcal{I}} x^\top \Sigma^{-1} x$, and thus minimizes the maximum of the variance of the predicted response over the design region $\mathcal{I}$.

Special relationships between some of optimal designs are provided by the equivalence theorem proved by Kiefer and Wolfowitz [1960] for real-valued outcome and later generalized for vector-valued outcome as well as for the polynomial regression model by Imhof [2000], Kiefer [1974], Krafft and Schaefer [1992], Soumaya et al. [2015].

Remark 3.2.1. In the multiresponse setting of dimension p (so that the response variable has p components), it is possible to vectorize the parameter matrix by modifying the linear equation as

$$\begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(p)} \end{pmatrix} = \begin{pmatrix} \mathscr{X} & & \\ & \ddots & \\ & & \mathscr{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}. \tag{3.11}$$

The vector on the left-hand side arises by stacking together n observations for each component $y^{(p)}$, $j = 1, \ldots, p$, of the response. Furthermore, $\operatorname{diag}(\mathscr{X}, \ldots, \mathscr{X})$ is an $np \times (r+1)p$ block-diagonal matrix built of the $n \times (r+1)$ dimensional design matrix $\mathscr{X}$; $\boldsymbol{\theta}_j \in \mathbb{R}^{r+1}$, $j = 1, \ldots, p$, are the parameters to be estimated, and ε_j, $j = 1, \ldots, p$, are n-dimensional random vectors. Using the vector representation like (3.11), we assume that for each component of the response the regression function is the same. In this setting, Chang [1994] proved that under the framework of approximate designs the D-optimal design in the multiresponse model is exactly the D-optimal design arising in the case of a univariate response. Kurotschka and Schwabe [1996] extended this reduction result for both exact and approximate designs for D, A, and E-optimality criteria and for more general Φ-optimality defined by Kiefer [1974].

In this thesis we extend the optimal design concepts for the i.i.d. data $(x_i, Y_i)_{i=1}^n$, where $x_i \in \mathcal{I}$ and Y_i is a compact convex set in $\mathbb{R}^p$. Considering linear regression technique for such data proposed by Beresteanu and Molinari [2008], we assume that for each sample $(x_i, y_i)_{i=1}^n$ with $(x_i, y_i) \in (x_i, Y_i)$, there exist a $(r+1) \times p$ matrix Θ, so that

$$\mathscr{Y} = \mathscr{X}\Theta + \mathscr{E}, \tag{3.12}$$

where

$$\mathscr{Y} = (y_1^\top, \dots, y_n^\top)^\top$$

is a matrix and the matrix $\mathscr{E} = (\varepsilon_1^\top, \dots, \varepsilon_n^\top)^\top$ consists of i.i.d. square integrable centered random vectors $\varepsilon_i = (\varepsilon_{i1}, \dots \varepsilon_{ip})^\top$ such that $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma_{jk}$ for $i = 1, \dots, n$ and $j, k = 1, \dots, p$ and $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{i'k}) = 0$ for $i \neq i'$. We assume that $\mathscr{E}$ does not depend on $\mathscr{X}$. The least square estimator of Θ is

$$\hat{\Theta} = \Sigma^{-1} \mathscr{X}^\top \mathscr{Y}.$$

Then $\mathbf{E}\hat{\Theta} = \Theta$ and

$$\mathrm{Cov}(\hat{\Theta}_{(j)}, \hat{\Theta}_{(k)}) = \sigma_{jk} \Sigma^{-1}, \quad j, k = 1, \dots, p,$$

where $\hat{\Theta}_{(k)}$ denotes the kth column of $\hat{\Theta}$. Therefore, all these least square estimators $\hat{\Theta}$'s based on arbitrarily selected samples from $(x_i, Y_i)_{i=1}^n$ form a family of matrices

$$\hat{\mathbf{\Theta}} = (\mathscr{X}^\top \mathscr{X})^{-1} \mathscr{X}^\top \left\{ \begin{pmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{pmatrix} : y_i \in Y_i \text{ for all } i \in \{1, \dots, n\} \right\}. \tag{3.13}$$

Note the $\hat{\mathbf{\Theta}}$ is the estimator of the unknown parameters based on the linear regression on the data $(x_i, Y_i)_{i=1}^n$.

In order to define the variance of $\hat{\mathbf{\Theta}}$, we consider products of all matrices in $\mathbf{\Theta}$ with a given $u \in \mathbb{S}^{p-1}$; and then the support function of the obtained random convex set in $\mathbb{R}^{r+1}$ in direction v from the unit sphere $\mathbb{S}^r$ in $\mathbb{R}^{r+1}$. In other words, we work with the variance

$$\mathrm{Var}_{\mathscr{X}}\, s(\hat{\mathbf{\Theta}} u, v) = \mathbf{E}_{\mathscr{X}} (s(\hat{\mathbf{\Theta}} u, v) - s(\mathbf{E}_{\mathscr{X}}(\hat{\mathbf{\Theta}} u), v))^2$$

of the support function of $\hat{\mathbf{\Theta}} u$ and aim to minimize it as the function of the design. Note that $\hat{\mathbf{\Theta}} u$ is a random convex set in $\mathbb{R}^{r+1}$ and $\mathbf{E}_{\mathscr{X}}$ is the expectation assuming that the design matrix is $\mathscr{X}$. Following the classical definitions of A, G and E-optimal designs, we define the objective function for these designs in the set-identified framework as

$$f^A(\mathscr{X}) = \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^r} \mathrm{Var}_{\mathscr{X}}\, s(\hat{\mathbf{\Theta}} u, v)\, dv du, \tag{3.14}$$

$$f^G(\mathscr{X}) = \max_{x \in \mathcal{I}} \int_{\mathbb{S}^{p-1}} \mathrm{Var}_{\mathscr{X}}\, s(\hat{\mathbf{\Theta}}^\top x, u)\, du = \max_{x \in \mathcal{I}} \int_{\mathbb{S}^{p-1}} \mathrm{Var}_{\mathscr{X}}\, s(\hat{\mathbf{\Theta}} u, x)\, du, \tag{3.15}$$

$$f^E(\mathscr{X}) = \max_{v \in \mathbb{S}^r} \max_{u \in \mathbb{S}^{p-1}} \mathrm{Var}_{\mathscr{X}}\, s(\hat{\mathbf{\Theta}} u, v). \tag{3.16}$$

Here the integrals over spheres are understood with respect to a finite rotation invariant measure (the Haar measure) and $\mathcal{I}$ is a compact subset of $\{1\} \times \mathbb{R}^r$. If $p = 1$ (in the case of interval-valued response) the integrals over $\mathbb{S}^{p-1}$ is the sum of the values of the support function at $+1$ and -1.

Working with L_1 norm, the unit sphere $\mathbb{S}^r$ becomes a cube in $\mathbb{R}^{r+1}$, denoted by $\mathbb{S}^r_{L_1}$, with vertices in $\{e_k : k \in \{1, \dots, (r+1)\}\} \cup \{-e_k : k \in \{1, \dots, (r+1)\}$, where e_k is a

vector with the kth element equal to 1 and other elements equal to 0. Thus, we propose the objective function for the MV-optimal design as

$$f^{MV}(\mathscr{X}) = \max_{v \in \mathbb{S}^r_{L_1}} \max_{u \in \mathbb{S}^{p-1}} \mathrm{Var}_{\mathscr{X}}\, s(\hat{\boldsymbol{\Theta}} u, v). \tag{3.17}$$

Denote $M(x) = \mathbf{E}(\mathbf{Y}|\boldsymbol{x} = x)$. Under rather mild assumptions, we are able to show that the optimal designs under the set-valued setting correspond to their classical counterparts.

Theorem 3.2.2 (Theorem 4.2 in Paper B)**.** *Assume that*

$$s(\mathbf{Y}, u) - s(M(\boldsymbol{x}), u) = \varepsilon(u), \quad u \in \mathbb{S}^{p-1}, \tag{3.18}$$

where ε is a random function on the unit sphere that does not depend on $\boldsymbol{x}$ and satisfies $\mathbf{E}\varepsilon(u) = 0$ and $\mathrm{Var}(\varepsilon(u)) = \sigma_u^2 < \infty$ for all $u \in \mathbb{S}^{p-1}$. Then the designs minimizing the objective functions defined in (3.14) and (3.15) correspond to the classical A and G-optimal designs.

Theorem 3.2.3 (Theorem 4.3 in Paper B)**.** *Assume that (3.18) holds with ε being a random function that does not depend on $\boldsymbol{x}$ and satisfying $\mathbf{E}\varepsilon(u) = 0$ and $\mathrm{Var}(\varepsilon(u)) = \mathrm{Var}(\varepsilon(-u)) = \sigma_u^2 < \infty$ for all $u \in \mathbb{S}^{p-1}$. Then the design minimizing the objective function (3.16) coincides with the classical E-optimal design.*

Theorem 3.2.4. *Keep the same assumption as in Theorem 3.2.3. Then the design minimizing the objective function (3.17) coincides with the classical MV-optimal design.*

The vectorization representation in (3.11) was not applied to multivariate set-valued outcome. Instead, we used the matrix representation (see (3.12)) because of the model assumption (3.18). Technically, the vectorization representation can also be used for the multivariate set-valued outcome to derive the optimal design and we denote corresponding parameter by $\hat{\boldsymbol{\Theta}}'$. It is a set with vectors in $\mathbb{R}^{(r+1)p}$ and the variance of $\hat{\boldsymbol{\Theta}}'$ in direction $u \in \mathbb{S}^{(r+1)p-1}$ can be expressed as

$$\mathrm{Var}_{\mathscr{X}}\, s(\hat{\boldsymbol{\Theta}}', u) = \mathbf{E}_{\mathscr{X}}(s(\hat{\boldsymbol{\Theta}}', u) - s(\mathbf{E}_{\mathscr{X}}(\hat{\boldsymbol{\Theta}}'), u))^2.$$

The objective function can be defined similarly to (3.14)–(3.17). To derive the results mentioned above, $\mathbf{Y}$ in the model assumption (3.18) should be replaced by

$$\mathbf{Y}' = \left\{ \begin{pmatrix} \boldsymbol{y}^{(1)} \\ \vdots \\ \boldsymbol{y}^{(p)} \end{pmatrix} : \boldsymbol{y} \in \mathbf{Y} \right\},$$

where $\boldsymbol{y}^{(j)}$ is a random vector replicating the jth component of $\boldsymbol{y}$ for $(r+1)$ times. Further, $M(\boldsymbol{x})$ is replaced by $\mathbf{E}(\mathbf{Y}')$, and $\mathbb{S}^{p-1}$ is replaced by $\mathbb{S}^{(r+1)p-1}$.

Remark 3.2.5. For the classic D-optimal design, we choose the design points so that the column of the confidence ellipsoid of the parameter is minimized. However, in the set-valued setting the parameter $\boldsymbol{\Theta}$ is a family of matrices. It is unclear what would be the "confidence ellipsoid" of a set. Therefore, using our approach we can not offer a a generalization of D-optimal designs.

3.3 Random sets: depth and outliers

While a boxplot is a useful tool to identify outliers in samples of real numbers, it is not straightforward to generalize it for samples in $\mathbb{R}^d$. For a sample of values of a random vector $\xi \in \mathbb{R}^d$, outliers are conventionally identified by depth functions and associated depth-trimmed regions, so that points lying outside these regions are considered as outliers (see Liu et al. [1999], Mosler [2002]). The depth function $\mathrm{D}(x, \xi)$ assigns to a point $x \in \mathbb{R}^d$ a certain level of centrality regarding to the distribution of ξ and takes values between 0 and 1. The higher the level of a point, the nearer is this point to the center of the distribution of ξ. The excursion set of ξ

$$\mathrm{D}^\alpha(\xi) = \{x \in \mathbb{R}^d : \mathrm{D}(x, \xi) \geq \alpha\}$$

is called a depth-trimmed region at level α. In statistical applications, the probability distribution of ξ is usually replaced by its empirical probability measure.

Zuo and Serfling [2000] postulate four desirable properties of a depth function, namely affine invariance, maximality at center, monotonicity relative to deepest point and vanishing at infinity. Further properties, such as subadditivity of depth-trimmed region and upper semicontinuity of the depth function have been discussed by Cascos [2010], Cascos and Molchanov [2007]. The subadditivity means that if $\mathrm{D}(z, \xi + \eta) \geq \alpha$ and so z is not an outlier for $\xi + \eta$, then it is possible to find two non-outliers from the support of ξ and η respectively so that $z = x + y$. A upper semicontinuity property of a depth-function means that its depth-trimmed region at any level α is closed.

There are several general approaches to construct depth-trimmed region, such as half-space depth (see Nagy et al. [2019], Rousseeuw and Ruts [1999], Tukey [1975]), simplicial depth (see Liu [1990]) and convex hull depth (see Cascos [2010]). The construction of these depth functions become more complicated if data consist of functions. Since a sample of functions is a too meagre set in a functional space, there is a danger that most of functions will be assigned depth zero. Various proposals of depth for functional data with its advantages and drawbacks were discussed by Gijbels and Nagy [2017], Kuelbs and Zinn [2013], López-Pintado and Romo [2009].

Further generalizations of the concept of depth to other data types have been elaborated by Pandolfo et al. [2018] for directional data (and so belonging to a nonlinear space), by Chen et al. [2018] and Paindaveine and Van Bever [2018] for matrix-valued data, and by Lafaye De Micheaux et al. [2020] for curves.

We continue the programme of exploring non-traditional data types and aim to construct suitable depth of a convex closed set with respect to the distribution of a random convex closed set $\mathbf{X}$. The depth function of a convex closed set F regarding to $\mathbf{X}$ is denoted by $\mathrm{D}(F, \mathbf{X})$.

A closed set $F \in \mathcal{F}_\mathcal{C}(\mathbb{R}^d)$ is said to belong to the support of $\mathbf{X}$ if $\mathbf{X}$ with a positive probability belongs to any open neighbourhood of F in the Fell topology. Furthermore, F belongs to the convex hull of the support of $\mathbf{X}$ if F is the limit of convex combinations $p_1F_1 + \cdots + p_nF_n$, where $n \geq 1$, $p_1, \ldots, p_n$ are nonnegative numbers that sum up to 1, and $F_1, \ldots, F_n$ belong to the support of $\mathbf{X}$. If F does not belong to the convex hull of the support of $\mathbf{X}$, we set $\mathrm{D}(F, \mathbf{X}) = 0$.

Translating the properties of depth functions into the set-valued framework, one comes up with the affine invariance property (Property (D1), Paper C), upper semicontinuity

with convergency in the context of Fell topology (Property (D2), Paper C) and

(D3) If $\mathbf{X}$ is deterministic, that is, $\mathbf{X} = L$ almost surely for an $L \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$, then $\mathrm{D}(F, L) = \mathbf{1}_{F=L}$.

According to the general classification scheme of Zuo and Serfling [2000], we can classify the depth functions for set-valued data into four types:

- *Type A* depth function is constructed as expectation of functionals $\phi(F; \mathbf{X}_1, \ldots, \mathbf{X}_j)$ of F and j i.i.d. copies of $\mathbf{X}$. The functional ϕ measures the closeness of F to the sample of sets.
- *Type B* depth function is defined by $(1 + \mathbf{E}\phi(F; \mathbf{X}_1, \ldots, \mathbf{X}_j))^{-1}$, where ϕ describes certain distance of F to the family of i.i.d. copies of $\mathbf{X}$.
- *Type C* depth function is defined as *Type B*, but ϕ measures the outlyingness of F to the chosen sets.
- *Type D* depth function relies on taking infimum of probabilities that $\mathbf{X}$ belongs to certain families of sets related to F.

In the following four subsections we shortly present depth functions of each type.

3.3.1 Band depth [Type A]

Originally, the concept of band depth was introduced for functional data, see López-Pintado and Romo [2009]. The band is defined as the family of functions with values lying between the pointwise minimum and maximum of j functions. The band depth is the probability that a given function lies in the band generated by j independent copies of a given random function.

Given the i.i.d. copies $\mathbf{X}_1, \ldots, \mathbf{X}_i$ of $\mathbf{X}$, we can consider the convex hull of $\cup_{i=1}^{j} \mathbf{X}_i$ and $\cap_{i=1}^{j} \mathbf{X}_i$ as "maximum" and "minimum" of these j sample sets, respectively. Therefore, the band depth of a convex closed set F can be defined as

$$\mathrm{BD}^j(F, \mathbf{X}) = \mathbf{P}\{\cap_{i=1}^{j} \mathbf{X}_i \subseteq F \subseteq \operatorname{conv} \cup_{i=1}^{j} \mathbf{X}_i\}.$$

It is obvious that such band depth for $j = 1$ is not informative. It becomes nontrivial for $j \leq 2$. As advised by López-Pintado and Romo [2009], we can combine the bands built out a varying number j of sets by taking their averages as

$$\overline{\mathrm{BD}}^J(F, \mathbf{X}) = \frac{1}{J-1} \sum_{j=2}^{J} \mathrm{BD}^j(F, \mathbf{X}),$$

where $2 \leq J$. It is recommended to choose $J = 3$. By choosing a larger J, it is difficult to detect F as an outlier, if F has a very peculiar shape but rather normal magnitude. On the other hand, too many sets F will have depth zero by choosing $J = 2$.

The empirical version of the band depth with data $X_1, \ldots, X_n$ drawn from $\mathbf{X}$ is the average of $\mathbf{1}\{\cap_{i=1}^{j} X_{k_i} \subseteq F \subseteq \operatorname{conv} \cup_{i=1}^{j} X_{k_i}\}$ with $1 \leq k_1 < \cdots < k_j \leq n$ over $\binom{n}{j}$ samples (see equation (23) in Paper C).

Another variant of band depth for functional data suggested by Cascos and Molchanov [2018] can also be generalized for the set-valued data. Indeed, each convex and closed set is uniquely described by its support function. The specific construction and example can be found on page 16 in Paper C.

3.3.2 Simplicial depth [Type B]

The simplicial depth for random vectors is defined as the probability that a point belongs to the convex hull of $(d+1)$ independent copies of this vector, see Liu [1990]. Following this idea, the convex combination of sets $X_1, \dots, X_j$ is the family of sets obtained as

$$p_1 X_1 + \dots + p_j X_j$$

for nonnegative $p_1, \dots, p_j$ that sum to one. However, the family of such convex combinations is a meagre set in $\mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$. Because of this, a direct generalization of the simplicial depth for random sets fails. Note that taking convex hull of the union of $X_1, \dots, X_j$ substantially differs from taking their convex combination.

Following the idea of type B depth functions from Zuo and Serfling [2000], it is possible to define the depth by letting

$$\mathrm{D}(F, \mathbf{X}) = \frac{\mathbf{E}\psi(\mathrm{conv}(\mathbf{X}_1 \cup \dots \cup \mathbf{X}_j))}{\mathbf{E}\psi(\mathrm{conv}(F \cup \mathbf{X}_1 \cup \dots \cup \mathbf{X}_j))}, \tag{3.19}$$

where ψ is a monotone functional on $\mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$. For instance, it is possible to let ψ be the Lebesgue measure on $\mathbb{R}^d$; this yields a generalization of the simplicial volume depth, see [Zuo and Serfling, 2000, Example 2.2]. Another possibility is to let ψ be the surface area.

The depth function (3.19) is not sensitive to small F (see Example 7.1 in Paper C). A possible correction is to replace the the numerator in equation (3.19) by $\mathbf{E}\psi(\mathrm{conv}(F \cap (\mathbf{X}_1 \cup \dots \cup \mathbf{X}_j)))$; this has been done in Paper C.

3.3.3 Depth based on nonlinear expectations [Type C]

The concept of sublinear expectation for random variables was initially brought to probability theory by Peng [2004, 2019]. Later on this concept was introduced for random sets by Molchanov and Mühlemann [2021]. They worked with two functions $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{U}}$ which are called the sub- and superlinear expectations with

$$\begin{aligned} \boldsymbol{\mathcal{E}}(\mathbf{X} + \mathbf{Y}) &\subseteq \boldsymbol{\mathcal{E}}(\mathbf{X}) + \boldsymbol{\mathcal{E}}(\mathbf{Y}) \\ \boldsymbol{\mathcal{U}}(\mathbf{X} + \mathbf{Y}) &\supseteq \boldsymbol{\mathcal{U}}(\mathbf{X}) + \boldsymbol{\mathcal{U}}(\mathbf{Y}), \end{aligned}$$

for all p-integrable random convex closed sets $\mathbf{X}$ and $\mathbf{Y}$. Other properties of these two functions are given in Section 4.1 of Paper C. Note that $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{U}}$ are set-valued.

Consider families of nonlinear expectations $\boldsymbol{\mathcal{U}}_\alpha(\mathbf{X})$ and $\boldsymbol{\mathcal{E}}_\alpha(\mathbf{X})$ for $\alpha \in (0, 1]$ such that $\boldsymbol{\mathcal{U}}_\alpha(\mathbf{X})$ becomes larger and $\boldsymbol{\mathcal{E}}_\alpha(\mathbf{X})$ becomes smaller with increasing $\alpha \in (0, 1]$. Let F belong to the convex hull of the support of $\mathbf{X}$. We propose to define the depth of F as

$$\mathrm{D}(F, \mathbf{X}) = \sup\{\alpha : \boldsymbol{\mathcal{U}}_\alpha(\mathbf{X}) \subseteq F \subseteq \boldsymbol{\mathcal{E}}_\alpha(\mathbf{X})\}. \tag{3.20}$$

The depth function constructed by (3.20) fulfills (D1)–(D3) (see Proposition 4.3 in Paper C) and a variation of the classic subadditivity property

(D4) If $\mathrm{D}(F, \mathbf{X}+\mathbf{Y}) \geq \alpha$, then there exist $F_1, F_2, F_1', F_2' \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$, such that $\mathrm{D}(F_1, \mathbf{X}) \geq \alpha$, $\mathrm{D}(F_1', \mathbf{X}) \geq \alpha$, $\mathrm{D}(F_2, \mathbf{Y}) \geq \alpha$, and $\mathrm{D}(F_2', \mathbf{Y}) \geq \alpha$, and

$$F_1 + F_2 \subseteq F \subseteq F_1' + F_2'.$$

Property (D4) is the set-valued version of subadditivity. Because not all convex sets are decomposable as sum of nontrivial summands (see Section 3.2 by Schneider [2014]), we only require that F is sandwiched between two summands instead satisfying an exact equation.

The depth-trimmed region $\mathrm{D}^{\alpha}(\mathbf{X})$ corresponding to (3.20) consists of convex closed sets sandwiched between $\boldsymbol{\mathcal{U}}_\alpha$ and $\boldsymbol{\mathcal{E}}_\alpha$.

We consider two basic constructions of $(\boldsymbol{\mathcal{U}}_\alpha, \boldsymbol{\mathcal{E}}_\alpha)$. The first construction is based on i.i.d. copies $\{\mathbf{X}_n, n \geq 1\}$ of $\mathbf{X}$. Fix a pair of nonlinear expectations $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{E}}$. For each $\alpha \in (0, 1]$, let

$$\begin{aligned}\boldsymbol{\mathcal{E}}_\alpha(\mathbf{X}) &= \boldsymbol{\mathcal{E}}(\mathrm{conv}(\mathbf{X}_1 \cup \cdots \cup \mathbf{X}_{[\alpha^{-1}]})) \\ \boldsymbol{\mathcal{U}}_\alpha(\mathbf{X}) &= \boldsymbol{\mathcal{U}}(\mathbf{X}_1 \cap \cdots \cap \mathbf{X}_{[\alpha^{-1}]}),\end{aligned}$$

where $[\alpha^{-1}]$ is the integer part of α^{-1}. The corresponding depth function is

$$\mathrm{D}(F, \mathbf{X}) = \Big(\min\{n : \boldsymbol{\mathcal{U}}(\mathbf{X}_1 \cap \cdots \cap \mathbf{X}_n) \subseteq F \subseteq \boldsymbol{\mathcal{E}}(\mathrm{conv}(\mathbf{X}_1 \cup \cdots \cup \mathbf{X}_n))\} \Big)^{-1}. \quad (3.21)$$

In particular, $\mathrm{D}(F, \mathbf{X}) = 1$ if $\boldsymbol{\mathcal{U}}(\mathbf{X}) \subseteq F \subseteq \boldsymbol{\mathcal{E}}(\mathbf{X})$. If $\mathbf{X}$ is a singleton and $\boldsymbol{\mathcal{E}}(\cdot) = \boldsymbol{\mathcal{U}}(\cdot) = \mathbf{E}(\cdot)$ is the (linear) selection expectation, (3.21) corresponds to the expected convex hull depth suggested by Cascos [2007].

Remark 3.3.1 (see also Example 4.9 in Paper C). In order to construct the empirical version of (3.21) and make it more intuitive, consider a sample of convex closed sets $X_1, \ldots, X_n$ drawn from a random closed set $\mathbf{X}$ and assume that $\boldsymbol{\mathcal{E}}(\mathbf{X}) = \boldsymbol{\mathcal{U}}(\mathbf{X}) = E(\mathbf{X})$ for convenience. For each $m \in \{1, \ldots, n\}$, we draw $\binom{n}{m}$ samples. Each sample consists of m observations drawn with replacement from $X_1, \ldots, X_n$, denoted by $X_{j_1}, \ldots, X_{j_m}$, where j denote the jth sample. Therefore, the depth of F with respect to a sample $X_1, \ldots, X_n$ is the m^{-1} for the smallest m such that

$$\frac{1}{\binom{n}{m}} \sum_{j=1}^{\binom{n}{m}} \mathrm{conv}(X_{j_1} \cap \cdots \cap X_{j_m}) \subseteq F \subseteq \frac{1}{\binom{n}{m}} \sum_{j=1}^{\binom{n}{m}} \mathrm{conv}(X_{j_1} \cup \cdots \cup X_{j_m}).$$

Another possible construction is based on average quantiles. Let $\mathcal{M}_\alpha$ with $\alpha \in (0, 1]$ be the family of random variables with values in $[0, \alpha^{-1}]$. Call

$$\begin{aligned}\boldsymbol{\mathcal{E}}_\alpha(\mathbf{X}) &= \mathrm{conv} \bigcup_{\gamma \in \mathcal{M}_\alpha, \mathbf{E}(\gamma)=1} \mathbf{E}(\gamma \mathbf{X}), \\ \boldsymbol{\mathcal{U}}_\alpha(\mathbf{X}) &= \bigcap_{\gamma \in \mathcal{M}_\alpha, \mathbf{E}(\gamma)=1} \mathbf{E}(\gamma \mathbf{X}),\end{aligned}$$

the *average quantile* nonlinear expectation. The reason for this name stems from the fact that

$$s(\boldsymbol{\mathcal{E}}_\alpha(\mathbf{X}), u) = \mathsf{e}_\alpha(s(\mathbf{X}, u)) \quad (3.22)$$

with

$$\mathtt{e}_\alpha(\beta) = \frac{1}{\alpha}\int_{1-\alpha}^{1} q_t(\beta)dt \tag{3.23}$$

being the average of the quantiles of $\beta \in L_1(\mathbb{R})$ at levels from $[1-\alpha, 1]$, see Föllmer and Schied [2004][Th. 4.47]. Note that the function $\mathtt{e}(\cdot) : L_1(\mathbb{R}) \to (-\infty, \infty]$ is a numerical sublinear expectation, see Peng [2019].

However, a variant of (3.22) is not possible in the superlinear case — the support function of $\boldsymbol{\mathcal{U}}_\alpha(\mathbf{X})$ is only dominated by $\mathtt{u}_\alpha(s(\mathbf{X}, u))$, where $\mathtt{u}_\alpha(\beta) = -\mathtt{e}_\alpha(-\beta)$ is a superlinear expectation with

$$\mathtt{u}_\alpha(\beta) = \frac{1}{\alpha}\int_{0}^{\alpha} q_t(\beta)dt.$$

Note that $\mathtt{u}_\alpha(\beta)$ is the average of lower quantiles of $\beta \in L_1(\mathbb{R})$. With this construction, we define the depth function of F as

$$\mathrm{D}(F, \mathbf{X}) = \sup\{\alpha : \mathtt{u}_\alpha(s(\mathbf{X}, u)) \le s(F, u) \le \mathtt{e}_\alpha(s(\mathbf{X}, u)), \text{for all } u \in \mathbb{S}^{d-1}\}. \tag{3.24}$$

The construction can be modified if $\mathbf{X}$ almost surely contains the origin. Then $\mathbf{X}$ is *star-shaped* and can be identified by its *radial function*

$$r(\mathbf{X}, u) = \sup\{t \in \mathbb{R} : xu \in \mathbf{X}\}, \quad t \neq 0.$$

This function has a nice property, namely

$$r(\boldsymbol{\mathcal{U}}_\alpha(\mathbf{X}), u) = \mathtt{u}_\alpha(r(\mathbf{X}, u)),$$

which overcomes the problem of using support function with

$$s(\boldsymbol{\mathcal{U}}_\alpha(\mathbf{X}), u) \le \mathtt{u}_\alpha(s(\mathbf{X}, u)).$$

In this case, we can replace (3.24) by

$$\begin{aligned}\mathrm{D}(F, \mathbf{X}) = \sup\{&\alpha : \mathtt{u}_\alpha(r(\mathbf{X}, u)) \le r(F, u) \text{ and} \\ &s(F, u) \le \mathtt{e}_\alpha(s(\mathbf{X}, u)), \text{for all } u \in \mathbb{S}^{d-1}\}.\end{aligned}$$

Example 3.3.2 (Examples 4.10 and 4.11 in Paper C). If $\mathbf{X} = \{\xi\}$ is a random variable with $\xi \in L_p(\mathbb{R}^d)$, then $\boldsymbol{\mathcal{E}}_\alpha(\{\xi\})$ is the convex closed set with the support function given by $\mathtt{e}_\alpha(\langle \xi, u\rangle)$ and it coincides with the zonoid-trimmed region of ξ introduced by Koshevoy and Mosler [1997], while $\boldsymbol{\mathcal{U}}_\alpha(\{\xi\}) = \emptyset$ for $\alpha < 1$ when ξ is not deterministic. Recall that the *zonoid-trimmed region* of ξ is the set

$$\mathrm{ZD}^\alpha(\xi) = \alpha^{-1}\{x \in \mathbb{R}^d : \ (\alpha, x) \in \mathbf{E}(\mathrm{co}(0, (1, \xi)))\} \tag{3.25}$$

obtained as the normalized section of the expectation $\mathbf{E}(\mathbf{Y})$ of the random closed convex set $\mathbf{Y}$, being the convex hull of the origin and the point $(1, \xi)$ in $\mathbb{R}^{d+1}$. The expectation $\mathbf{E}(\mathbf{Y})$ is termed the lift zonoid of ξ. The corresponding depth concept is called the zonoid depth of ξ, see Koshevoy and Mosler [1997] and Mosler [2002]. The lift zonoid concept was extended for general random convex sets by Diaye et al. [2018]. By considering $\mathbf{X}$ is a random interval (see Example 4.1 in Diaye et al. [2018]), the proposed depth-trimmed region of $\mathbf{X}$ is equivalent to

$$\mathrm{D}^\alpha(\mathbf{X}) = \alpha^{-1}\{a \in \mathbb{R} : a \times u \le \mathtt{e}_\alpha(s(\mathbf{X}, u)), \text{for } u \in [-1, 1]\}.$$

This proposal is very similar to our proposal (3.24) but only giving the upper bound.

3.3.4 Half-space depth [Type D]

Since the support function $s(F,u)$, $u \in \mathbb{S}^{d-1}$, uniquely identifies each closed convex set F, it is possible to use the concept of half-space depth for functional data suggested by Kuelbs and Zinn [2013]. Then the half-space depth of $F \in \mathcal{F}_{\mathcal{C}}(\mathbb{R}^d)$ with respect to $\mathbf{X}$ is defined as

$$\mathrm{HD}(F,\mathbf{X}) = \min(\mathrm{HD}_+(F,\mathbf{X}), \mathrm{HD}_-(F,\mathbf{X})),$$

where

$$\mathrm{HD}_+(F,\mathbf{X}) = \inf_{u\in\mathbb{S}^{d-1}} \mathbf{P}\{s(\mathbf{X},u) \geq s(F,u)\}, \tag{3.26}$$

$$\mathrm{HD}_-(F,\mathbf{X}) = \inf_{u\in\mathbb{S}^{d-1}} \mathbf{P}\{s(\mathbf{X},u) \leq s(F,u)\}. \tag{3.27}$$

Given a sample $X_1,\dots,X_n$ drawn from $\mathbf{X}$, one can replace the probabilities by the sample averages of $\mathbf{1}\{s(X_i,u) \geq s(F,u)\}$ to construct the empirical versions of HD_+ and HD_-.

Using this definition working in a general functional space, most of the function will have zero depth. However, it is not the case for support functions of random convex compact sets.

Theorem 3.3.3 (Theorem 5.3 in Paper C)**.** *If* $\mathbf{X}$ *is a random convex compact set, then* $\mathrm{HD}_+(F,\mathbf{X}) > 0$ *(respectively,* $\mathrm{HD}_-(F,\mathbf{X}) > 0$*) for all* F *from the support of* $\mathbf{X}$ *such that* $\mathbf{P}\{s(F,u) \leq s(\mathbf{X},u)\} > 0$ *(respectively,* $\mathbf{P}\{s(F,u) \geq s(\mathbf{X},u)\} > 0$*) for all* u*. Furthermore, the infima in* (3.26) *and* (3.27) *are attained.*

Another concept called half-region depth and its variation for functional data (see López-Pintado and Romo [2011], Kuelbs and Zinn [2015]) can also be generalized to the set-valued data, see equation (19) and (20) in Paper C.

3.3.5 Other set-valued data

In Example 9.1 of Paper C we considered a data set, where observations are random cumulative distribution functions, equivalently, a sample of random measures. Using the fact that lift zonoid uniquely identifies a probability distribution, each distribution function can be represented by its lift zonoid which is a compact convex set in $\mathbb{R}^2$. Then we can use any of previously introduced depth-based methods to identify outliers for samples of random measures.

Generally, sets have special features that may be taken into account in statistical procedures, such as position, size and shape, see Dryden and Mardia [1997]. However, in some cases not all these factors are relevant. For example, the locations and orientations of particles, like stones are not relevant for their statistical analysis, see Stoyan and Stoyan [1994], Stoyan and Molchanov [1997]. On the other hand, position of a partially identified variable (e.g., salary interval from a questionnaire) is important for the analysis. In Section 8 of Paper C, we propose a depth function factoring out location, if locations are irrelevant for the statistical analysis. In Section 10.2 we applied our depth functions to a sample of particles. A comparison with the visual perception shows that the half-region depth is too sensitive to outliers, while the average quantile depth, band depth and modified half-space depth show the best performance.

Bibliography

K. Adusumilli and T. Otsu. Empirical likelihood for random sets. *J. Amer. Statist. Assoc.*, 112(519):1064–1075, 2017.

A. C. Atkinson, A. N. Donev, and R. D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, Oxford, 2007.

A. Beresteanu and F. Molinari. Asymptotic properties for a class of partially identified models. *Econometrica*, 76:763–814, 2008.

A. Blanco-Fernández, N. Corral, and G. González-Rodrí guez. Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comput. Statist. Data Anal.*, 55(9):2568–2578, 2011.

A. Blanco-Fernández, A. Colubi, and G. González-Rodríguez. Confidence sets in a linear regression model for interval data. *J. Statist. Plann. Inference*, 142:1320–1329, 2012.

A. Blanco-Fernández, A. Colubi, and M. García-Bárzana. A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables. *Inform. Sci.*, 247:109–122, 2013a.

A. Blanco-Fernández, A. Colubi, and G. González-Rodríguez. Linear regression analysis for interval-valued data based on set arithmetic: A review. In C. Borgelt, M. Á. Gil, J. M. Sousa, and M. Verleysen, editors, *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, pages 19–31. Springer, Berlin, Heidelberg, 2013b.

C. Bontemps, T. Magnac, and E. Maurin. Set identified linear models. *Econometrica*, 80:1129–1155, 2012.

I. Cascos. The expected hull trimmed regions of a sample. *Comp. Statist.*, 22:557–569, 2007.

I. Cascos. Data depth: multivariate statistics and geometry. In W. S. Kendall and I. Molchanov, editors, *New Perspectives in Stochastic Geometry*, pages 398–426. Oxford University Press, Oxford, 2010.

I. Cascos and I. Molchanov. Multivariate risks and depth-trimmed regions. *Finance Stoch.*, 11:373–397, 2007.

I. Cascos and I. Molchanov. Band depths based on multiple time instances. In E. Gil, E. Gil, J. Gil, and M.-A. Gil, editors, *The Mathematics of the Uncertain. A tribute to Pedro Gil*, pages 67–78. Springer, 2018.

A. Chandrasekhar, V. Chernozhukov, F. Molinari, and P. Schrimpf. Inference for best linear approximations to set identified functions. CeMMAP Working Paper CWP 43/12, 2012.

S. I. Chang. Some properties of multiresponse D-optimal designs. *J. Math. Anal. Appl.*, 184:256–262, 1994.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.*, 46(5):1932–1960, 2018.

F. Ciliberto and E. Tamer. Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828, 2009.

I. Couso and D. Dubois. Statistical reasoning with set-valued information: ontic vs. epistemic views. *Internat. J. Approx. Reason.*, 55(7):1502–1518, 2014.

P. Diamond. Least squares fitting of compact set-valued data. *J. Math. Anal. Appl.*, 147 (2):351–362, 1990.

M.-A. Diaye, G. A. Koshevoy, and I. Molchanov. Lift expectations of random sets and their applications. *Statist. Prob. Lett.*, 145:110–117, 2018.

I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. Wiley, Chichester, 1997.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, 21(1):196–216, 1993.

J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, 20(4):2008–2036, 1992.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996.

G. Fisher. Linear regression problems in case of interval-identied data. Master's thesis, University of Bern, Bern, 2010.

H. Föllmer and A. Schied. *Stochastic finance*. Walter de Gruyter & Co., Berlin, 2 edition, 2004.

I. Gijbels and S. Nagy. On a general definition of depth for functional data. *Statist. Sci.*, 32(4):630–639, 2017.

M. A. Gil, M. T. López-García, M. A. Lubiano, and M. Montenegro. Regression and correlation analyses of a linear relation between random intervals. *Test*, 10:183–201, 2001.

R. D. Gill, M. J. van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In D. Y. Lin and T. R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294, New York, NY, 1997. Springer US.

G. González-Rodríguez, Á. Blanco, N. Corral, and A. Colubi. Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification*, 1(1):67–81, 2007.

D. F. Heitjan. Ignorability in general incomplete-data models. *Biometrika*, 81(4):701–708, 1994.

D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *Ann. Statist.*, 19(4): 2244–2253, 1991.

L. Imhof. Optimum designs for a multiresponse regression model. *J. Multivariate Anal.*, 72:120–131, 2000.

M. Jacroux. On the MV-optimality of chemical balance weighing designs. *Calcutta Statist. Assoc. Bull.*, 32:143–151, 1983.

H. Kaido. Asymptotically efficient estimation of weighted average derivatives with an interval censored variable. *Econometric Theory*, 2016. Forthcoming.

J. Kiefer. General equivalence theory for optimum designs (approximate theory). *Ann. Statist.*, 2:849–879, 1974.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Can. J. Math.*, 12:363–366, 1960.

A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea, New York, 1950.

G. Koshevoy and K. Mosler. Zonoid trimming for multivariate distributions. *Ann. Statist.*, 25:1998–2017, 1997.

O. Krafft and M. Schaefer. *D*-optimal designs for a multivariate regression model. *J. Multivariate Anal.*, 42:130–140, 1992.

J. Kuelbs and J. Zinn. Concerns with functional depth. *ALEA Lat. Am. J. Probab. Math. Stat.*, 10(2):831–855, 2013.

J. Kuelbs and J. Zinn. Half-region depth for stochastic processes. *J. Multivariate Anal.*, 142:86–105, 2015.

V. G. Kurotschka and R. Schwabe. The reduction of design problems for multivariate experiments to univariate possibilities and their limitations. In E. Brunner and M. Denker, editors, *Research Developments in Probability and Statistics*, pages 193–204. VSP, Utrecht, 1996.

P. Lafaye De Micheaux, P. Mozharovskyi, and M. Vimond. Depth for curve data and applications. *J. Am. Stat. Assoc.*, page to appear, 2020.

R. Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18: 405–414, 1990.

R. Y. Liu, J. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *Ann. Statist.*, 27:783–858, 1999.

J. López-Fidalgo, B. Torsney, and R. Ardanuy. Mv-optimization in weighted linear regression. In *Proceedings of MODA*, number 5, pages 39–50. Springer Verlag, 1998.

S. López-Pintado and J. Romo. On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104:718–734, 2009.

S. López-Pintado and J. Romo. A half-region depth for functional data. *Comput. Statist. Data Anal.*, 55:1679–1695, 2011.

T. Maatouk. *Some application of nonparametric regression with constrained data*. PhD thesis, University of Glasgow, Glasgow, 2003.

C. Manski. *Identification for Prediction and Decision*. Harvard University Press, 2007. ISBN 9780674026537.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer Verlag, New York, 2003.

C. F. Manski and E. Tamer. Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70:519–546, 2002.

G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

I. Molchanov. *Theory of Random Sets*. Springer, London, 2 edition, 2017.

I. Molchanov and F. Molinari. *Random Sets in Econometrics*. Cambridge University Press, Cambridge, 2018.

I. Molchanov and A. Mühlemann. Nonlinear expectations of random sets. *Finance Stoch.*, 25:5–41, 2021.

K. Mosler. *Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach*, volume 165 of *Lect. Notes Stat.* Springer, Berlin, 2002.

S. Nagy, C. Schütt, and E. Werner. Halfspace depth and floating body. *Stat. Surv.*, 13: 52–118, 2019.

D. Paindaveine and G. Van Bever. Halfspace depths for scatter, concentration and shape matrices. *Ann. Statist.*, 46(6B):3276–3307, 2018.

G. Pandolfo, D. Paindaveine, and G. Porzio. Distance-based depths for directional data. *Canad. J. Statist.*, 46(4):593–609, 2018.

S. Peng. Nonlinear expectations, nonlinear evaluations and risk measures. In *Stochastic Methods in Finance*, volume 1856 of *Lecture Notes in Math.*, pages 165–253. Springer, Berlin, 2004.

S. Peng. *Nonlinear Expectations and Stochastic Calculus under Uncertainty*. Springer, Berlin, 2019.

P. Rousseeuw and I. Ruts. The depth function of a population distribution. *Metrika*, 49: 213–244, 1999.

R. Schneider. *Convex Bodies. The Brunn–Minkowski Theory*. Cambridge University Press, Cambridge, 2 edition, 2014.

G. Schollmeyer and T. Augustin. Statistical modeling under partial identification: distinguishing three types of identification regions in regression analysis with interval data. *Internat. J. Approx. Reason.*, 56(part B):224–248, 2015.

S. D. Silvey. *Optimal Design*. Chapman and Hall, London, 1980.

B. Sinova, A. Colubi, M. Á. Gil, and G. González-Rodríguez. Interval arithmetic-based simple linear regression between interval data: discussion and sensitivity analysis on the choice of the metric. *Inform. Sci.*, 199:109–124, 2012.

M. Soumaya, N. Gaffke, and R. Schwabe. Optimal design for multivariate observations in seemingly unrelated linear models. *J. Multivariate Anal.*, 142:48–56, 2015.

D. Stoyan and I. S. Molchanov. Set-valued means of random particles. *J. Math. Imaging Vision*, 7(2):111–121, 1997.

D. Stoyan and H. Stoyan. *Fractals, Random Shapes and Point Fields*. Wiley, Chichester, 1994.

E. Tamer. Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1):147–165, 2003. ISSN 00346527, 1467937X.

B. Torsney and J. López-Fidalgo. Mv-optimization in simple linear regression. In *Proceedings of MODA*, number 4, pages 57–69. Springer Verlag, 1995.

J. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*, pages 523–531. 1975.

Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Stat.*, 28: 461–482, 2000.

Appendix

Paper A

International Journal of Approximate Reasoning 128 (2021) 129–150

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Local regression smoothers with set-valued outcome data ☆

Qiyu Li [a,b], Ilya Molchanov [a], Francesca Molinari [c,*], Sida Peng [d]

[a] Department of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
[b] Swiss Group for Clinical Cancer Research (SAKK), Switzerland
[c] Department of Economics, Cornell University, United States of America
[d] Office of Chief Economist, Microsoft, United States of America

ARTICLE INFO

Article history:
Received 19 September 2018
Received in revised form 9 January 2020
Accepted 7 October 2020
Available online 17 October 2020

Keywords:
Local regression smoothers
Set valued outcome data
Random sets
Support function

ABSTRACT

This paper proposes a method to conduct local linear regression smoothing in the presence of set-valued outcome data. The proposed estimator is shown to be consistent, and its mean squared error and asymptotic distribution are derived. A method to build error tubes around the estimator is provided, and a small Monte Carlo exercise is conducted to confirm the good finite sample properties of the estimator. The usefulness of the method is illustrated on a novel dataset from a clinical trial to assess the effect of certain genes' expressions on different lung cancer treatments outcomes.

1. Introduction

Statistical analysis has traditionally contended with problems of data imprecision due to limits in the measuring instruments and to measurement error, as well as with missing data, data coarsening and grouping. Geostatistical analysis and mathematical morphology have contended with observational frameworks where the outcome of interest is a two or three dimensional set-valued object, e.g. a tumor or a grain. The common denominator of these challenging data-frameworks is the presence of set-valued data. Within the social sciences in particular, collection of data in the form of sets, especially intervals, has become increasingly widespread. For example, the Health and Retirement Study is one of the first surveys where, in order to reduce item nonresponse, income data is collected from respondents in the form of brackets, with degenerate (singleton) intervals for individuals who opt to fully report their income (see, e.g. [1]). To reduce response burden, the Occupational Employment Statistics (OES) program at the Bureau of Labor Statistics collects wage data from employers as intervals, and uses these data to construct estimates for wage and salary workers in 22 major occupational groups and 801 detailed occupations. Privacy concerns often motivate providing public use tax data as the number of tax payers in each of a finite number of cells. In the medical field, due to ethical and cost reasons, time-to-event measurements are not collected on a continuous scale, but at pre-specified time intervals.

The partial identification literature in econometrics (e.g., [2]) has addressed the question of what can be learned about functionals of probability distributions, when some of the variables are only known to belong to (random) sets and no assumptions are imposed on the distribution of the true variables within these sets. We take the identification results of this literature as our point of departure. Our contribution is to provide statistical results on local linear regression smoothing

☆ We are grateful to the Editor, the Area Editor, and two anonymous referees for comments that helped us substantially improve the paper. Molinari gratefully acknowledges support from NSF grant SES1824375.

* Corresponding author.
E-mail address: fm72@cornell.edu (F. Molinari).

https://doi.org/10.1016/j.ijar.2020.10.005
0888-613X/

when the outcome data is set-valued and the regressors are exactly measured. Our paper relaxes the textbook setting (e.g., [3]) of nonparametric regression – where regressors and outcome data $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, are precisely measured – by assuming that $\boldsymbol{y}_i$ is only known to belong to an observed set $\boldsymbol{Y}_i$. In other words, we deal with an independently and identically distributed sample of observations for the pair $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$ composed of a random vector $\boldsymbol{x}_i$ in $\mathbb{R}^m$ and a random convex compact set $\boldsymbol{Y}_i$ in $\mathbb{R}^d$. Independence and identical distribution for random sets and measurability of $\boldsymbol{Y}$ are notions made precise in Appendix D, while in Section 2 we explain that the distribution of $\boldsymbol{Y}$ can be characterized as a *belief function*. The true (however unobservable) outcome associated with $\boldsymbol{x}$ is a random vector $\boldsymbol{y}$ that almost surely takes values in $\boldsymbol{Y}$. Our goal is to provide a nonparametric regression estimator for the expectation conditional on $\boldsymbol{x}$ of each random vector $\boldsymbol{y} \in \boldsymbol{Y}$. One can think of such expectation as the first-order moment of the belief function generated by $\boldsymbol{Y}$ conditional on $\boldsymbol{x}$.

For a given tuple $(\boldsymbol{x}, \boldsymbol{y})$ that almost surely belongs to $\{\boldsymbol{x}\} \times \boldsymbol{Y}$, we denote by $m(x) = \mathbf{E}[\boldsymbol{y}|\boldsymbol{x} = x]$ the regression function for the chosen $(\boldsymbol{x}, \boldsymbol{y})$. Each choice of $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. gives rise to a function m and we denote by $\mathcal{M}$ the family of all regression functions generated in this manner. We let $M(x) \equiv \{m(x) : m \in \mathcal{M}\}$ and we observe that

$$M(x) = \mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x] = \Big\{\mathbf{E}[\boldsymbol{y}|\boldsymbol{x} = x] : \boldsymbol{y} \in \boldsymbol{Y} \text{ a.s.}\Big\}$$

is the conditional selection expectation of $\boldsymbol{Y}$, see [4, Sec. 2.1.6] and Section 2.

For example, consider the empirically relevant case that $d = 1$ and $\boldsymbol{Y} = [\boldsymbol{y}_\mathrm{L}, \boldsymbol{y}_\mathrm{U}]$ for two random variables $\boldsymbol{y}_\mathrm{L}, \boldsymbol{y}_\mathrm{U}$ such that $\mathbf{P}(\boldsymbol{y}_\mathrm{L} \leq \boldsymbol{y}_\mathrm{U}) = 1$. Then

$$M(x) = \Big[\mathbf{E}[\boldsymbol{y}_\mathrm{L}|\boldsymbol{x} = x], \mathbf{E}[\boldsymbol{y}_\mathrm{U}|\boldsymbol{x} = x]\Big]. \tag{1}$$

Our proposal is to estimate $M(x)$ as a weighted sum of the sets $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, with weights defined as in the local linear estimation literature.[1] The development of our technical results directly builds on classic references such as [5] and [6], and is closely related to [7] and [3].

For the case that $d = 1$, inspection of equation (1) might suggest to report an estimator given by the interval between a local constant or local linear regression of $\boldsymbol{y}_\mathrm{L}$ on $\boldsymbol{x}$ and one of $\boldsymbol{y}_\mathrm{U}$ on $\boldsymbol{x}$. Alternatively, it might suggest to report a local constant or local linear regression of the interval midpoint, $\tilde{\boldsymbol{y}} = (\boldsymbol{y}_\mathrm{L} + \boldsymbol{y}_\mathrm{U})/2$, and of the interval width, $\boldsymbol{w} = \boldsymbol{y}_\mathrm{U} - \boldsymbol{y}_\mathrm{L}$, on $\boldsymbol{x}$. While both in finite sample and asymptotically these approaches are equivalent to what we propose for the case of a local constant regression, for the case of local linear regression equivalence breaks down in finite sample. The difference is important: we show in Remark 3.1 below that the alternative estimators just described may lead to a finite sample bias understating the width of $M(x)$ and are therefore unpalatable. For example, such estimators might be empty or a singleton in finite sample even though $M(x)$ is an interval of strictly positive width in population. In contrast, the estimator that we propose does not suffer from this problem, although it does have an asymptotic bias term similar to that of point identified local linear regression estimators.

Our approach is the first contribution in the literature to local regression smoothing when the set-valued outcome variable is in $\mathbb{R}^d$ with $d > 1$. We derive the asymptotic properties of our estimator and extend results from [8] to obtain pointwise confidence bands that asymptotically cover the functional of interest with probability $1 - \alpha$. We report the results of Monte Carlo simulations with interval-valued $\boldsymbol{Y}$ and with $\boldsymbol{Y}$ being a ball randomly placed on the plane that support our theoretical findings.

We also demonstrate the usefulness of our approach with an empirical illustration that uses a novel dataset from a clinical trial on non-small-cell lung cancer patients, to study the relationship between time to tumor progression and specific gene expression measures.

Related literature. Within the partial identification literature, there is a large body of work analyzing regression with interval-valued data. [9] consider models where one variable (either outcome or covariate) is observed as intervals and all others are perfectly measured, and provide identification results for nonparametric as well as parametric models in this setting. [8] introduce to the partial identification literature the use of random set theory and provide results on identification and inference on best linear prediction parameters (ordinary least squares) when the outcome variable is interval-valued and the regressors are perfectly measured. [10] extend the familiar Sargan test for overidentifying restrictions to the setting studied by [8]. [11] extend [8]'s approach to cover best linear approximation of any function $f(x)$ that is known to lie within two identified bounding functions. [12] proposes an estimator for weighted average derivatives of conditional mean and conditional quantile functionals when either the outcome variable or a regressor is interval-valued. [13] propose empirical likelihood methods for random sets to conduct inference in the class of problems analyzed by [8]. All these papers focus exclusively on the case that the set-valued outcome data is in $\mathbb{R}$.

In contrast, our approach leverages the theory of random sets to propose a set-valued local linear regression estimator for conditional set-valued expectations with $\boldsymbol{Y} \subset \mathbb{R}^d, d \geq 1$, and to establish its asymptotic properties. This estimator is novel in the literature, and so are our results establishing its consistency and asymptotic distribution.

[1] We comment on the case of local constant (Nadaraya–Watson) estimator in Appendix C.

The method that we propose differs significantly from other approaches in the statistical literature; see [14] for a discussion bridging this literature with partial identification. In particular, our proposal is distinct from the large and closely related literature that posits parametric models for set-valued data. In these models tools from interval arithmetic are used to build analogs of the classic linear regression model for perfectly measured data, e.g. by assuming that $\mathbf{E}[\boldsymbol{Y}_i|\boldsymbol{x}_i] = A\boldsymbol{x}_i + B$, where A and B are intervals. See e.g. [15], [16], [17], and [18] among others for a discussion of least squares analysis of this and related models. [19] proposes nonparametric smoothing for this model, by applying weighted least squares to the interval data and then using the resulting intercept as the estimator. [20] discuss various interpretations of set-valued data. Compared to this literature, we leave the conditional set-valued expectation completely unspecified, and nonparametrically estimate all regression functions compatible with the interval-valued data.

Finally, our proposal is distinct from the literature on data coarsening, e.g. [21], [22] and [23]. In that literature, the key assumption of "coarsening at random" requires that for any possible value A of the random set $\boldsymbol{Y}$ and a random vector $\boldsymbol{y}$ that almost surely belongs to $\boldsymbol{Y}$, the conditional probability $\mathbf{P}(\boldsymbol{Y} = A|\boldsymbol{y} = y_0)$ does not depend on $y_0 \in A$. This assumption restricts directly the conditional distribution of the random set $\boldsymbol{Y}$, whereas we leave this distribution completely unrestricted.

Structure of the paper. In Section 2 we set up our notation and we briefly review local linear regression with singleton data. Our method implicitly applies it to each tuple $(\boldsymbol{x}, \boldsymbol{y}) : (\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. In Section 3 we propose our estimator and in Section 4 derive its asymptotic properties. In Section 5 we describe a cross-validation method for bandwidth selection, and we extend the methods proposed by [8] to test a hypothesis about the conditional expectation (evaluated at x_0) and to build pointwise error bands with prespecified asymptotic coverage. In Section 6 we report the results of Monte Carlo experiments and in Section 7 the results of our empirical illustration. Section 8 concludes. All technical proofs are collected in Appendix A. Throughout we consider the case that the regressors $\boldsymbol{x}$ are random variables (random design case). In keeping with the tradition in the statistics literature (e.g., [3]), we also report in Appendix B the case of deterministic design (nonstochastic explanatory variables). Appendix C briefly discusses the local constant regression case. Appendix D reports some basic facts in convex geometry and random set theory that we use throughout the paper. We refer to [4] for a thorough account of random sets theory. Appendix E provides additional simulation results.

2. Notation and preliminaries

We begin with listing our notation. We use boldface capital letters $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ to denote random compact convex sets, normal font capital letters X, Y, Z and A, B, C to denote deterministic compact convex sets, boldface lower case letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ to denote random vectors or random variables, and normal font lowercase letters x, y, z to denote deterministic vectors. For $x \in \mathbb{R}$, we denote the positive and negative parts of x respectively by $x^+ = \max(0, x)$ and $x^- = -\min(0, x)$. We let $(\Omega, \mathfrak{F}, \mathbf{P})$ denote a nonatomic probability space on which all random vectors and random sets that we work with are defined, where Ω is the space of elementary events equipped with σ-algebra $\mathfrak{F}$ and probability measure $\mathbf{P}$. We denote the Euclidean space by $\mathbb{R}^d$, and equip it with the Euclidean norm (which is denoted by $\|\cdot\|$). We denote by $\mathcal{K}(\mathbb{R}^d)$ the collection of compact subsets of $\mathbb{R}^d$ and by $\mathcal{K}_{\mathcal{C}}(\mathbb{R}^d)$ the family of non-empty compact convex sets, also called convex bodies. We let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ denote the unit sphere in $\mathbb{R}^d$.

We assume that $\boldsymbol{Y}$ is a random set in $\mathbb{R}^d$ taking almost surely compact and convex values. In terms of measurability requirements, this amounts to

$$\{\omega : \boldsymbol{Y}(\omega) \cap K \neq \emptyset\} \in \mathfrak{F} \quad \forall K \in \mathcal{K}(\mathbb{R}^d). \tag{2}$$

The probabilities $\mathbf{P}(\boldsymbol{Y} \subseteq K)$, $K \in \mathcal{K}(\mathbb{R}^d)$, called the *containment functional* of $\boldsymbol{Y}$, fully characterize the distribution of $\boldsymbol{Y}$, [e.g.,4, Thm. 1.8.9]. As function of K, these probabilities are special cases of the *belief functions*, see [24] and more recently [25] and [26]. While general belief functions do not necessarily satisfy regularity conditions specific for the containment functional, the containment functionals are exactly semicontinuous belief functions. Then $\boldsymbol{Y}$ describes the possible regions where a true value lies, and hence represents the ambiguity embedded in the observations, and coincides with the multivalued mapping Γ in [24].

To set the stage for local regression smoothing, we recall the standard construction of the local polynomial estimators for singleton-valued outcomes, see e.g. [6]. Suppose one is interested in estimating $\mathbf{E}(\boldsymbol{y}_i|\boldsymbol{x}_i = x_0)$ based on observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, where x_0 is a given value on the support of $\boldsymbol{x}$ (e.g., a particular level of the gene expression measure in our empirical study). Then one fits a p-th order local model

$$\boldsymbol{y}_i = \theta_0(x_0) + \theta_1(x_0)(\boldsymbol{x}_i - x_0) + \cdots + \theta_p(x_0)(\boldsymbol{x}_i - x_0)^p + \varepsilon_i,$$

using the regressor $\boldsymbol{x}_i - x_0$ (rather than $\boldsymbol{x}_i$) so that the intercept equals $\mathbf{E}(\boldsymbol{y}_i|\boldsymbol{x}_i = x_0)$. In this expression, the coefficients θ are written as a function of x_0 to emphasize that they change with the evaluation point (and this is what makes the model "local"); to simplify notation, such dependence is suppressed henceforth. The local polynomial estimator of order p is then obtained by minimizing the weighted least squares

$$\sum_{i=1}^{n}\Big(\boldsymbol{y}_i-\theta_0-\theta_1(\boldsymbol{x}_i-x_0)-\cdots-\theta_p(\boldsymbol{x}_i-x_0)^p\Big)^2 K\Big(\frac{\boldsymbol{x}_i-x_0}{h_n}\Big) \tag{3}$$

with respect to $\theta_0,\dots,\theta_p$. The kernel function $K(\cdot)$ is a nonnegative integrable function and the tuning parameter h_n is the bandwidth. As it is typically done, we assume that $h_n\to 0$ and $nh_n\to\infty$ as $n\to\infty$. The following condition on the kernel function is imposed throughout this paper.

Assumption A *(Kernel function).* The kernel $K(z)$, $z\in\mathbb{R}$, is a nonnegative function bounded above by $K_{\max}<\infty$, with compact support $[-c_K,c_K]$ for some $c_K\in(0,\infty)$, and satisfying

$$\int K(z)\,dz=1,\qquad \int zK(z)\,dz=0.$$

Denote $\mathrm{Var}_K=\int z^2K(z)\,dz$.

The normalization conditions on K are standard, while the compact support ensures that observations sufficiently far (compared to the order of the bandwidth) from the current point do not influence the estimator at this point, see also Appendix B.

Solving explicitly the weighted least squares minimization problem in (3), one obtains the minimizer $\hat\theta$, and the first entry of it, the intercept $\hat\theta_0$, is the estimate of $m(x_0)$. This estimator can be written as

$$\hat{\boldsymbol{m}}(x_0)=\sum_{i=1}^{n}\ell_i(x_0)\boldsymbol{y}_i, \tag{4}$$

where

$$\ell_i(x_0)=\frac{1}{nh_n}u^\top(0)\mathcal{B}_{nx_0}^{-1}u\Big(\frac{\boldsymbol{x}_i-x_0}{h_n}\Big)\boldsymbol{\kappa}_{in},$$

$$u(z)=\Big(1,z,z^2/2!,\dots,z^p/p!\Big)^\top,$$

$$\mathcal{B}_{nx_0}=\frac{1}{nh_n}\sum_{i=1}^{n}u\Big(\frac{\boldsymbol{x}_i-x_0}{h_n}\Big)u^\top\Big(\frac{\boldsymbol{x}_i-x_0}{h_n}\Big)\boldsymbol{\kappa}_{in},$$

with $\boldsymbol{\kappa}_{in}=K\big(\frac{\boldsymbol{x}_i-x_0}{h_n}\big)$. Note that $\ell_i(x_0)$, $i=1,\dots,n$, sum up to one, and write

$$\boldsymbol{s}_j=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\kappa}_{in}(\boldsymbol{x}_i-x_0)^j,\qquad j=0,1,\dots$$

It is easy to see that $\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2\geq 0$, and that the right-hand side of (4) is linear in the response variables, since the weights do not depend on the $\boldsymbol{y}_i$'s.

If $p=0$ (local constant regression), $\hat{\boldsymbol{m}}(x_0)$ is the Nadaraya-Watson estimator with $\ell_i(x_0)=\boldsymbol{\kappa}_{in}/(n\boldsymbol{s}_0)$. If $p=1$ (local linear regression), then

$$\ell_i(x_0)=\frac{\boldsymbol{\kappa}_{in}}{n}\frac{\boldsymbol{s}_2-(\boldsymbol{x}_i-x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2}. \tag{5}$$

Our goal is to extend the local linear regression framework to set-valued outcomes: we propose an analog to estimator (4) with $p=1$ and $\ell_i(x_0)$ as given in (5), for the case that instead of knowing the exact value of $\boldsymbol{y}$, it is only assumed that $\boldsymbol{y}$ almost surely belongs to a random set $\boldsymbol{Y}$. In this case $\boldsymbol{y}$ is said to be a (measurable) *selection* of $\boldsymbol{Y}$. Distributions of all selections of $\boldsymbol{Y}$ can be identified with the probability measures from the core of the belief function generated by $\boldsymbol{Y}$, that is, probability measures dominating the belief function. The pair $(\boldsymbol{x},\boldsymbol{y})$ is a selection of $\{\boldsymbol{x}\}\times\boldsymbol{Y}$, a random closed set in $I\times\mathbb{R}^d$ with I the support of $\boldsymbol{x}$. This framework can alternatively be described as associating with each value of the explanatory variable $\boldsymbol{x}$ a belief function describing the (conditional) distribution of $\boldsymbol{Y}$.

Whereas in the standard case of singleton-valued outcomes one observes singleton-valued data $(\boldsymbol{x}_i,\boldsymbol{y}_i)$, $i=1,\dots,n$, in our framework the observations are set-valued, $(\boldsymbol{x}_i,\boldsymbol{Y}_i)$, $i=1,\dots,n$. As a result, our estimators are also set-valued, and in order to assess their properties, we need to define square loss for sets, so as to formalize consistency results and the notion of mean squared error. To do so, and to provide a computationally tractable estimator, we exploit the duality between convex sets and their *support function* (see, e.g., Chapter 13 in [27], and (D.2) in Appendix D). The support function of $\boldsymbol{Y}$ in direction $v\in\mathbb{S}^{d-1}$ is given by $s(\boldsymbol{Y},v)\equiv\sup_{\boldsymbol{y}\in\boldsymbol{Y}}v^\top y$, and can be used to define the *width function* of $\boldsymbol{Y}$ in direction $v\in\mathbb{S}^{d-1}$, $w(\boldsymbol{Y},v)\equiv s(\boldsymbol{Y},v)+s(\boldsymbol{Y},-v)$ (see Appendix D). We assume that $\boldsymbol{Y}$ is integrably bounded, that is, $\|\boldsymbol{Y}\|=\sup_{\boldsymbol{y}\in\boldsymbol{Y}}\|y\|$ is

integrable (Assumption B in the next section provides sufficient conditions guaranteeing that this is the case), and since $|s(\boldsymbol{Y}, v)| \leq \|\boldsymbol{Y}\|$ for all v from the unit sphere, this implies that the support function is integrable. It is possible to show that $\mathbf{E}s(\boldsymbol{Y}, v) = s(\mathbf{E}\boldsymbol{Y}, v)$ [see4, Theorem 2.1.35], i.e. the expected support function is the support function of a convex body $\mathbf{E}\boldsymbol{Y}$, which in turn is called the *expectation* of $\boldsymbol{Y}$. This expectation equals the set of values $\mathbf{E}\boldsymbol{y}$ for all random vectors $\boldsymbol{y}$ such that $\boldsymbol{y} \in \boldsymbol{Y}$ a.s.

Similarly, for given x it is possible to define the *conditional expectation*

$$\mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x] = \left\{\mathbf{E}[\boldsymbol{y}|\boldsymbol{x} = x] : \boldsymbol{y} \in \boldsymbol{Y} \text{ a.s.}\right\},$$

and also in this case it holds that $\mathbf{E}[s(\boldsymbol{Y}, v)|\boldsymbol{x} = x] = s(\mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x], v)$ [see, e.g.,4, Sec. 2.1.6]. The set $\mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x]$ is the object of interest in this paper, and one can think of it as the first-order moment of the belief function generated by $\boldsymbol{Y}$ conditional on $\boldsymbol{x}$.

To simplify the exposition, henceforth we assume that $\boldsymbol{x}$ is a scalar random variable and that I is an interval, $I \subset \mathbb{R}$. Our results apply, subject only to modification in notation and convergence rates (as in the point identified case), with vector-valued $\boldsymbol{x}$ provided the real-valued bandwidth is replaced by a matrix-valued one.

The family of support functions of all non-empty compact convex subsets in $\mathbb{R}^d$ is a subset of the family of continuous functions on the unit sphere $\mathbb{S}^{d-1}$. In particular, the Hausdorff metric between compact convex sets equals the uniform (L_∞) distance between their support functions, see e.g. [28, Lemma 1.8.14]. For our purposes, it is convenient to endow the family of continuous functions on the unit sphere with the L_2-metric, so that the distance between two non-empty compact convex sets A_1 and A_2 is given by

$$L(A_1, A_2) = \left(\int_{\mathbb{S}^{d-1}} (s(A_1, v) - s(A_2, v))^2 \, dv \right)^{\frac{1}{2}}. \tag{6}$$

The integration is performed with respect to the uniform measure on $\mathbb{S}^{d-1}$. If $d = 1$, the integral turns into the sum of two terms for $v = 1$ and $v = -1$. The distance to the empty set is assigned to be infinite.

In Section 3, we employ this distance to define the mean square error of our estimator. This distance differs from the standard Hausdorff distance used in the related literature in partial identification and in the standard laws of large numbers and central limit theorems for Minkowski averages of random sets. However, under our assumptions the result of Theorem 3 in [29] yields that these two metrics define the same topology, and so the consistency with respect to the L_2-distance implies consistency with respect to the L_∞-distance. At the same time, use of the L_2-distance is particularly well suited to analyze properties of estimators based on least squares minimization.

3. Nonparametric smoothing for random sets

In the following we assume that $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, is a sample of i.i.d. realizations of $(\boldsymbol{x}, \boldsymbol{Y})$ as defined in Appendix D, where $\boldsymbol{Y}$ satisfies Assumption B introduced below. This i.i.d. assumption is consistent with many collection processes of set-valued data, such as, e.g., the use of unfolding brackets in the Health and Retirement Study, in the Occupational Employment Statistics survey of the Bureau of Labor Statistics, and in the empirical application that we present in Section 7. We relate it to the typical i.i.d. assumption for singleton-valued data following our statement of Assumption B below.

When the outcome data is set-valued, it is necessary to obtain an estimator for the collection of conditional expectations $\mathbf{E}[\boldsymbol{y}|\boldsymbol{x} = x]$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s. This can be accomplished by repeating the procedure in the previous section for all selections of $\{\boldsymbol{x}\} \times \boldsymbol{Y}$. Computationally this is easily achieved by taking the weighted Minkowski average of the $\boldsymbol{Y}_i$ data (see Appendix D for a formal definition of Minkowski sum):

$$\hat{\boldsymbol{M}}(x_0) = \sum_{i=1}^{n} \ell_i(x_0) \boldsymbol{Y}_i. \tag{7}$$

For $p = 0$ we obtain a local constant set-valued regression estimator; the choice $p = 1$ yields a local linear set-valued regression estimator. Note that (7) is also the Fréchet mean of the observed values $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ in the metric given by (6), see [30] and Sec. 2.2.5 in [4].

The estimator in (7) yields a convex set, therefore we can characterize its properties by working with its support function (see (D.2) in Appendix D and Chapter 13 of [27]). To simplify notation, in what follows we omit the argument x_0 in $\ell_i(x_0)$ and write shortly ℓ_i, unless the dependence on x_0 is essential. By representing the difference of its positive and negative parts as $\ell_i = \ell_i^+ - \ell_i^-$, and using that $s(-A, v) = s(A, -v)$ for a convex compact set A and its centrally symmetric set $-A = \{-x : x \in A\}$, we arrive at

$$s(\hat{\boldsymbol{M}}(x_0), v) = s\Big(\sum_{i=1}^{n} \big(\ell_i^+ - \ell_i^-\big)\boldsymbol{Y}_i, v\Big) = \sum_{i=1}^{n} \ell_i^+ s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- s(\boldsymbol{Y}_i, -v)$$
$$= \sum_{i=1}^{n} (\ell_i + \ell_i^-) s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- s(\boldsymbol{Y}_i, -v) = \sum_{i=1}^{n} \ell_i s(\boldsymbol{Y}_i, v) + \sum_{i=1}^{n} \ell_i^- w(\boldsymbol{Y}_i, v).$$

A key feature of the above estimator is that it averages the support function of the set $\boldsymbol{Y}_i$ in direction $+v$ when $\ell_i > 0$, and in direction $-v$ when $\ell_i < 0$. In doing so we guarantee that the estimator is always *non-empty* for any n, a highly desirable feature in light of Assumption B.

Remark 3.1. When $d = 1$ and $\boldsymbol{Y} = [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ with $\mathbf{P}(\boldsymbol{y}_{\mathrm{U}} \geq \boldsymbol{y}_{\mathrm{L}}) = 1$, one might consider two estimators as alternatives to $\hat{\boldsymbol{M}}(x_0)$. One is given by

$$\hat{\boldsymbol{N}}(x_0) = \left[\sum_{i=1}^{n} \ell_i \boldsymbol{y}_{i\mathrm{L}}, \sum_{i=1}^{n} \ell_i \boldsymbol{y}_{i\mathrm{U}}\right].$$

The other is obtained by regressing the midpoint ($\tilde{\boldsymbol{y}}$) and the width ($\boldsymbol{w}$) of the interval $[\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ on $\boldsymbol{x}$ and letting

$$\hat{\boldsymbol{O}}(x_0) = \left[\sum_{i=1}^{n} \ell_i \tilde{\boldsymbol{y}}_i - \sum_{i=1}^{n} \ell_i \frac{\boldsymbol{w}_i}{2}, \sum_{i=1}^{n} \ell_i \tilde{\boldsymbol{y}}_i + \sum_{i=1}^{n} \ell_i \frac{\boldsymbol{w}_i}{2}\right].$$

Standard arguments in [5] yield that $\hat{\boldsymbol{N}}(x_0)$ and $\hat{\boldsymbol{O}}(x_0)$ are consistent estimators of

$$M(x_0) = \mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x_0] = \Big[\mathbf{E}[\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x} = x_0], \mathbf{E}[\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x} = x_0]\Big]$$

with respect to the L_2-distance. However, these estimators can have large finite sample bias, and even be empty (with asymptotically vanishing probability), as illustrated in the following example. Suppose that for i with $\ell_i > 0$, $\boldsymbol{y}_{i\mathrm{L}} = \boldsymbol{y}_{i\mathrm{U}}$; and for i with $\ell_i < 0$, $\boldsymbol{y}_{i\mathrm{U}} > \boldsymbol{y}_{i\mathrm{L}}$.[2] Then

$$\sum_{i=1}^{n} \ell_i \boldsymbol{y}_{i\mathrm{L}} = \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{i\mathrm{L}} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{i\mathrm{L}} = \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{i\mathrm{U}} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{i\mathrm{L}}$$
$$> \sum_{i=1}^{n} \ell_i^+ \boldsymbol{y}_{i\mathrm{U}} - \sum_{i=1}^{n} \ell_i^- \boldsymbol{y}_{i\mathrm{U}} = \sum_{i=1}^{n} \ell_i \boldsymbol{y}_{i\mathrm{U}},$$

and $\hat{\boldsymbol{N}}(x_0)$ is empty. One can similarly show that $\hat{\boldsymbol{O}}(x_0)$ is empty. Similarly empty estimators may result even if $\boldsymbol{y}_{i\mathrm{U}} > \boldsymbol{y}_{i\mathrm{L}}$ whenever $\ell_i > 0$, depending on the realizations of $\boldsymbol{y}_{i\mathrm{L}}$ and $\boldsymbol{y}_{i\mathrm{U}}$, see Fig. 1 for $\hat{\boldsymbol{N}}(x_0)$. Even if one censors $\boldsymbol{w}_i = 0$ if $\ell_i < 0$, the resulting estimator may still in finite sample significantly understate the width of $M(x_0)$.

While the example in Remark 3.1 might appear stylized, it highlights a finite sample problem that can easily occur in practice with interval-valued data, but does not affect the corresponding estimators in the singleton-valued case. The reason is that in the singleton case, local regression smoothers are weighted averages of the observed outcomes. That is also the case for our estimator, $\hat{\boldsymbol{M}}(x_0)$, which averages the *sets* $\boldsymbol{Y}_i$ and indeed is always non-empty. On the other hand, $\hat{\boldsymbol{N}}(x_0)$ and $\hat{\boldsymbol{O}}(x_0)$ average specific *selections* of $\boldsymbol{Y}_i$ (e.g., the extreme points), without recognizing that the sign of the weight may affect which selection is extreme in a given direction.

Throughout the paper we assume $I = \mathbb{R}$ and we impose the following restrictions on the observed and theoretical responses and on the density function of $\boldsymbol{x}$.

Assumption B *(Observed responses).*

(i) Let $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, $i = 1, \ldots, n$, be a sample of i.i.d. realizations of $(\boldsymbol{x}, \boldsymbol{Y})$, $i = 1, \ldots, n$. Conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, are non-empty random compact convex sets.
(ii) $\boldsymbol{Y}_i \subset \xi_i + B$ a.s. for square integrable random vectors ξ_i, $i = 1, \ldots, n$, and a deterministic compact set B that is the same for all i.

[2] While the example is provided for the case $d = 1$, similar constructions can be obtained also when $d \geq 2$.

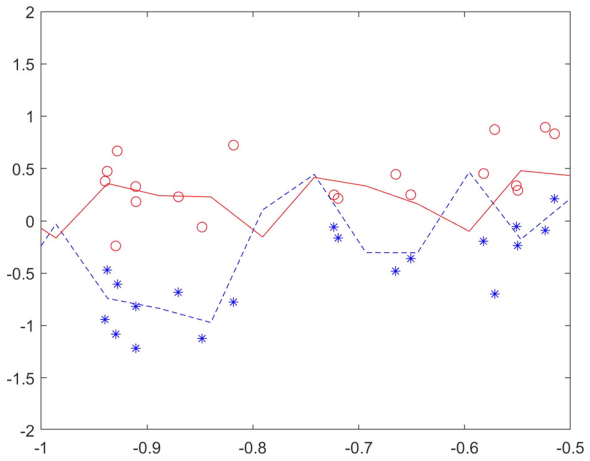

Fig. 1. Possible emptiness of the estimator $\hat{\boldsymbol{N}}(x_0)$. Stars: $(x_i, \boldsymbol{y}_{i\mathrm{L}})$; Circles: $(x_i, \boldsymbol{y}_{i\mathrm{U}})$; dashed line: $\sum_{i=1}^n \ell_i \boldsymbol{y}_{i\mathrm{L}}$; solid line: $\sum_{i=1}^n \ell_i \boldsymbol{y}_{i\mathrm{U}}$.

Define

$$\varepsilon_i(v) \equiv s(\boldsymbol{Y}_i, v) - s(M(\boldsymbol{x}_i), v), \quad v \in \mathbb{S}^{d-1}. \tag{8}$$

By Assumption B, $\varepsilon_i(\cdot)$, $i = 1, \dots, n$, are i.i.d. copies of a square integrable random function $\varepsilon(v)$, $v \in \mathbb{S}^{d-1}$, such that $\mathbf{E}[\varepsilon_i(v)|\boldsymbol{x}_i] = 0$ $\boldsymbol{x}_i$-a.s. for all $v \in \mathbb{S}^{d-1}$. The square integrability follows from the inequality,

$$\varepsilon_i(v) \le s(B, v) + |\xi_i^\top v| + |\eta_i^\top v|,$$

where η_i is a square integrable selection of $M(\boldsymbol{x}_i)$. This selection exists in view of Assumption B(ii) and can be chosen as the point of $M(\boldsymbol{x}_i) = \mathbf{E}(\boldsymbol{Y}_i|\boldsymbol{x}_i) \subset \mathbf{E}(\xi_i|\boldsymbol{x}_i) + B$ nearest to $\mathbf{E}(\xi_i|\boldsymbol{x}_i)$. Note that ε does not admit a geometric interpretation as the support function of a random set.

Denote by $C(v, u) = \mathbf{E}[\varepsilon(v)\varepsilon(u)]$ the covariance function of ε and let $\sigma^2_{\max}$ be the supremum of $C(v, v) = \mathbf{E}[\varepsilon(v)^2]$ over all v from the unit sphere. Assumption B(ii) guarantees that $\boldsymbol{Y}_i$ is uniformly integrably bounded, and implies that the diameters of all $\boldsymbol{Y}_i$'s are bounded by a deterministic constant. Hence, the ambiguity range is limited to belong to a deterministic set, and $\sigma^2_{\max}$ is finite.

It is worth to compare our random sampling assumption with the standard one for singleton-valued variables. In that context, one has $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \varepsilon_i$, and $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are assumed i.i.d., and as a consequence ε_i are i.i.d. In our context, we assume that $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$ are i.i.d., and as a consequence $\varepsilon_i(v)$ are i.i.d.

In dimension $d = 1$, we have $s(\boldsymbol{Y}_i, 1) = \boldsymbol{y}_{i\mathrm{U}}$, $s(\boldsymbol{Y}_i, -1) = -\boldsymbol{y}_{i\mathrm{L}}$, and Part (i) of Assumption B requires that $\boldsymbol{y}_{i\mathrm{L}} = \mathbf{E}[\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x}] - \varepsilon_i(-1)$, $\boldsymbol{y}_{i\mathrm{U}} = \mathbf{E}[\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x}] + \varepsilon_i(1)$ with $\varepsilon_i(1) + \varepsilon_i(-1) \ge -(\mathbf{E}[\boldsymbol{y}_{\mathrm{U}}|\boldsymbol{x}] - \mathbf{E}[\boldsymbol{y}_{\mathrm{L}}|\boldsymbol{x}])$ almost surely. The latter condition replicates the requirement that $\mathbf{P}(\boldsymbol{y}_{\mathrm{U}} \ge \boldsymbol{y}_{\mathrm{L}}) = 1$.

Next, we require the conditional expectation of $\mathbf{E}[\boldsymbol{Y}|\boldsymbol{x}]$ to have a sufficiently smooth support function, thereby allowing for standard expansions used in obtaining the asymptotic properties of the local linear estimator.

Assumption C *(Theoretical response function).* The function $M(x)$, $x \in \mathbb{R}$, is such that $s(M(x), v)$ admits a second derivative $s''(M(x), v)$ in x, uniformly bounded for all $v \in \mathbb{S}^{d-1}$.

In dimension $d = 1$, Assumption C means the second order differentiability of the end-points of the interval-valued function $M(x)$. Finally, we assume that the common density f of the independent design points satisfies the following condition, which is similar to that imposed in Condition 1(ii) of [5] with singleton responses. This is a standard condition in nonparametric regression; it guarantees that the design points are not too concentrated in some areas.

Assumption D *(Density).* The density f is strictly positive at x_0 and belongs to the family $\mathcal{H}(1, \gamma)$ of Lipschitz functions with constant $\gamma > 0$, that is,

$$|f(x') - f(x'')| \le \gamma |x' - x''|$$

for all $x', x'' \in \mathbb{R}$.

We measure the quality of $\hat{\boldsymbol{M}}(x_0)$ as set-valued estimator of $M(x_0)$ by the quadratic loss function defined in (6),

$$L(\hat{\boldsymbol{M}}(x_0), M(x_0))^2 = \int_{\mathbb{S}^{d-1}} (s(\hat{\boldsymbol{M}}(x_0), v) - s(M(x_0), v))^2 \, dv.$$

The mean squared error (MSE) of the estimator is then the expectation of $L(\hat{\boldsymbol{M}}(x_0), M(x_0))^2$. A classic bias-variance decomposition yields

$$\mathrm{MSE}(x_0) = \int_{\mathbb{S}^{d-1}} b^2_{x_0}(v)\, dv + \int_{\mathbb{S}^{d-1}} \sigma^2_{x_0}(v)\, dv,$$

where $b^2_{x_0}(v)$ and $\sigma^2_{x_0}(v)$ are squared bias and variance, given by

$$b^2_{x_0}(v) = \mathbf{E}\Big(\mathbf{E}[s(\hat{\boldsymbol{M}}(x_0), v)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] - s(M(x_0), v)\Big)^2,$$
$$\sigma^2_{x_0}(v) = \mathbf{E}\Big(s(\hat{\boldsymbol{M}}(x_0), v) - s(\mathbf{E}[\hat{\boldsymbol{M}}(x_0)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], v)\Big)^2.$$

Because $\mathbf{E}[\boldsymbol{Y}_i|\boldsymbol{x}_i] = M(\boldsymbol{x}_i)$, we have

$$\mathbf{E}[s(\hat{\boldsymbol{M}}(x_0), v)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = \sum_{i=1}^n \ell_i s(M(\boldsymbol{x}_i), v) + \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i), v).$$

Rearranging the terms, we arrive at

$$b^2_{x_0}(v) = \mathbf{E}\Big(\sum_{i=1}^n \ell_i (s(M(\boldsymbol{x}_i), v) - s(M(x_0), v)) + \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i), v)\Big)^2 \tag{9}$$

and

$$\sigma^2_{x_0}(v) = \mathbf{E}\Big(\sum_{i=1}^n \ell_i (s(\boldsymbol{Y}_i, v) - s(M(\boldsymbol{x}_i), v)) + \sum_{i=1}^n \ell_i^- (w(\boldsymbol{Y}_i, v) - w(M(\boldsymbol{x}_i), v))\Big)^2.$$

By Assumption B, the variance can be expressed as

$$\sigma^2_{x_0}(v) = \mathbf{E}\Big(\sum_{i=1}^n \ell_i \varepsilon_i(v) + \sum_{i=1}^n \ell_i^- (\varepsilon_i(v) + \varepsilon_i(-v))\Big)^2. \tag{10}$$

Differently from the classical case with singleton responses $\boldsymbol{y}_i$, the *negative* parts of the weights in (9) play an essential role with set-valued responses. This is because while the difference between $s(M(\boldsymbol{x}_i), v)$ and $s(M(x_0), v)$ is small when $\boldsymbol{x}_i$ is close to x_0, the width $w(M(\boldsymbol{x}_i), v)$ does not vanish as $\boldsymbol{x}_i$ becomes closer to x_0. Thus, the bias increases by a constant and may not tend to zero if some weights are negative and not close to zero. Much of our asymptotic analysis is concerned with establishing the asymptotic behavior of these negative weights.

The methodology that we propose for local linear regression smoothing can be applied also in the case of local polynomial regression models with $p \geq 2$. In this case, however, extra care is required to show that the negative weights are asymptotically negligible.

4. Asymptotic properties of the set-valued estimators

In the local linear regression setting, negative weights may appear in (9) and hence affect the bias in the case of set-valued data. Following [5], in order to avoid zero in the denominator of the local linear estimator, we redefine ℓ_i by letting

$$\ell_i = \frac{\boldsymbol{\kappa}_{in}}{n} \frac{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}}. \tag{11}$$

We use $\mathcal{O}$ and $\mathcal{O}$ to denote the deterministic order of magnitude uniformly in $f \in \mathcal{H}(1, \gamma)$. For a sequence $\{\boldsymbol{z}_n, n \geq 1\}$ of random variables determined through the design points and the observations, write $\boldsymbol{z}_n = \mathcal{O}_r(a_n)$ if

$$\sup_{f \in \mathcal{H}(1,\gamma)} \mathbf{E}|\boldsymbol{z}_n|^r = \mathcal{O}(a_n^r).$$

The notation $\mathcal{o}_r(a_n)$ is defined similarly. We then have $\mathcal{O}_r(a_n)\mathcal{O}_r(b_n) = \mathcal{O}_{r/2}(a_n b_n)$, and

$$\boldsymbol{z}_n = \mathbf{E}\boldsymbol{z}_n + \mathcal{O}_r(\mathbf{E}|\boldsymbol{z}_n - \mathbf{E}\boldsymbol{z}_n|^r)^{1/r}.$$

To determine the contribution to the bias resulting from the negative weights, we first derive the expected sum of the squared weights ℓ_i^2. Proofs of the following results are given in Appendix A.

Proposition 4.1. *Let $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Under Assumptions A and D,*

$$\mathbf{E}\sum_{i=1}^n \ell_i^2 = \frac{1}{nh_n f(x_0)} \int K^2(z)\,dz + \mathcal{O}\Big(\frac{1}{nh_n}\Big). \tag{12}$$

Next, we obtain the second moment of the sum of the negative weights.

Proposition 4.2. *Let $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Under Assumptions A and D, for sufficiently large r,*

$$\mathbf{E}\Big(\sum_{i=1}^n \ell_i^-\Big)^2 = \frac{1}{h_n}\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big).$$

With this result in hand, we can derive the mean squared error of our estimator. As the mean squared error converges to zero as n increases to infinity, this result yields consistency of our estimator as well as its rate of convergence.

Theorem 4.3. *Under Assumptions A, B, C, and D, if $h_n = cn^{-\beta}$ with $0 < \beta < 1$ and a constant $c > 0$, the mean squared error of the local linear estimator (7) is*

$$\mathrm{MSE}(x_0) = \frac{h_n^4(\mathrm{Var}_K)^2}{4} \int\limits_{\mathbb{S}^{d-1}} s''(M(x_0), v)^2\,dv + \frac{\int_{\mathbb{S}^{d-1}} C(v,v)\,dv}{nh_n f(x_0)} \int K^2(z)\,dz + \mathcal{O}\left(h_n^4 + \frac{1}{nh_n}\right).$$

We conclude this section by deriving a limit theorem for the support function of the estimators as processes on the unit sphere. In turn, this limit theorem can be used to build error tubes for the estimator as explained in Section 5. Let $\zeta(v)$, $v \in \mathbb{S}^{d-1}$, be a centered Gaussian process on the unit sphere with the covariance

$$\mathbf{E}[\zeta(v)\zeta(u))] = \frac{C(v,u)}{f(x_0)} \int K(z)^2\,dz. \tag{13}$$

Theorem 4.4. *Assume that $h_n = cn^{-\beta}$ with $0 < \beta < 1$, and fix $x_0 \in I$. Under Assumptions A, B, C, and D, the stochastic process*

$$\sqrt{nh_n}\left(s(\hat{\boldsymbol{M}}(x_0), v) - s(M(x_0), v) - h_n^2\frac{1}{2}s''(M(x_0), v)\,\mathrm{Var}_K\right)$$

constructed using the local linear estimator in (7) converges in distribution in the space of continuous functions on $\mathbb{S}^{d-1}$ with the uniform metric to the Gaussian process ζ.

5. Cross-validation and error tubes

Cross-validation. In the classical setting, where the observation pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are real-valued, one typically chooses the bandwidth h_n to minimize the leave-one-out cross-validation score, defined as

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y}_i - \hat{\boldsymbol{m}}_{(-i)}(\boldsymbol{x}_i))^2,$$

where $\hat{\boldsymbol{m}}_{(-i)}(x) = \sum_{j=1}^n \boldsymbol{y}_j \ell_{j,(-i)}(x)$ and

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if } j = i, \\ \frac{\ell_j(x)}{\sum_{k\neq i}\ell_k(x)} & \text{if } j \neq i. \end{cases}$$

This procedure assigns weight zero to $\boldsymbol{x}_i$ and renormalizes the other weights to sum to one.

Following the same idea, we define the cross-validation score for the set-valued responses $\boldsymbol{Y}_i$ as

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^n \int\limits_{\mathbb{S}^{d-1}} (s(\boldsymbol{Y}_i, v) - s(\hat{\boldsymbol{M}}_{(-i)}(\boldsymbol{x}_i), v))^2\,dv, \tag{14}$$

where $\hat{\boldsymbol{M}}_{(-i)}(x) = \sum_{j=1}^{n} \boldsymbol{Y}_j \ell_{j,(-i)}(x)$. If one is interested in a specific projection in direction v, the above expression simplifies by removing the integral.

If $\boldsymbol{Y}_i = [\boldsymbol{y}_{i\mathrm{L}}, \boldsymbol{y}_{i\mathrm{U}}] \subset \mathbb{R}$, (14) turns into

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\Big((\boldsymbol{y}_{i\mathrm{L}} - \hat{\boldsymbol{M}}_{(-iL)}(\boldsymbol{x}_i))^2 + (\boldsymbol{y}_{i\mathrm{U}} - \hat{\boldsymbol{M}}_{(-iU)}(\boldsymbol{x}_i))^2\Big), \tag{15}$$

where $\hat{\boldsymbol{M}}_{(-iL)}(\boldsymbol{x}_i)$ and $\hat{\boldsymbol{M}}_{(-iU)}(\boldsymbol{x}_i)$ denote the lower and upper bounds of $\hat{\boldsymbol{M}}_{(-i)}(\boldsymbol{x}_i)$. We denote by $h_{n,\mathrm{CV}}$ the bandwidth that minimizes (15) (or (14), depending on the application).

Error tubes. The optimal bandwidth which minimizes the MSE in Theorem 4.3 is $h_{n,\mathrm{mse}} = Cn^{-1/5}$, with some constant C that does not depend on n. However, such a choice of bandwidth implies $nh_n^5 \not\to 0$ and the leading bias term in Theorem 4.4 does not vanish, as in the classical case for singleton-valued outcomes. Similarly to that case, one can use undersmoothing as an approach to bias reduction. In Section 6 we illustrate the impact of undersmoothing on the error tubes that we describe next.

Rather than undersmooth, we propose to report statistical uncertainty in our estimates in the form of pointwise error tubes – an analog of error bands for singleton-valued data. Specifically, for each value x_0 considered we propose to report the set

$$\hat{\mathcal{C}}(x_0) = \hat{\boldsymbol{M}}(x_0) + \frac{c_\alpha}{\sqrt{nh_n}}B, \tag{16}$$

where $B = \{b : \|b\| \le 1\}$ is the unit ball. In (16) c_α is chosen so that

$$\mathbf{P}\left(\max_{v:\,\|v\|=1}\{\zeta(v)\}_+ > c_\alpha\right) = \alpha, \tag{17}$$

where ζ is the centered Gaussian process with covariance kernel (13), see Theorem 4.4. The critical value c_α can be obtained by simulation, or can be estimated using the bootstrap. Validity of the bootstrap can be formally established as in Proposition 2.1 of [8] [see also31, Theorem 4.13]. It follows from Theorem 4.4 that

$$\lim_{n\to\infty}\mathbf{P}\Big(\max_{v:\,\|v\|=1}\{s(\hat{\boldsymbol{M}}(x_0), v) - s(M(x_0), v) - h_n^2\frac{1}{2}s''(M(x_0), v)\,\mathrm{Var}_K - s(\hat{\mathcal{C}}(x_0), v)\}_+ = 0\Big) \ge 1-\alpha. \tag{18}$$

If one is interested in a specific projection in direction v, a valid error band for $s(M(x_0), v)$ is obtained by replacing (16) with

$$\left[s(\hat{\boldsymbol{M}}(x_0), v) - \frac{c_{\alpha,v}}{\sqrt{nh_n}}, s(\hat{\boldsymbol{M}}(x_0), v) + \frac{c_{\alpha,v}}{\sqrt{nh_n}}\right], \tag{19}$$

where $c_{\alpha,v}$ is obtained as in (17) replacing the maximization over v with $\|v\| = 1$ by a fixed direction v.

Existing methods of bias correction (other than undersmoothing, the effect of which we are already investigating in our Monte Carlo exercise) could be extended to the case of set-valued outcomes. However, we do not report such findings here,[3] because any form of bias reduction may result in an empty estimator, which we regard as an undesirable feature as discussed in Remark 3.1.

6. Monte Carlo simulations

We perform a simulation study for the case that $d = 1$ and for the case that $d = 2$. In the first case, we use the following data generating process (DGP1):

$$\boldsymbol{y}_{\mathrm{L}} = 0.90 + 1.27\boldsymbol{x} + 5.18\boldsymbol{x}^2 - \varepsilon_L$$
$$\boldsymbol{y}_{\mathrm{U}} = 0.90 + 1.27\boldsymbol{x} + 10.18\boldsymbol{x}^2 + \varepsilon_U,$$

with $\boldsymbol{x}$ drawn from a Beta distribution with support shifted to be $[-1, 1]$ and with shape parameters $(2, 4)$, and ε_L and ε_U drawn independently from a Uniform distribution on $[0, 1]$. We let the sample size $n = 200, 500, 1000, 2000$. For values of $x_0 = -0.4, 0, 0.2, 0.4$ we evaluate the coverage probability of the error tubes in equation (16).

We compare different implementations of the error tubes, and in Table 1 we report: (i) coverage probability of the true set $M(x_0)$ by the error tube (meaning that the true set is a subset of the tube) in (16) computed using the cross-validation

[3] Although these are available from the authors upon request.

Table 1
Coverage probability at 95% nominal level using cross-validation for DGP1.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.8315	0.8245	0.9055	0.9695
	0	0.8855	0.8550	0.8565	0.9515
	0.2	0.9330	0.9270	0.9865	0.9980
	0.4	0.9270	0.9040	0.9255	0.9875
500	-0.4	0.8580	0.8485	0.9300	0.9790
	0	0.9245	0.9095	0.9710	0.9920
	0.2	0.9240	0.9200	0.9710	0.9950
	0.4	0.9340	0.9145	0.9180	0.9760
1000	-0.4	0.8910	0.8760	0.9430	0.9845
	0	0.9035	0.8935	0.9360	0.9830
	0.2	0.9230	0.9210	0.9570	0.9890
	0.4	0.9225	0.9125	0.9125	0.9760
2000	-0.4	0.8820	0.8710	0.9450	0.9835
	0	0.9020	0.8915	0.9390	0.9870
	0.2	0.9320	0.9125	0.9525	0.9900
	0.4	0.9335	0.9170	0.9635	0.9915

bandwidth (column 3); (ii) coverage probability as in (18), with the error tube in (16) computed using the cross-validation bandwidth (column 4); (iii) same exercise as in (i) but using undersmoothed bandwidths (columns 5 and 6). The results are based on 200 Monte Carlo replications.

In these simulations, the asymptotic bias does not affect the ability of the error tube in (16) to cover the true set $M(x_0)$ compared to $\mathbf{E}[\hat{\boldsymbol{M}}(x_0)]$, see columns (3) and (4) of the table. If we undersmooth the bandwidth, the confidence interval enlarges substantially and coverage of the true set becomes conservative. In Appendix E (Tables E.4 and E.5) we report the results of two additional simulation studies that vary the expressions for $\mathbf{E}(\boldsymbol{y}_L|\boldsymbol{x})$ and $\mathbf{E}(\boldsymbol{y}_U|\boldsymbol{x})$, as well as the distribution of ε_L (to be Beta(2,2) instead of Uniform(0,1)). Qualitatively the results are similar to what we report here.

We also perform a simulation study for the case that $d = 2$ with the following data generating process (DGP2):

$$\boldsymbol{Y} = \begin{bmatrix} 0.90 + 1.27\boldsymbol{x} + 10.18\boldsymbol{x}^2 \\ 0.60 - 1.00\boldsymbol{x} - 5.18\boldsymbol{x}^2 \end{bmatrix} + B_\xi,$$

where B_ξ is a ball of radius 1 centered at the random vector ξ, and ξ is uniformly distributed on the unit ball in $\mathbb{R}^2$. As in the previous simulation, $\boldsymbol{x}$ is drawn from a Beta distribution with support shifted to be $[-1, 1]$ and with shape parameters $(2, 4)$. We let the sample size $n = 200, 500, 1000, 2000$. For values of $x_0 = -0.4, 0, 0.2, 0.4$ we evaluate the coverage probability of the error bands in equation (19) for $v = (1, 0)$, $v = (1, 1)/\sqrt{2}$, and $v = (0, 1)$. To conserve space, we report the results for $v = (1, 0)$ in Table 2 here, and for $v = (1, 1)/\sqrt{2}$ and $v = (0, 1)$, respectively, in Tables E.6 and E.7 in Appendix E. Overall the results are qualitatively similar to those reported for DGP1: once the bandwidth is undersmoothed and sample size is sufficiently large, coverage becomes valid.

7. Empirical application

We demonstrate the usefulness of our approach with an empirical illustration that studies the association between cancer treatment outcomes and certain gene expression measures.

A key outcome of interest in cancer treatment research is the progression-free survival (PFS), which is defined as the time measured in months from baseline until tumor progression or death (whichever occurs first). Tumor progression is defined as an increase in the diameter of the tumor lesions of 20% compared with the smallest diameters of all previous tumor assessments or the appearance of new lesions, as measured by CT-scans or MRIs (this is called RECIST criterion in the medical literature, see [32]). However, due to ethical and cost constraints, CT-scans and MRIs cannot be performed daily, but rather scheduled every 3 to 6 months. Hence, the PFS of patients can only be measured by intervals (with the true PFS falling between the last assessment without tumor progression and the assessment with progression), and no information is available on the distribution of true PFS within the interval. In contrast, the PFS of patients who died without tumor progression is measured exactly.

The question that we focus on in this paper is part of a subproject of the Swiss Cancer Research Group (SAKK) 19/09 for anti-cancer treatment regimens described in [33]. This subproject is concerned with finding, out of a total of 202 investigated genes, those whose baseline expression affects patient's PFS differently in two treatment arms described below. Genes expression is evaluated by isolating RNA from baseline tumor tissue sections and processing it for gene expression analysis

Table 2
Coverage probability at 95% nominal level using cross-validation for DGP2 with $v = (1, 0)$.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.8290	0.8395	0.8960	0.9725
	0	0.8515	0.8760	0.8525	0.9530
	0.2	0.9290	0.9360	0.9840	0.9985
	0.4	0.9085	0.9290	0.9220	0.9835
500	-0.4	0.8580	0.8665	0.9345	0.9805
	0	0.9195	0.9275	0.9745	0.9960
	0.2	0.9260	0.9325	0.9730	0.9945
	0.4	0.9210	0.9270	0.8965	0.9675
1000	-0.4	0.8830	0.8910	0.9315	0.9820
	0	0.9055	0.9125	0.9330	0.9785
	0.2	0.9210	0.9255	0.9425	0.9875
	0.4	0.9325	0.9345	0.9120	0.9725
2000	-0.4	0.8805	0.8835	0.9495	0.9875
	0	0.8900	0.8985	0.9355	0.9860
	0.2	0.9220	0.9300	0.9490	0.9915
	0.4	0.9270	0.9360	0.9595	0.9900

using the Nanostring nCounter® System (Nanostring Technologies), including 6 housekeeping genes.[4] The gene expression measure that we report and use for our analysis is the $\log_2$ of the output of Nanostring.

It is worth mentioning that classical statistical methods of survival analysis, such as Cox regression or the accelerated failure time model, can also be applied to this data (and we do so below). These models are typically implemented with a parametric or semi-parametric specification of the hazard rate to construct the likelihood function. For example, the Cox proportional hazard model [34] assumes a hazard rate that is constant over time, and the resulting survival data follow a Markovian process; the accelerated failure time model posits an acceleration factor that is constant over time. The probability of censoring can then be calculated based on the functional form assumption. For example, the PFS variable in our example is usually treated as an interval censored data, for which one can construct the likelihood function and obtain point identified estimates of the model's parameters, and then back out the implied conditional expectation of the treatment outcome given gene expression. In contrast, our method provides a consistent estimator of the set of admissible values for the conditional expectation of treatment outcome given gene expression, as well as $1 - \alpha$ pointwise confidence bands for it as in (16), without making any assumption on how PFS is distributed over the measured intervals that it is known to belong to, nor how it is related to the genes, as these assumptions may fail to hold in a given application.[5]

We use a novel dataset that follows 132 patients who were accrued between November 2010 and July 2014 to the SAKK 19/09 clinical trial for anti-cancer treatment regimens described in [33]. These patients are affected by advanced non-squamous non-small cell lung cancer and present an epidermal growth factor receptor (EGFR) of the wild type. Excluding 3 patients with protocol violations, 77 patients were treated with the drug Bevacizumab plus chemotherapy (C1) and 52 were treated with chemotherapy alone (C2). The question of interest of the SAKK 19/09 subproject that we revisit in this section is whether the gene expression of PTGS2 (COX2) at baseline affects differently patient's PFS in the two treatment arms. The gene PTGS2 (COX2) is frequently expressed in lung cancer patients and the drug Bevacizumab directly interacts with the COX2 pathway. One speculates that in patients with a high expression of COX2 the tumor cells are predominately dependent on this signaling pathway for proliferation and the use of Bevacizumab has a more pronounced effect. Vice-versa, if COX2 is only expressed at a low level, this could reflect a tumor that is not dependent on this inflammatory pathway and therefore the use of Bevacizumab is not beneficial. Another gene of interest (whose effect on cancer treatment efficacy has not been previously analyzed) is CDC25A, which is a key regulator of the cells cycles. One speculates that overexpression of gene CDC25A is associated with a poorer prognosis with regard to its biological role.

In our analysis, $\boldsymbol{y} =$ PFS, $\boldsymbol{y}_L$ is the time of the last assessment without tumor progression, and $\boldsymbol{y}_U$ is the time of the assessment with tumor progression. Table 3 reports descriptive statistics for these data. The sample used for the analysis is constituted by 99 patients, from which four were excluded because they were still alive at the last follow up (and therefore for these patients $\boldsymbol{y}_{iU} = \infty$). Of the sample used for our analysis, 58 patients were treated following protocol C1, and 37 following protocol C2. Because durations are non-negative by definition while local linear regression smoothers may yield negative predictions, we work with the natural logarithm of our data, adjusted as follows

$$\tilde{\boldsymbol{y}}_{\mathrm{k}} = \ln(\boldsymbol{y}_{\mathrm{k}} + 0.033), \quad \mathrm{k} = \mathrm{L}, \mathrm{U}$$

[4] See https://www.nanostring.com for a description of this method.
[5] [35] point out that individual heterogeneity and hazard rate cannot be jointly non-parametrically point identified.

Table 3
Descriptive statistics for interval-valued PFS and genes PTGS2 and CDC25A; $\boldsymbol{y}$ denotes the progression-free survival (time from baseline until tumor progression or death), $\boldsymbol{y}_L$ is last assessment without tumor progression, and $\boldsymbol{y}_U$ is the assessment with tumor progression.

variable	mean	stdErr	max	min	N
$\boldsymbol{y}_L$	7.62	9.08	52.40	0	95
$\boldsymbol{y}_U$	9.25	9.65	55.16	0.23	95
CDC25A	7.23	2.76	14.22	0	95
PTGS2	8.66	1.90	13.37	2.86	95

where we add 0.033 because for some individuals $\boldsymbol{y}_{\mathrm{L}} = 0$. The choice of 0.033 is motivated by the unit of measure for $\boldsymbol{y}$, which is months: following the convention in the medical literature, we add one day (approximately 0.033 months).

The results of the analysis are reported in the top panels of Fig. 2 for the gene PTGS2 (COX2), with panel A reporting the results using the Accelerated Failure Time (AFT) model, and Panel B reporting our set-valued local linear regression estimator. The bottom panels of Fig. 2 report the results for the gene CDC25A, with panel C reporting the results using the AFT model, and Panel D reporting our set-valued local linear regression estimator.

We first comment on the comparison between the standard AFT model and our set-valued estimator in terms of the shape of the predicted conditional PFS. For the PTGS2 (COX2) gene, the patterns are similar, although we uncover a more markedly nonlinear relation (especially for treatment C1). For the gene CDC25A, the pattern uncovered by the AFT method and our method are similar for treatment C2, but for treatment C1 we uncover a remarkably more nonlinear relationship.

The results of the AFT analysis suggest that the use of Bevacizumab in cancer treatment is quite beneficial for patients with moderate to relatively high (6-10) expression of gene PTGS2 (COX2), although the benefit seems to taper off at extremely high levels of the gene. Similarly, at medium to high levels (6-12) of expression of gene CDC25A the use of Bevacizumab seems beneficial, while at low levels of the gene the two treatment arm's effects are not significantly different. Our results, however, suggest that these findings might result from the functional form assumptions: for the gene PTGS2 (COX2) we find that for patients with moderate to relatively high (6-10) levels of the gene the set-valued estimates are consistent with a beneficial effect of Bevacizumab, but the confidence bands overlap, suggesting that the difference is not statistically significant. For the gene CDC25A we find that for CDC25A levels between 9 and 10, Bevacizumab is (statistically significantly) beneficial, but not at other levels of gene expression.

We note, however, that the results of this analysis are retrospective. To confirm the medical findings, a prospective randomized clinical trial needs to be carried out. We also highlight a drawback of our method: it is not able to handle survival data censored on the right, where the observations become half-lines unbounded on the right. In our example such observations have been eliminated.

8. Conclusions

This paper has introduced local linear regression smoothing for set-valued data. We have established consistency of the set-valued estimator, derived its mean squared error, and its (pointwise) asymptotic distribution. We have extended the cross-validation method for bandwidth selection to the case of set-valued local linear regression, and examined the finite sample properties of our estimator in a Monte Carlo exercise. We have illustrated the usefulness of our method in an empirical illustration studying the effect of gene expression on cancer therapy outcomes.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Appendix A. Proofs of main results

Proof of Proposition 4.1. Our proof builds on [5, Eqs. (6.4), (6.6) and (6.13)]. Since the kernel is assumed to have a compact support, we have $\int z^{2r} K(z)dz < \infty$ for all $r \geq 0$. For any integer $r \geq 1$,

$$\boldsymbol{s}_j = \mathbf{E}\boldsymbol{s}_j + h_n^{j+1}\mathcal{O}_r\big(1/\sqrt{nh_n}\big), \quad j = 0, 1, 2, \tag{A.1}$$

as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$. The expectations of $\boldsymbol{s}_j$ can be calculated as follows:

$$\mathbf{E}\boldsymbol{s}_0 = h_n \int K(z) f(zh_n + x_0)\, dz = h_n \int K(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n[f(x_0) + \mathcal{O}(h_n)],$$

$$\mathbf{E}\boldsymbol{s}_1 = h_n^2 \int zK(z) f(zh_n + x_0)\, dz = h_n^2 \int zK(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n^2\mathcal{O}(h_n),$$

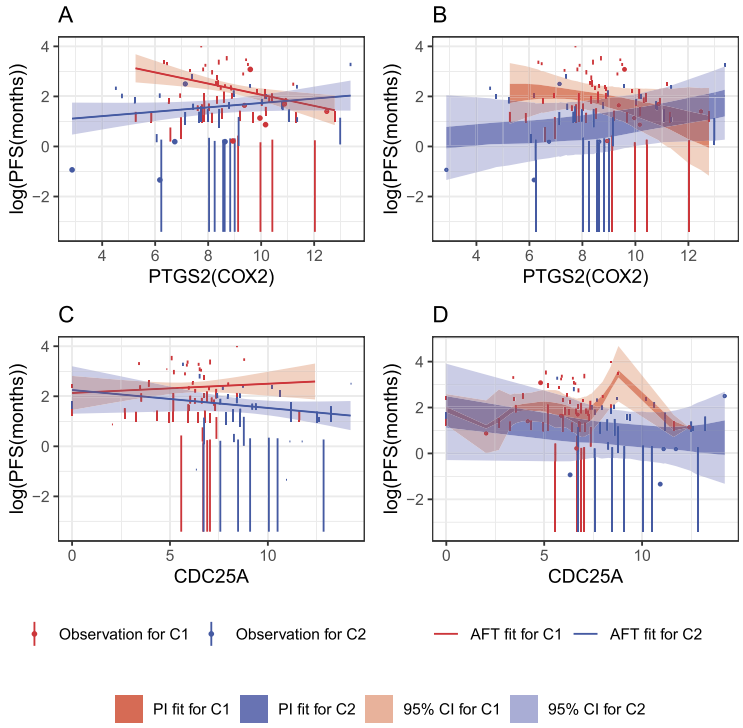

Fig. 2. Results of the analysis for the genes PTGS2 and CDC25A ($\log_2$ of the Nanostring output).

$$\mathbf{E}s_2 = h_n^3 \int z^2 K(z) f(zh_n + x_0)\, dz = h_n^3 \int z^2 K(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n^3 (f(x_0) \operatorname{Var}_K + \mathcal{O}(h_n)).$$

In view of (A.1), for an integer $r \geq 1$,

$$s_j = h_n^{j+1} \left(f(x_0) \int z^j K(z)\, dz + \mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}}) \right), \quad j = 0, 1, 2. \tag{A.2}$$

Thus,

$$s_0 = h_n f(x_0)(1 + \mathcal{O}_r(1)), \tag{A.3}$$

$$s_1 = h_n^2 \mathcal{O}_r(1), \tag{A.4}$$

$$s_2 = h_n^3 f(x_0) \operatorname{Var}_K (1 + \mathcal{O}_r(1)). \tag{A.5}$$

It is easy to see that

$$\sum_{i=1}^{n} \ell_i = \frac{s_2 s_0 - s_1^2}{s_2 s_0 - s_1^2 + n^{-4}}.$$

Moreover, for a sufficiently large r,

$$\frac{h_n^4}{s_2 s_0 - s_1^2 + n^{-4}} = \frac{1}{f(x_0)^2 \operatorname{Var}_K} + \mathcal{O}_r(1), \tag{A.6}$$

cf. [5, Eq. (6.6)]. In view of (A.3), (A.4), and (A.5),

$$s_2 s_0 - s_1^2 = h_n^4 f(x_0)^2 \operatorname{Var}_K (1 + \mathcal{O}_r(1)). \tag{A.7}$$

By (11),

$$\sum_{i=1}^{n} \ell_i^2 = \frac{\sum_{i=1}^{n} \kappa_{in}^2 (s_2 - (x_i - x_0) s_1)^2}{n^2 (s_2 s_0 - s_1^2 + n^{-4})^2} = \frac{s_2^2 s_0^*}{n(s_2 s_0 - s_1^2 + n^{-4})^2} + \frac{(-2 s_2 s_1 s_1^* + s_1^2 s_2^*)}{n(s_2 s_0 - s_1^2 + n^{-4})^2}, \tag{A.8}$$

where

$$s_j^* = \frac{1}{n}\sum_{i=1}^{n} \kappa_{in}^2 (x_i - x_0)^j = h_n^{j+1}\left(f(x_0)\int z^j K^2(z)\,dz + \mathcal{O}_r(1)\right), \quad j = 0, 1, 2.$$

Furthermore, (A.2) implies that

$$s_2^2 s_0^* = h_n^7 f^3(x_0)(\mathrm{Var}_K)^2 \int K^2(z)\,dz + h_n^7 \mathcal{O}_{r/2}(1).$$

Combining this with (A.6) and letting $r = 4$, we obtain

$$\begin{aligned}\mathbf{E}\left(\frac{s_2^2 s_0^*}{n(s_2 s_0 - s_1^2 + n^{-4})^2}\right) &= \frac{h_n^7 f^3(x_0)(\mathrm{Var}_K)^2 \int K^2(z)\,dz}{n h_n^8 f^4(x_0)(\mathrm{Var}_K)^2} + \frac{h_n^7}{n h_n^8}\mathcal{O}(1) \\ &= \frac{\int K^2(z)\,dz}{n h_n f(x_0)} + \mathcal{O}\left(\frac{1}{n h_n}\right).\end{aligned}$$

Since $\int zK(z)\,dz = 0$,

$$-2 s_2 s_1 s_1^* = h_n^7 (f(x_0)\,\mathrm{Var}_K + \mathcal{O}_8(1))\mathcal{O}_8(1)\left(f(x_0)\int z^j K^2(z)\,dz + \mathcal{O}_4(1)\right) = h_n^7 \mathcal{O}_2(1).$$

Analogously, $s_1^2 s_2^* = h_n^7 \mathcal{O}_2(1)$. Both these terms are as small as the minor term of $s_2^2 s_0^*$. Therefore, (A.8) is dominated by its first term, whence (12) holds. □

Proof of Proposition 4.2. By definition, $\ell_i < 0$ if and only if $s_2 - (x_i - x_0)s_1 < 0$. Hence,

$$\begin{aligned}\mathbf{E}\left(\sum_{i=1}^{n} \ell_i^-\right)^2 &= \mathbf{E}\left(\sum_{i=1}^{n} -\ell_i \mathbf{1}\{s_2 - (x_i - x_0)s_1 < 0\}\right)^2 \le n\mathbf{E}\left(\sum_{i=1}^{n} \ell_i^2 \mathbf{1}\{s_2 - (x_i - x_0)s_1 < 0\}\right) \\ &\le n\mathbf{E}\left(\sum_{i=1}^{n} \ell_i^2 \mathbf{1}\{s_2 < c_K h_n |s_1|\}\right) = n\mathbf{E}\left(\mathbf{1}\{s_2 < c_K h_n |s_1|\}\sum_{i=1}^{n} \ell_i^2\right) \\ &\le n\sqrt{\mathbf{P}(s_2 < c_K h_n |s_1|)}\left(\mathbf{E}\left(\sum_{i=1}^{n} \ell_i^2\right)^2\right)^{1/2}, \end{aligned} \tag{A.9}$$

where the second inequality relies on Assumption A and the last one follows from the Chebyshev inequality. Using (A.2), we have, for an integer $r \ge 1$,

$$\begin{aligned} s_1 &= h_n^2\left(\mathcal{O}(h_n) + \mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right), \\ s_2 &= h_n^3\left(f(x_0)\,\mathrm{Var}_K + \mathcal{O}(h_n) + \mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right). \end{aligned}$$

Hence,

$$\begin{aligned}&\mathbf{P}(s_2 < c_K h_n |s_1|) \qquad\qquad (\text{A.10}) \\ &\quad \le \mathbf{P}\left(f(x_0)\,\mathrm{Var}_K + \mathcal{O}(h_n) + \mathcal{O}_r\left(1/\sqrt{n h_n}\right) < |\mathcal{O}(h_n)| + \left|\mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right|\right) \\ &\quad = \mathbf{P}\left(f(x_0)\,\mathrm{Var}_K < |\mathcal{O}(h_n)| + \left|\mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right|\right). \qquad (\text{A.11})\end{aligned}$$

For sufficiently large n, there exists a ξ with $0 < \xi < f(x_0)\,\mathrm{Var}_K$ so that $|\mathcal{O}(h_n)| \le \xi$ for all sufficiently large n. Building on (A.11), the Markov inequality and the definition of $\mathcal{O}_r(a_n)$ yield that

$$\begin{aligned}\mathbf{P}(s_2 < c_K h_n |s_1|) &\le \mathbf{P}\left(f(x_0)\,\mathrm{Var}_K < \xi + \left|\mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right|\right) \\ &= \mathbf{P}\left(\left|\mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right| > f(x_0)\,\mathrm{Var}_K - \xi\right) \\ &\le \frac{\sup_{f\in\mathcal{H}(1,\gamma)} \mathbf{E}\left|\mathcal{O}_r\left(1/\sqrt{n h_n}\right)\right|^r}{(f(x_0)\,\mathrm{Var}_K - \xi)^r} = \frac{c_r\left(1/\sqrt{n h_n}\right)^r}{(f(x_0)\,\mathrm{Var}_K - \xi)^r}\end{aligned}$$

for a positive constant c_r. Therefore,

$$\mathbf{P}(\boldsymbol{s}_2 < c_K h_n |\boldsymbol{s}_1|) = \mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big). \tag{A.12}$$

From the proof of Proposition 4.1 with $r = 8$, squaring and taking expectation,

$$\mathbf{E}\Big(\sum_{i=1}^n \ell_i^2\Big)^2 = \frac{1}{n^2 h_n^2}\left(\int K^2(z)dz\right)^2 (1 + \mathcal{O}(1)). \tag{A.13}$$

Substituting (A.12) and (A.13) into (A.9),

$$\mathbf{E}\Big(\sum_{i=1}^n \ell_i^-\Big)^2 \le \frac{1}{h_n}\int K^2(z)dz\sqrt{1+\mathcal{O}(1)}\,\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big),$$

which converges to 0 by choosing a sufficiently large r. □

Proof of Theorem 4.3. The squared bias can be written as

$$b_{x_0}^2(v) = \mathbf{E}[(b_1 + b_2)^2],$$

for $b_1 = \sum_{i=1}^n \ell_i(s(M(\boldsymbol{x}_i), v) - s(M(x_0), v))$ and $b_2 = \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i), v)$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)(s(M(\boldsymbol{x}_i), v) - s(M(x_0), v)) \\
&\qquad = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)(s(M(\boldsymbol{x}_i), v) - s(M(x_0), v) + s'(M(x_0), v)(\boldsymbol{x}_i - x_0)) \\
&\qquad = h_n^6 f(x_0)\,\mathrm{Var}_K\, a_n + \mathcal{O}_4(h_n^6),
\end{aligned}$$

where

$$a_n = h_n^{-3}\mathbf{E}\left(s(M(\boldsymbol{x}), v) - s(M(x_0), v) - s'(M(x_0), v)(\boldsymbol{x} - x_0)K\Big(\frac{\boldsymbol{x} - x_0}{h_n}\Big)\right).$$

By (A.6), and using the definition of $\mathcal{O}_r$, we have

$$\mathbf{E}b_1^2 = \mathbf{E}\left(\frac{\frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)(m_v(\boldsymbol{x}_i) - m_v(x_0))}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}}\right)^2 = \left(\frac{U_n}{f(x_0)}\right)^2 h_n^4 + \mathcal{O}(h_n^4),$$

where, taking a Taylor expansion,

$$U_n = h_n^{-2}\left(\frac{1}{2}s''(M(x_0), v)\,\mathrm{Var}_K\, f(x_0)h_n^2 + \mathcal{O}(h_n^2)\right).$$

Therefore,

$$\mathbf{E}b_1^2 = \frac{1}{4}s''(M(x_0), v)^2(\mathrm{Var}_K)^2 h_n^4 + \mathcal{O}(h_n^4), \tag{A.14}$$

cf. the proof of [5, Theorem 3].

By Proposition 4.2,

$$\mathbf{E}b_2^2 \le w_{\max}^2\mathbf{E}\Big(\sum_{i=1}^n \ell_i^-\Big)^2 = \frac{1}{h_n}\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big), \tag{A.15}$$

where $w_{\max}$ is a finite deterministic bound on the width of $M(\boldsymbol{x})$ in any direction $v \in \mathbb{S}^{d-1}$ resulting from Assumption B. By the Cauchy-Schwarz inequality, (A.15) and (A.14),

$$\mathbf{E}(b_1 b_2) \le \sqrt{\mathbf{E}b_1^2\mathbf{E}b_2^2} = \frac{1}{2}\left(s''(M(x_0), v)^2(\mathrm{Var}_K)^2 h_n^4 + \mathcal{O}(h_n^4)\right)^{1/2} h_n^{-1/2}\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^{r/2}\Big),$$

which, for sufficiently large r and given that $h_n = cn^{-\beta}$, is of a smaller order than h_n^4. Thus,

$$\int_{\mathbb{S}^{d-1}} b_{x_0}^2(v)\,dv = \frac{1}{4}\int_{\mathbb{S}^{d-1}} s''(M(x_0), v)^2\,dv(\mathrm{Var}_K)^2 h_n^4 + \mathcal{O}\left(h_n^4 + \frac{1}{nh_n}\right). \tag{A.16}$$

Now we bound the variance of the estimator splitting (10) into the sum of three terms. By Proposition 4.1, the first term is

$$\mathbf{E}\Big(\sum_{i=1}^{n}\ell_i\varepsilon_i(v)\Big)^2=\mathbf{E}\sum_{i=1}^{n}\ell_i^2C(v,v)=\frac{1}{nh_nf(x_0)}C(v,v)\int K^2(z)\,dz+\mathcal{O}\left(\frac{1}{nh_n}\right).$$

The second term is

$$\mathbf{E}\sum_{1\le i<j\le n}\ell_i\ell_j^-\varepsilon_i(v)(\varepsilon_j(v)+\varepsilon_j(-v))=0.$$

Finally, consider

$$\begin{aligned}\mathbf{E}\Big(\sum_{i=1}^{n}\ell_i^-(\varepsilon_i(v)+\varepsilon_i(-v))\Big)^2&=(C(v,v)+2C(v,-v)+C(-v,-v))\mathbf{E}\sum_{i=1}^{n}(\ell_i^-)^2\\&\le 4\sigma_{\max}^2\mathbf{E}\sum_{i=1}^{n}(\ell_i^-)^2\le 4\sigma_{\max}^2\mathbf{E}\Big(\sum_{i=1}^{n}\ell_i^-\Big)^2\\&=4\sigma_{\max}^2h_n^{-1}\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big).\end{aligned}$$

For a large r, $(nh_n)^{(-r/2)}$ is of smaller order than $(nh_n)^{-1}$. Hence,

$$\int_{\mathbb{S}^{d-1}}\sigma_{x_0}^2(v)\,dv=\frac{1}{nh_nf(x_0)}\int_{\mathbb{S}^{d-1}}C(v,v)\,dv\int K^2(z)\,dz+\mathcal{O}\left(\frac{1}{nh_n}\right),$$

and the result follows by adding (A.16) to it. □

Proof of Theorem 4.4. It suffices to establish the convergence of one-dimensional distributions; the weak convergence of finite dimensional distributions follows from the Cramér–Wold device, and the functional convergence is established by bounding the Lipschitz constants of the processes as in [4, Theorem 3.2.1].

First, decompose

$$\begin{aligned}s(\hat{\boldsymbol{M}}(x_0),v)-s(M(x_0),v)&=\sum_{i=1}^{n}\ell_is(\boldsymbol{Y}_i,v)+\sum_{i=1}^{n}\ell_i^-w(\boldsymbol{Y}_i,v)-s(M(x_0),v)\\&=\sum_{i=1}^{n}\ell_is(M(\boldsymbol{x}_i),v)+\sum_{i=1}^{n}\ell_i\varepsilon_i(v)+\sum_{i=1}^{n}\ell_i^-w(\boldsymbol{Y}_i,v)-s(M(x_0),v).\end{aligned}\tag{A.17}$$

By Proposition 4.2, noticing that the L_2-convergence implies the convergence in probability, and choosing r large enough, we have that

$$\sum_{i=1}^{n}\ell_i^-w(\boldsymbol{Y}_i,v)\le w_{\max}\sum_{i=1}^{n}\ell_i^-=\mathcal{O}_p\big(1/\sqrt{nh_n}\big).$$

Using a Taylor expansion,

$$s(M(\boldsymbol{x}_i),v)=s(M(x_0),v)+(\boldsymbol{x}_i-x_0)s'(M(x_0),v)+\frac{1}{2}(\boldsymbol{x}_i-x_0)^2s''(M(x_0),v)+R(x_0,\boldsymbol{x}_i,v),$$

where the remainder term $R(x_0,\boldsymbol{x}_i,v)$ is of a smaller order than $\frac{1}{2}(\boldsymbol{x}_i-x_0)^2s''(M(x_0),v)$. Since the local linear estimator satisfies $\sum_{i=1}^{n}\ell_i(\boldsymbol{x}_i-x_0)=0$, we have

$$\begin{aligned}&\sum_{i=1}^{n}\ell_is(M(\boldsymbol{x}_i),v)+\sum_{i=1}^{n}\ell_i\varepsilon_i(v)-s(M(x_0),v)\\&=\sum_{i=1}^{n}\ell_i(s(M(\boldsymbol{x}_i),v)-s(M(x_0),v))-\frac{n^{-4}}{\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2+n^{-4}}s(M(x_0),v)+\sum_{i=1}^{n}\ell_i\varepsilon_i(v)\\&=\sum_{i=1}^{n}\ell_i\left(\frac{1}{2}(\boldsymbol{x}_i-x_0)^2s''(M(x_0),v)+R(x_0,\boldsymbol{x}_i,v)+\varepsilon_i(v)\right)-\frac{n^{-4}}{\boldsymbol{s}_2\boldsymbol{s}_0-\boldsymbol{s}_1^2+n^{-4}}s(M(x_0),v).\end{aligned}$$

Since for a sequence of $\{Z_n, n \geq 1\}$ of square integrable random variables

$$Z_n = \mathbf{E}Z_n + \mathcal{O}_p(\sqrt{\operatorname{Var} Z_n}),$$

(A.2) yields that

$$\boldsymbol{s}_j = h_n^{j+1} f(x_0) \int z^j K(z)\, dz\, (1 + \mathcal{O}_p(1)), \quad j = 0, 1, 2, 3. \tag{A.18}$$

By (A.7) and since $nh_n \to \infty$, we have

$$\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4} = h_n^4 \operatorname{Var}_K f^2(x_0)\, (1 + \mathcal{O}_p(1)). \tag{A.19}$$

Therefore,

$$\frac{n^{-4}}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}} s(M(x_0), v) = \mathcal{O}_p\left(n^{-4}h_n^{-4}\right) = \mathcal{O}_p\left(n^{-3}h_n^{-3}\right).$$

Combining (A.18) and (A.19), we have

$$\begin{aligned}
&\sum_{i=1}^n \ell_i \left(\frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v) + R(x_0, \boldsymbol{x}_i, v) + \varepsilon_i(v)\right) \\
&= \left(\frac{1}{2}(\boldsymbol{s}_2^2 - \boldsymbol{s}_3\boldsymbol{s}_1) s''(M(x_0), v) + \frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)\varepsilon_i(v)\right)(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^{-1} \\
&= \frac{1}{2}\operatorname{Var}_K s''(M(x_0), v) h_n^2 (1 + \mathcal{O}_p(1)) + \frac{1}{nh_n f(x_0)} \sum_{i=1}^n \boldsymbol{\kappa}_{in}\varepsilon_i(v)(1 + \mathcal{O}_p(1)).
\end{aligned} \tag{A.20}$$

By the central limit theorem,

$$\frac{1}{\sqrt{nh_n}} \sum_{i=1}^n \boldsymbol{\kappa}_{in}\varepsilon_i \tag{A.21}$$

converges in distribution to the centered normal random variable with variance equal to that of $\zeta(v)$. The combination of (A.17), (A.19), (A.20) and (A.21) yields the result. □

Appendix B. Deterministic design points

When the design points $\boldsymbol{x}_i = x_i$, $i = 1, \ldots, n$, are deterministic,[6] (9) turns into

$$b_{x_0}^2(v) = \left(\sum_{i=1}^n \ell_i (s(M(x_i), v) - s(M(x_0), v)) + \sum_{i=1}^n \ell_i^- w(M(x_i), v)\right)^2.$$

Since $K(\cdot)$ has compact support in $[-c_K, c_K]$, we have $\ell_i = 0$ if $|x_i - x_0| > c_K h_n$. It is easy to see that all weights are nonnegative if and only if

$$\sum \kappa_{in} \left(\frac{x_i - x_0}{h_n}\right)^2 \geq \left|\sum \kappa_{in} \frac{x_i - x_0}{h_n}\right|.$$

This assumption means that the sample rescaled around each point to lie in the range $[-1, 1]$ has the variance that dominates the absolute value of the expectation. For this, the rescaled points should be sufficiently balanced on the left and on the right of x_0. The assumption can be alternatively expressed as

$$\frac{s_2}{h_n^3} \geq c_K \left|\frac{s_1}{h_n^2}\right|.$$

It holds when $s_1/h_n^2 \to 0$ as $n \to \infty$.

By a direct computation, it is possible to show that, in the regular design case, the weights are nonnegative for all n.

[6] Because with deterministic design $\boldsymbol{x}_i = x_i$, $i = 1, \ldots, n$, $\boldsymbol{s}_j$, $j = 0, 1, 2$ and $\boldsymbol{\kappa}_{in}$, $i = 1, \ldots, n$ are also deterministic and we write $\boldsymbol{s}_j = s_j$ and $\boldsymbol{\kappa}_{in} = \kappa_{in}$.

Proposition B.1. *Consider the local linear setting with uniform kernel supported on* $[-c_K, c_K]$ *and equally spaced (regular) design points* $x_1, \dots, x_n$ *on a bounded interval* I. *If* $1/n \le c_K h_n \le 1$, *then* $\ell_i(x_0) \ge 0$ *for all* i, n *and each*

$$x_0 \in I_n = \{x \in I : [x - c_K h_n, x + c_K h_n] \subset I\}.$$

In case of deterministic design points in a bounded interval I, the following assumptions are often imposed; they appear as (LP1)-(LP2) in [3].

Assumption E *(Design points).* The design points $x_1, \dots, x_n$ are such that:

(i) There exists $\lambda_0 > 0$ such that all eigenvalues of $\mathcal{B}_{nx_0}$ are greater than or equal to λ_0 for all sufficiently large n and all $x_0 \in I$.
(ii) There exists $a_0 > 0$ such that, for any interval $J \subset I$ and all $n > 1$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{x_i \in J} \le a_0 \max(\mathrm{Leb}(J)/\mathrm{Leb}(I), 1/n),$$

where $\mathrm{Leb}(\cdot)$ denotes the Lebesgue measure.

We impose the following assumption on the response function.

Assumption F *(Theoretical response function).* The function $M(x)$, $x \in I$, is defined on a bounded closed interval $I \subset \mathbb{R}$, and there exists $\gamma > 0$ such that, for all $v \in \mathbb{S}^{d-1}$, the derivative of $s(M(x), v)$ with respect to x is Lipschitz with constant γ.

The following result is similar to [3, Prop. 1.13] in the singleton-valued data framework.

Proposition B.2. *If* $x_0 \in I_n$, $\ell_i \ge 0$ *for all* i, *and Assumptions A, B, E and F are satisfied, then*

$$|b_{x_0}(v)| \le c_K^2 C_* \gamma h_n^2, \qquad \sigma_{x_0}^2(v) \le \frac{\sigma_{\max}^2 C_*^2}{n h_n}$$

for sufficiently large n *and* $h_n \ge 1/(2n)$, *where the constant* C_* *depends only on* λ_0, a_0 *and* K_{max}.

Proposition B.2 implies

$$\mathrm{MSE}(x_0) \le c_K^4 C_*^2 \gamma^2 h_n^4 + \frac{\sigma_{\max}^2 C_*^2}{n h_n}.$$

Therefore, the upper bound is minimized for the bandwidth given by

$$h_n^* = \left(\frac{\sigma_{\max}^2}{4 c_K^4 \gamma^2} \right)^{\frac{1}{5}} n^{-\frac{1}{5}},$$

and the following result holds.

Theorem B.3. *If the bandwidth is chosen to be* $h_n = \alpha n^{-\frac{1}{5}}$ *for* $\alpha > 0$ *and Assumptions A, B, E hold, then*

$$\limsup_{n \to \infty} \sup_{x_0 \in I_n} \mathbf{E}[n^{\frac{2}{5}} L(\hat{\boldsymbol{M}}(x), M(x))] \le C_1 < \infty,$$

uniformly over all response functions satisfying Assumption F, where L *is the loss function given by (6),* C_1 *is a constant depending only on* γ, a_0, λ_0, $\sigma_{\max}^2$, $K_{\max}$ *and* α.

Appendix C. Local constant regression

In the local constant case, the weights $\ell_i = \boldsymbol{\kappa}_{in}/(n\boldsymbol{s}_0)$ are always nonnegative. Then the estimator $\hat{\boldsymbol{M}}(x_0)$ can be constructed as the convex set whose support functions is obtained by calculating the Nadaraya–Watson estimator for the sample $s(\boldsymbol{Y}_i, v)$, $i = 1, \dots, n$, in each particular direction v. In other words, $\hat{\boldsymbol{M}}(x_0)$ is the sum of the observed sets $\boldsymbol{Y}_i$ multiplied by nonnegative coefficients ℓ_i. Therefore, the bias and variance of the set-valued local constant estimator can be obtained similarly to the singleton-valued data case. For this, it suffices to assume that the function $s(M(x), v)$ is Lipschitz in x with the same constant for all v, which is equivalent to requiring that $M(x)$, $x \in I$, is Lipschitz in the Hausdorff metric.

Appendix D. Basic definitions from random set theory

A *random compact set* $\boldsymbol{Y}$ is a map from $(\Omega, \mathfrak{F}, \mathbf{P})$ to $\mathcal{K}(\mathbb{R}^d)$ such that

$$\{\omega : \boldsymbol{Y}(\omega) \cap K \neq \emptyset\} \in \mathfrak{F}, \tag{D.1}$$

for each compact set $K \subset \mathbb{R}^d$.

Random sets $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are said to be independent and identically distributed if

$$\mathbf{P}(\boldsymbol{Y}_1 \cap K_1 \neq \emptyset, \ldots, \boldsymbol{Y}_n \cap K_n \neq \emptyset) = \prod_{i=1}^{n} \mathbf{P}(\boldsymbol{Y}_i \cap K_i \neq \emptyset),$$

for all $K_1, \ldots, K_n \in \mathcal{K}(\mathbb{R}^d)$ and $\mathbf{P}(\boldsymbol{Y}_i \cap K \neq \emptyset) = \mathbf{P}(\boldsymbol{Y}_j \cap K \neq \emptyset)$ for all $i \neq j \in \{1, \ldots, n\}$ and $K \in \mathcal{K}(\mathbb{R}^d)$.

We define the Minkowski sum of two compact sets A_1 and A_2 in $\mathbb{R}^d$ elementwise as

$$A + B = \{x + y : x \in A,\ y \in B\}.$$

We let $cA = \{cx : x \in A\}$ denote the scaling of A by $c \in \mathbb{R}$. Given a compact convex set (a *convex body*) $A \subset \mathbb{R}^d$, the *support function* of A is

$$s(A, v) = \sup_{a \in A} v^\top a, \qquad v \in \mathbb{R}^d,$$

where $v^\top a$ denotes the scalar product. If A is convex, its support function uniquely identifies A, because

$$A = \bigcap_{v \in \mathbb{S}^{d-1}} \{a \in \mathbb{R}^d : v^\top a \leq s(A, v)\}. \tag{D.2}$$

Because $s(tA, v) = ts(A, v)$ for $t \geq 0$, the support function is often restricted to $v \in \mathbb{S}^{d-1}$. Note that

$$s(A_1 + A_2, v) = s(A_1, v) + s(A_2, v).$$

The width function of A is defined by

$$w(A, v) = s(A, v) + s(A, -v) = w(A, -v), \qquad v \in \mathbb{S}^{d-1},$$

and it is easy to see that the width function is nonnegative. If $d = 1$, then A is a closed interval in $\mathbb{R}$, and the unit sphere $\mathbb{S}^{d-1} = \{-1, 1\}$ consists of two points. In this case, the width function is the length of the interval.

A *random convex compact set* $\boldsymbol{Y}$ is a map from $(\Omega, \mathfrak{F}, \mathbf{P})$ to $\mathcal{K}_{\mathcal{C}}(\mathbb{R}^d)$ satisfying equation (D.1). Its measurability is equivalent to the fact that $s(\boldsymbol{Y}, v)$ is a random variable for each $v \in \mathbb{R}^d$.

Appendix E. Additional simulation results

Table E.4
Coverage probability at 95% nominal level using cross-validation for a modified DGP1 with $\boldsymbol{y}_{\mathrm{L}} = 0.90 + 1.27\boldsymbol{x} + 10.18\boldsymbol{x}^2 - \epsilon_L$, $\boldsymbol{y}_{\mathrm{U}} = 0.90 + 1.27\boldsymbol{x} + 10.18\boldsymbol{x}^2 + \epsilon_U$, and $\epsilon_L, \epsilon_U \sim^{i.i.d.} Uniform[0, 1]$.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.8630	0.8540	0.9165	0.9690
	0	0.8965	0.8865	0.8790	0.9520
	0.2	0.9465	0.9405	0.9825	0.9980
	0.4	0.9330	0.9215	0.9200	0.9745
500	-0.4	0.8705	0.8595	0.9290	0.9755
	0	0.9460	0.9410	0.9760	0.9935
	0.2	0.9315	0.9280	0.9655	0.9910
	0.4	0.9415	0.9320	0.9260	0.9800
1000	-0.4	0.9070	0.9040	0.9525	0.9855
	0	0.8990	0.8985	0.9175	0.9695
	0.2	0.9205	0.9160	0.9425	0.9760
	0.4	0.8965	0.8940	0.9090	0.9570
2000	-0.4	0.8970	0.8925	0.9440	0.9820
	0	0.9305	0.9290	0.9585	0.9865
	0.2	0.9230	0.9215	0.9425	0.9815
	0.4	0.8925	0.8935	0.9040	0.9600

Table E.5
Coverage probability at 95% nominal level using cross-validation for a modified DGP1 with $\boldsymbol{y}_{\mathrm{L}} = 0.90 + 1.27\boldsymbol{x} - \epsilon_L$, $\boldsymbol{y}_{\mathrm{U}} = 0.90 + 1.27\boldsymbol{x} + 10.18\boldsymbol{x}^2 + \epsilon_U$, $\epsilon_L \sim Beta(2,2)$ and $\epsilon_U \sim Uniform(0,1)$.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.6510	0.8945	0.9515	0.9865
	0	0.6610	0.9125	0.9050	0.9770
	0.2	0.7495	0.9600	0.9920	0.9995
	0.4	0.7210	0.9565	0.9680	0.9960
500	-0.4	0.6255	0.8945	0.9575	0.9875
	0	0.7200	0.9445	0.9870	0.9995
	0.2	0.7355	0.9605	0.9825	0.9985
	0.4	0.7155	0.9575	0.9525	0.9880
1000	-0.4	0.6345	0.9175	0.9660	0.9955
	0	0.6485	0.9330	0.9625	0.9895
	0.2	0.6960	0.9580	0.9715	0.9945
	0.4	0.7025	0.9535	0.9545	0.9870
2000	-0.4	0.6195	0.9255	0.9710	0.9935
	0	0.6290	0.9360	0.9610	0.9905
	0.2	0.6605	0.9500	0.9750	0.9935
	0.4	0.6755	0.9600	0.9785	0.9955

Table E.6
Coverage probability at 95% nominal level using cross-validation for DGP2 with $v = (1,1)/\sqrt{2}$.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.8225	0.9475	0.9490	0.9870
	0	0.8150	0.9370	0.9400	0.9820
	0.2	0.7825	0.9170	0.9330	0.9835
	0.4	0.7310	0.9020	0.9265	0.9815
500	-0.4	0.8445	0.9495	0.9635	0.9890
	0	0.7655	0.9195	0.9525	0.9895
	0.2	0.7385	0.9150	0.9410	0.9830
	0.4	0.6820	0.8745	0.9475	0.9890
1000	-0.4	0.8230	0.9500	0.9595	0.9895
	0	0.7945	0.9350	0.9455	0.9825
	0.2	0.7270	0.9185	0.9580	0.9900
	0.4	0.6830	0.8645	0.9290	0.9825
2000	-0.4	0.7965	0.9440	0.9480	0.9900
	0	0.7925	0.9430	0.9390	0.9860
	0.2	0.7485	0.9370	0.9390	0.9845
	0.4	0.7370	0.9250	0.9515	0.9890

Table E.7
Coverage probability at 95% nominal level using cross-validation for DGP2 with $v = (0,1)$.

sample size	x_0	Coverage of $M(x_0)$	Coverage of $\mathbf{E}(\hat{M}(x_0))$	Coverage of $M(x_0)$ with $h = 1/2h_{n,CV}$	Coverage of $M(x_0)$ with $h = 1/3h_{n,CV}$
200	-0.4	0.8395	0.9450	0.9485	0.9875
	0	0.8085	0.9160	0.9230	0.9765
	0.2	0.7815	0.9090	0.9445	0.9840
	0.4	0.7405	0.8945	0.9310	0.9820
500	-0.4	0.8020	0.9395	0.9530	0.9875
	0	0.7995	0.9330	0.9545	0.9905
	0.2	0.7550	0.9210	0.9380	0.9805
	0.4	0.7215	0.9025	0.9495	0.9875
1000	-0.4	0.8175	0.9485	0.9560	0.9905
	0	0.7900	0.9405	0.9420	0.9870
	0.2	0.7290	0.9345	0.9535	0.9865
	0.4	0.7070	0.8830	0.9415	0.9895
2000	-0.4	0.7945	0.9440	0.9475	0.9895
	0	0.7935	0.9430	0.9395	0.9860
	0.2	0.7495	0.9375	0.9400	0.9845
	0.4	0.7355	0.9245	0.9515	0.9890

References

[1] F.T. Juster, R. Suzman, An overview of the health and retirement study, J. Hum. Resour. 30 (Supplement) (1995) S7–S56.
[2] C.F. Manski, Partial Identification of Probability Distributions, Springer, New York, 2003.
[3] A.B. Tsybakov, Introduction to Nonparametric Estimation, Springer, New York, 2009.
[4] I. Molchanov, Theory of Random Sets, 2nd edition, Springer, London, 2017.
[5] J. Fan, Local linear regression smoothers and their minimax efficiencies, Ann. Stat. 21 (1993) 196–216, https://doi.org/10.1214/aos/1176349022.
[6] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Applications, Chapman & Hall, London, 1996.
[7] J. Fan, I. Gijbels, Variable bandwidth and local linear regression smoothers, Ann. Stat. 20 (1992) 2008–2036, https://doi.org/10.1214/aos/1176348900.
[8] A. Beresteanu, F. Molinari, Asymptotic properties for a class of partially identified models, Econometrica 76 (2008) 763–814.
[9] C.F. Manski, E. Tamer, Inference on regressions with interval data on a regressor or outcome, Econometrica 70 (2002) 519–546.
[10] C. Bontemps, T. Magnac, E. Maurin, Set identified linear models, Econometrica 80 (2012) 1129–1155.
[11] A. Chandrasekhar, V. Chernozhukov, F. Molinari, P. Schrimpf, Inference for best linear approximations to set identified functions, CeMMAP Working Paper CWP 43/12, 2012.
[12] H. Kaido, Asymptotically efficient estimation of weighted average derivatives with an interval censored variable, Econom. Theory 33 (2017) 1218–1241.
[13] K. Adusumilli, T. Otsu, Empirical likelihood for random sets, J. Am. Stat. Assoc. 112 (519) (2017) 1064–1075.
[14] G. Schollmeyer, T. Augustin, Statistical modeling under partial identification: distinguishing three types of identification regions in regression analysis with interval data, Int. J. Approx. Reason. 56 (part B) (2015) 224–248, https://doi.org/10.1016/j.ijar.2014.07.003.
[15] P. Diamond, Least squares fitting of compact set-valued data, J. Math. Anal. Appl. 147 (1990) 351–362, https://doi.org/10.1016/0022-247X(90)90353-H.
[16] M.A. Gil, M.T. López-García, M.A. Lubiano, M. Montenegro, Regression and correlation analyses of a linear relation between random intervals, Test 10 (2001) 183–201.
[17] G. González-Rodríguez, Á. Blanco, N. Corral, A. Colubi, Least squares estimation of linear regression models for convex compact random sets, Adv. Data Anal. Classif. 1 (2007) 67–81.
[18] B. Sinova, A. Colubi, M.Á. Gil, G. González-Rodríguez, Interval arithmetic-based simple linear regression between interval data: discussion and sensitivity analysis on the choice of the metric, Inf. Sci. 199 (2012) 109–124, https://doi.org/10.1016/j.ins.2012.02.040.
[19] T. Maatouk, Some application of nonparametric regression with constrained data, Ph.D. thesis, University of Glasgow, Glasgow, 2003.
[20] I. Couso, D. Dubois, Statistical reasoning with set-valued information: ontic vs. epistemic views, Int. J. Approx. Reason. 55 (2014) 1502–1518, https://doi.org/10.1016/j.ijar.2013.07.002.
[21] D.F. Heitjan, D.B. Rubin, Ignorability and coarse data, Ann. Stat. 19 (4) (1991) 2244–2253.
[22] D.F. Heitjan, Ignorability in general incomplete-data models, Biometrika 81 (4) (1994) 701–708.
[23] R.D. Gill, M.J. van der Laan, J.M. Robins, Coarsening at random: characterizations, conjectures, counter-examples, in: D.Y. Lin, T.R. Fleming (Eds.), Proceedings of the First Seattle Symposium in Biostatistics, Springer, New York, 1997, pp. 255–294.
[24] H.T. Nguyen, On random sets and belief functions, J. Math. Anal. Appl. 65 (1978) 531–542.
[25] I. Couso, D. Dubois, L. Sánchez, Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables, Springer, Cham, 2014.
[26] M. Grabisch, Set Functions, Games and Capacities in Decision Making, Springer, Cham, 2016.
[27] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.
[28] R. Schneider, Convex Bodies. The Brunn–Minkowski Theory, 2nd edition, Cambridge University Press, Cambridge, 2014.
[29] R.A. Vitale, L_p metrics for compact, convex sets, J. Approx. Theory 45 (1985) 280–287.
[30] H. Le, A. Kume, The Fréchet mean shape and the shape of the means, Adv. Appl. Probab. 32 (2000) 101–113.
[31] I. Molchanov, F. Molinari, Random Sets in Econometrics, Cambridge University Press, Cambridge, 2018.
[32] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij, New response evaluation criteria in solid tumours: revised recist guideline (version 1.1), Eur. J. Cancer 45 (2009) 228–247.
[33] O. Gautschi, S.I. Rothschild, Q. Li, K. Matter-Walstra, A. Zippelius, D.C. Betticher, M. Früh, R.A. Stahel, R. Cathomas, D. Rauch, M. Pless, S. Peters, P. Froesch, T. Zander, M. Schneider, C. Biaggi, N. Mach, A.F. Ochsenbein, Swiss Group for Clinical Cancer Research, Bevacizumab plus pemetrexed versus pemetrexed alone as maintenance therapy for patients with advanced nonsquamous non-small-cell lung cancer: update from the Swiss group for clinical cancer research (SAKK) 19/09 trial, Clin. Lung Cancer 18 (2017) 303–309.
[34] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. B 43 (2) (1972) 187–220.
[35] J. Heckman, B. Singer, The identifiability of the proportional hazard model, Rev. Econ. Stud. 51 (2) (1984) 231–241.

Paper B

Journal of Statistical Planning and Inference 203 (2019) 215–223

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimal design for multivariate multiple linear regression with set-identified response

Qiyu Li [a,b,*], Ilya Molchanov [a]

[a] University of Bern, Institute of Mathematical Statistics and Actuarial Science, Sidlerstrasse 5, CH-3012 Bern, Switzerland
[b] Swiss Group for Clinical Cancer Research (SAKK), Effingerstrasse 35, CH-3008 Bern, Switzerland

ARTICLE INFO

Article history:
Received 11 May 2018
Received in revised form 13 February 2019
Accepted 2 April 2019
Available online 11 April 2019

Keywords:
Design
Partially identified model
Random convex set
Regression
Set-identified response

ABSTRACT

We consider the partially identified regression model with set-identified responses, where the estimator is the set of the least square estimators obtained for all possible choices of points sampled from set-identified observations. We address the issue of determining the optimal design for this case and show that, for objective functions mimicking those for several classical optimal designs, their set-identified analogues coincide with the optimal designs for point-identified real-valued responses.

1. Introduction

Consider the basic regression model

$$y_i = x_i^\top \theta + \varepsilon_i, \quad i = 1, \dots, n,$$

where the design points $x_1, \dots, x_n$ belong to $\mathbb{R}^{r+1}$ called the *design space*, y_i, $i = 1, \dots, n$, are observed real-valued responses, θ is a vector of $(r+1)$ unknown numerical parameters, and $\varepsilon_1, \dots, \varepsilon_n$ are independent identically distributed (i.i.d.) centred random variables with variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$. This setting includes the classical multivariate linear model, and also other models, like the quadratic one that appears if

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i1}^2.$$

The basic problem in the theory of optimal design for regression models aims to identify the locations of design points $x_1, \dots, x_n$ which ensure the best properties of the unbiased estimator $\hat\theta$ of θ. As the objective function to minimize, one can choose, e.g., the sum of the variances of the components of $\hat\theta$ (the criterion function for the *A*-optimal design) or the largest variance of $a^\top \hat\theta$ over all unit vectors a (which yields the *E*-optimal design). Further optimality criteria lead to a multitude of other optimal designs, see Atkinson et al. (2007) and Silvey (1980).

In this paper we consider the situation when the possibly multivariate response y is set-identified, so instead of observing $y_1, \dots, y_n$, the statistician is only given sets $Y_1, \dots, Y_n$ that contain the true observations. It is assumed that the specific points $y_i \in Y_i$ are chosen by a completely unknown selection mechanism which is not a subject to statistical modelling. In this *partially identified* setting, it is not possible to come up with a single-valued estimator for θ. We follow

* Corresponding author at: University of Bern, Institute of Mathematical Statistics and Actuarial Science, Sidlerstrasse 5, CH-3012 Bern, Switzerland.
E-mail addresses: qiyu.li@stat.unibe.ch (Q. Li), ilya.molchanov@stat.unibe.ch (I. Molchanov).

https://doi.org/10.1016/j.jspi.2019.04.003
0378-3758/

the approach advocated by Beresteanu and Molinari (2008) who suggested considering all possible points (selections) $y_i \in Y_i$, $i = 1, \dots, n$, fitting to them the linear regression model in order to obtain particular (least squares) estimator $\hat{\theta}$ and, finally, use the set of all estimators $\hat{\theta}$ obtained in this way as the estimator for the set-identified regression, see also Molchanov and Molinari (2018). The most important special case arises if the observations $Y_1, \dots, Y_n$ are intervals on the line; then one talks about interval regression, see also Blanco-Fernández et al. (2013), Diamond (1990) for an alternative approach based on the interval arithmetics. The main reason of having interval-identified data is variability and uncertainty. For example, the temperature on a certain day is typically reported by weather forecasts as an interval between the lowest and the highest temperature. This interval represents the variability of the temperature. In social surveys, salaries of respondents are usually reported as intervals. Another example in the field of oncology is the time to recurrence of a tumour. The recurrence status of a patient is assessed by imaging techniques such as a CT scan at every visit, which is not scheduled every day but rather every two or three months. Therefore, we only know that recurrence occurs between two visits but not its exact time point. In this case, the data of time to recurrence are also interval-identified. In case of several interval-identified responses, the obtained multiple response is set-identified by a parallelepiped or its subset determined by the imposed constraints.

In this paper, we address the issue of optimal design in the partially identified least squares setting of Beresteanu and Molinari (2008). The crucial issue is to properly handle the variance of the estimated parameters; unlike the expectation, the variance of random sets is rather poorly understood, see Molchanov (2017).

In Section 2 we introduce the notation used throughout the paper and recall some definitions and results from random set theory. This is followed by Section 3, where we recall the classical A-, G- and E-optimal designs with point-identified data. In Section 4, we introduce the objective functions for the set-identified setting and prove that the corresponding optimal designs coincide with the classical A-, G- and E-optimal design under some assumptions on the model structure. As a corollary, we deduce that the A-, G-, and E-optimal multiresponse designs in the multiresponse point-identified setting coincide with their classical analogues; this extends the result of Chang (1994) derived for D-optimal designs.

2. Random convex sets and their expectation

We use $\|\cdot\|$ to denote the Euclidean norm. Let $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$ denote the unit sphere in $\mathbb{R}^d$. If $d = 1$, then the sphere consists of two points $\{-1, 1\}$. The family of non-empty compact convex sets (also called convex bodies) in $\mathbb{R}^d$ is denoted by $\mathcal{K}(\mathbb{R}^d)$. The *support function* of $K \in \mathcal{K}(\mathbb{R}^d)$ is defined as

$$s(K, v) = \max_{y \in K} v^\top y, \qquad v \in \mathbb{S}^{d-1},$$

so that $s(K, v)$ is the signed length of the projection of K onto the line with direction v. If $K = [a, b]$, then $s(K, 1) = b$ and $s(K, -1) = -a$.

The support function identifies uniquely the corresponding convex compact set and satisfies

$$\begin{aligned} s(tK, v) &= ts(K, v), \quad t > 0, \\ s(-K, v) &= s(K, -v), \\ s(K_1 + K_2, v) &= s(K_1, v) + s(K_2, v), \end{aligned}$$

where $-K = \{-x : x \in K\}$ is the centrally symmetric set to K, and

$$K_1 + K_2 = \{x + y : x \in K_1, y \in K_2\}$$

is the Minkowski sum of two convex bodies K_1 and K_2.

Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be a nonatomic probability space, where all random vectors and random sets are defined. The map $\boldsymbol{Y} : \Omega \mapsto \mathcal{K}(\mathbb{R}^d)$ is called a random convex body, if $\{\omega \in \Omega : \boldsymbol{Y}(\omega) \cap A \neq \emptyset\} \in \mathfrak{F}$ for every compact set A in $\mathbb{R}^d$. A random vector $\boldsymbol{y}$ in $\mathbb{R}^d$ is called a *selection* of $\boldsymbol{Y}$ if $\boldsymbol{y}(\omega) \in \boldsymbol{Y}(\omega)$ for almost all $\omega \in \Omega$. We denote this as $\boldsymbol{y} \in \boldsymbol{Y}$ a.s.

We assume throughout that $\boldsymbol{Y}$ is *integrably bounded*, that is, $\|\boldsymbol{Y}\| = \sup\{\|\boldsymbol{y}\| : \boldsymbol{y} \in \boldsymbol{Y}\}$ is an integrable random variable. In this case, all selections of $\boldsymbol{Y}$ are integrable and the *expectation* $\mathbf{E}\boldsymbol{Y}$ is defined as the set of $\mathbf{E}\boldsymbol{y}$ for all selections $\boldsymbol{y}$ of $\boldsymbol{Y}$. Equivalently, $\mathbf{E}\boldsymbol{Y}$ is the convex body that satisfies

$$\mathbf{E}s(\boldsymbol{Y}, v) = s(\mathbf{E}\boldsymbol{Y}, v), \quad v \in \mathbb{S}^{d-1}.$$

If $\boldsymbol{Y} = [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ is the interval then $\mathbf{E}\boldsymbol{Y} = [\mathbf{E}\boldsymbol{y}_{\mathrm{L}}, \mathbf{E}\boldsymbol{y}_{\mathrm{U}}]$.

Similarly, the conditional expectation of $\boldsymbol{Y}$ given a random vector (or matrix) $\boldsymbol{x}$ is defined as

$$\mathbf{E}(\boldsymbol{Y}|\boldsymbol{x}) = \big\{\mathbf{E}(\boldsymbol{y}|\boldsymbol{x}) : \boldsymbol{y} \in \boldsymbol{Y} \text{ a.s.}\big\},$$

and then $\mathbf{E}(s(\boldsymbol{Y}, v)|\boldsymbol{x}) = s(\mathbf{E}(\boldsymbol{Y}|\boldsymbol{x}), v)$ a.s.

3. Classical optimal designs in the multiresponse setting

Consider i.i.d. sample (x_i, y_i), $i = 1, \dots, n$, where $y_i = (y_{i1}, \dots, y_{ip})^\top \in \mathbb{R}^p$ designates response and $x_i = (1, x_{i1}, \dots x_{ir})^\top \in \mathbb{R}^{r+1}$ is the vector composed of explanatory variables. Let

$$\mathscr{X} = (x_1^\top, \dots, x_n^\top)^\top = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nr} \end{pmatrix}$$

be the *design matrix* with n rows and $r+1$ columns. Collect all the responses in the matrix

$$\mathscr{Y} = (y_1^\top, \dots, y_n^\top)^\top = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \vdots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix}.$$

Consider the regression model

$$\mathscr{Y} = \mathscr{X}\Theta + \mathscr{E},$$

where

$$\Theta = \begin{pmatrix} \theta_{01} & \cdots & \theta_{0p} \\ \theta_{11} & \cdots & \theta_{1p} \\ \vdots & \vdots & \vdots \\ \theta_{r1} & \cdots & \theta_{rp} \end{pmatrix}$$

is the matrix of unknown parameters, and the matrix $\mathscr{E} = (\varepsilon_1^\top, \dots, \varepsilon_n^\top)^\top$ consists of i.i.d. square integrable centred random vectors $\varepsilon_i = (\varepsilon_{i1}, \dots \varepsilon_{ip})^\top$ such that $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma_{jk}$ for $i = 1, \dots, n$ and $j, k = 1, \dots, p$ and $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{i'k}) = 0$ for $i \neq i'$.

Assume that the model has full rank, meaning that

$$\Sigma = \mathscr{X}^\top \mathscr{X}$$

is invertible. We stress that Σ depends on $\mathscr{X}$. The least square estimator of Θ is

$$\hat{\Theta} = \Sigma^{-1} \mathscr{X}^\top \mathscr{Y}.$$

Then $\mathbf{E}\hat{\Theta} = \Theta$ and

$$\mathrm{Cov}(\hat{\Theta}_{(j)}, \hat{\Theta}_{(k)}) = \sigma_{jk} \Sigma^{-1}, \quad j, k = 1, \dots, p,$$

where $\hat{\Theta}_{(k)}$ denotes the kth column of $\hat{\Theta}$.

The sum of variances of all elements in the matrix $\hat{\Theta}$ is

$$\sum_{k=1}^{p} \sigma_{kk} \mathrm{Tr}(\Sigma^{-1}),$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. The A-optimal design minimizes this sum of variances and so can be obtained by minimizing $\mathrm{Tr}(\Sigma^{-1})$ over all designs. Depending on the framework, the designs mentioned here could be exact n-trial designs or approximate ones. Note that the objective function for the A-optimality does not depend on the dimension of the response variable. The objective function can be equivalently written as

$$\sum_{j=1}^{r+1} \frac{1}{\lambda_j},$$

where $\lambda_1, \dots, \lambda_{r+1}$ are the eigenvalues of Σ.

The E-optimal design is chosen to minimize the variance of the least well estimated contrast $a^\top(\hat{\Theta}_{(1)}, \dots, \hat{\Theta}_{(p)})^\top$ under the constraint $\|a\| = a^\top a = 1$. This objective function could be expressed as the maximum element on the diagonal of Σ^{-1}, which is also known as the MV-optimal designed introduced by Jacroux (1983). The E-optimality criterion can be equivalently expressed as minimization of $\max(\lambda_1^{-1}, \dots, \lambda_{r+1}^{-1})$.

The variance of the response at certain $x \in \{1\} \times \mathbb{R}^r$ could be expressed as

$$\mathrm{Var}(\hat{\mathbf{y}}(x)) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix} x^\top \Sigma^{-1} x.$$

The design which minimizes the maximum of the variance of the predicted response over an arbitrary design region $\mathcal{I} \subset \{1\} \times \mathbb{R}^r$ is called G-optimal. The corresponding objective function is $\max_{x \in \mathcal{I}} x^\top \Sigma^{-1} x$.

4. Optimal designs for set-identified response

Assume that the response is set-identified, and the statistician observes compact convex sets $Y_1, \dots, Y_n$ in $\mathbb{R}^p$, where possible responses $y_1, \dots, y_n$ take their values. The explanatory variables are assumed to be point-identified. Following Beresteanu and Molinari (2008) and given the i.i.d. data $(x_i, Y_i)_{i=1}^n$, the least square estimators of the regression coefficients Θ form the *family of matrices*

$$\hat{\Theta} = (\mathscr{X}^\top \mathscr{X})^{-1} \mathscr{X}^\top \left\{ \begin{pmatrix} y_1^\top \\ \vdots \\ y_n^\top \end{pmatrix} : y_i \in Y_i \right\} = \Sigma^{-1} \sum_{i=1}^n \operatorname{diag}(x_i) \boldsymbol{G}_i, \tag{1}$$

where throughout this paper $\operatorname{diag}(\cdot)$ of a vector denotes the diagonal matrix built from this vector, and $\boldsymbol{G}_i$ is the set of $(r+1) \times p$ matrices with

$$\boldsymbol{G}_i = \left\{ \begin{pmatrix} y_i^\top \\ \vdots \\ y_i^\top \end{pmatrix} : y_i \in Y_i \right\}, \quad i = 1, \dots, n.$$

Denote by $\mathbf{E}_{\mathscr{X}}$ the expectation assuming that the design matrix is $\mathscr{X}$. Note that $\hat{\Theta}$ is a set of matrices, each of them is a least square estimator for a certain sample of responses $y_1, \dots, y_n$ arbitrarily selected from $Y_1, \dots, Y_n$. In order to define its variance, we consider products of all matrices with a given $u \in \mathbb{S}^{p-1}$; and then the support function of the obtained random convex set in $\mathbb{R}^{r+1}$ in direction v from the unit sphere $\mathbb{S}^r$ in $\mathbb{R}^{r+1}$. In other words, we work with the variance

$$\operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}u, v) = \mathbf{E}_{\mathscr{X}} (s(\hat{\Theta}u, v) - s(\mathbf{E}_{\mathscr{X}}(\hat{\Theta}u), v))^2$$

of the support function of $\hat{\Theta}u$ and aim to minimize it as function of the design. Note that $\hat{\Theta}u$ is a random convex set in $\mathbb{R}^{r+1}$, and its expectation is defined in Section 2.

Following the classical definitions of *A*-, *G*- and *E*-optimal designs, we define the objective function for these designs in the set-identified framework as

$$f^A(\mathscr{X}) = \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^r} \operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}u, v)\, dv du, \tag{2}$$

$$f^G(\mathscr{X}) = \max_{x \in \mathcal{I}} \int_{\mathbb{S}^{p-1}} \operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}^\top x, u)\, du = \max_{x \in \mathcal{I}} \int_{\mathbb{S}^{p-1}} \operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}u, x)\, du, \tag{3}$$

$$f^E(\mathscr{X}) = \max_{u \in \mathbb{S}^{p-1}} \max_{v \in \mathbb{S}^r} \operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}u, v). \tag{4}$$

Here the integrals over spheres are understood with respect to a finite rotation invariant measure (the Haar measure) and $\mathcal{I}$ is a compact subset of $\{1\} \times \mathbb{R}^r$.

Example 4.1 (*Univariate Set-Identified Response*)**.** If $p = 1$ and $\boldsymbol{Y} = [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$ (like in the examples mentioned in the introduction), then $\hat{\Theta}$ is a family of $(r+1) \times 1$ matrices, equivalently, vectors in $\mathbb{R}^{r+1}$. In this case, $u = \pm 1$ and the integrals (or maximum) with respect to u in the objective functions reduce to the sum (or maximum) over $u = \pm 1$ of the variances of the support function of the predicted response at $v \in \mathbb{S}^r$.

We denote $M(x) = \mathbf{E}(\boldsymbol{Y} | \boldsymbol{x} = x)$ and $m(x) = \mathbf{E}(\boldsymbol{y} | \boldsymbol{x} = x)$, and also $M_i = M(x_i)$ and $m_i = m(x_i)$.

Theorem 4.2. *Assume that*

$$s(\boldsymbol{Y}, u) - s(M(\boldsymbol{x}), u) = \varepsilon(u), \quad u \in \mathbb{S}^{p-1}, \tag{5}$$

where ε is a random function on the unit sphere that does not depend on $\boldsymbol{x}$ and satisfies $\mathbf{E}\varepsilon(u) = 0$ and $\operatorname{Var}(\varepsilon(u)) = \sigma_u^2 < \infty$ for all $u \in \mathbb{S}^{p-1}$. Then the designs minimizing the objective functions defined in (2) and (3) correspond to the classical A- and G-optimal designs.

Remark 4.3. Condition (5) is a modelling assumption. In case of random intervals $\boldsymbol{Y} = [\boldsymbol{y}_{\mathrm{L}}, \boldsymbol{y}_{\mathrm{U}}]$, it means that

$$\begin{cases} \boldsymbol{y}_{\mathrm{L}} = \mathbf{E}(\boldsymbol{y}_{\mathrm{L}} | \boldsymbol{x}) - \varepsilon(-1), \\ \boldsymbol{y}_{\mathrm{U}} = \mathbf{E}(\boldsymbol{y}_{\mathrm{U}} | \boldsymbol{x}) + \varepsilon(1), \end{cases}$$

for a centred random vector $(\varepsilon(-1), \varepsilon(1))$ such that

$$\varepsilon(1) + \varepsilon(-1) \geq \mathbf{E}(\boldsymbol{y}_{\mathrm{L}} | \boldsymbol{x}) - \mathbf{E}(\boldsymbol{y}_{\mathrm{U}} | \boldsymbol{x}) \quad \text{a.s.}$$

The latter condition replicates the requirement that $\mathbf{P}(\boldsymbol{y}_{\mathrm{U}} > \boldsymbol{y}_{\mathrm{L}}) = 1$.

Proof of Theorem 4.2. First, consider the A-optimal design. Using (1) and the additivity of support function as function of convex bodies,

$$\begin{aligned} & s(\hat{\boldsymbol{\Theta}}u, v) - s(\mathbf{E}_{\mathscr{X}}(\hat{\boldsymbol{\Theta}}u), v) \\ &\quad = \sum_{i=1}^{n} \left\{ s\left(\Sigma^{-1}\operatorname{diag}(x_i)\boldsymbol{G}_i u, v\right) - s\left(\Sigma^{-1}\operatorname{diag}(x_i)\mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_i u), v\right)\right\} \\ &\quad = \sum_{i=1}^{n} \left\{ s\left(\boldsymbol{G}_i u, \operatorname{diag}(x_i)\Sigma^{-1}v\right) - s\left(\mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_i u), \operatorname{diag}(x_i)\Sigma^{-1}v\right)\right\}. \end{aligned}$$

Denote $\tilde{v}_i = \operatorname{diag}(x_i)\Sigma^{-1}v$ and

$$\delta_i(u, v) = s\left(\boldsymbol{G}_i u, \tilde{v}_i\right) - s\left(\mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_i u), \tilde{v}_i\right).$$

Then

$$\begin{aligned} f^A(\mathscr{X}) &= \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^r} \mathbf{E}_{\mathscr{X}} \Big(\sum_{i=1}^{n} \delta_i(u, v)\Big)^2 dvdu \\ &= \sum_{i=1}^{n} \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^r} \mathbf{E}_{\mathscr{X}} \delta_i^2(u, v) dvdu + \sum_{i\neq i'} \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^r} \mathbf{E}_{\mathscr{X}}(\delta_i(u, v)\delta_{i'}(u, v)) dvdu. \end{aligned}$$

By expressing $\boldsymbol{G}_i u$ and $\mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_i u)$ as

$$\begin{aligned} \boldsymbol{G}_i u &= \left\{ \begin{pmatrix} y_i^\top u \\ \vdots \\ y_i^\top u \end{pmatrix} : y_i \in Y_i \right\}, \\ \mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_i u) &= \left\{ \begin{pmatrix} m_i^\top u \\ \vdots \\ m_i^\top u \end{pmatrix} : m_i \in M_i \right\}, \end{aligned}$$

we have

$$\delta_i(u, v) = \max_{y_i\in Y_i} \begin{pmatrix} y_i^\top u \\ \vdots \\ y_i^\top u \end{pmatrix}^\top \tilde{v}_i - \max_{m_i\in M_i} \begin{pmatrix} m_i^\top u \\ \vdots \\ m_i^\top u \end{pmatrix}^\top \tilde{v}_i = \max_{y_i\in Y_i} y_i^\top u\tilde{v}_i^\top \boldsymbol{e} - \max_{m_i\in M_i} m_i^\top u\tilde{v}_i^\top \boldsymbol{e},$$

where $\boldsymbol{e}$ is the $(r+1)$-dimensional vector with all entries equal to one. Then

$$\begin{aligned} \mathbf{E}_{\mathscr{X}} \delta_i^2(u, v) &= \mathbf{E}_{\mathscr{X}} \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \tilde{v}_i^\top \boldsymbol{e} \Big(\max\{y_i^\top u : y_i \in Y_i\} - \max\{m_i^\top u : m_i \in M_i\} \Big) \\ &\qquad + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \tilde{v}_i^\top \boldsymbol{e} \Big(\min\{y_i^\top u : y_i \in Y_i\} - \min\{m_i^\top u : m_i \in M_i\} \Big) \Big]^2 \\ &= (\tilde{v}_i^\top \boldsymbol{e})^2 \mathbf{E}_{\mathscr{X}} \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \big(s(Y_i, u) - s(M_i, u) \big) + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \big(-s(Y_i, -u) + s(M_i, -u) \big) \Big]^2 \\ &= (\tilde{v}_i^\top \boldsymbol{e})^2 \mathbf{E}_{\mathscr{X}} \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \varepsilon_i(u) - \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \varepsilon_i(-u) \Big]^2 \\ &= (\tilde{v}_i^\top \boldsymbol{e})^2 \mathbf{E}_{\mathscr{X}} \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \varepsilon_i(u)^2 + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \varepsilon_i(-u)^2 \Big] \\ &= (\tilde{v}_i^\top \boldsymbol{e})^2 \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \sigma_u^2 + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \sigma_{-u}^2 \Big]. \end{aligned} \tag{6}$$

Since

$$\begin{aligned} \mathbf{E}_{\mathscr{X}} \delta_i^2(u, v) + \mathbf{E}_{\mathscr{X}} \delta_i^2(-u, v) &= (\tilde{v}_i^\top \boldsymbol{e})^2 \Big[\mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \sigma_u^2 + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \sigma_{-u}^2 + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}\geq 0\}} \sigma_{-u}^2 + \mathbf{1}_{\{\tilde{v}_i^\top \boldsymbol{e}<0\}} \sigma_u^2 \Big] \\ &= (\tilde{v}_i^\top \boldsymbol{e})^2 (\sigma_u^2 + \sigma_{-u}^2), \end{aligned}$$

we have

$$\begin{aligned}\int_{\mathbb{S}^{p-1}}\int_{\mathbb{S}^r}\mathbf{E}_{\mathscr{X}}\delta_i^2(u,v)\,dvdu &= \frac{1}{2}\int_{\mathbb{S}^{p-1}}\int_{\mathbb{S}^r}\left(\mathbf{E}_{\mathscr{X}}\delta_i^2(u,v)+\mathbf{E}_{\mathscr{X}}\delta_i^2(-u,v)\right)dvdu \\ &= \frac{1}{2}\int_{\mathbb{S}^r}\left((\Sigma^{-1}\operatorname{diag}(x_i)\boldsymbol{e})^\top v\right)^2 dv\int_{\mathbb{S}^{p-1}}(\sigma_u^2+\sigma_{-u}^2)du \\ &= \frac{1}{2}\|\Sigma^{-1}x_i\|^2\int_{\mathbb{S}^r}\left(w^\top v\right)^2\,dv\int_{\mathbb{S}^{p-1}}(\sigma_u^2+\sigma_{-u}^2)du,\end{aligned} \tag{7}$$

where $w=\Sigma^{-1}x_i/\|\Sigma^{-1}x_i\|$ is a unit vector in $\mathbb{R}^{r+1}$. Note that

$$\int_{\mathbb{S}^r}(w^\top v)^2\,dv=\mathbf{E}\left(\frac{Z_j^2}{\sum_{j=1}^{r+1}Z_j^2}\right), \tag{8}$$

where $i\in\{1,\dots,r+1\}$ and $(Z_1,\dots,Z_{r+1})^\top$ is multivariate standard normal.

Taking the sum on the right-hand side of (8) over $j=1,\dots,r+1$ and noticing that this sum is one, we obtain

$$\int_{\mathbb{S}^r}\left(w^\top v\right)^2\,dv=\frac{1}{r+1},$$

whence

$$\int_{\mathbb{S}^{p-1}}\int_{\mathbb{S}^r}\mathbf{E}_{\mathscr{X}}\delta_i^2(u,v)\,dvdu=\frac{1}{2(r+1)}\|\Sigma^{-1}x_i\|^2\int_{\mathbb{S}^{p-1}}(\sigma_u^2+\sigma_{-u}^2)du.$$

Since ε_i are centred i.i.d., we have

$$\mathbf{E}_{\mathscr{X}}(\delta_i(u,v)\delta_j(u,v))=0,\quad i\neq j. \tag{9}$$

Thus,

$$f^A(\mathscr{X})=\frac{1}{2(r+1)}\sum_{i=1}^n\|\Sigma^{-1}x_i\|^2\int_{\mathbb{S}^{p-1}}(\sigma_u^2+\sigma_{-u}^2)du, \tag{10}$$

so that the A-optimal design minimizes $\sum_{i=1}^n\|\Sigma^{-1}x_i\|^2$. This sum can be expressed as

$$\begin{aligned}\sum_{i=1}^n\|\Sigma^{-1}x_i\|^2 &= \sum_{i=1}^n x_i^\top\Sigma^{-2}x_i=\sum_{i=1}^n\operatorname{Tr}\left(x_i^\top\Sigma^{-2}x_i\right) \\ &= \sum_{i=1}^n\operatorname{Tr}\left(\Sigma^{-2}x_ix_i^\top\right)=\operatorname{Tr}\left(\sum_{i=1}^n\Sigma^{-2}x_ix_i^\top\right) \\ &= \operatorname{Tr}\left(\Sigma^{-2}\sum_{i=1}^n x_ix_i^\top\right)=\operatorname{Tr}\left(\Sigma^{-1}\right)=\sum_{j=1}^{r+1}\frac{1}{\lambda_j},\end{aligned}$$

where λ_j is the jth eigenvalue of Σ.

The design minimizing this expression is exactly the A-optimal one for the case of real-valued responses.

In order to prove the statement concerning the G-optimal design, note that

$$s(\hat{\boldsymbol{\Theta}}^\top x,u)=\sum_{i=1}^n s(\boldsymbol{G}_i{}^\top\Sigma^{-1}\operatorname{diag}(x_i)x,u)=\sum_{i=1}^n s(\boldsymbol{G}_iu,\operatorname{diag}(x_i)\Sigma^{-1}x),$$

where $\boldsymbol{G}_i{}^\top$ is the family of all transposed matrices from $\boldsymbol{G}_i$. Denote $\xi_i(x)=\operatorname{diag}(x_i)\Sigma^{-1}x$ and

$$\Delta_i(u,x)=s(\boldsymbol{G}_iu,\xi_i(x))-s(\mathbf{E}_{\mathscr{X}}(\boldsymbol{G}_iu),\xi_i(x)).$$

Then

$$\begin{aligned}\operatorname{Var}_{\mathscr{X}}s(\hat{\boldsymbol{\Theta}}^\top x,u)\,du &= \mathbf{E}_{\mathscr{X}}\left(\sum_{i=1}^n\Delta_i(u,x)\right)^2 \\ &= \sum_{i=1}^n\mathbf{E}_{\mathscr{X}}(\Delta_i^2(u,x))+\sum_{i\neq j}\mathbf{E}_{\mathscr{X}}(\Delta_i(u,x)\Delta_j(u,x)) \\ &= \sum_{i=1}^n\mathbf{E}_{\mathscr{X}}(\Delta_i^2(u,x))\end{aligned}$$

by the same argument as in (9). Following the derivation of (6), we get

$$\Delta_i(u,x) = \max_{y_i \in Y_i} y_i^\top u \xi_i(x)^\top \boldsymbol{e} - \max_{m_i \in M_i} m_i^\top u \xi_i(x)^\top \boldsymbol{e}$$

and

$$\mathbf{E}_{\mathscr{X}} \Delta_i^2(u,x) = (\xi_i(x)^\top \boldsymbol{e})^2 [\mathbf{1}_{\{\xi_i(x)^\top \boldsymbol{e} \ge 0\}} \sigma_u^2 + \mathbf{1}_{\{\xi_i(x)^\top \boldsymbol{e} < 0\}} \sigma_{-u}^2].$$

The idea used to obtain (7) is also applicable here, namely,

$$\begin{aligned}\int_{\mathbb{S}^{p-1}} \mathbf{E}_{\mathscr{X}}(\Delta_i^2(u,x))\,du &= \frac{1}{2}\int_{\mathbb{S}^{p-1}} \left(\mathbf{E}_{\mathscr{X}} \Delta_i^2(u,x) + \mathbf{E}_{\mathscr{X}} \Delta_i^2(-u,x)\right) du \\ &= \frac{1}{2}(\xi_i(x)^\top \boldsymbol{e})^2 \int_{\mathbb{S}^{p-1}} (\sigma_u^2 + \sigma_{-u}^2)\,du.\end{aligned}$$

Therefore,

$$f^G(\mathscr{X}) = \max_{x \in \mathcal{I}} \sum_{i=1}^n (\xi_i(x)^\top \boldsymbol{e})^2 \frac{1}{2} \int_{\mathbb{S}^{p-1}} (\sigma_u^2 + \sigma_{-u}^2)\,du. \tag{11}$$

Furthermore,

$$\begin{aligned}\sum_{i=1}^n (\xi_i(x)^\top \boldsymbol{e})^2 &= \sum_{i=1}^n (x^\top \Sigma^{-1} \operatorname{diag}(x_i)\boldsymbol{e})^2 = \sum_{i=1}^n (x^\top \Sigma^{-1} x_i)^2 \\ &= \sum_{i=1}^n (x_i^\top \Sigma^{-1} x)^2 = \sum_{i=1}^n \left[(x_i^\top \Sigma^{-1} x)^\top x_i^\top \Sigma^{-1} x\right] \\ &= \sum_{i=1}^n \left[x^\top \Sigma^{-1} x_i x_i^\top \Sigma^{-1} x\right] = x^\top \Sigma^{-1} \left(\sum_{i=1}^n x_i x_i^\top\right) \Sigma^{-1} x \\ &= x^\top \Sigma^{-1} x. \end{aligned} \tag{12}$$

Combine (11) and (12) to see that the design minimizing $f^G(\mathscr{X})$ minimizes $\max_{x \in \mathcal{I}} x^\top \Sigma^{-1} x$, which corresponds to the objective function of the classical G-optimal design. □

Remark 4.4. The Equivalence Theorem by Kiefer and Wolfowitz (1960) establishes that the approximate design which is G-optimal is also D-optimal in the case of point-identified univariate response. By letting $\mathcal{I}$ be a singleton in (11), it is immediately seen that the classical D-optimal design minimizes $\operatorname{Var}_{\mathscr{X}} s(\hat{\boldsymbol{\Theta}}^\top x, u)$ for each given x.

Now consider the case of E-optimal designs.

Theorem 4.5. *Assume that (5) holds with ε being a random function that does not depend on $\boldsymbol{x}$ and satisfying $\mathbf{E}\varepsilon(u) = 0$ and $\operatorname{Var}(\varepsilon(u)) = \operatorname{Var}(\varepsilon(-u)) = \sigma_u^2 < \infty$ for all $u \in \mathbb{S}^{p-1}$. Then the design minimizing the objective function (4) coincides with the classical E-optimal design.*

Proof. Similarly to the case of the A-optimal design, Eq. (6) can be written as

$$\mathbf{E}_{\mathscr{X}} \delta_i^2(u,v) = (\tilde{v}_i^\top \boldsymbol{e})^2 \sigma_u^2 \tag{13}$$

due to the assumption that $\operatorname{Var}(\varepsilon(u)) = \operatorname{Var}(\varepsilon(-u)) = \sigma_u^2$. The expression of variance can be simplified by using (13) and (9), so that

$$f^E(\mathscr{X}) = \max_{u \in \mathbb{S}^{p-1}} \max_{v \in \mathbb{S}^r} \sum_{i=1}^n \mathbf{E}_{\mathscr{X}} \delta_i^2(u,v) = \max_{u \in \mathbb{S}^{p-1}} \sigma(u)^2 \max_{v \in \mathbb{S}^r} \sum_{i=1}^n (\tilde{v}_i^\top \boldsymbol{e})^2. \tag{14}$$

Thus, using the same approach of developing (12), the E-optimal design in the set-identified setting minimizes the maximum over $v \in \mathbb{S}^r$ of

$$\sum_{i=1}^n (\tilde{v}_i^\top \boldsymbol{e})^2 = v^\top \Sigma^{-1} v.$$

Finally, observe that this maximum is the maximal eigenvalue of Σ^{-1}, that is,

$$\max_{v \in \mathbb{S}^r} v^\top \Sigma^{-1} v = \max_{j \in (1,\dots,r+1)} \frac{1}{\lambda_j}. \quad \square$$

Remark 4.6. If $\boldsymbol{Y} = \{\boldsymbol{y}\}$ is a singleton in $\mathbb{R}^p$, we are in the situation of the multiresponse design, and (5) holds with $s(\boldsymbol{Y}, u) = y^\top u$ and $\varepsilon(-u) = -\varepsilon(u)$. Then $\hat{\boldsymbol{\Theta}} = \{\hat{\Theta}\}$ is a singleton, and

$$\operatorname{Var}_{\mathscr{X}} s(\hat{\Theta}u, v) = \mathbf{E}_{\mathscr{X}}[((\hat{\Theta} - \mathbf{E}_{\mathscr{X}}\hat{\Theta})u)^\top v]^2.$$

For a matrix A,

$$\int_{\mathbb{S}^r} ((Au)^\top v)^2\, dv = c_r \|Au\|^2$$

with a constant c_r depending only on dimension r and $\max_{v\in\mathbb{S}^r}((Au)^\top, v)^2 = \|Au\|^2$. Therefore, the objective functions of these designs in the multiresponse setting are given by

$$f^A(\mathscr{X}) = c_r \int_{\mathbb{S}^{p-1}} \mathbf{E}_{\mathscr{X}} \|(\hat{\Theta} - \mathbf{E}_{\mathscr{X}}\hat{\Theta})u\|^2 du,$$
$$f^G(\mathscr{X}) = c_{p-1} \max_{x\in\mathcal{I}} \mathbf{E}_{\mathscr{X}} \|(\hat{\Theta} - \mathbf{E}_{\mathscr{X}}\hat{\Theta})^\top x\|^2,$$
$$f^E(\mathscr{X}) = \max_{u\in\mathbb{S}^{p-1}} \mathbf{E}_{\mathscr{X}} \|(\hat{\Theta} - \mathbf{E}_{\mathscr{X}}\hat{\Theta})u\|^2.$$

By Theorem 4.2, the multiresponse A- and G-optimal designs coincide with their univariate response analogues, the same is the case for E-optimal designs, since the condition of Theorem 4.5 is automatically satisfied.

In the case of univariate responses, $\hat{\Theta}$ becomes a vector $\hat{\theta}$, $u = \pm 1$, and so the objective functions f^A and f^G become $\mathbf{E}\|\hat{\theta} - \mathbf{E}\hat{\theta}\|^2$ and f^G is the maximum of $\mathbf{E}((\hat{\theta} - \mathbf{E}\hat{\theta})^\top x)^2$ (up to dimension-dependent constants).

5. Discussion

The choice of objective functions in our setting is explained by the lack of a standard definition of the variance for random sets, see Molchanov (2017). In the set-identified multiple response setting, the estimated parameter is a (convex) family $\hat{\boldsymbol{\Theta}}$ of matrices, which is a convex set in dimension $(r+1)\times p$. Then $s(\hat{\boldsymbol{\Theta}}u, v)$ can be interpreted as the support function of $\hat{\boldsymbol{\Theta}}$ in direction $u \otimes v$ understood in the tensor space $\mathbb{R}^p \times \mathbb{R}^{r+1}$. Then $f^A(\mathscr{X})$ corresponds to the expectation of the squared L_2 distance between the support functions of $\hat{\boldsymbol{\Theta}}$ and its expectation, see Vitale (1985) for a study of this distance between convex sets. Hence, the A-optimal design aims to minimize the expected square distance between $\hat{\boldsymbol{\Theta}}$ and its expectation. The E-optimal design minimizes the L_∞ distance between the variance of the support function and zero, which is the support function of the origin. In case of the objective function (3) for the G-optimal design, the metric is mixed — the L_2 distance for one component of the tensor product $\mathbb{R}^p \times \mathbb{R}^{r+1}$ and the L_∞ distance for the other one.

In the classic linear univariate response setting with normal errors, the D-optimal design minimizes the confidence ellipsoid of parameter $\theta \in \mathbb{R}^{r+1}$

$$\{\theta : (\theta - \hat{\theta})^\top \Sigma^{-1}(\theta - \hat{\theta}) \le c\}$$

for a constant c, where $\hat{\theta}$ is the least square estimator of θ. The volume of this ellipsoid is proportional to $\det(\Sigma)^{-1/2}$, so that the objective function for D-optimal design becomes $\det(\Sigma)^{-1}$.

In the multiresponse setting of dimension p, it is possible to vectorize the parameter matrix by modifying the linear equation as

$$\begin{pmatrix} Y_1 \\ \\ Y_p \end{pmatrix} = \begin{pmatrix} \mathscr{X} & & \\ & \ddots & \\ & & \mathscr{X} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}. \tag{15}$$

The vector on the left-hand side arises by stacking together n observations for each component Y_j, $j = 1, \dots, p$, of the response. Furthermore, $\operatorname{diag}(\mathscr{X}, \dots, \mathscr{X})$ is an $np \times (r+1)p$ block-diagonal matrix built of the $n \times (r+1)$ dimensional design matrix $\mathscr{X}$; $\theta_j \in \mathbb{R}^{r+1}$, $j = 1, \dots, p$, are the estimated parameters, and ε_j, $j = 1, \dots, p$, are n-dimensional random vectors. Using the vector representation (15), Chang (1994) proved that under the framework of approximate designs the D-optimal design in the multiresponse model is exactly the D-optimal design arising in the case of a univariate response. Kurotschka and Schwabe (1996) extended this reduction result for both exact and approximate designs for D-, A-, E-optimality criteria and for more general Φ-optimality defined by Kiefer (1974).

However, it is not possible to come up with an analogue of (15) for multivariate set-identified responses. In this case, one estimates the support function of $\boldsymbol{\Theta} = \mathbf{E}\hat{\boldsymbol{\Theta}}$, which is an infinite-dimensional parameter. Still, if one aims to reduce the integrated variance of $s(\hat{\boldsymbol{\Theta}}u, x)$ for each x, then the optimal design is the classical D-optimal one, see Remark 4.4.

It is worth to mention that our results in this paper are also applicable for multivariate polynomial regression models with set-identified response. In the classical setting of point-identified responses, Krafft and Schaefer (1992) determined the approximate D-optimal design for the polynomial regression model and obtained a partial result for exact n-point D-optimal designs complemented later by Imhof Imhof (2000) with a conjecture on G-optimum.

Our setting is restricted to the case of responses identified to belong to convex sets. In the non-convex setting, even the estimation of parameters is poorly understood not to mention the optimal design issues. In this case, the least square

estimator (1) involves taking the sum of possibly non-convex sets. However, due to the convexification effect of Minkowski sums (see Molchanov (2017, Sec. 3.1.1)), the estimator $\hat{\Theta}$ asymptotically becomes a convex set, and so such an estimator neglects the non-convexity of observations. Besides, we did not consider the case where the design matrix for each component of the response may be different. This was thoroughly studied by Soumaya et al. (2015) for the approximate *D*-optimal design with point-identified response.

Finally, note that our results are applicable only in the framework of Beresteanu and Molinari (2008); the optimal design issues in the interval regression setting of Blanco-Fernández et al. (2013) based on interval arithmetics do not fall into our scope of investigation.

Acknowledgements

The authors are grateful to the two anonymous referees for careful reading of the manuscript and supplying stimulating comments.

References

Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford University Press, Oxford.
Beresteanu, A., Molinari, F., 2008. Asymptotic properties for a class of partially identified models. Econometrica 76, 763–814.
Blanco-Fernández, A., Colubi, A., García-Bárzana, M., 2013. A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables. Inform. Sci. 247, 109–122.
Chang, S.I., 1994. Some properties of multiresponse *D*-optimal designs. J. Math. Anal. Appl. 184, 256–262.
Diamond, P., 1990. Least squares fitting of compact set-valued data. J. Math. Anal. Appl. 147, 351–362.
Imhof, L., 2000. Optimum designs for a multiresponse regression model. J. Multivariate Anal. 72, 120–131.
Jacroux, M., 1983. On the MV-optimality of chemical balance weighing designs. Calcutta Statist. Assoc. Bull. 32, 143–151.
Kiefer, J., 1974. General equivalence theory for optimum designs (approximate theory). Ann. Statist. 2, 849–879.
Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canad. J. Math. 12, 363–366.
Krafft, O., Schaefer, M., 1992. *D*-optimal Designs for a multivariate regression model. J. Multivariate Anal. 42, 130–140.
Kurotschka, V.G., Schwabe, R., 1996. The reduction of design problems for multivariate experiments to univariate possibilities and their limitations. In: Brunner, E., Denker, M. (Eds.), Research Developments in Probability and Statistics. VSP, Utrecht, pp. 193–204.
Molchanov, I., 2017. Theory of Random Sets, second ed. Springer, London.
Molchanov, I., Molinari, F., 2018. Random Sets in Econometrics. Cambridge University Press, Cambridge.
Silvey, S.D., 1980. Optimal Design. Chapman and Hall, London.
Soumaya, M., Gaffke, N., Schwabe, R., 2015. Optimal design for multivariate observations in seemingly unrelated linear models. J. Multivariate Anal. 142, 48–56.
Vitale, R.A., 1985. L_p Metrics for compact, convex sets. J. Approx. Theory 45, 280–287.

Paper C

Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. 2021 doi: 10.1111/anzs.12326

Depth and outliers for samples of sets and random sets distributions

Ignacio Cascos[1] Qiyu Li[2] and Ilya Molchanov[2*]

Universidad Carlos III de Madrid and the University of Bern

Summary

We suggest several constructions suitable to define the depth of set-valued observations with respect to a sample of convex sets or with respect to the distribution of a random closed convex set. With the concept of a depth, it is possible to determine if a given convex set should be regarded an outlier with respect to a sample of convex closed sets. Some of our constructions are motivated by the known concepts of half-space depth and band depth for function-valued data. A novel construction derives the depth from a family of non-linear expectations of random sets. Furthermore, we address the role of positions of sets for evaluation of their depth. Two case studies concern interval regression for Greek wine data and detection of outliers in a sample of particles.

Key words: depth; half-space depth; outliers; random set; set-valued data; sublinear expectation.

1. Introduction

Statistical data in the form of sets or images are relevant in many fields of research. In a large number of applications in biology, microscopy and image analysis, the observed sets are non-convex, see Chiu *et al.* (2013). In some cases, for instance in statistics of particles (see Stoyan & Stoyan 1994) and in partially identified models in econometrics (see Molchanov & Molinari 2018), the data consist of convex sets, which is the setting of this paper. In the simplest one-dimensional case, observations are given by intervals, for example, data on daily price ranges in finance, imprecise measurements, salary brackets in econometrics, to name a few sources, see Beresteanu & Molinari (2008), Blanco-Fernández, Colubi & González-Rodríguez (2012), Blanco-Fernández, Corral & González-Rodríguez (2011), Manski & Tamer (2002) and Yang *et al.* (2016) and references therein. A substantial body of these works focuses on regression with interval responses and sometimes also interval regressors.

This paper aims to explore possible ways to identify outliers in samples of general convex sets. This obviously includes the case of random points, typical in multivariate

*Author to whom correspondence should be addressed.
[1]Department of Statistics, Universidad Carlos III de Madrid, Av. Universidad 30, Leganés (Madrid), 28911, Spain.
[2]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, Bern, 3012, Switzerland. e-mail: qiyu.li@stat.unibe.ch (Q.L.); ilya.molchanov@stat.unibe.ch (I.M.)

Acknowledgements. The author gratefully acknowledges advice from Francesca Molinari. Any remaining errors or omissions are all their fault. IM was partially supported by Swiss National Science Foundation Grant 200021_175584.

statistics. For multivariate samples (of points), the outliers are conventionally identified using depth functions and associated depth-trimmed regions, so that sample points lying outside these regions are regarded outliers, see Liu, Parelius & Singh (1999) or Mosler (2002). Decreasing transformations of depth functions serve as outlyingness functions, see Dang & Serfling (2010).

There are several general approaches to construct multivariate depth-trimmed regions. The historically first and most popular one is the Tukey depth (also called the half-space depth): the depth of a point x with respect to a probability distribution in Euclidean space is the smallest probability content among all half-spaces having x on the boundary, see Tukey (1975) and Rousseeuw & Ruts (1999). Various interpretations of this depth can be found in the recent survey by Nagy, Schütt & Werner (2019). Hamel & Kostner (2018) discussed the quantile-based multivariate depth in relation to a partial order on the space. Further depth notions (simplicial depth, convex hull depth, etc.) are based on assessing the location of a point with respect to an i.i.d. sample from a given distribution, see Liu (1990) and Cascos (2010). General properties of statistical depth functions have been analysed in Zuo & Serfling (2000), where further references can be found.

Extensions of the concept of depth to the functional data setting go back to the work of Fraiman & Muniz (2001). However, the construction of depth often causes difficulties when the data consist of functions. The main problem lies in the fact that a sample of functions is a too meagre set in a functional space and so most functions have depth zero in relation to the sampled ones. Difficulties with the half-space depth of functional data are discussed in Dutta, Ghosh & Chaudhuri (2011) and Kuelbs & Zinn (2013). This degeneracy of the half-space depth can be overcome by considering band depth, see López-Pintado & Romo (2009). The key idea is to replace taking the convex hull of functions with their envelopes determined by pointwise minima and maxima of functions. An alternative way to deal with the degeneracy problem is by considering the infimum of the half-space depth over the domain of the function, obtaining the so-called infimal depth, see Mosler (2013) and Gijbels & Nagy (2015). Various concepts of the depth in the functional setting are discussed by Gijbels & Nagy (2017).

Further generalisations of the concept of depth to other data types have been elaborated by Pandolfo, Paindaveine & Porzio (2018) for directional data (and so belonging to a non-linear space), by Chen, Gao & Ren (2018) and Paindaveine & Van Bever (2018) for matrix-valued data, and by Lafaye De Micheaux, Mozharovskyi & Vimond (2020) for curves. Already in the context of set-valued data discussed in this manuscript, Whitaker, Mirzargar & Kirby (2013) extended the band constructions of López-Pintado & Romo (2009) by considering intersections and unions of sets. The current paper continues this programme of exploring non-traditional data types and presents several constructions suitable to define depth of a convex set with respect to a sample of sets or with respect to the distribution of a random convex closed set. In this way, it is possible to identify outliers as sets of low depth. Some of our constructions are applicable for closed convex sets, which are not necessarily bounded. Unlike points and functions, sets are visually perceived objects – it is very difficult to compete with human perception in identifying outliers. Still, the suggested tools might outperform the human eye for sets in higher-dimensional spaces. Deriving deep properties of the introduced depth functions is a challenging task, left for future work.

Section 2 recalls main concepts of a random convex closed set and its expectation. General properties of depth functions for sets are discussed in Section 3. Such depth functions are functionals $\mathrm{D}(F, \boldsymbol{X})$ of a convex closed set F and the distribution of random convex closed set $\boldsymbol{X}$. In statistical applications, the theoretical distribution of $\boldsymbol{X}$ is replaced by its empirical version.

Section 4 presents a concept of the depth based on set-valued non-linear expectations. In the general classification scheme suggested by Zuo & Serfling (2000), this definition corresponds to Type C depth functions. In this framework, one associates with the distribution of $\boldsymbol{X}$ and an $\alpha \in (0, 1]$ a family $\mathcal{F}_\alpha(\boldsymbol{X})$ of convex closed sets in $\mathbb{R}^d$, which becomes richer as α decreases. The depth of F is the largest α such that $F \in \mathcal{F}_\alpha(\boldsymbol{X})$, so that $\mathcal{F}_\alpha(\boldsymbol{X})$ becomes the depth-trimmed region at level α. As we see in Section 4, the families $\mathcal{F}_\alpha(\boldsymbol{X})$ can be conveniently chosen to be all convex closed sets sandwiched between $\mathcal{U}_\alpha(\boldsymbol{X})$ and $\mathcal{E}_\alpha(\boldsymbol{X})$, which are convex sets determined by $\boldsymbol{X}$ such that $\mathcal{U}_\alpha(\boldsymbol{X})$ becomes larger (as a set) and $\mathcal{E}_\alpha(\boldsymbol{X})$ becomes smaller as α increases between 0 and 1. These two functions $\mathcal{U}_\alpha(\boldsymbol{X})$ and $\mathcal{E}_\alpha(\boldsymbol{X})$ are termed set-valued non-linear expectations. In the special case of singleton sets and a particular choice of the non-linear expectations, this definition corresponds to the expected convex hull depth suggested by Cascos (2007) and the zonoid depth introduced by Koshevoy & Mosler (1997), see also Mosler (2002).

Sets can be represented as functions: the most obvious representation uses their indicator functions, while in the convex case one typically represents sets as support functions. Section 5 describes the half-space depth concept applied to support functions. Such a depth function is of Type D of Zuo & Serfling (2000); the depth of a set F is the infimum of probabilities that $\boldsymbol{X}$ belongs to certain families of sets related to F. It is related to the infimal depth studied by Gijbels & Nagy (2015).

Type A depth functions $\mathrm{D}(F, \boldsymbol{X})$ are constructed as expectations of functionals $\psi(F; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)$ of F and a fixed number j of i.i.d. copies of $\boldsymbol{X}$. The functional ψ measures the closeness of F to the sample of sets. An important example of this construction is motivated by the band depth for functions, see Section 6.

Type B depth functions are defined as $(1 + \mathrm{E}\psi(F; \boldsymbol{X}_1, \ldots, \boldsymbol{X}_j))^{-1}$, where ψ describes a kind of distance of F to the sample of sets $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j$. A variant of this construction for random convex sets is presented in Section 7. Further concepts of depth of sets can be elaborated by considering the space of closed convex sets a metric space and following the general approach by Mizera (2002).

Sets can be described in terms of their locations, shapes and sizes. While locations of sets are usually essential in econometric applications, they are irrelevant in statistics of particles. Section 8 considers a possibility of factoring out the location effect. It is considerably more complicated to eliminate effects of arbitrary rotations and/or scaling; this non-trivial (even for samples of point tuples, see Kendall *et al.* 1999) question is left outside the scope of the current work.

Section 9 defines depth of random integrable probability measures by associating them with special convex sets called lift zonoids. This is important in view of several recent works concerning statistics in the Wasserstein space, which is the space of integrable probability measures, see, for example, Bigot, Cazelles & Papadakis (2019), Cazelles *et al.* (2018) and Zemel & Panaretos (2019). In view of this, it is often desirable to consider samples of empirical probability measures (e.g. collections of empirical cumulative distribution functions) and identify outliers in such samples.

Section 10 presents two case studies. The first one concerns a Greek wine data (see Kallithrakaa *et al.* 2001) with the aim to fit regression lines to interval-valued responses. The second one identifies outliers in a sample of particles analysed by Stoyan & Molchanov (1997).

2. Random convex sets and their mean values

A *random closed set* $\boldsymbol{X}$ in Euclidean space is a measurable map from a probability space $(\Omega, \mathcal{F}, \Pr)$ to the family $\mathcal{F}$ of closed sets in $\mathbb{R}^d$. The measurability condition requires that $\{\omega : \boldsymbol{X}(\omega) \cap K \neq \varnothing\} \in \mathcal{F}$ for all compact sets K in $\mathbb{R}^d$. The empty set is closed and compact. The random closed set $\boldsymbol{X}$ is said to be *convex* if it almost surely takes values from the family $\operatorname{co}\mathcal{F}$ of convex closed sets; $\boldsymbol{X}$ is *compact* convex if $\boldsymbol{X}$ almost surely belongs to the family $\operatorname{co}\mathcal{K}$ of convex compact sets. Non-empty compact convex sets are also called convex bodies. A set $F \in \operatorname{co}\mathcal{F}$ is said to belong to the *support* of $\boldsymbol{X}$ if $\boldsymbol{X}$ with a positive probability belongs to any open neighbourhood of F in the Fell topology. Recall that a sequence $\{F_n, n \geqslant 1\}$ converges to $F \in \mathcal{F}$ in the Fell topology if $F_n \cap G \neq \varnothing$ for all sufficiently large n and any open set G that hits F, and $F_n \cap K = \varnothing$ for all sufficiently large n and any compact set K that misses F, see Molchanov (2017, Appendix C).

A random vector ξ in $\mathbb{R}^d$ is said to be a *selection* of $\boldsymbol{X}$ if $\xi \in \boldsymbol{X}$ almost surely, that is, $\xi(\omega)$ belongs to $\boldsymbol{X}(\omega)$ for almost all ω. The family of all selections of $\boldsymbol{X}$ is denoted by $L^0(\boldsymbol{X})$.

For $p \in [1, \infty]$, a random closed set $\boldsymbol{X}$ is said to be *p-integrable* (simply, integrable if $p = 1$) if the family $L^p(\boldsymbol{X})$ of its p-integrable (essentially bounded if $p = \infty$) selections is not empty. Note that $L^p(\boldsymbol{X})$ is the intersection of $L^0(\boldsymbol{X})$ with the family $L^p(\mathbb{R}^d)$ of p-integrable random vectors. If $\boldsymbol{X}$ is integrable, its *selection expectation* is defined by

$$\mathrm{E}(\boldsymbol{X}) = \operatorname{cl}\big\{\mathrm{E}(\xi) : \xi \in L^1(\boldsymbol{X})\big\}.$$

The closure on the right-hand side is not needed if $\boldsymbol{X}$ is a subset of a centred ball with integrable radius. If the underlying probability space is non-atomic, the expectation of $\boldsymbol{X}$ and the expectation of its closed convex hull $\operatorname{conv}\boldsymbol{X}$ are the same, see Molchanov (2017, Sec. 2.1).

Convex closed sets are uniquely identified by their *support functions*

$$h(F, u) = \sup\{\langle u, x\rangle : x \in F\}, \quad u \in \mathbb{R}^d,$$

where $\langle u, x\rangle$ is the scalar product in $\mathbb{R}^d$. The support function may take the value ∞; it takes the value $-\infty$ only if F is empty. For a random convex closed set $\boldsymbol{X}$, the support function $h(\boldsymbol{X}, u)$ is a random function of $u \in \mathbb{R}^d$, and

$$\mathrm{E}(h(\boldsymbol{X}, u)) = h(\mathrm{E}(\boldsymbol{X}), u), \quad u \in \mathbb{R}^d,$$

if $\boldsymbol{X}$ is integrable.

Distances between convex bodies can be defined using L^p-distances between their support functions restricted for u belonging to the unit sphere $\mathbb{S}^{d-1}$. Most importantly, the L^∞-distance between support functions

$$\rho_H(K, L) = \|h(K, \cdot) - h(L, \cdot)\|_\infty = \sup_{u \in \mathbb{S}^{d-1}} |h(K, u) - h(L, u)|$$

is called the *Hausdorff distance* between K and L from $\operatorname{co}\mathcal{K}$. Furthermore,

$$\|K\| = \rho_H(K, \{0\}) = \sup_{u \in \mathbb{S}^{d-1}} |h(K, u)|$$

is called the norm of K. If $K = \{x\}$ is a singleton, then $\|K\| = \|x\|$ is the Euclidean norm of x.

The L^p-distance between the support functions yields the metric on $\operatorname{co}\mathcal{K}$ given by

$$\rho_p(K, L) = \left(\int_{u \in \mathbb{S}^{d-1}} |h(K, u) - h(L, u)|^p du \right)^{1/p}, \quad 1 \leqslant p < \infty.$$

Other distances between sets arise by taking L^p-distances between their (truncated) distance functions, see Baddeley (1992). Recall that the distance function of a set A is defined by $d(x, A) = \inf\{\|x - y\| : y \in A\}$.

Closed sets in $\mathbb{R}^d$ can be added elementwisely (in the Minkowski sense), so that

$$F_1 + F_2 = \operatorname{cl}\{x + y : x \in F_1, y \in F_2\}.$$

The closure on the right-hand side is not needed if at least one of the summands is bounded (and so is compact). In particular, $F + a = \{x + a : x \in F\}$, for $a \in \mathbb{R}^d$.

3. General depth functions and depth-trimmed regions for set-valued observations

Classical depth functions for random vectors associate to each point x in $\mathbb{R}^d$ its depth $\mathrm{D}(x, \xi)$ in relation to the distribution of a random vector ξ, see Zuo & Serfling (2000). The *depth function* is assumed to take values between 0 and 1, and can be equivalently represented in terms of its upper level sets

$$\mathrm{D}^\alpha(\xi) = \{x \in \mathbb{R}^d : \mathrm{D}(x, \xi) \geqslant \alpha\},$$

called the *depth-trimmed regions*. A point x with $\mathrm{D}(x, \xi) < \alpha$ and so lying outside $\mathrm{D}^\alpha(\xi)$ is considered an outlier at level α; then α is chosen to be rather small. For larger αs, the depth-trimmed region can be used to describe the centre of the probability distribution. It should be noted that some definitions of depth functions impose integrability assumptions on ξ. In the empirical variant of the depth, the probability distribution of ξ is usually replaced by the empirical probability measure.

If $\boldsymbol{X}$ is a random closed set, its depth function $\mathrm{D}(F, \boldsymbol{X})$ is a function of a closed set F. For random vectors ξ, all points outside the convex hull of the support of ξ are usually assigned zero depth. For random sets, we let $\mathrm{D}(F, \boldsymbol{X}) = 0$ if F does not belong to the *convex hull of the support* of $\boldsymbol{X}$, that is, for F that cannot be represented as a limit (in the Fell topology) of the sums $p_1F_1 + \cdots + p_nF_n$ for $n \geqslant 1$, with closed sets $F_1, \ldots, F_n$ from the support of $\boldsymbol{X}$ and non-negative weights $p_1, \ldots, p_n$ that sum up to one. In particular, if $\boldsymbol{X}$ is a random convex closed set, then all non-convex sets F are of zero depth.

Translating some general properties of depth-trimmed regions of random vectors postulated in Zuo & Serfling (2000) into the set-valued framework, one comes up with the following list.

(D1) Affine invariance:

$$\mathrm{D}(AF + b, A\boldsymbol{X} + b) = \mathrm{D}(F, \boldsymbol{X})$$

for all non-singular $d \times d$ matrices A and $b \in \mathbb{R}^d$.

(D2) Upper semicontinuity:

$$\mathrm{D}(F,\boldsymbol{X}) \geqslant \limsup_{n\to\infty} \mathrm{D}(F_n,\boldsymbol{X}),$$

if $F_n \to F$ as $n\to\infty$ in the Fell topology on $\operatorname{co}\mathcal{F}$.

(D3) If $\boldsymbol{X}$ is deterministic, that is, $\boldsymbol{X}=L$ a.s. for $L\in\operatorname{co}\mathcal{F}$, then $\mathrm{D}(F,L)=\mathbf{1}(F=L)$.

For a random convex closed set $\boldsymbol{X}$, the depth-trimmed region

$$\mathrm{D}^{\alpha}(\boldsymbol{X}) = \{F\in\operatorname{co}\mathcal{F}\colon \mathrm{D}(F,\boldsymbol{X})\geqslant\alpha\} \tag{1}$$

is a family of convex closed sets. The following result is a straightforward reformulation of the properties (D1) and (D2).

Proposition 3.1. *The properties of the depth imply the following properties of the depth-trimmed regions*

(i) *Affine equivariance:* $\mathrm{D}^{\alpha}(A\boldsymbol{X}+b)=\{AF+b\colon F\in\mathrm{D}^{\alpha}(\boldsymbol{X})\}$ *for all non-singular* $d\times d$ *matrices* A *and* $b\in\mathbb{R}^d$.

(ii) *Closedness: the family* $\mathrm{D}^{\alpha}(\boldsymbol{X})$ *is closed in the Fell topology on* $\operatorname{co}\mathcal{F}$.

4. Depth based on non-linear expectations

4.1. Non-linear expectations of random convex closed sets

The systematic study of sublinear expectations in probability theory was initiated by Peng (2004), see also the recent monograph Peng (2019). For random variables, sublinear expectations are deeply related to constructions of depth-trimmed regions, see Example 4.10 and Cascos & Molchanov (2007).

A set-valued generalisation of non-linear expectations elaborated in Molchanov & Mühlemann (2021) relies on working with two functions, one $\mathcal{E}$ being subadditive and the other $\mathcal{U}$ being superadditive for the conventional set inclusion. These functions are called sublinear and superlinear expectations; they are defined on p-integrable random closed sets, and in the following we fix $p\in[1,\infty]$. In the set-valued setting, it is not possible to pass from a sublinear expectation to a superlinear one by changing the sign (corresponding to the central symmetry transform for sets) – separate treatments of them are necessary.

Definition 4.1. A *sublinear set-valued* expectation is a function $\mathcal{E}$ defined on p-integrable random convex closed sets in $\mathbb{R}^d$ with values in the family $\operatorname{co}\mathcal{F}$ and such that

(i) for each deterministic $a\in\mathbb{R}^d$,

$$\mathcal{E}(\boldsymbol{X}+a)=\mathcal{E}(\boldsymbol{X})+a;$$

(ii) $\mathcal{E}(F)\supseteq F$ for all deterministic $F\in\operatorname{co}\mathcal{F}$;

(iii) $\mathcal{E}(\boldsymbol{X})\subseteq\mathcal{E}(\boldsymbol{Y})$ if $\boldsymbol{X}(\omega)\subseteq\boldsymbol{Y}(\omega)$ for almost all ω;

(iv) $\mathcal{E}(c\boldsymbol{X})=c\mathcal{E}(\boldsymbol{X})$ for all $c>0$;

(v) $\mathcal{E}$ is subadditive, that is,

$$\mathcal{E}(\boldsymbol{X}+\boldsymbol{Y})\subseteq\mathcal{E}(\boldsymbol{X})+\mathcal{E}(\boldsymbol{Y}), \tag{2}$$

for all p-integrable random convex closed sets $\boldsymbol{X}$ and $\boldsymbol{Y}$.

A *superlinear set-valued* expectation $\mathcal{U}$ satisfies the same properties with the exception of (ii) replaced by $\mathcal{U}(F)\subseteq F$ and (v), where (2) is replaced by the supperadditivity property

$$\mathcal{U}(\boldsymbol{X}+\boldsymbol{Y}) \supseteq \mathcal{U}(\boldsymbol{X}) + \mathcal{U}(\boldsymbol{Y}). \tag{3}$$

A non-linear expectation is said to be *law-determined* (often called law invariant) if it depends only on the distribution of $\boldsymbol{X}$. A non-linear expectation is called *constant preserving* if it preserves deterministic sets from $\operatorname{co}\mathcal{F}$, for example, $\mathcal{E}(F) = F$ for all $F \in \operatorname{co}\mathcal{F}$. In the following all non-linear expectations are assumed to be law-determined and constant preserving. Additionally, assume that all non-linear expectations are *affine equivariant*, for example, $\mathcal{E}(A\boldsymbol{X}) = A\mathcal{E}(\boldsymbol{X})$ for all non-singular $d \times d$ matrices A.

Two non-linear expectations $\mathcal{U}$ and $\mathcal{E}$ are said to form a *dual pair* if $\mathcal{U}(\boldsymbol{X}) \subseteq \mathcal{E}(\boldsymbol{X})$ for all p-integrable random convex closed sets $\boldsymbol{X}$. Note that $\mathcal{U}(\boldsymbol{X})$ is allowed to take empty values. The superlinear expectation is consistently extended for random sets which are empty with a positive probability (and so are not p-integrable) by letting it to be empty in such cases.

Numerous examples of sublinear and superlinear expectations are obtained by letting

$$\mathcal{E}(\boldsymbol{X}) = \operatorname{conv} \bigcup_{\gamma \in \mathcal{M}, \mathrm{E}(\gamma)=1} \mathrm{E}(\gamma \boldsymbol{X}) \tag{4}$$

and

$$\mathcal{U}(\boldsymbol{X}) = \bigcap_{\gamma \in \mathcal{M}, \mathrm{E}(\gamma)=1} \mathrm{E}(\gamma \boldsymbol{X}), \tag{5}$$

where $\mathcal{M}$ is a convex subset of the family $L^q(\mathbb{R}_+)$ consisting of q-integrable non-negative random variables with $1/p + 1/q = 1$. The set $\mathcal{M}$ is chosen to be closed in the $\sigma(L^q, L^p)$-topology, that is, in the weak-star topology on $L^q(\mathbb{R}_+)$. Note that $\gamma\boldsymbol{X} = \{\gamma x : x \in \boldsymbol{X}\}$, which is $\boldsymbol{X}$ scaled by γ.

The sublinear expectation $\mathcal{E}(\boldsymbol{X})$ given by (4) is the convex closed set satisfying

$$h(\mathcal{E}(\boldsymbol{X}), u) = \mathrm{e}(h(\boldsymbol{X}, u)), \quad u \in \mathbb{R}^d, \tag{6}$$

where $\mathrm{e} : L^p(\mathbb{R}) \to (-\infty, \infty]$ is a numerical sublinear expectation, see Peng (2019). Such an equality may be violated in the superlinear case – the support function of $\mathcal{U}(\boldsymbol{X})$ is only dominated by the superlinear expectation of $h(\boldsymbol{X}, u)$.

4.2. Depth defined using a general parametric family

Definition 4.2. A family $(\mathcal{U}_\alpha, \mathcal{E}_\alpha)$, $\alpha \in (0, 1]$, of dual pairs of non-linear expectations is said to form a *parametric family* if, for each p-integrable random closed set $\boldsymbol{X}$, $\mathcal{U}_\alpha(\boldsymbol{X})$ is increasing and $\mathcal{E}_\alpha(\boldsymbol{X})$ is decreasing (in the sense of set inclusions) as functions of $\alpha \in (0, 1]$.

Two basic constructions of parametric families are suggested in Sections 4.3 and 4.4. Fix a parametric family $(\mathcal{U}_\alpha, \mathcal{E}_\alpha)$, $\alpha \in (0, 1]$, of dual pairs of constant preserving law-determined affine equivariant non-linear expectations.

Let F belong to the convex hull of the support of $\boldsymbol{X}$. Define a depth function of such F by letting

$$\mathrm{D}(F, \boldsymbol{X}) = \sup\left\{\alpha \in (0, 1] : \mathcal{U}_\alpha(\boldsymbol{X}) \subseteq F \subseteq \mathcal{E}_\alpha(\boldsymbol{X})\right\} \tag{7}$$

and $\sup \varnothing = 0$. If $\mathcal{U}_\alpha = \varnothing$ is the trivial (always empty) superlinear expectation, then the above definition of the depth is also applicable; it reduces to the only inclusion for the sublinear

expectation. This definition of depth can be regarded as of Type C depth according to the classification from Zuo & Serfling (2000). The depth-trimmed region $\mathrm{D}^{\alpha}(\boldsymbol{X})$ consists of convex closed sets from the convex hull of the support of $\boldsymbol{X}$ which contain $\mathcal{U}_{\alpha}(\boldsymbol{X})$ and are subsets of $\mathcal{E}_{\alpha}(\boldsymbol{X})$.

Proposition 4.3. *The depth function constructed by (7) using a parametric family $(\mathcal{U}_{\alpha}, \mathcal{E}_{\alpha})$ of non-linear expectations satisfies conditions (D1)–(D3).*

Proof. (D1) It suffices to note that

$$\begin{aligned}\mathrm{D}(AF+b, A\boldsymbol{X}+b) &= \sup\bigl\{\alpha : \mathcal{U}_{\alpha}(A\boldsymbol{X}+b) \subseteq AF+b \subseteq \mathcal{E}_{\alpha}(A\boldsymbol{X}+b)\bigr\} \\ &= \mathrm{D}(F, \boldsymbol{X}).\end{aligned}$$

Here we have used the properties (i) and (iv) of non-linear expectations and the additionally imposed affine equivariance.

(D2) Let $F_n \to F$ in the Fell topology, and let $\mathrm{D}(F_n, \boldsymbol{X}) = \alpha_n$, $n \geqslant 1$, with $\alpha = \limsup \alpha_n$. For each $\varepsilon > 0$, there exists a subsequence $\{n_k\}$ such that $\alpha_{n_k} \geqslant \alpha - \varepsilon$, and the monotonicity property implies

$$\mathcal{U}_{\alpha-\varepsilon}(\boldsymbol{X}) \subseteq \mathcal{U}_{\alpha_{n_k}}(\boldsymbol{X}) \subseteq F_{n_k} \subseteq \mathcal{E}_{\alpha_{n_k}}(\boldsymbol{X}) \subseteq \mathcal{E}_{\alpha-\varepsilon}(\boldsymbol{X}).$$

Hence $\mathrm{D}(F, \boldsymbol{X}) \geqslant \alpha - \varepsilon$ for all $\varepsilon > 0$.

(D3) follows from the constant preserving property of the chosen non-linear expectations.

Remark 4.4. Standard properties of depth functions for random vectors have been further augmented in Cascos & Molchanov (2007) and Cascos (2010) by assuming that $\mathrm{D}^{\alpha}(\xi+\eta) \subseteq \mathrm{D}^{\alpha}(\xi) + \mathrm{D}^{\alpha}(\eta)$ for any two random vectors ξ and η from the domain of definition of the depth function. In other words, this means that if $\mathrm{D}(z, \xi+\eta) \geqslant \alpha$ and so z is not an outlier for $\xi+\eta$, then it is possible to represent z as the sum of two non-outliers for ξ and η, respectively. Then, if ξ is a random variable, $\inf \mathrm{D}^{\alpha}(\xi)$ is a superadditive homogeneous function of ξ. By changing the sign, one obtains a subadditive antimonotonic function of ξ, which is influenced by the lower tail of ξ and is usually called a *risk measure*, see for example, Delbaen (2002). This explains a connection between risk measures of random variables and depth-trimmed regions in dimension one, see Cascos & Molchanov (2007). However, such a connection fails in higher dimensions, since changing the sign does not alter the direction of inclusion for sets.

A variant of the above subadditivity property holds for depth functions defined using non-linear expectations of random sets. If $\mathrm{D}(F, \boldsymbol{X}+\boldsymbol{Y}) \geqslant \alpha$, equivalently, $F \in \mathrm{D}^{\alpha}(\boldsymbol{X}+\boldsymbol{Y})$, then

$$F_1 + F_2 \subseteq F \subseteq F_1' + F_2',$$

for some $F_1, F_1' \in \mathrm{D}^{\alpha}(\boldsymbol{X})$ and $F_2, F_2' \in \mathrm{D}^{\alpha}(\boldsymbol{Y})$. Indeed, it suffices to let $F_1 = \mathcal{U}_{\alpha}(\boldsymbol{X})$, $F_1' = \mathcal{E}_{\alpha}(\boldsymbol{X})$, and define F_2, F_2' in the same way for $\boldsymbol{Y}$.

Remark 4.5. The above construction of depth also applies without assuming sub- and superadditivity of the set-valued functions $\mathcal{U}_\alpha$ and $\mathcal{E}_\alpha$. All properties (D1)–(D3) hold in this case. However, we do not pursue this in the current work.

Example 4.6. Assume that $\mathcal{U}(\boldsymbol{X}) = F_{\boldsymbol{X}}$ is the set of *fixed points* of $\boldsymbol{X}$, which is the set of $x \in \mathbb{R}^d$ such that $\Pr(x \in \boldsymbol{X}) = 1$. Furthermore, let $\mathcal{E}(\boldsymbol{X}) = R_{\boldsymbol{X}}$ be the *range* of $\boldsymbol{X}$, which is the set of all $x \in \mathbb{R}^d$ such that $\boldsymbol{X}$ hits any neighbourhood of x with positive probability, the set $R_{\boldsymbol{X}}$ is sometimes called the support of $\boldsymbol{X}$. Each parametric family of dual pairs $(\mathcal{U}_\alpha, \mathcal{E}_\alpha)$ can be consistently extended for $\alpha = 0$ by letting $\mathcal{U}_0(\boldsymbol{X}) = F_{\boldsymbol{X}}$ and $\mathcal{E}_0(\boldsymbol{X}) = R_{\boldsymbol{X}}$. These non-linear expectations are defined for all random closed sets; they form the dual pair of the smallest superlinear and the largest sublinear expectations. Thus, $F_X \subseteq \mathcal{U}_\alpha(\boldsymbol{X})$ and $\mathcal{E}_\alpha(\boldsymbol{X}) \subseteq R_{\boldsymbol{X}}$ for all $\alpha \in (0, 1]$, hence, all convex closed sets F of a strictly positive depth should satisfy $F_{\boldsymbol{X}} \subseteq F \subseteq R_{\boldsymbol{X}}$.

4.3. Parametric families constructed using i.i.d. copies

Let $\{\boldsymbol{X}_m, m \geqslant 1\}$ be i.i.d. copies of $\boldsymbol{X}$. For each $\alpha \in (0, 1]$,

$$\mathcal{E}^{\cup}_\alpha(\boldsymbol{X}) = \mathcal{E}(\mathrm{conv}(\boldsymbol{X}_1 \cup \cdots \cup \boldsymbol{X}_{[\alpha^{-1}]})) \tag{8}$$

is a sublinear expectation, and

$$\mathcal{U}^{\cap}_\alpha(\boldsymbol{X}) = \mathcal{U}(\boldsymbol{X}_1 \cap \cdots \cap \boldsymbol{X}_{[\alpha^{-1}]}) \tag{9}$$

is a superlinear expectation, where $[\alpha^{-1}]$ is the integer part of α^{-1}. The intersection of random sets on the right-hand side of (9) may be empty with a positive probability; in this case the superlinear expectation is also set to be empty. If $(\mathcal{U}, \mathcal{E})$ is a dual pair, then $\mathcal{U}^{\cap}_\alpha$ and $\mathcal{E}^{\cup}_\alpha$ also form a dual pair.

For the parametric family given by (8) and (9), we have

$$\mathrm{D}(F, \boldsymbol{X}) = \max\left\{m^{-1} : \mathcal{U}(\boldsymbol{X}_1 \cap \cdots \cap \boldsymbol{X}_m) \subseteq F \subseteq \mathcal{E}(\mathrm{conv}(\boldsymbol{X}_1 \cup \cdots \cup \boldsymbol{X}_m))\right\}. \tag{10}$$

In particular, $\mathrm{D}(F, \boldsymbol{X}) = 1$ if $\mathcal{U}(\boldsymbol{X}) \subseteq F \subseteq \mathcal{E}(\boldsymbol{X})$. The empirical variant of this depth function is obtained by resampling m values from the set of n realisations of $\boldsymbol{X}$ and treating the intersections and the convex hull of the unions of these values as realisations of $\boldsymbol{X}_1 \cap \cdots \cap \boldsymbol{X}_m$ and $\mathrm{conv}(\boldsymbol{X}_1 \cup \cdots \cup \boldsymbol{X}_m)$, respectively.

Remark 4.7. The above definition (10) yields depth functions with a discrete range of values. It is possible to obtain the depth with the whole range of values in $(0, 1]$ using 'smooth' parametric families of non-linear set-valued expectations constructed as follows. Let N_λ denote a geometrically distributed random variable with parameter $\lambda \in (0, 1]$, that is, $\Pr(N_\lambda = k) = \lambda(1 - \lambda)^{k-1}$, $k \geqslant 1$. The depth is defined as in (10) by replacing m^{-1} with λ and $\boldsymbol{X}_m$ with $\boldsymbol{X}_{N_\lambda}$ for a sequence $\{\boldsymbol{X}_m, m \geqslant 1\}$ of independent copies of $\boldsymbol{X}$, which are also independent of N_λ.

Example 4.8. (Random singletons). Let $\boldsymbol{X} = \{\xi\}$ be a random singleton with $\xi \in L^p(\mathbb{R}^d)$. The superlinear expectation of a singleton is either a singleton or is empty, and it is additive on the subfamily of random vectors in $L^p(\mathbb{R}^d)$ where it is not empty, see Molchanov &

Mühlemann (2021). Furthermore, if ξ is not deterministic, the intersection of at least two independent copies of $\boldsymbol{X}=\{\xi\}$ is empty with a positive probability. Since possible values of $\boldsymbol{X}$ are singletons and convex combinations of singletons are also singletons, non-singleton sets are assigned zero depth, and for a singleton $F=\{x\}$ we have

$$\mathrm{D}(\{x\},\{\xi\})=\max\left\{m^{-1}: x\in\mathcal{E}(\mathrm{conv}\{\xi_1,\dots,\xi_m\})\right\}.$$

If $\mathcal{E}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{X})$ is the expectation, we recover the expected convex hull depth, whose empirical version was considered in Cascos (2007). Then $\mathcal{E}(\mathrm{conv}\{\xi_1,\dots,\xi_m\})$ is the expected random polytope, see Fresen & Vitale (2014) for the relevant asymptotic results.

Example 4.9. Assume that $\boldsymbol{X}$ is integrable, and let the dual pairs $(\mathcal{U}_\alpha^{\cap},\mathcal{E}_\alpha^{\cup})$ be derived from $\mathcal{U}(\boldsymbol{X})=\mathcal{E}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{X})$. Then the only closed set F of depth one is the expectation of $\boldsymbol{X}$. The depth of a convex closed set F from the support of $\boldsymbol{X}$ is m^{-1} for the smallest m such that

$$\mathrm{E}(\boldsymbol{X}_1\cap\dots\cap\boldsymbol{X}_m)\subseteq F\subseteq\mathrm{E}(\mathrm{conv}(\boldsymbol{X}_1\cup\dots\cup\boldsymbol{X}_m)),$$

where $\boldsymbol{X}_1,\dots,\boldsymbol{X}_m$ are i.i.d. copies of $\boldsymbol{X}$. In order to handle the empirical variant of this construction, consider a sample of convex closed sets $F_1,\dots,F_n$. The corresponding random convex closed set $\boldsymbol{X}$ is assumed to equally likely take any of the values $F_1,\dots,F_n$. Then $\boldsymbol{X}_1\cup\dots\cup\boldsymbol{X}_m$ is distributed as $F_{i_1}\cup\dots\cup F_{i_m}$, where $i_1,\dots,i_m$ are i.i.d. random variables equally likely taking values $1,\dots,m$. A similar construction applies to the intersection. Note that $F_1\cap\dots\cap F_n$ is the set of fixed points for the empirical distribution and $\mathrm{conv}(F_1\cup\dots\cup F_n)$ is the range.

If $\boldsymbol{X}=[\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$ is a random interval and $[\boldsymbol{x}_{\mathrm{L}i},\boldsymbol{x}_{\mathrm{U}i}]$, $i=1,\dots,n$, are its independent copies, then the depth of $F=[a,b]$ is m^{-1} for the smallest m such that

$$\begin{aligned}\mathrm{E}\min(\boldsymbol{x}_{\mathrm{L}1},\dots,\boldsymbol{x}_{\mathrm{L}m})\leqslant a\leqslant{}&\mathrm{E}\max(\boldsymbol{x}_{\mathrm{L}1},\dots,\boldsymbol{x}_{\mathrm{L}m})\\ &\leqslant\mathrm{E}\min(\boldsymbol{x}_{\mathrm{U}1},\dots,\boldsymbol{x}_{\mathrm{U}m})\leqslant b\leqslant\mathrm{E}\max(\boldsymbol{x}_{\mathrm{U}1},\dots,\boldsymbol{x}_{\mathrm{U}m}),\end{aligned}$$

if $\mathrm{E}\max(\boldsymbol{x}_{\mathrm{L}1},\dots,\boldsymbol{x}_{\mathrm{L}m})\leqslant\mathrm{E}\min(\boldsymbol{x}_{\mathrm{U}1},\dots,\boldsymbol{x}_{\mathrm{U}m})$, and

$$\mathrm{E}\min(\boldsymbol{x}_{\mathrm{L}1},\dots,\boldsymbol{x}_{\mathrm{L}m})\leqslant a\leqslant b\leqslant\mathrm{E}\max(\boldsymbol{x}_{\mathrm{U}1},\dots,\boldsymbol{x}_{\mathrm{U}m}),$$

otherwise.

4.4. Parametric families constructed using weights: average quantiles

Fix a dual pair $(\mathcal{U},\mathcal{E})$ of non-linear expectations. Let $\mathcal{M}_\alpha$ with $\alpha\in(0,1]$ be a parametric family of random variables γ in $L^0(\mathbb{R}_+)$ which is decreasing in α, that is, $\mathcal{M}_\alpha\supseteq\mathcal{M}_{\alpha'}$ if $\alpha\leqslant\alpha'$, and such that $\mathcal{E}(\gamma\boldsymbol{X})$ and $\mathcal{U}(\gamma\boldsymbol{X})$ are well defined for a p-integrable random closed convex set $\boldsymbol{X}$. Define

$$\begin{aligned}\mathcal{E}_\alpha(\boldsymbol{X})&=\mathrm{conv}\bigcup_{\gamma\in\mathcal{M}_\alpha,\mathrm{E}(\gamma)=1}\mathcal{E}(\gamma\boldsymbol{X}),\\ \mathcal{U}_\alpha(\boldsymbol{X})&=\bigcap_{\gamma\in\mathcal{M}_\alpha,\mathrm{E}(\gamma)=1}\mathcal{U}(\gamma\boldsymbol{X}).\end{aligned}$$

In particular, the parametric variants of (4) and (5) arise by letting $\mathcal{E}(\boldsymbol{X})=\mathcal{U}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{X})$.

The most important case arises when $\mathcal{M}_\alpha$ is the family of all random variables with values in $[0,\alpha^{-1}]$ for $\alpha\in(0,1]$. Using this family $\mathcal{M}_\alpha$ with $\mathcal{E}(\boldsymbol{X})=\mathcal{U}(\boldsymbol{X})=\mathrm{E}(\boldsymbol{X})$, we arrive at

a parametric family of sublinear and superlinear expectations $\mathcal{E}_\alpha$ and $\mathcal{U}_\alpha$ called the *average quantile* ones. The reason for this name stems from the fact that

$$h(\mathcal{E}_\alpha(\boldsymbol{X}), u) = \mathtt{e}_\alpha(h(\boldsymbol{X}, u)) \tag{11}$$

with

$$\mathtt{e}_\alpha(\beta) = \frac{1}{\alpha}\int_{1-\alpha}^{1} q_t(\beta)dt, \tag{12}$$

being the average of the quantiles of $\beta \in L^1(\mathbb{R})$ at levels in $[1-\alpha, 1]$, see Föllmer & Schied (2004, Th. 4.47) for this fact derived for the risk measure $\mathtt{e}_\alpha(-\beta)$. Furthermore, $\mathcal{U}_\alpha(\boldsymbol{X})$ is the largest convex set whose support function is dominated by $\mathtt{u}_\alpha(h(\boldsymbol{X}, u))$, where $\mathtt{u}_\alpha(\beta) = -\mathtt{e}_\alpha(-\beta)$ is a numerical superlinear expectation, equivalently,

$$\mathtt{u}_\alpha(\beta) = \frac{1}{\alpha}\int_{0}^{\alpha} q_t(\beta)dt \tag{13}$$

is the average of lower quantiles of $\beta \in L^1(\mathbb{R})$. Note that this construction may result in superlinear expectation being empty on some random sets. With this construction, the depth-trimmed region $\mathrm{D}^\alpha(\boldsymbol{X})$ is the family of convex closed sets sandwiched between the intersection and the convex hull of the union of the sets $\mathrm{E}(\gamma\boldsymbol{X})$ for $\gamma \in \mathcal{M}_\alpha$.

Example 4.10. (Singletons). If $\boldsymbol{X} = \{\beta\}$ with $\beta \in L^1(\mathbb{R})$, then

$$\mathcal{E}_\alpha(\{\beta\}) = [\mathtt{u}_\alpha(\beta), \mathtt{e}_\alpha(\beta)],$$

while $\mathcal{U}_\alpha(\{\beta\}) = \varnothing$ for $\alpha < 1$ whenever β is not deterministic. If $\boldsymbol{X} = \{\xi\}$ for an integrable random vector ξ in $\mathbb{R}^d$, then $\mathcal{E}_\alpha(\{\xi\})$ is the convex closed set with the support function given by $\mathtt{e}_\alpha(\langle \xi, u\rangle)$. In this case, $\mathcal{U}_\alpha(\{\xi\}) = \varnothing$ for $\alpha < 1$ whenever ξ is not deterministic, and $\mathcal{E}_\alpha(\{\xi\})$ equals the zonoid-trimmed region of ξ introduced in Koshevoy & Mosler (1997). Recall that the *zonoid-trimmed region* of ξ is the set

$$\mathrm{ZD}^\alpha(\xi) = \alpha^{-1}\left\{x \in \mathbb{R}^d : (\alpha, x) \in \mathrm{E}(\mathrm{co}(0, (1, \xi)))\right\} \tag{14}$$

obtained as the rescaled section of the expectation $\mathrm{E}(\boldsymbol{Y})$ of the random closed convex set $\boldsymbol{Y}$, being the convex hull of the origin and the point $(1, \xi)$ in $\mathbb{R}^{d+1}$. The expectation $\mathrm{E}(\boldsymbol{Y})$ is termed the *lift zonoid* of ξ. The corresponding depth concept is called the zonoid depth of ξ, see Koshevoy & Mosler (1997) and Mosler (2002).

Example 4.11. (Lift-expectation). The lift zonoid concept was extended for general random convex sets by Diaye, Koshevoy & Molchanov (2018). Assume that $\mathrm{E}\|\boldsymbol{X}\| < \infty$; in this case $\boldsymbol{X}$ is called integrably bounded. Let $\boldsymbol{Y}$ be the convex hull of the origin in $\mathbb{R}^{d+1}$ and the set obtained as the Cartesian product $\{1\} \times \boldsymbol{X}$ (so to say, the uplifted $\boldsymbol{X}$). Since $\boldsymbol{Y}$ is integrably bounded, its expectation $\mathrm{E}(\boldsymbol{Y})$ is a convex body in $\mathbb{R}^{d+1}$ called the *lift-expectation* of $\boldsymbol{X}$ and denoted by $\hat{Z}_{\boldsymbol{X}}$. If the support function $h(\boldsymbol{X}, u)$ has a non-atomic distribution for all u, then

$$\mathcal{E}_\alpha(\boldsymbol{X}) = \alpha^{-1}\left\{x \in \mathbb{R}^d : (\alpha, x) \in \hat{Z}_{\boldsymbol{X}}\right\}$$

is a sublinear expectation and (11) holds.

Example 4.12. (Random intervals). Let $\boldsymbol{X}=[\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$ be a random interval on $\mathbb{R}$. Then $\mathcal{E}_\alpha(\boldsymbol{X})=[\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{L}}),\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{U}})]$, and $\mathcal{U}_\alpha(\boldsymbol{X})=[\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{L}}),\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{U}})]$ if $\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{L}})\leqslant\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{U}})$ and is empty otherwise. Hence, the depth of an interval $F=[a,b]$, see (7), is the largest α such that

$$\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{L}})\leqslant a\leqslant\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{L}})\leqslant\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{U}})\leqslant b\leqslant\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{U}}),$$

if $\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{L}})\leqslant\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{U}})$, and

$$\mathtt{u}_\alpha(\boldsymbol{x}_{\mathrm{L}})\leqslant a\leqslant b\leqslant\mathtt{e}_\alpha(\boldsymbol{x}_{\mathrm{U}}),$$

otherwise. The interval $[\mathrm{E}(\boldsymbol{x}_{\mathrm{L}}),\mathrm{E}(\boldsymbol{x}_{\mathrm{U}})]$ has depth 1.

In order to illustrate the intervals $\mathcal{E}_\alpha(\boldsymbol{X})$ and $\mathcal{U}_\alpha(\boldsymbol{X})$ obtained for a specific random set $\boldsymbol{X}=[\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$ and value of α, Figure 1 represents them together with the lift-expectation of $\boldsymbol{X}$, which was introduced in Example 4.11. The centrally symmetric lower almond-shaped set is the lift zonoid of $\boldsymbol{x}_{\mathrm{L}}$, denoted by $\hat{Z}_{\boldsymbol{x}_{\mathrm{L}}}$, while the upper almond-shaped set is the lift zonoid of $\boldsymbol{x}_{\mathrm{U}}$, denoted by $\hat{Z}_{\boldsymbol{x}_{\mathrm{U}}}$, and the whole shaded region is the lift-expectation of $\boldsymbol{X}$. The zonoid-trimmed regions at level α of $\boldsymbol{x}_{\mathrm{L}}$ and $\boldsymbol{x}_{\mathrm{U}}$, see (14), are the intervals obtained after scaling by α^{-1} the projection on the second coordinate of the intersection of the vertical line $x=\alpha$ with the corresponding lift zonoid. For example, the vertical line $x=0.4$ in Figure 1 represents of $\alpha\,\mathrm{ZD}^\alpha(\boldsymbol{x}_{\mathrm{L}})$ and $\alpha\,\mathrm{ZD}^\alpha(\boldsymbol{x}_{\mathrm{U}})$ for $\alpha=0.4$. The set $\mathcal{E}_\alpha(\boldsymbol{X})$ is α^{-1} times the projection on the second coordinate of the intersection of the vertical line $x=\alpha$ with the lift expectation of $\boldsymbol{X}$, while the set $\mathcal{U}_\alpha(\boldsymbol{X})$ is α^{-1} times the projection on the second coordinate of the intersection of the vertical line $x=\alpha$ with the lift-expectation of $\boldsymbol{X}$ and not in the lift zonoids of either of $\boldsymbol{x}_{\mathrm{L}}$ and $\boldsymbol{x}_{\mathrm{U}}$. See the vertical line $x=0.8$ in the chart for the representation of $\alpha\mathcal{E}_\alpha(\boldsymbol{X})$ and $\alpha\mathcal{U}_\alpha(\boldsymbol{X})$ for $\alpha=0.8$.

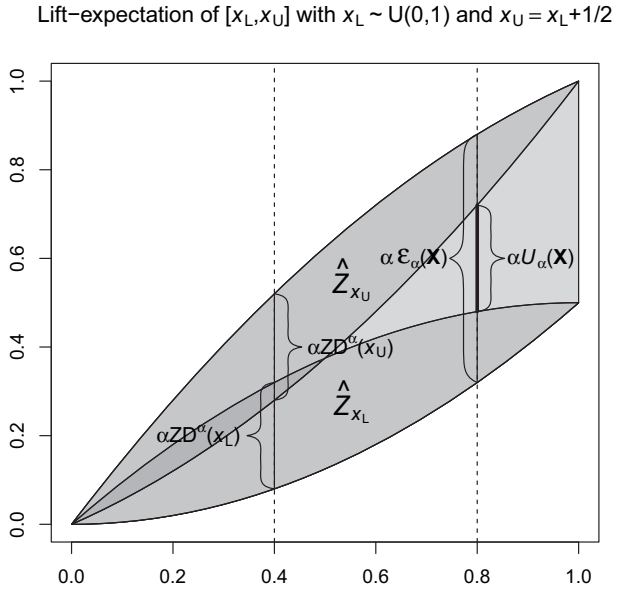

Figure 1. Lift expectation of the random interval $\boldsymbol{X}=[\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$, where $\boldsymbol{x}_{\mathrm{L}}$ is uniformly distributed on the unit interval $[0,1]$ and $\boldsymbol{x}_{\mathrm{U}}=\boldsymbol{x}_{\mathrm{L}}+1/2$.

5. Half-space depth

Since each random compact convex set $\boldsymbol{X}$ is uniquely associated with its support function $h(\boldsymbol{X},u)$ defined on the unit sphere $\mathbb{S}^{d-1}$, it is possible to apply known concepts of depth for functional data in order to quantify the depth of sets.

Following the half-space functional depth concept of Kuelbs & Zinn (2013), the *half-space depth* of $F \in \operatorname{co}\mathcal{F}$ with respect to $\boldsymbol{X}$ is defined as

$$\mathrm{HD}(F,\boldsymbol{X}) = \min(\mathrm{HD}_+(F,\boldsymbol{X}), \mathrm{HD}_-(F,\boldsymbol{X})), \tag{15}$$

where

$$\mathrm{HD}_+(F,\boldsymbol{X}) = \inf_{u\in\mathbb{S}^{d-1}} \Pr(h(\boldsymbol{X},u) \geqslant h(F,u)), \tag{16}$$

$$\mathrm{HD}_-(F,\boldsymbol{X}) = \inf_{u\in\mathbb{S}^{d-1}} \Pr(h(\boldsymbol{X},u) \leqslant h(F,u)). \tag{17}$$

It is easy to see that $\mathrm{HD}(F,\boldsymbol{X})$ satisfies the properties (D1)–(D3). Furthermore, $\mathrm{HD}_+(F,\boldsymbol{X}) \geqslant \Pr(F \subseteq \boldsymbol{X})$, and $\mathrm{HD}_-(F,\boldsymbol{X}) \geqslant \Pr(\boldsymbol{X} \subseteq F)$. If $\boldsymbol{X} = \{\xi\}$ is a random singleton and $F = \{x\}$, both HD_+ and HD_- equal the Tukey's half-space depth of x with respect to the distribution of ξ.

Example 5.1. Let $\boldsymbol{X}$ be the random ball B_β of radius β centred at the origin. For $F = B_r$, we have

$$\mathrm{HD}(B_r,\boldsymbol{X}) = \min(\Pr(\beta \geqslant r), \Pr(\beta \leqslant r)).$$

The deepest ball $F = B_r$ has the radius r, being the (assuming unique) median of β.

Example 5.2. Let $\boldsymbol{X} = [\boldsymbol{x}_{\mathrm{L}}, \boldsymbol{x}_{\mathrm{U}}]$ be a random interval on the line. Then

$$\mathrm{HD}([a,b],\boldsymbol{X}) = \min\left(\Pr(\boldsymbol{x}_{\mathrm{U}} \geqslant b), \Pr(\boldsymbol{x}_{\mathrm{U}} \leqslant b), \Pr(\boldsymbol{x}_{\mathrm{L}} \leqslant a), \Pr(\boldsymbol{x}_{\mathrm{L}} \geqslant a)\right).$$

While for random elements in Banach spaces (in particular, for stochastic processes) the half-space depth assigns zero values to most of reference points (functions), this is not the case for the above defined half-space depth. The reason is that, when compact convex sets are substituted by their support functions as in (16) and (17), the depth considered here becomes an infimal depth, see Gijbels & Nagy (2015), which typically does not degenerate.

Theorem 5.3. *If $\boldsymbol{X}$ is a random convex compact set, then $\mathrm{HD}_+(F,\boldsymbol{X}) > 0$ (respectively, $\mathrm{HD}_-(F,\boldsymbol{X}) > 0$) for all F from the support of $\boldsymbol{X}$ such that $\Pr(h(F,u) \leqslant h(\boldsymbol{X},u)) > 0$ (respectively, $\Pr(h(F,u) \geqslant h(\boldsymbol{X},u)) > 0$), for all u. Furthermore, the infima in (16) and (17) are attained.*

Proof. Assume that $\mathrm{HD}_+(F,\boldsymbol{X}) = 0$, that is, there exists a sequence $\{u_n, n \geqslant 1\}$ on the unit sphere such that

$$\Pr(h(\boldsymbol{X},u_n) \geqslant h(F,u_n)) \to 0, \quad \text{as } n \to \infty.$$

Without loss of generality assume that $u_n \to u$ as $n \to \infty$. For each $F \in \operatorname{co}\mathcal{K}$, its support function $h(F,u)$ is Lipschitz with the Lipschitz constant being the norm of F, see (Schneider 2014, Lemma 1.8.12). Hence,

$$\Pr(h(\boldsymbol{X},u)-\delta_n\|\boldsymbol{X}\|\geqslant h(F,u_n)+\delta_n\|F\|)\to 0,\quad \text{as } n\to\infty,$$

where $\delta_n=\|u_n-u\|$. Therefore, for all $c>0$,

$$\Pr(h(\boldsymbol{X},u)\geqslant h(F,u)+\delta_n(\|F\|+c),\|\boldsymbol{X}\|\leqslant c)\to 0,\quad \text{as } n\to\infty.$$

By letting c increase to infinity, we have that

$$\Pr(h(\boldsymbol{X},u)\geqslant h(F,u))=0,$$

contrary to the assumption. The statement for $\mathrm{HD}_-(F,\boldsymbol{X})$ has a similar proof.

The finite sample version of the half-space depth involves the empirical variants of HD_+ and HD_-, for example,

$$\mathrm{HD}_+(F,\{X_1,\ldots,X_n\})=\inf_{u\in\mathbb{S}^{d-1}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(h(X_i,u)\geqslant h(F,u)),$$

where $X_1,\ldots,X_n$ are sample values of $\boldsymbol{X}$.

Since the values of the support function at u and $-u$ determine the width of the set, it is possible to modify (16) (and (17) as well) by letting

$$\mathrm{HD}'_+(F,\boldsymbol{X})=\inf_{u\in\mathbb{S}^{d-1}}\Pr(h(\boldsymbol{X},u)\geqslant h(F,u),h(\boldsymbol{X},-u)\geqslant h(F,-u)).$$

A variant of the half-space depth for functional data is the *half-region depth* introduced by López-Pintado & Romo (2011) and further studied by Kuelbs & Zinn (2015). In case of random sets, it is defined by

$$\mathrm{HRD}(F,\boldsymbol{X})=\min\Big(\Pr(F\subseteq\boldsymbol{X}),\Pr(F\supseteq\boldsymbol{X})\Big). \tag{18}$$

Note that $\mathrm{HRD}(F,\boldsymbol{X})\leqslant\min\big(\mathrm{HD}_+(F,\boldsymbol{X}),\mathrm{HD}_-(F,\boldsymbol{X})\big)$. Like $\mathrm{HD}_-(F,\boldsymbol{X})$, this depth returns zero value if $\Pr(F\subseteq\boldsymbol{X})$ vanishes. This is the case for most F if $\boldsymbol{X}$ is 'thin', like singletons or other lower-dimensional sets in Euclidean space. To avoid this phenomenon, it is possible to enlarge $\boldsymbol{X}$ by taking its ε-envelope (which is the set of all points within distance ε to $\boldsymbol{X}$). This is akin to smoothing procedures in density estimation theory, and ε plays the role of a bandwidth. A similar 'smoothing' construction in the context of the functional band depth was discussed by Gijbels & Nagy (2015, Sec. 3.2).

The *modified half-region depth* is defined by

$$\begin{aligned}\mathrm{MHR}(F,\boldsymbol{X})=\min\Big(&\int_{\mathbb{S}^{d-1}}\Pr(h(F,u)\leqslant h(\boldsymbol{X},u))\mu(du),\\&\int_{\mathbb{S}^{d-1}}\Pr(h(F,u)\geqslant h(\boldsymbol{X},u))\mu(du)\Big),\end{aligned} \tag{19}$$

where μ is the normalised Haar measure on the unit sphere (i.e. the surface area measure). It is also possible to treat symmetrically the values of the support function at opposite directions by letting

$$\begin{aligned}\mathrm{MHR}'(F,\boldsymbol{X})=\min\Big(&\int_{\mathbb{S}^{d-1}}\Pr(h(F,u)\leqslant h(\boldsymbol{X},u),h(F,-u)\leqslant h(\boldsymbol{X},-u))\mu(du),\\&\int_{\mathbb{S}^{d-1}}\Pr(h(F,u)\geqslant h(\boldsymbol{X},u),h(F,-u)\geqslant h(\boldsymbol{X},-u))\mu(du)\Big).\end{aligned} \tag{20}$$

6. Band depth for sets and set-valued functions

6.1. Samples of convex sets

López-Pintado & Romo (2009) introduced the concept of a band depth for functional data. The *band* generated by j functions is defined as the family of functions with values lying between the pointwise minimum and pointwise maximum of these j functions. The band depth is the probability that a given function lies in the band generated by j independent copies of a random function. The band depth avoids the problem of having too many functions of depth zero.

Following López-Pintado & Romo (2009), the band generated by $F_1,\dots,F_j \in \operatorname{co}\mathcal{F}$ is the family of sets $F \in \operatorname{co}\mathcal{F}$ such that

$$\min_{1\leqslant i\leqslant j} h(F_i,u) \leqslant h(F,u) \leqslant \max_{1\leqslant i\leqslant j} h(F_i,u), \quad u\in\mathbb{S}^{d-1}.$$

While the right-hand side is a support function, namely, that of $\operatorname{conv}(F_1\cup\dots\cup F_j)$, the left-hand side is not necessarily a support function. Since López-Pintado & Romo (2009) did not specifically consider the case of convex functions, it is reasonable to adjust their definition of the band so that the lower bound becomes the largest convex function dominated by the pointwise minimum of $h(F_i,u)$, $i=1,\dots,j$. This function is called the subdifferential of this minimum, see Rockafellar (1970). This largest convex function is the support function of $F_1\cap\dots\cap F_j$.

With such an adjustment, the band generated by $F_1,\dots,F_j\in\operatorname{co}\mathcal{F}$ is defined as

$$\mathrm{B}^j(F_1,\dots,F_j)=\{F\in\operatorname{co}\mathcal{F}\colon \cap_{i=1}^j F_i\subseteq F\subseteq \operatorname{conv}\cup_{i=1}^j F_i\}.$$

This band is almost identical to the band proposed by Whitaker, Mirzargar & Kirby (2013), except for the fact that we consider here the case of convex sets, and so use the closed convex hull of the union of the generating sets.

The population version of the band depth is then

$$\mathrm{BD}^j(F,\boldsymbol{X})=\Pr\bigl(\cap_{i=1}^j \boldsymbol{X}_i\subseteq F\subseteq \operatorname{conv}\cup_{i=1}^j \boldsymbol{X}_i\bigr), \tag{21}$$

where $\boldsymbol{X}_1,\dots,\boldsymbol{X}_j$ are i.i.d. copies of $\boldsymbol{X}$. Property (D1) is easy to see, (D2) is a consequence of the fact that the convergence in the Fell topology preserves the inclusion relations, and (D3) is evident. For a sample $F_1,\dots,F_n\in\operatorname{co}\mathcal{F}$, the empirical version of the band depth is given by

$$\mathrm{BD}^j(F;F_1,\dots,F_n)=\binom{n}{j}^{-1}\sum_{1\leqslant k_1<\dots<k_j\leqslant n}\mathbf{1}\bigl(\cap_{i=1}^j F_{k_i}\subseteq F\subseteq\operatorname{conv}\cup_{i=1}^j F_{k_i}\bigr). \tag{22}$$

It describes the proportion of j-bands containing F. Since the empirical band depth is a symmetric statistics of $F_1,\dots,F_n$, it is possible to use limit theorems for U-statistics to obtain limit theorems for the empirical band depth.

A modified band depth for sets can be built from (21) and (22) adapting the construction of the (functional) modified band depth of López-Pintado & Romo (2009). The required adaptation has already been used here to build the modified half-region depth for sets, see (19), and its symmetrised version, see (20), from the half-region depth given in (18). Specifically, the modified band depth for sets is the average fraction of surface area of the unit sphere such that, when evaluated on any of its points, the support function of a set F

is sandwiched between the support functions of the intersection and union of j sets. In its symmetrised version, it is the average fraction of the surface area of the unit sphere such that, when evaluated on any of its points u, the support function of a set F is sandwiched between the support functions of the intersection and union of j sets, and the same is the case for $-u$.

In order to overcome consistency issues with the band depth of Whitaker, Mirzargar & Kirby (2013), Nagy (2017) presented an alternative modified band depth, which does not use support functions and is thus applicable for general (non-convex) compact sets.

If $j=1$, then $\mathrm{BD}^1(F,\boldsymbol{X})=\Pr(F=\boldsymbol{X})$ which vanishes for most F and so is not informative. This depth function becomes non-trivial for $j\geqslant 2$. Following López-Pintado & Romo (2009), it is advisable to combine the band depths built using a varying number j of functions by taking their averages as

$$\overline{\mathrm{BD}}^J(F,\boldsymbol{X})=\frac{1}{J-1}\sum_{j=2}^{J}\mathrm{BD}^j(F,\boldsymbol{X}), \tag{23}$$

where $2\leqslant J$. It is recommended to use $J=3$ for this type of band depth due to various reasons. By choosing larger J, the band depth increases, and so F with a very peculiar shape but rather normal magnitude comparing with other sets from the sample might attain a higher depth. This makes identification of F as a shape outlier rather difficult. In contrast, for $J=2$ too many sets F will be of depth zero.

Remark 6.1. The band depth can be constructed for distance functions $d(x,\boldsymbol{X})=\inf\{\|x-y\|: y\in\boldsymbol{X}\}$ generated by random closed sets. For this, one calculates the band depth of the function $d(x,F)$ with respect to functions $d(x,\boldsymbol{X}_i)$, $i=1,\dots,n$, obtained using i.i.d. copies $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ of $\boldsymbol{X}$. This construction makes it possible to introduce the depth for not necessarily convex random closed sets.

A refined variant of the band depth suggested by Cascos & Molchanov (2018) relies on considering tuples of values for the functions. Fix $m\geqslant 1$ and $F_1,\dots,F_j\in\operatorname{co}\mathcal{F}$. In terms of support functions, F belongs to the m-band if, for all $u_1,\dots,u_m$ from the unit sphere, the vector $(h(F,u_1),\dots,h(F,u_m))$ belongs to the convex hull in $\mathbb{R}^m$ of the vectors $(h(F_i,u_1),\dots,h(F_i,u_m))$, $i=1,\dots,j$, that is,

$$\min_{1\leqslant i\leqslant j}\sum_{k=1}^{m}h(F_i,u_k)v_k\leqslant\sum_{k=1}^{m}h(F,u_k)v_k\leqslant\max_{1\leqslant i\leqslant j}\sum_{k=1}^{m}h(F_i,u_k)v_k, \tag{24}$$

for every $v=(v_1,\dots,v_m)\in\mathbb{R}^m$. Note that the left inequality can be derived from the right one by altering the sign of v.

Imposing (24) only for every $v\in\mathbb{R}^m_+$ results in a larger band called the *positive* m-band. It admits a nice geometrical interpretation in terms of inclusions of intersections and unions of sets,

$$\mathrm{B}^j_{m,+}(F_1,\dots,F_j)=\left\{F\in\operatorname{co}\mathcal{F}:\cap_{i=1}^{j}F_i^{\times m}\subseteq F^{\times m}\subseteq\operatorname{conv}\cup_{i=1}^{j}F_i^{\times m}\right\},$$

where $F^{\times m}$ is the convex set in $\mathbb{R}^{dm}$ obtained as the Cartesian product of m factors, all being F. Indeed, $h(F^{\times m},(u_1,\dots,u_m))=\sum_{i=1}^{m}h(F,u_i)$ for every $(u_1,\dots,u_m)\in\mathbb{R}^{dm}$.

Example 6.2. Let $\boldsymbol{X}=[\boldsymbol{x}_{\mathrm{L}},\boldsymbol{x}_{\mathrm{U}}]$ be a random interval on $\mathbb{R}$, and let $[\boldsymbol{x}_{\mathrm{L}1},\boldsymbol{x}_{\mathrm{U}1}],\dots,[\boldsymbol{x}_{\mathrm{L}j},\boldsymbol{x}_{\mathrm{U}j}]$ be its j independent copies. Then the band depth of the interval $[a,b]$ (with $a\leqslant b$) becomes

$$\begin{aligned}\Pr(\min_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{L}i}\leqslant a,b\leqslant\max_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{U}i},\ \max_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{L}i}>\min_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{U}i})\\ +\Pr(\min_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{L}i}\leqslant a\leqslant\max_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{L}i}\leqslant\min_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{U}i}\leqslant b\leqslant\max_{1\leqslant i\leqslant j}\boldsymbol{x}_{\mathrm{U}i}).\end{aligned}$$

The two-band depth becomes

$$\Pr((a,b)\in\operatorname{conv}\{(\boldsymbol{x}_{\mathrm{L}1},\boldsymbol{x}_{\mathrm{U}1}),\dots,(\boldsymbol{x}_{\mathrm{L}j},\boldsymbol{x}_{\mathrm{U}j})\}).$$

The positive two-band depth becomes

$$\begin{aligned}&\Pr([\boldsymbol{x}_{\mathrm{L}k},\boldsymbol{x}_{\mathrm{U}k}]\subseteq[a,b]\subseteq[\boldsymbol{x}_{\mathrm{L}l},\boldsymbol{x}_{\mathrm{U}l}]\text{ for some }1\leqslant k,l\leqslant j\\ &\text{or }(a,b)\in\operatorname{conv}\{(\boldsymbol{x}_{\mathrm{L}1},\boldsymbol{x}_{\mathrm{U}1}),\dots,(\boldsymbol{x}_{\mathrm{L}j},\boldsymbol{x}_{\mathrm{U}j})\}).\end{aligned}$$

6.2. Samples of set-valued functions

The concept of a band can be naturally extended to handle samples of *set-valued functions*. Such functions also appear as sections of a convex set parameterised by one of the coordinates, see Example 9.1. Consider set-valued functions $F_i(t)$, for $i=1,\dots,j$, where the argument for simplicity is assumed to belong to $[0,1]$. The band formed by these functions is the family of set-valued functions F such that

$$\bigcap_{i=1}^{j}F_i(t)\subseteq F(t)\subseteq\operatorname{co}\bigcup_{i=1}^{j}F_i(t),\quad t\in[0,1]. \tag{25}$$

Note that the left-hand side may be void, and then becomes irrelevant for the corresponding t.

Figure 2(left) represents the interval-valued functions given by the monthly averages of daily minimum and maximum temperatures over the decade 2007–2016 in Boulder (Colorado). The shades of grey at each point reflect the percentage of the intervals where this observed temperature lies on a particular day. Figure 2(right) represents the interval band of these interval-valued functions, which consists of all interval-valued functions whose upper and lower end-points respectively lie in the upper and lower shaded regions.

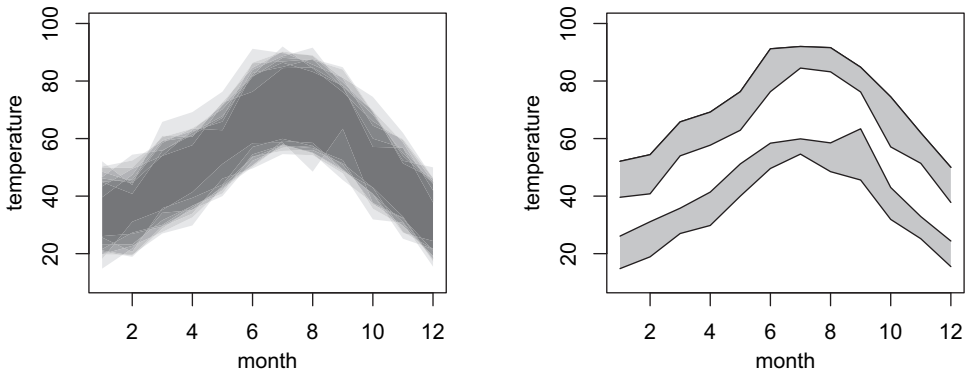

Figure 2. Temperature ranges in Boulder (Colorado) over 2007–2016 (left) and their band (right).

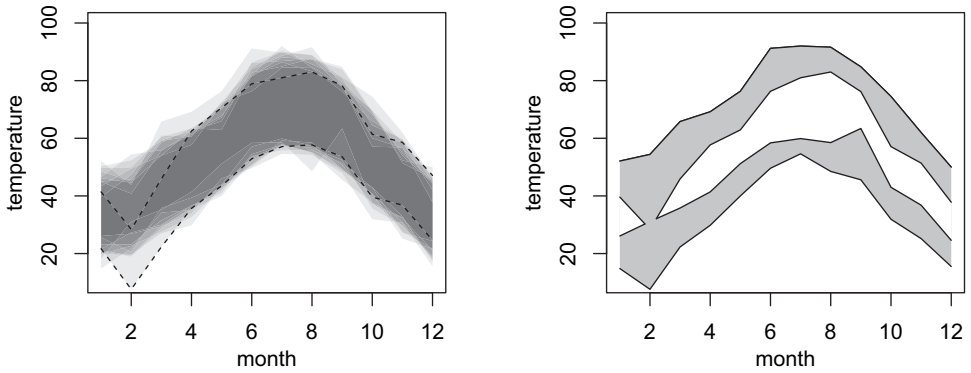

Figure 3. Temperature ranges in Boulder (Colorado) over 2007–2016 and 1899 (left) and their band (right).

Figure 3 has been produced similarly to Figure 2, but the temperatures of the year 1899 have been considered together with those from the decade 2007–2016. In order to highlight the interval-valued function associated with the 1899 temperatures, its end-points are represented as dashed lines in Figure 3(left). In particular, one notices the extremely low temperatures recorded in February 1899. Figure 3(right) represents the interval band of the 11 interval-valued functions, constituted again by all interval-valued functions whose upper and lower end-points respectively lie in the upper and lower shaded regions. Observe that in February there is no separation between the shaded regions. The reason is that the average of the daily maximum temperatures in February 1899 is below the average of the daily minimum temperatures in February during some other year over the decade 2007–2016, and thus the only restrictions imposed by the band in February appear in the form of an upper bound for the average maximum temperature and a lower one for the average minimum temperature.

7. Simplicial depths for sets

The simplicial depth for random vectors is defined as the probability that a point belongs to the convex hull of $(d+1)$ independent copies of this vector, see Liu (1990). Following this idea, the convex combination of sets $F_1,\dots,F_j$ is the family of sets obtained as

$$p_1F_1+\cdots+p_jF_j$$

for non-negative $p_1,\dots,p_j$ that sum to one. However, the family of such convex combinations is a finite-dimensional subset of $\operatorname{co}\mathcal{F}$; only few convex sets are representable as convex combinations of given convex sets. Because of this, a direct generalisation of the simplicial depth for random sets fails. Note that taking convex hull of the union of $F_1,\dots,F_j$ substantially differs from taking their convex combination.

Following the idea of type B depth functions from Zuo & Serfling (2000), it is possible to define the depth function of a convex compact set $\boldsymbol{X}$ by letting

$$\mathrm{D}(F,\boldsymbol{X})=\frac{\mathrm{E}\psi(\operatorname{conv}(F\cap(\boldsymbol{X}_1\cup\cdots\cup\boldsymbol{X}_j)))}{\mathrm{E}\psi(\operatorname{conv}(F\cup\boldsymbol{X}_1\cup\cdots\cup\boldsymbol{X}_j))}, \tag{26}$$

where ψ is a monotonic functional on $\operatorname{co}\mathcal{K}$ which does not vanish on non-empty convex sets, and $F \in \operatorname{co}\mathcal{K}$. In order to ensure (D1), it is possible to assume that $\psi(AF) = g(A)\psi(F)$, where $g(A)$ is a function of the matrix A. For (D2), we impose that ψ is continuous with respect to the convergence of convex compact sets in the Hausdorff metric. Property (D3) holds if ψ is strictly monotone on $\operatorname{co}\mathcal{K}$. For instance, it is possible to let ψ be the Lebesgue measure on $\mathbb{R}^d$; this yields a generalisation of the simplicial volume depth, see Zuo & Serfling (2000, Example 2.2).

A similar construction was recently suggested by Staerman, Mozharovskyi & Clémençon (2020) in view of assessing the depth for a sample of curves. An empirical variant of this depth function is defined by replacing the expectations with U-statistics constructed by replacing $\boldsymbol{X}_1 \cup \cdots \cup \boldsymbol{X}_j$ with $X_{i_1} \cup \cdots \cup X_{i_j}$, where $X_{i_1}, \ldots, X_{i_j}$ are sampled from the realisations $X_1, \ldots, X_n$ of $\boldsymbol{X}$.

Example 7.1. Let $\boldsymbol{X} = [\xi, \xi + 1] \subseteq \mathbb{R}$, where ξ is a uniform random variable in the unit interval. Let ψ be the Lebesgue measure. Then $\mathrm{D}(\{1\}, \boldsymbol{X}) = 0$, since the singleton $\{1\}$ has Lebesgue measure 0, while for $j = 1$, the interval $[1/2, 3/2]$ has depth $3/5$.

8. Splitting location and shape effects

A rather specific feature of samples of sets relates to the rôle of positions, sizes and shapes of sets in the statistical procedures. This is a well-known issue in the statistical theory of shape, see Dryden & Mardia (1997). For samples of sets representing particles, like stones or sand grains, see, for example, Stoyan & Stoyan (1994) and Stoyan & Molchanov (1997), the locations and orientations of particles are irrelevant for their statistical analysis. On the other hand, positions of convex sets arising from econometric applications are usually very relevant for the statistical analysis. In the context of outlier detection in functional data, Febrero, Galeano & González-Manteiga (2008) pointed out that both of location and shape are relevant, since most outliers are curves that are either significantly far from the average of the process or have a different shape than the rest of curves.

In order to leverage the effects of the location and shape in the detection of set-outliers, we adapt the principle that a point is a multivariate outlier if it is a univariate outlier for at least one of its projections. Namely, we define the location-scale depth as the minimum of a (multivariate) location depth (for points chosen from the sets) and a set depth (for the sample of translated sets).

Let $\boldsymbol{X}$ be a random convex closed set. The *location-shape depth* of $F \in \operatorname{co}\mathcal{F}$ is defined by letting

$$\mathrm{D}_{\mathrm{ls}}(F, \boldsymbol{X}) = \sup_{x \in F, \xi \in L^0(\boldsymbol{X})} \min(\mathrm{D}(x, \{\xi\}), \mathrm{D}(F - x, \boldsymbol{X} - \xi)). \tag{27}$$

This concept of depth relies on the choice of a point x in F and a selection ξ of $\boldsymbol{X}$ that ensures that x is deep with respect to the distribution of ξ and the 'centred' F is deep with respect to the 'centred' $\boldsymbol{X}$. Note that $\mathrm{D}(x, \{\xi\})$ can be defined by specialising the underlying depth function for sets, being singletons. Alternatively, any standard multivariate depth can be chosen for this purpose.

Example 8.1. For convex sets on the line, the shape refers to the width of intervals. Consider the random interval $\boldsymbol{X} = [0, \xi]$ with ξ following an exponential distribution of rate $\lambda = 1$.

Consider the average quantile depth from Section 4.4, which becomes the zonoid depth if applied to singletons. Clearly, the depth of $F=[-1,1]$ with respect to $[0,\xi]$ is 0, since $\boldsymbol{X}$ does not contain negative numbers and so the set $[-1,1]$ does not belong to the convex hull of the support of $\boldsymbol{X}$. Nevertheless, we can take $x=1$ as a reference point from $[-1,1]$ and ξ as selection of $\boldsymbol{X}=[0,\xi]$. Since the mean of ξ is 1, the zonoid depth of $x=1$ with respect to ξ is one, and $F-x$ becomes $[-2,0]$, while $\boldsymbol{X}-\xi=[-\xi,0]$. Finally, we have $\mathrm{D}_{\mathrm{ls}}(F,\boldsymbol{X})\geqslant \mathrm{D}([-2,0],[-\xi,0])=e^{-1}$.

Example 8.2. Consider now random sets in the plane. Let $C=[0,1]^2$ be the unit square and let η and ζ be two independent uniform random variables in the unit interval. Define the random set $\boldsymbol{X}=(\eta,\eta)+\zeta C$ and consider the deterministic set $F=-(0.25,0.25)+0.5C$, being the square with two opposite corners at $(-0.25,-0.25)$ and $(0.25,0.25)$. Clearly, F does not lie in the convex hull of the support of $\boldsymbol{X}$, so its depth is 0. Take now $\xi=(\eta+\zeta,\eta+\zeta)$ which is a selection of $\boldsymbol{X}$ (it is actually its upper right corner) and $x=(0.25,0.25)\in F$ (also its upper right corner). Now $\mathrm{D}(x,\xi)>0$, in fact, for the zonoid depth it equals $9/128$, and $F-x=-0.5C$, $\boldsymbol{X}-\xi=-\zeta C$, so $\mathrm{D}(F-x,\boldsymbol{X}-\xi)=\mathrm{D}(0.5C,\zeta C)$. For the average quantile depth of Section 4.4, we conclude that $\mathrm{D}(0.5C,\zeta C)=0.5$ calculated as the zonoid depth of 0.5 with respect to ζ. Hence $\mathrm{D}_{\mathrm{ls}}(F,\boldsymbol{X})\geqslant 9/128$.

An advantage of the definition (27) is that the translated random set $\boldsymbol{X}-\xi$ contains the origin. In this case, another description of sets using functions is available. Assume that $\boldsymbol{X}$ almost surely contains the origin. Since $\boldsymbol{X}$ is convex, it is also *star-shaped* and so is identified by its *radial function*

$$r(\boldsymbol{X},u)=\sup\{t:\, tu\in\boldsymbol{X}\},\quad u\neq 0.$$

The *radial sum* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the set whose radial function equals the sum of the radial functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this case, it is useful to consider radially superlinear expectation, which satisfies (3) with the radial sum replacing the Minkowski sum on the left-hand side. The dual pair of non-linear expectations is then given by

$$h(\mathcal{E}(\boldsymbol{X}),u)=\mathfrak{e}(h(\boldsymbol{X},u)),\quad r(\mathcal{U}(\boldsymbol{X}),u)=\mathfrak{u}(r(\boldsymbol{X},u))$$

for a dual pair $(\mathfrak{u},\mathfrak{e})$ of numerical non-linear expectations. For the latter, one usually chooses the exact dual pair given by $\mathfrak{u}(\beta)=-\mathfrak{e}(-\beta)$ for $\beta\in L^p(\mathbb{R})$. Note that the sublinear expectation is applied to the support function, while the superlinear one to the radial function of $\boldsymbol{X}$. The so defined set $\mathcal{U}(\boldsymbol{X})$ is star-shaped, but not necessarily convex.

Example 8.3. Let $\mathfrak{u}(\beta)=\mathfrak{e}(\beta)=\mathrm{E}(\beta)$. Following the approach of Section 4.3, the depth of F with respect to $\boldsymbol{X}$ that almost surely contain the origin is m^{-1} for the smallest m such that

$$h(F,u)\leqslant \mathrm{E}\,\max(h(\boldsymbol{X}_1,u),\ldots,h(\boldsymbol{X}_m,u))$$

and

$$r(F,u)\geqslant \mathrm{E}\,\min(r(\boldsymbol{X}_1,u),\ldots,r(\boldsymbol{X}_m,u))$$

for all u from the unit sphere.

Example 8.4. By using the approach based on average quantiles, the depth of F with respect to $\boldsymbol{X}$ that almost surely contain the origin is the largest $\alpha\in(0,1]$ such that $h(F,u)\leqslant \mathfrak{e}_\alpha(h(\boldsymbol{X},u))$

and $r(F,u) \geqslant \mathtt{u}_\alpha(r(\boldsymbol{X},u))$ for all u from the unit sphere, where $\mathtt{e}_\alpha$ and $\mathtt{u}_\alpha$ are defined in (12) and (13).

9. Depth of random measures

An integrable probability measure μ on $\mathbb{R}^d$ is uniquely identified by a convex set $\hat{Z}_\mu$ in $\mathbb{R}^{d+1}$ called the *lift zonoid* of μ, see Koshevoy & Mosler (1998) and Mosler (2002). The set $\hat{Z}_\mu$ is the expectation of the random segment in $\mathbb{R}^{d+1}$ with one end-point at the origin and the other at $(1,\xi)$, where ξ is distributed according to μ.

If μ is a *random* integrable probability measure, then the expectation of the segment should be conditional upon μ, and $\hat{Z}_\mu$ becomes a random convex closed set in $\mathbb{R}^{d+1}$. Then all concepts of depth for random sets are applicable in this setting in order to identify outliers in a sample of probability measures.

Example 9.1. We use here notions of depth for sets and curves to compare (empirical) distributions. Consider a parametric distribution model, and several samples taken from it, each of them corresponding to a different value of the parameter. Among the values of the parameter, there are some outliers, and the goal is to detect them. In order to do so, we will study the empirical cumulative distribution functions together with a notion of functional depth and the empirical lift zonoids together with a notion of depth for sets.

We have simulated 30 samples of 100 observations, each from the Beta distribution $B(p,q)$ with parameters (p,q) given by a sample of random points uniformly distributed on the square $[0.5,1.5]^2$, see the 30 bullets in Figure 4a. Then we simulated four further samples of Beta distributions with parameters equal to those identified with asterisks in Figure 4a. Observe that the four asterisks enclose the 30 bullets in their convex hull. In order to detect possible outliers in our set of samples, we considered two alternative procedures. First, we built the empirical cdfs for the samples grouped in five bins of equal width, see Figure 4b. The second approach was to approximate the lift zonoid of each of the samples by grouping each data set in five bins of equal size ($100/5 = 20$ observations each). In Figure 4c, the lift zonoids of all 34 samples are represented (each sample was shifted 0.5 units to the left for better visualisation of the lift zonoids).

The detection of outliers in the set of empirical cdfs was performed by computing the band depth for $j = 3$ of each cdf with respect to the sample of curves. In the case of the lift zonoids, for each of them, the set-valued function $F(t)$ that represents the projection on the last d coordinates of the intersection of the lift zonoid $\hat{Z}_\mu$ with the hyperplane whose first coordinate is t was built. Finally, the bands for $j = 3$ were built as in (25), see Figure 4d for an example of a band (generated by three arbitrary lift zonoids, none of which corresponds to an outlier in the set of parameters), and the band depth was computed.

For the particular simulated data set used in Figure 4, the band depths of the lift zonoids achieved the identification of the four outliers which are the only four lift zonoids attaining the minimum depth, while the band depth of the empirical cdfs achieved the identification of three out of the four outliers which together with other four empirical cdfs attained the minimum depth. They failed to identify the sample from the Beta distribution with shape parameters $p = 0.8$ and $q = 0.4$. Showing these results we do not mean that one procedure is better than the other, but instead present an application of depth notions for curves and sets that describe distributions (cdfs and lift zonoids).

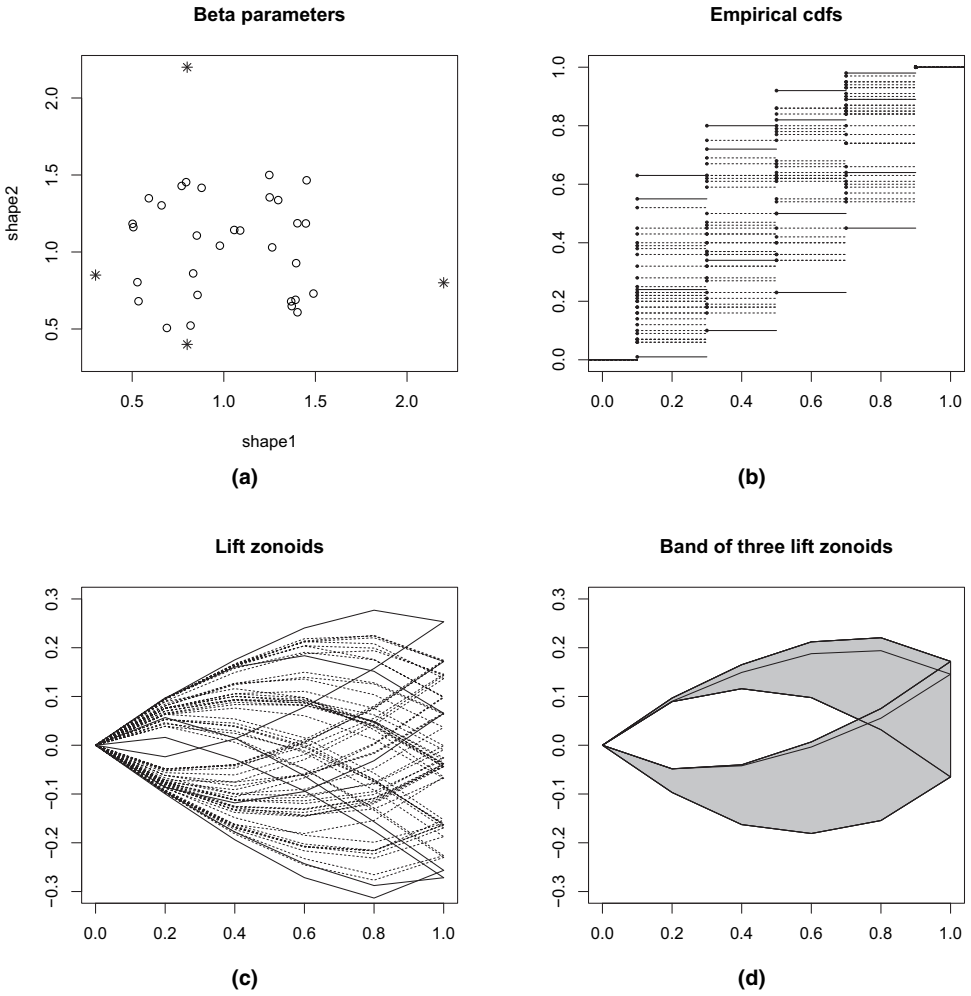

Figure 4. A total of 34 samples of Beta $B(p,q)$ distributions were simulated. The scatter plot in (a) represents the 34 pairs of parameters (p,q) among which there are 4 outliers marked as asterisks. All 34 data sets are summarised by their empirical cdfs in (b) and their lift zonoids in (c). The solid lines in (b) and (c) represent the samples drawn from the outlying parameters. The shaded region in (d) corresponds to the band generated by three arbitrary lift zonoids.

Both band depths for $j=3$ were computed using *brute force* algorithms, with complexity $O(kn^3)$, where n is the sample size (in this case 34 curves), and k the number of points at which each scalar or interval-valued function is evaluated (in this case 5).

10. Examples

10.1. Interval regression

Consider interval regression setting with data given by $(x_i,[y_{\mathrm{L}i},y_{\mathrm{U}i}])$, $i=1,\ldots,n$, so that the response values are given by intervals. Such data appear in the econometric analysis of wages, which are often reported as intervals, see Beresteanu & Molinari (2008) and

Molchanov & Molinari (2018). In this case, the convex set formed by all intercepts and slopes compatible with the model is obtained applying ordinary least squares to all possible samples (x_i, y_i), where $y_i \in [y_{\mathrm{L}i}, y_{\mathrm{U}i}]$, $i = 1,\ldots,n$, see Beresteanu & Molinari (2008). Such a set of intercepts and slopes is a zonotope, which is a convex set given by the Minkowski sum of a number of line segments, see Schneider (2014, Sec. 3.5).

As in the simple linear regression setting, we build the $n \times 2$ design matrix X whose first column is filled with 1s, while the second column contains the observed values of the explanatory variable x. The zonotope that contains all possible intercepts and slopes is

$$\left\{(X^\top X)^{-1}X^\top y : y = (y_1,\ldots,y_n)^\top \text{ with } y_i \in [y_{\mathrm{L}i}, y_{\mathrm{U}i}],\ i = 1,\ldots,n\right\}.$$

The set of all possible adjusted response values corresponds to the multiplication of the design matrix X by all elements of the set of intercepts and slopes, and finally the residual errors are obtained as

$$\left\{(I_n - X(X^\top X)^{-1}X^\top)y : y = (y_1,\ldots,y_n)^\top \text{ with } y_i \in [y_{\mathrm{L}i}, y_{\mathrm{U}i}],\ i = 1,\ldots,n\right\},$$

where I_n is the $n \times n$ identity matrix. Observe that, while the set of intercepts and slopes lies in the plane, these last two sets lie in $\mathbb{R}^n$, each coordinate corresponding to one of the observations. Nevertheless, for each observation $(x_i, [y_{\mathrm{L}i}, y_{\mathrm{U}i}])$ we are interested in its associated residual, and that corresponds to the projection of the set of residual errors on the ith coordinate, which is an interval.

With the interval residuals, we can study the presence of outliers in the original data set by computing the depth of each residual with respect to the sample of all of them. In the subsequent example we will also show how to determine which observations are more influential by deleting each individual observation from the data set and comparing the zonotope of intercepts and slopes obtained for each of such subsamples with that of the original sample.

Example 10.1. Kallithrakaa *et al.* (2001) presented data on 33 Greek wines for classification purposes. They consider several interval-valued instrumental variables together with other point-valued variables obtained from sensory analysis. In Figure 5a, we present in x-axis the astringency evaluation of each of the wines, while the y-axis corresponds to the caffeic acid (interval-valued). The number to the right of each segment is its wine code, while the dashed straight line is the ordinary least squares regression line that predicts the mid-point of the caffeic acid interval with information from the astringency.

Figure 5b shows the interval residual for each wine versus the wine code. The dashed vertical lines separate each group of five wines in order to facilitate the identification of the wine codes. Finally, the number to the right of each segment represents the ranking of the residual with respect to the 1-band depth for $j = 2$ described in detail in Example 6.2. The outermost intervals are associated with the smallest values. They correspond to wine codes 12 and 16, while wines 10 and 28 come next among the less deep wines. The complexity of the algorithm to compute the 1-band depth for intervals when $j = 2$ is $O(n^2)$ for a sample of n intervals.

Once the outliers have been detected, we conclude our regression analysis studying the influence of each observation in the set of intercepts and slopes. In Figure 5c, the region in grey is the set of all intercepts and slopes obtained for the complete data set.

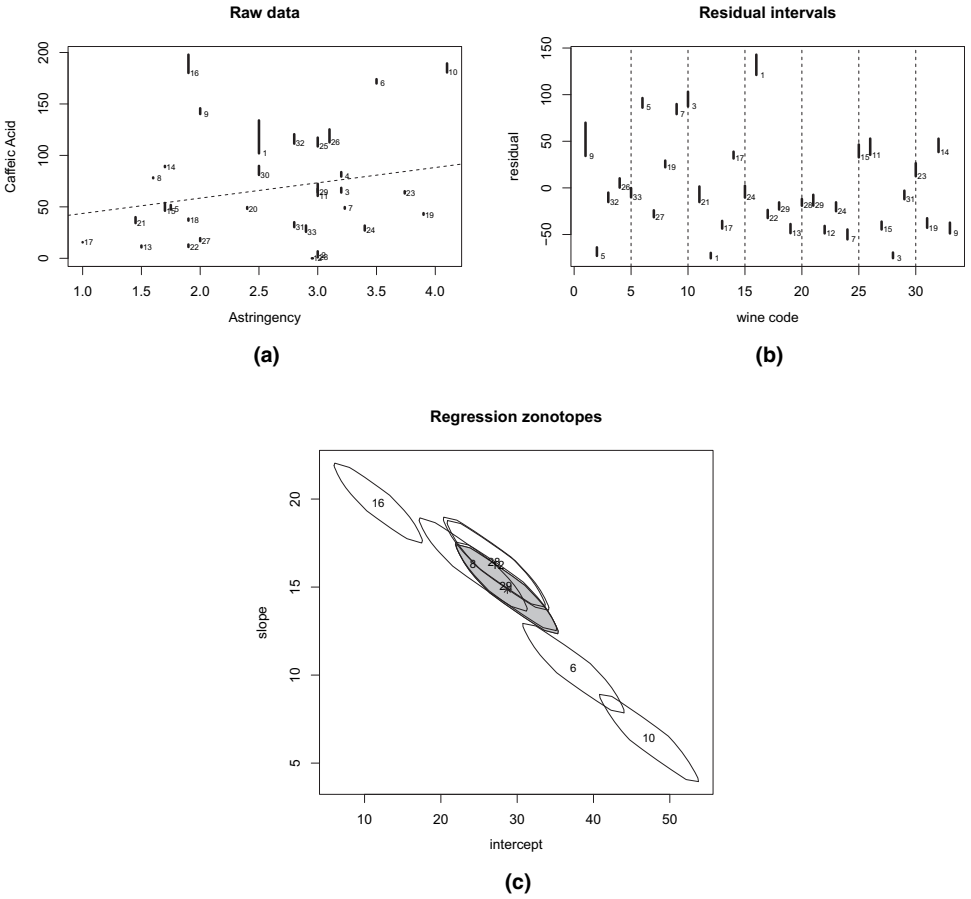

Figure 5. Regression analysis for explanatory variable Astringency (point-valued) and response variable Caffeic Acid (interval-valued) of 33 Greek wines. Raw data are presented in (a), interval-valued residuals in (b) and the set of all intercepts and slopes for the complete data set as well as after deleting some individual wines from the data set in (c).

The point marked on it corresponds to the intercept and slope of the regression line that predicts the mid-point of the caffeic acid interval with information from Astringency. The remaining seven sets have been obtained after deleting a single observation from the data set. The code of the deleted wine is represented by a number in the centre of the region. Specifically, the sets presented here correspond to wine codes 6, 8, 10, 12, 16, 28, and 29. Wines 10, 16 and 6 have been selected because they are the most influential ones in our regression model, while the other four (including wine 12 which was marked as an outlier) have been selected to show the standard behaviour of most wines from this data set.

In conclusion, we should be particularly concerned with wines 16 and 10. Wine 16 is an influential outlier, while wine 10 is a not-so-obvious outlier which strongly influences the regression model.

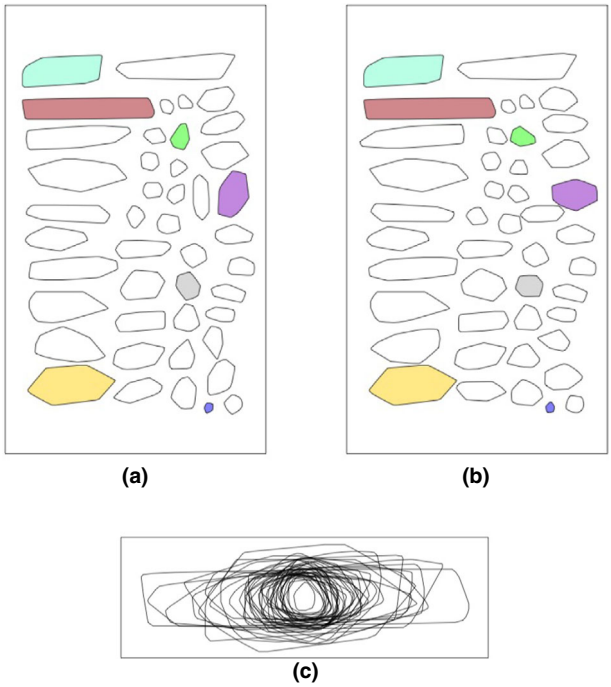

Figure 6. Sample of particles (a) and their optimal rotations (b). Shaded particles from top to bottom have numbers 1, 2, 15, 31, 35, 10, 43. The contours of the centred superimposed particles from (b) are shown in (c) (enlarged).

10.2. Sample of particles

Consider the sample of 44 planar images of particles taken from the collection of sand particles analysed in Stoyan & Molchanov (1997), see Figure 6a. Each particle was placed so that its centre of mass is located at the origin. Then all particles have been rotated following the iterative algorithm described in Stoyan & Molchanov (1997): it aims to minimise the L^2-distance between the indicator function of a particle and the average of indicator functions of all other particles. The L^2-distance between two indicator functions was evaluated over 250,000 grid points uniformly located in the smallest square, which covers all centred particles. The optimal rotation of each particle was searched over a sequence of angles with common difference 0.1π in $[0, 2\pi)$. The rotated particles are shown in Figure 6b.

The half-region depth and the band depth have been computed by explicit evaluation of the intersections of convex hulls of the particles as binary images. All other depths have been evaluated using the support and radial functions of the particles as functions on $[0, 2\pi]$ discretised using a mesh of 60 equidistant points. The average quantile depth was computed as described in Example 8.4 using the parametric family of sublinear expectations constructed with average quantiles. For all depths involving support functions, such as half-space, modified half-region, half-region symmetrised and average quantile depth, the same sequence of angles as in searching for the optimal rotation was used. To check the inclusions of the intersections in the half-space and band depths, we used the function `gCovers` from

Table 1. Depths of 7 marked particles.

Particle number	1	2	10	15	31	35	43
Half-space, (15)	0.000	0.000	0.000	0.093	0.093	0.256	0.000
Half-region, (18)	0.000	0.000	0.000	0.093	0.047	0.186	0.000
Modified half-region, (19)	0.126	0.025	0.106	0.282	0.296	0.414	0.000
Half-region symmetrised, (20)	0.110	0.020	0.094	0.234	0.276	0.378	0.000
Band (22), $j=2$	0.000	0.000	0.000	0.062	0.055	0.097	0.000
Band (23), $J=3$	0.000	0.000	0.000	0.112	0.089	0.189	0.000
Average quantile, Example 8.4	0.000	0.000	0.000	0.305	0.165	0.595	0.000

R package rgeos, see Bivand & Rundel (2020). See Table 1 for the values of this 7 notions of depth computed for the 7 marked particles.

The symmetrised version of the modified half-region depth returns similar results to its non-symmetric version, since many particles are rather close to being centrally symmetric. All suggested depth concepts identify particles 2 and 43 as outliers. For other particles, the half-region depth may be regarded as far too sensitive, while the average quantile depth, band depth and modified half-space depth show the best performance comparing to visual perception of outliers.

11. Acknowledgements

The authors are honoured to contribute this work to the special issue devoted to Professor Adrian Baddeley and hope that it combines the theory, statistical methods and their implementation in a way that has been promoted and appreciated by him.

The authors are grateful to the two referees for careful reading of this work, correcting mistakes, encouragements and suggesting references to a number of related papers.

References

BADDELEY, A.J. (1992). Errors in binary images and an L^p version of the Hausdorff metric. *Nieuw Archief voor Wiskunde* **10**, 157–183.

BERESTEANU, A. & MOLINARI, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica* **76**, 763–814.

BIGOT, J., CAZELLES, E. & PAPADAKIS, N. (2019). Penalization of barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* **51**, 2261–2285.

BIVAND, R. & RUNDEL, C. (2020). *rgeos: Interface to Geometry Engine - Open Source ('GEOS')*. Available from URL: https://CRAN.R-project.org/package=rgeos. R package version 0.5-5.

BLANCO-FERNÁNDEZ, A., COLUBI, A. & GONZÁLEZ-RODRÍGUEZ, G. (2012). Confidence sets in a linear regression model for interval data. *Journal of Statistical Planning and Inference* **142**, 1320–1329.

BLANCO-FERNÁNDEZ, A., CORRAL, N. & GONZÁLEZ-RODRÍGUEZ, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis* **55**, 2568–2578.

CASCOS, I. (2007). The expected hull trimmed regions of a sample. *Computational Statistics* **22**, 557–569.

CASCOS, I. (2010). Data depth: multivariate statistics and geometry. In *New Perspectives in Stochastic Geometry*, eds. W.S. Kendall and I. Molchanov, pp. 398–426. Oxford: Oxford University Press.

CASCOS, I. & MOLCHANOV, I. (2007). Multivariate risks and depth-trimmed regions. *Finance and Stochastics* **11**, 373–397.

CASCOS, I. & MOLCHANOV, I. (2018). Band depths based on multiple time instances. In *The Mathematics of the Uncertain. A tribute to Pedro Gil*, eds. E. Gil, E. Gil, J. Gil and M.A. Gil, pp. 67–78. Cham: Springer.

CAZELLES, E., SEGUY, V., BIGOT, J., CUTURI, M. & PAPADAKIS, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing* **40**, B429–B456.

CHEN, M., GAO, C. & REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Annals of Statistics* **46**, 1932–1960.

CHIU, S., STOYAN, D., KENDALL, W. & MECKE, J. (2013). *Stochastic Geometry and its Applications*, 3rd edn. London: Wiley.

DANG, X. & SERFLING, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference* **140**, 198–213.

DELBAEN, F. (2002). Coherent risk measures on general probability spaces. In *Advances in Finance and Stochastics*, eds. K. Sandmann and P.J. Schönbucher, pp. 1–37. Berlin: Springer.

DIAYE, M.A., KOSHEVOY, G.A. & MOLCHANOV, I. (2018). Lift expectations of random sets and their applications. *Statistics & Probability Letters* **145**, 110–117.

DRYDEN, I.L. & MARDIA, K.V. (1997). *Statistical Shape Analysis*. Chichester: Wiley.

DUTTA, S., GHOSH, A.K. & CHAUDHURI, P. (2011). Some intriguing properties of Tukey's half-space depth. *Bernoulli* **17**, 1420–1434.

FEBRERO, M., GALEANO, P. & GONZÁLEZ-MANTEIGA, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. *Environmetrics* **19**, 331–345.

FÖLLMER, H. & SCHIED, A. (2004). *Stochastic Finance: an Introduction in Discrete Time*, 2nd edn. Berlin: Walter de Gruyter & Co.

FRAIMAN, R. & MUNIZ, G. (2001). Trimmed means for functional data. *Test* **10**, 419–440.

FRESEN, D.J. & VITALE, R.A. (2014). Concentration of random polytopes around the expected convex hull. *Electronic Communications in Probability* **19**, 8.

GIJBELS, I. & NAGY, S. (2015). Consistency of non-integrated depths for functional data. *Journal of Multivariate Analysis* **140**, 259–282.

GIJBELS, I. & NAGY, S. (2017). On a general definition of depth for functional data. *Statistical Science* **32**, 630–639.

HAMEL, A.H. & KOSTNER, D. (2018). Cone distribution functions and quantiles for multivariate random variables. *Journal of Multivariate Analysis* **167**, 97–113.

KALLITHRAKAA, S., ARVANITOYANNIS, I., KEFALASA, P., EL-ZAJOULIA, A., SOUFLEROS, E. & PSARRA, E. (2001). Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin. *Food Chemistry* **73**, 501–514.

KENDALL, D.G., BARDEN, D., CARNE, T.K. & LE, H. (1999). *Shape and Shape Theory*. Chichester: John Wiley & Sons.

KOSHEVOY, G. & MOSLER, K. (1997). Zonoid trimming for multivariate distributions. *Annals of Statistics* **25**, 1998–2017.

KOSHEVOY, G.A. & MOSLER, K. (1998). Lift zonoids, random convex hulls and the variability of random vectors. *Bernoulli* **4**, 377–399.

KUELBS, J. & ZINN, J. (2013). Concerns with functional depth. *ALEA - Latin American Journal of Probability and Mathematical Statistics* **10**, 831–855.

KUELBS, J. & ZINN, J. (2015). Half-region depth for stochastic processes. *Journal of Multivariate Analysis* **142**, 86–105.

LAFAYE DE MICHEAUX, P., MOZHAROVSKYI, P. & VIMOND, M. (2020). Depth for curve data and applications. *Journal of the American Statistical Association*, to appear.

LIU, R.Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics* **18**, 405–414.

LIU, R.Y., PARELIUS, J. & SINGH, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics* **27**, 783–858.

LÓPEZ-PINTADO, S. & ROMO, J. (2009). On the concept of depth for functional data. *Journal of the American Statistics Association* **104**, 718–734.

LÓPEZ-PINTADO, S. & ROMO, J. (2011). A half-region depth for functional data. *Computational Statistics & Data Analysis* **55**, 1679–1695.

MANSKI, C.F. & TAMER, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica* **70**, 519–546.

MIZERA, I. (2002). On depth and deep points: a calculus. *The Annals of Statistics* **30**, 1681–1736.

MOLCHANOV, I. (2017). *Theory of Random Sets*, 2nd edn. London: Springer.

MOLCHANOV, I. & MOLINARI, F. (2018). *Random Sets in Econometrics*. Cambridge: Cambridge University Press.

MOLCHANOV, I. & MÜHLEMANN, A. (2021). Nonlinear expectations of random sets. *Finance and Stochastics* **25**, 5–41.

MOSLER, K. (2002). *Multivariate Dispersion, Central Regions and Depth. The Lift Zonoid Approach, Lecture Notes in Statistics*, vol. 165. Berlin: Springer.

MOSLER, K. (2013). Depth statistics. In *Robustness and Complex Data Structures*, pp. 17–34. Heidelberg: Springer.

NAGY, S. (2017). Integrated depth for measurable functions and sets. *Statistics & Probability Letters* **123**, 165–170.

NAGY, S., SCHÜTT, C. & WERNER, E. (2019). Halfspace depth and floating body. *Statistics Surveys* **13**, 52–118.

PAINDAVEINE, D. & VAN BEVER, G. (2018). Halfspace depths for scatter, concentration and shape matrices. *The Annals of Statistics* **46**, 3276–3307.

PANDOLFO, G., PAINDAVEINE, D. & PORZIO, G.C. (2018). Distance-based depths for directional data. *Canadian Journal of Statistics* **46**, 593–609. doi:10.1002/cjs.11479.

PENG, S. (2004). Nonlinear expectations, nonlinear evaluations and risk measures. In *Stochastic Methods in Finance, Lecture Notes in Mathematics*, vol. 1856, pp. 165–253. Berlin: Springer.

PENG, S. (2019). *Nonlinear Expectations and Stochastic Calculus under Uncertainty*. Berlin: Springer.

ROCKAFELLAR, R.T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

ROUSSEEUW, P. & RUTS, I. (1999). The depth function of a population distribution. *Metrika* **49**, 213–244.

SCHNEIDER, R. (2014). *Convex Bodies. The Brunn–Minkowski Theory*, 2nd edn. Cambridge: Cambridge University Press.

STAERMAN, G., MOZHAROVSKYI, P. & CLÉMENÇON, S. (2020). The area of the convex hull of sampled curves: a robust functional statistical depth measure. *Proceedings of Machine Learning Research* **108**, 570–579.

STOYAN, D. & MOLCHANOV, I.S. (1997). Set-valued means of random particles. *Journal of Mathematical Imaging and Vision* **7**, 111–121.

STOYAN, D. & STOYAN, H. (1994). *Fractals, Random Shapes and Point Fields*. Chichester: Wiley.

TUKEY, J. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974)*, Vol. 2, pp. 523–531.

WHITAKER, R.T., MIRZARGAR, M. & KIRBY, R.M. (2013). Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics* **19**, 2713–2722.

YANG, W., HAN, A., HONG, Y. & WANG, S. (2016). Analysis of crisis impact on crude oil prices: a new approach with interval time series modelling. *Quantitative Finance* **16**, 1917–1928.

ZEMEL, Y. & PANARETOS, V.M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25**, 932–976.

ZUO, Y. & SERFLING, R. (2000). General notions of statistical depth function. *The Annals of Statistics* **28**, 461–482.

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Li Qiyu

Registration Number: 05-123-294

Study program: PhD in Statistics

Bachelor ☐ Master ☐ Dissertation ☑

Title of the thesis: Set-valued Data: Regression, Design and Outliers

Supervisor: Prof. Dr. Ilya Molchanov

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, 28. January 2021

Place/Date

Signature