Local Regression Smoothers with Set-Valued Outcome Data

Qiyu Li[*], Ilya Molchanov[†], Francesca Molinari[‡], Sida Peng[§]

September 17, 2018

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.

Keywords: Local regression smoothers; set valued outcome data; random sets; support function.

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

[*] Department of Mathematical Statistics and Actuarial Science, University of Bern, and Swiss Group for Clinical Cancer Research (SAKK).

[†] Department of Mathematical Statistics and Actuarial Science, University of Bern.

[‡] Department of Economics, Cornell University. Financial support through NSF grant SES-1824375 is gratefully acknowledged.

[§] Microsoft Research.

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. Juster and Suzman (1995)). 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., Manski (2003)) has addressed the question of what can be learned about functionals of probability distributions of interest, 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 when the outcome data is set-valued and the regressors are exactly measured. Specifically, the paper relaxes the textbook setting (e.g., Tsybakov (2009)) 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 $(\boldsymbol{x}, \boldsymbol{Y})$ composed of a random vector $\boldsymbol{x}$ in $\mathbb{R}^m$ and a random convex compact set $\boldsymbol{Y}$ in $\mathbb{R}^d$. Here $\boldsymbol{Y}$ is assumed to be measurable in a sense made precise in Section 2. 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 $\boldsymbol{y} \in \boldsymbol{Y}$. 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. Additionally, we let $M(x) = \{m(x) : m \in \mathcal{M}\}$ 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 Molchanov (2017, Sec. 2.1.6) and Section 2 below. 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 Fan (1993) and Fan and Gijbels (1996), and is closely related to Fan and Gijbels (1992) and Tsybakov (2009).

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

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

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 Beresteanu and Molinari (2008) 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 $\boldsymbol{Y}$ 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 tumor time to 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. Manski and Tamer (2002) 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. Beresteanu and Molinari (2008) 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. Bontemps et al. (2012) extend the familiar Sargan test for overidentifying restrictions to the setting studied by Beresteanu and Molinari (2008). Chandrasekhar et al. (2012) extend Beresteanu and

Molinari (2008)'s approach to cover best linear approximation of any function $f(x)$ that is known to lie within two identified bounding functions. Kaido (2017) 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. Adusumilli and Otsu (2017) propose empirical likelihood methods for random sets to conduct inference in the class of problems analyzed by Beresteanu and Molinari (2008). 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.

The method that we propose differs significantly from other approaches in the statistical literature; see Schollmeyer and Augustin (2015) 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. Diamond (1990), Gil et al. (2001), González-Rodríguez et al. (2007), and Sinova et al. (2012) among others for a discussion of least squares analysis of this and related models. Maatouk (2003) proposes nonparametric smoothing for this model, by applying weighted least squares to the interval data and then using the resulting intercept as the estimator. Couso and Dubois (2014) 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. Heitjan and Rubin (1991), Heitjan (1994) and Gill et al. (1997). 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 Beresteanu and Molinari (2008) 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., Tsybakov (2009)), 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 Molchanov (2017) for a thorough account of random sets theory.

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 convex compact set in $\mathbb{R}^d$. We denote by $s(\boldsymbol{Y}, v) = \sup_{y \in \boldsymbol{Y}} v^\top y$ and $w(\boldsymbol{Y}, v) = s(A, v) + s(A, -v)$, respectively, its support function and width function in direction v (see Appendix D). Assume that $\boldsymbol{Y}$ is integrably bounded, that is, $\|\boldsymbol{Y}\| = \sup_{y \in \boldsymbol{Y}} \|y\|$ is integrable. Since $|s(\boldsymbol{Y}, v)| \leq \|\boldsymbol{Y}\|$ for all v from the unit sphere, the support function is integrable and $\mathbf{E}s(\boldsymbol{Y}, v) = s(\mathbf{E}\boldsymbol{Y}, v)$, 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.; in this case $\boldsymbol{y}$ is said to be a (measurable) *selection* of $\boldsymbol{Y}$.

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

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

Also in this case it holds that $\mathbf{E}[s(\boldsymbol{Y}, v)|\boldsymbol{x} = x] = s(\mathbf{E}[\boldsymbol{Y}|\boldsymbol{x} = x], v)$.

To simplify the exposition, henceforth we assume that $\boldsymbol{x}$ is a scalar random variable taking values in 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.

In our analysis, the true but unobservable outcome associated with $\boldsymbol{x} \in I$ is a random vector $\boldsymbol{y}$ that almost surely takes values in $\boldsymbol{Y}$, so that $\boldsymbol{y}$ is a measurable selection of $\boldsymbol{Y}$. The pair $(\boldsymbol{x}, \boldsymbol{y})$ is a selection of $\{\boldsymbol{x}\} \times \boldsymbol{Y}$, a random closed set in $I \times \mathbb{R}^d$.

We first focus on a specific selection $(\boldsymbol{x}, \boldsymbol{y}) \in \{\boldsymbol{x}\} \times \boldsymbol{Y}$ a.s.. Such selection is associated with a function $m(x) = \mathbf{E}[\boldsymbol{y}|\boldsymbol{x} = x]$, and the estimator for this function can be obtained from the classical approach. In particular, the local polynomial estimator of order p based on observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, \ldots, n$, is obtained by minimizing the weighted least squares

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

with respect to $\theta_0, \ldots, \theta_p$. 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 $\mathrm{Var}_K = \int z^2 K(z)\, dz$.

Solving explicitly the weighted least squares minimization problem of (2), one obtains the minimizer $\hat{\theta}$, and the first entry of it, the intercept $\hat{\theta}_0$, is used to estimate $m(x_0)$. This estimator can be written as

$$\hat{\boldsymbol{m}}(x_0) = \sum_{i=1}^{n} \ell_i(x_0)\boldsymbol{y}_i, \tag{3}$$

where

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

$$u(z) = \big(1, z, z^2/2!, \ldots, z^p/p!\big)^\top,$$

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

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

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

It is easy to see that $\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2 \geq 0$.

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

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

Our goal is to extend the local linear regression framework to set-valued outcomes. In other words, we aim to propose an analog of estimator (3) when $p = 1$ and $\boldsymbol{Y}$ is set-valued. In order to do so, we need to define square loss for sets, so as to formalize consistency results and the notion of mean squared error. 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. 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

$$L(A_1, A_2) = \left(\int_{\mathbb{S}^{d-1}} (s(A_1, v) - s(A_2, v))^2 \, dv \right)^{\frac{1}{2}}. \tag{4}$$

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

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 Vitale (1985) 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. 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 $\{\boldsymbol{x}\} \times \boldsymbol{Y}$. We show that computationally this is easily achieved by taking the following Minkowski average (see Appendix D) of the $\boldsymbol{Y}_i$ data:

$$\hat{\boldsymbol{M}}(x_0) = \sum_{i=1}^{n} \ell_i(x_0) \boldsymbol{Y}_i. \tag{5}$$

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 (5) is also the Fréchet mean of the observed values $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ in the metric given by (4), see Le and Kume (2000) and Molchanov (2017, Sec. 2.2.5).

The estimator in (5) yields a convex set, therefore we can characterize its properties by working with its support function (see (37) in Appendix D and Chapter 13 of Rockafellar (1970)). 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

$$
\begin{aligned}
s(\hat{\boldsymbol{M}}(x_0), v) &= s\Big(\sum_{i=1}^n \left(\ell_i^+ - \ell_i^-\right) \boldsymbol{Y}_i, v\Big) = \sum_{i=1}^n \ell_i^+ s(\boldsymbol{Y}_i, v) + \sum_{i=1}^n \ell_i^- s(\boldsymbol{Y}_i, -v) \\
&= \sum_{i=1}^n (\ell_i + \ell_i^-) s(\boldsymbol{Y}_i, v) + \sum_{i=1}^n \ell_i^- s(\boldsymbol{Y}_i, -v) = \sum_{i=1}^n \ell_i s(\boldsymbol{Y}_i, v) + \sum_{i=1}^n \ell_i^- w(\boldsymbol{Y}_i, v).
\end{aligned}
$$

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, alternative 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

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

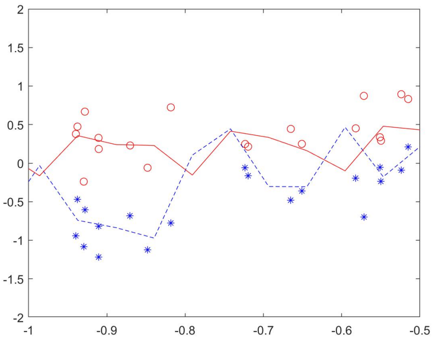

Figure 1: Possible emptiness of the estimator $\hat{\boldsymbol{N}}(x_0)$. Blue dashed line: $\sum_{i=1}^n \ell_i \boldsymbol{y}_{i\text{L}}$; red solid line: $\sum_{i=1}^n \ell_i \boldsymbol{y}_{i\text{U}}$.

Standard arguments in Fan (1993) yield that $\hat{\boldsymbol{N}}(x_0)$ and $\hat{\boldsymbol{O}}(x_0)$ are consistent estimators of

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

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 $\ell_i > 0$, $\boldsymbol{y}_{i\text{L}} = \boldsymbol{y}_{i\text{U}}$; and for i with $\ell_i < 0$, $\boldsymbol{y}_{i\text{U}} > \boldsymbol{y}_{i\text{L}}$.[2] Then

$$\begin{aligned}\sum_{i=1}^n \ell_i \boldsymbol{y}_{i\text{L}} &= \sum_{i=1}^n \ell_i^+ \boldsymbol{y}_{i\text{L}} - \sum_{i=1}^n \ell_i^- \boldsymbol{y}_{i\text{L}} = \sum_{i=1}^n \ell_i^+ \boldsymbol{y}_{i\text{U}} - \sum_{i=1}^n \ell_i^- \boldsymbol{y}_{i\text{L}} \\ &> \sum_{i=1}^n \ell_i^+ \boldsymbol{y}_{i\text{U}} - \sum_{i=1}^n \ell_i^- \boldsymbol{y}_{i\text{U}} = \sum_{i=1}^n \ell_i \boldsymbol{y}_{i\text{U}},\end{aligned}$$

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 $\boldsymbol{y}_{i\text{U}} > \boldsymbol{y}_{i\text{L}}$ whenever $\ell_i > 0$, depending on the realizations of $\boldsymbol{y}_{i\text{L}}$ and $\boldsymbol{y}_{i\text{U}}$, see Figure 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 $M(x_0)$.

Throughout the paper we assume $I = \mathbb{R}$ and we impose the following restrictions on the

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

observed and theoretical responses and on the density function of $\boldsymbol{x}$.

Assumption B (Observed responses)**.** *Conditionally on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, the observations $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$, are non-empty random compact convex sets such that*

(i) *$s(\boldsymbol{Y}_i, v) = s(M(\boldsymbol{x}_i), v) + \varepsilon_i(v)$, $v \in \mathbb{S}^{d-1}$, where $\varepsilon_i(\cdot)$, $i = 1, \dots, n$, 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 $v \in \mathbb{S}^{d-1}$. Assume that*

$$\sigma^2_{\max} = \max_{v \in \mathbb{S}^{d-1}} \mathbf{E}[\varepsilon(v)^2] < \infty$$

and denote the covariance function of ε by $C(v, u) = \mathbf{E}[\varepsilon(v)\varepsilon(u)]$.

(ii) *$\boldsymbol{Y}_i \subset \xi_i + B$ a.s. for integrable random vectors ξ_i, $i = 1, \dots, n$, and a deterministic compact set B that is the same for all i.*

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 that $\varepsilon_i(1) + \varepsilon_i(-1) \geq -(\mathbf{E}[\boldsymbol{y}_\mathrm{U}|\boldsymbol{x}] - \mathbf{E}[\boldsymbol{y}_\mathrm{L}|\boldsymbol{x}])$ almost surely. The latter condition replicates the requirement that $\mathbf{P}(\boldsymbol{y}_\mathrm{U} \geq \boldsymbol{y}_\mathrm{L}) = 1$. Note that ε does not admit a geometric interpretation as the support function of a random set. Part (ii) of Assumption B 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.

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 $v \in \mathbb{S}^{d-1}$.*

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 Fan (1993) with singleton responses.

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,*

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

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 (4),

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

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

$$\text{MSE}(x_0) = \int_{\mathbb{S}^{d-1}} b_{x_0}^2(v) \, dv + \int_{\mathbb{S}^{d-1}} \sigma_{x_0}^2(v) \, dv,$$

where $b_{x_0}^2(v)$ and $\sigma_{x_0}^2(v)$ are squared bias and variance, given by

$$\begin{aligned} b_{x_0}^2(v) &= \mathbf{E}\Big(\mathbf{E}[s(\hat{\boldsymbol{M}}(x_0), v) | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] - s(M(x_0), v)\Big)^2, \\ \sigma_{x_0}^2(v) &= \mathbf{E}\Big(s(\hat{\boldsymbol{M}}(x_0), v) - s(\mathbf{E}[\hat{\boldsymbol{M}}(x_0) | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], v)\Big)^2. \end{aligned}$$

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

$$b_{x_0}^2(v) = \mathbf{E}\Big(\sum_{i=1}^n \ell_i(s(M(\boldsymbol{x}_i),v) - s(M(x_0),v)) + \sum_{i=1}^n \ell_i^- w(M(\boldsymbol{x}_i),v)\Big)^2 \tag{6}$$

and

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

By Assumption B, the variance can be expressed as

$$\sigma_{x_0}^2(v) = \mathbf{E}\Big(\sum_{i=1}^n \ell_i\varepsilon_i(v) + \sum_{i=1}^n \ell_i^- (\varepsilon_i(v) + \varepsilon_i(-v))\Big)^2. \tag{7}$$

Differently from the classical case with singleton responses $\boldsymbol{y}_i$, the *negative* parts of the weights in (6) 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 cases, 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 (6) and hence affect the bias in the case of set-valued data. Following Fan (1993), in order to avoid zero in the denominator of the local linear estimator, we redefine ℓ_i by letting

$$\ell_i = \frac{\boldsymbol{\kappa}_{in}}{n}\frac{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}}. \tag{8}$$

We use 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 $o_r(a_n)$ 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.

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

$$\mathbf{E}\sum_{i=1}^n \ell_i^2 = \frac{1}{nh_n f(x_0)}\int K^2(z)\,dz + o\Big(\frac{1}{nh_n}\Big). \tag{9}$$

Proof. See Appendix A. □

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,*

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

Proof. See Appendix A. □

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* (5) *is*

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

Proof. See Appendix A. □

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

$$\mathbf{E}[\zeta(v)\zeta(u))] = \frac{C(v,u)}{f(x_0)}\int K(z)^2\,dz. \tag{10}$$

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 local the linear estimator in (5) *converges in distribution in the space of continuous functions on* $\mathbb{S}^{d-1}$ *with the uniform metric to the Gaussian process* ζ*.*

Proof. See Appendix A. □

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

$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{S}^{d-1}} (s(\boldsymbol{Y}_i, v) - s(\hat{\boldsymbol{M}}_{(-i)}(\boldsymbol{x}_i), v))^2 \, dv, \tag{11}$$

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

If $\boldsymbol{Y}_i = [\boldsymbol{y}_{i\mathrm{L}}, \boldsymbol{y}_{i\mathrm{U}}] \subset \mathbb{R}$, (11) turns into

$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} \Big(\boldsymbol{y}_{i\mathrm{L}} - \hat{\boldsymbol{M}}_{(-iL)}(\boldsymbol{x}_i))^2 + (\boldsymbol{y}_{i\mathrm{U}} - \hat{\boldsymbol{M}}_{(-iU)}(\boldsymbol{x}_i))^2 \Big), \tag{12}$$

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 (12) (or (11), 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

$$\hat{\mathcal{C}}(x_0) = \hat{\boldsymbol{M}}(x_0) + \frac{c_\alpha}{\sqrt{nh_n}} B, \tag{13}$$

where $B = \{b : \ \|b\| \leq 1\}$ is the unit ball. In (13) c_α is chosen so that

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

where ζ is the centered Gaussian process with covariance kernel (10), 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 Beresteanu and Molinari (2008) (see also Molchanov and Molinari, 2018, Theorem 4.13). It follows from Theorem 4.4 that

$$\lim_{n\to\infty} \mathbf{P}\Big(\max_{v:\ \|v\|=1} \{ s(\hat{\boldsymbol{M}}(x_0), v) - s(M(x_0), v) \\ - h_n^2 \frac{1}{2} s''(M(x_0), v) \operatorname{Var}_K - s(\hat{\mathcal{C}}(x_0), v) \}_+ = 0 \Big) \geq 1 - \alpha. \tag{14}$$

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.

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

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

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

6 Monte Carlo Simulations

We perform a simulation study with the following data generating process:

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

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

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 $x_0 = 0, 0.2, 0.4, 0.6$ we evaluate the coverage probability of the error tubes in equation (13).

We compare different implementations of the error tubes, and in Table 1 we report: (i) coverage probability of the true set $\boldsymbol{M}(x_0)$ by the error tube (meaning that the true set is a subset of the tube) in (13) computed using the cross-validation bandwidth (column 3); (ii) coverage probability as in (14), with the error tube in (13) 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 (13) 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.

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 Eisenhauer et al. (2009)). 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 Gautschi et al. (2017). 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 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.

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 (13), 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.

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 Gautschi et al. (2017). 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

[4] See `https://www.nanostring.com` for a description of this method.

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.

Table 2: 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

Table 2 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 $\boldsymbol{y}_{i\mathrm{U}} = \infty$). Of the sample used for our analysis, 58 patients were treated following protocol C1, and 37 following protocol C2.

The results of the analysis are reported in Figure 2 for the gene PTGS2 (COX2) and in Figure 3 for the gene CDC25A. The results suggest that the use of Bevacizumab in cancer treatment is quite beneficial for patients with moderate to relatively high 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 of expression of gene CDC25A the use of Bevacizumab seems highly beneficial, while at extremely low or high levels of the gene chemotherapy alone appears to be more effective. 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.

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.

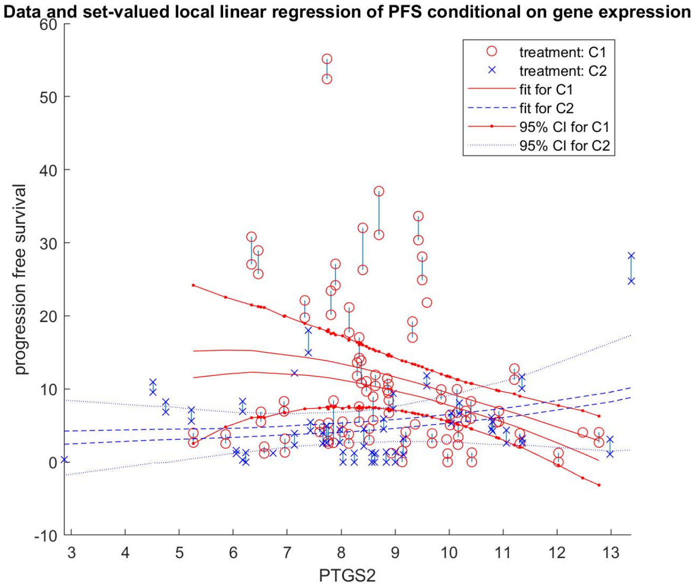

Figure 2: Results of the analysis for the gene PTGS2 ($\log_2$ of the Nanostring output)

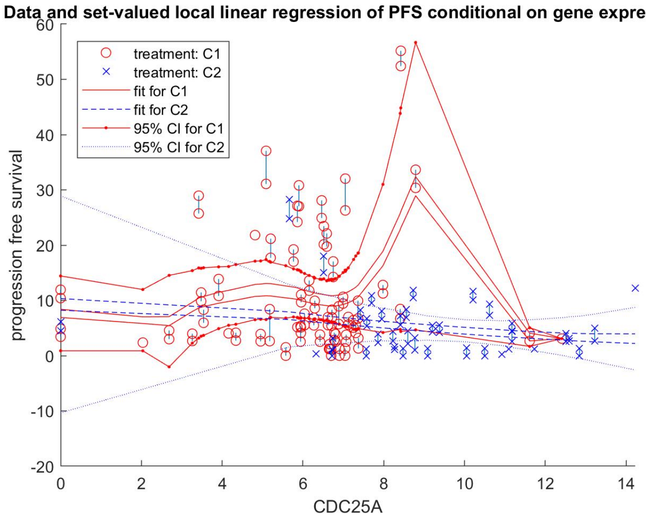

Figure 3: Results of the analysis for the gene CDC25A ($\log_2$ of the Nanostring output)

A Proofs of Main Results

Proof of Proposition 4.1. Our proof builds on (Fan, 1993, 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 $r \geq 1$,

$$\boldsymbol{s}_j = \mathbf{E}\boldsymbol{s}_j + h_n^{j+1}\mathcal{O}_r\big(1/\sqrt{nh_n}\big), \quad j = 0, 1, 2, \tag{15}$$

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

$$\begin{aligned}
\mathbf{E}\boldsymbol{s}_0 &= h_n \int K(z) f(zh_n + x_0)\, dz = h_n \int K(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n[f(x_0) + \mathcal{O}(h_n)],\\
\mathbf{E}\boldsymbol{s}_1 &= h_n^2 \int zK(z) f(zh_n + x_0)\, dz = h_n^2 \int zK(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n^2\mathcal{O}(h_n),\\
\mathbf{E}\boldsymbol{s}_2 &= h_n^3 \int z^2K(z) f(zh_n + x_0)\, dz = h_n^3 \int z^2K(z)(f(x_0) + \mathcal{O}(h_n))\, dz = h_n^3(f(x_0)\operatorname{Var}_K + \mathcal{O}(h_n)).
\end{aligned}$$

In view of (15), for an integer $r \geq 1$,

$$\boldsymbol{s}_j = h_n^{j+1}\left(f(x_0)\int z^j K(z)\,dz + \mathcal{O}_r(h_n + \frac{1}{\sqrt{nh_n}})\right), \quad j = 0, 1, 2. \tag{16}$$

Thus,

$$\boldsymbol{s}_0 = h_n f(x_0)(1 + o_r(1)), \tag{17}$$

$$\boldsymbol{s}_1 = h_n^2 o_r(1), \tag{18}$$

$$\boldsymbol{s}_2 = h_n^3 f(x_0)\operatorname{Var}_K(1 + o_r(1)). \tag{19}$$

It is easy to see that

$$\sum_{i=1}^n \ell_i = \frac{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}}.$$

Moreover, for a sufficiently large r,

$$\frac{h_n^4}{\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4}} = \frac{1}{f(x_0)^2\operatorname{Var}_K} + o_r(1), \tag{20}$$

cf. (Fan, 1993, Eq. (6.6)). In view of (17), (18), and (19),

$$\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 = h_n^4 f(x_0)^2\operatorname{Var}_K(1 + o_r(1)). \tag{21}$$

By (8),

$$\sum_{i=1}^n \ell_i^2 = \frac{\sum_{i=1}^n \boldsymbol{\kappa}_{in}^2(\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1)^2}{n^2(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2} = \frac{\boldsymbol{s}_2^2\boldsymbol{s}_0^*}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2} + \frac{(-2\boldsymbol{s}_2\boldsymbol{s}_1\boldsymbol{s}_1^* + \boldsymbol{s}_1^2\boldsymbol{s}_2^*)}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}, \tag{22}$$

where

$$\boldsymbol{s}_j^* = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{in}^2(x_i - x_0)^j = h_n^{j+1}\left(f(x_0)\int z^j K^2(z)\,dz + o_r(1)\right), \quad j = 0, 1, 2.$$

Furthermore, (16) implies that

$$\boldsymbol{s}_2^2\boldsymbol{s}_0^* = h_n^7 f^3(x_0)(\mathrm{Var}_K)^2 \int K^2(z)\,dz + h_n^7\mathcal{O}_{r/2}(1).$$

Combining this with (20) and letting $r = 4$, we obtain

$$\begin{aligned}\mathbf{E}\left(\frac{\boldsymbol{s}_2^2\boldsymbol{s}_0^*}{n(\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^2}\right) &= \frac{h_n^7 f^3(x_0)(\mathrm{Var}_K)^2 \int K^2(z)\,dz}{nh_n^8 f^4(x_0)(\mathrm{Var}_K)^2} + \frac{h_n^7}{nh_n^8}\mathcal{O}(1)\\ &= \frac{\int K^2(z)\,dz}{nh_n f(x_0)} + \mathcal{O}\left(\frac{1}{nh_n}\right).\end{aligned}$$

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

$$-2\boldsymbol{s}_2\boldsymbol{s}_1\boldsymbol{s}_1^* = h_n^7(f(x_0)\,\mathrm{Var}_K + \mathcal{O}_8(1))\mathcal{O}_8(1)(f(x_0)\int z^j K^2(z)\,dz + \mathcal{O}_4(1)) = h_n^7\mathcal{O}_2(1).$$

Analogously, $\boldsymbol{s}_1^2\boldsymbol{s}_2^* = h_n^7\mathcal{O}_2(1)$. Both these terms are as small as the minor term of $\boldsymbol{s}_2^2\boldsymbol{s}_0^*$. Therefore, (22) is dominated by its first term, whence (9) holds. □

Proof of Proposition 4.2. By definition, $\ell_i < 0$ if and only if $\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1 < 0$. Hence,

$$\begin{aligned}\mathbf{E}\Bigg(\sum_{i=1}^n \ell_i^-\Bigg)^2 &= \mathbf{E}\Bigg(\sum_{i=1}^n -\ell_i\mathbf{1}\{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1 < 0\}\Bigg)^2 \le n\mathbf{E}\Bigg(\sum_{i=1}^n \ell_i^2\mathbf{1}\{\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1 < 0\}\Bigg)\\ &\le n\mathbf{E}\Bigg(\sum_{i=1}^n \ell_i^2\mathbf{1}\{\boldsymbol{s}_2 < c_K h_n|\boldsymbol{s}_1|\}\Bigg) = n\mathbf{E}\Bigg(\mathbf{1}\{\boldsymbol{s}_2 < c_K h_n|\boldsymbol{s}_1|\}\sum_{i=1}^n \ell_i^2\Bigg)\\ &\le n\sqrt{\mathbf{P}(\boldsymbol{s}_2 < c_K h_n|\boldsymbol{s}_1|)}\Bigg(\mathbf{E}\Bigg(\sum_{i=1}^n \ell_i^2\Bigg)^2\Bigg)^{1/2},\end{aligned}\tag{23}$$

where the second inequality relies on Assumption A and the last one follows from the Chebyshev inequality. Using (16), we have, for an integer $r \geq 1$,

$$\begin{aligned}
\boldsymbol{s}_1 &= h_n^2\Big(\mathcal{O}(h_n) + \mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big),\\
\boldsymbol{s}_2 &= h_n^3\Big(f(x_0)\operatorname{Var}_K + \mathcal{O}(h_n) + \mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbf{P}(\boldsymbol{s}_2 < c_K h_n |\boldsymbol{s}_1|) \qquad (24)\\
&\quad \leq \mathbf{P}\Big(f(x_0)\operatorname{Var}_K + \mathcal{O}(h_n) + \mathcal{O}_r\big(1/\sqrt{nh_n}\big) < |\mathcal{O}(h_n)| + \Big|\mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big|\Big)\\
&\quad = \mathbf{P}\Big(f(x_0)\operatorname{Var}_K < |\mathcal{O}(h_n)| + \Big|\mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big|\Big). \qquad (25)
\end{aligned}$$

For sufficiently large n, there exist a ξ with $0 < \xi < f(x_0)\operatorname{Var}_K$ so that $|\mathcal{O}(h_n)| \leq \xi$ for all sufficiently large n. Building on (25), the Markov inequality and the definition of $\mathcal{O}_r(a_n)$ yield that

$$\begin{aligned}
\mathbf{P}(\boldsymbol{s}_2 < c_K h_n |\boldsymbol{s}_1|) &\leq \mathbf{P}\Big(f(x_0)\operatorname{Var}_K < \xi + \Big|\mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big|\Big)\\
&= \mathbf{P}\Big(\Big|\mathcal{O}_r\big(1/\sqrt{nh_n}\big)\Big| > f(x_0)\operatorname{Var}_K - \xi\Big)\\
&\leq \frac{\sup_{f\in\mathcal{H}(1,\gamma)} \mathbf{E}\left|\mathcal{O}_r\big(1/\sqrt{nh_n}\big)\right|^r}{(f(x_0)\operatorname{Var}_K - \xi)^r} = \frac{c_r\big(1/\sqrt{nh_n}\big)^r}{(f(x_0)\operatorname{Var}_K - \xi)^r}
\end{aligned}$$

for a positive constant c_r. Therefore,

$$\mathbf{P}(\boldsymbol{s}_2 < c_K h_n |\boldsymbol{s}_1|) = \mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big). \qquad (26)$$

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

$$\mathbf{E}\Big(\sum_{i=1}^n \ell_i^2\Big)^2 = \frac{1}{n^2h_n^2}\Big(\int K^2(z)dz\Big)^2(1 + o(1)). \qquad (27)$$

Substituting (26) and (27) into (23),

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

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

Proof of Theorem4.3. 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

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

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 (20), 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,

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

Therefore,

$$\mathbf{E}b_1^2 = \frac{1}{4}s''(M(x_0), v)^2(\mathrm{Var}_K)^2 h_n^4 + o(h_n^4), \tag{28}$$

cf. the proof of (Fan, 1993, Theorem 3).

By Proposition 4.2,

$$\mathbf{E}b_2^2 \leq w_{\max}^2 \mathbf{E}\Big(\sum_{i=1}^n \ell_i^-\Big)^2 = \frac{1}{h_n}\mathcal{O}\Big(\big(1/\sqrt{nh_n}\big)^r\Big), \tag{29}$$

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, (29) and (28),

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

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

$$\int_{\mathbb{S}^{d-1}} b_{x_0}^2(v)\, dv = \frac{1}{4}\int_{\mathbb{S}^{d-1}} s''(M(x_0), v)^2\, dv (\mathrm{Var}_K)^2 h_n^4 + o\left(h_n^4 + \frac{1}{nh_n}\right). \tag{30}$$

Now we bound the variance of the estimator splitting (7) into the sum of three terms. By Proposition 4.1, the first term is

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

The second term is

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

Finally, consider

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

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

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

and the result follows by adding (30) 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 Molchanov (2017, Theorem 3.2.1).

First, decompose

$$\begin{aligned}s(\hat{\boldsymbol{M}},v)-s(M(x_0),v) &= \sum_{i=1}^{n}\ell_i s(\boldsymbol{Y}_i,v)+\sum_{i=1}^{n}\ell_i^- w(\boldsymbol{Y}_i,v)-s(M(x_0),v)\\ &= \sum_{i=1}^{n}\ell_i s(M(\boldsymbol{x}_i),v)+\sum_{i=1}^{n}\ell_i\varepsilon_i(v)+\sum_{i=1}^{n}\ell_i^- w(\boldsymbol{Y}_i,v)-s(M(x_0),v). \qquad (31)\end{aligned}$$

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

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

Using a Taylor expansion,

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

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

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

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

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

(16) yields that

$$\boldsymbol{s}_j = h_n^{j+1} f(x_0) \int z^j K(z)\, dz \; (1 + o_p(1)), \quad j = 0,1,2,3. \tag{32}$$

By (21) and since $nh_n \to \infty$, we have

$$\boldsymbol{s}_2\boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4} = h_n^4 \operatorname{Var}_K f^2(x_0) \; (1 + o_p(1)). \tag{33}$$

Therefore,

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

Combining (32) and (33), we have

$$\begin{aligned}
&\sum_{i=1}^{n} \ell_i \left(\frac{1}{2}(\boldsymbol{x}_i - x_0)^2 s''(M(x_0), v) + R(x_0, \boldsymbol{x}_i, v) + \varepsilon_i(v) \right) \\
&= \left(\frac{1}{2}(\boldsymbol{s}_2^2 - \boldsymbol{s}_3 \boldsymbol{s}_1) s''(M(x_0), v) + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\kappa}_{in} (\boldsymbol{s}_2 - (\boldsymbol{x}_i - x_0)\boldsymbol{s}_1) \varepsilon_i(v) \right) (\boldsymbol{s}_2 \boldsymbol{s}_0 - \boldsymbol{s}_1^2 + n^{-4})^{-1} \\
&= \frac{1}{2} \operatorname{Var}_K s''(M(x_0), v) h_n^2 (1 + o_p(1)) + \frac{1}{n h_n f(x_0)} \sum_{i=1}^{n} \boldsymbol{\kappa}_{in} \varepsilon_i(v)(1 + o_p(1)). \qquad (34)
\end{aligned}$$

By the central limit theorem,

$$\frac{1}{\sqrt{n h_n}} \sum_{i=1}^{n} \boldsymbol{\kappa}_{in} \varepsilon_i \qquad (35)$$

converges in distribution to the centered normal random variable with variance equal to that of $\zeta(v)$. The combination of (31), (33), (34) and (35) yields the result. □

B Deterministic design points

When the design points $\boldsymbol{x}_i = x_i$, $i = 1, \ldots, n$, are deterministic[5], (6) 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

[5] 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 $\boldsymbol{\kappa}_{in} = \kappa_{in}$.

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.

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* $1/n \leq c_K h_n \leq 1$*, then* $\ell_i(x_0) \geq 0$ *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 Tsybakov (2009).

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* $n > 1$*,*

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

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 (Tsybakov, 2009, Prop. 1.13) in the singleton-valued data framework.

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

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

for sufficiently large n and $h_n \geq 1/(2n)$.

Proposition B.2 implies

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

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*

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

uniformly over all response functions satisfying Assumption F, where C_1 is a constant depending only on γ, a_0, λ_0, $\sigma_{\max}^2$, $K_{\max}$ and α.

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.

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

$$\{\omega : \ \boldsymbol{Y}(\omega) \cap K \neq \emptyset\} \in \mathfrak{F}, \tag{36}$$

for each compact set $K \subset \mathbb{R}^d$. Random sets $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ are said to be independently and identically distributed if

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

for all $K_1, \dots, K_n \in \mathcal{K}(\mathbb{R}^d)$ and

$$\mathbf{P}(\boldsymbol{Y}_i \cap K \neq \emptyset) = \mathbf{P}(\boldsymbol{Y}_j \cap K \neq \emptyset),$$

for all $i \neq j \in \{1, \dots, n\}$ 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

$$A = \bigcap_{v \in \mathbb{S}^{d-1}} \{a \in \mathbb{R}^d : v^\top a \leq s(A, v)\}. \tag{37}$$

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 (36). Its measurability is equivalent to the fact that $s(\boldsymbol{Y}, v)$ is a random variable for each $v \in \mathbb{R}^d$.

References

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

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

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

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

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

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

Eisenhauer, E., 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, and J. Verweij (2009). New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1). *European Journal of Cancer 45*, 228 – 247.

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

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

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

Gautschi, O., 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, and Swiss Group for Clinical Cancer Research (2017). 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*, 303–309.

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

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

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

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

Heitjan, D. F. and D. B. Rubin (1991, 12). Ignorability and coarse data. *Ann. Statist. 19*(4), 2244–2253.

Juster, F. T. and R. Suzman (1995). An overview of the health and retirement study. *Journal of Human Resources 30 (Supplement)*, S7–S56.

Kaido, H. (2017). Asymptotically efficient estimation of weighted average derivatives with an interval censored variable. *Econometric Theory 33*, 12181241.

Le, H. and A. Kume (2000). The Fréchet mean shape and the shape of the means. *Adv. Appl. Probab. 32*, 101–113.

Maatouk, T. (2003, September). *Some application of nonparametric regression with constrained data.* Ph. D. thesis, University of Glasgow, Glasgow.

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

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

Molchanov, I. (2017). *Theory of Random Sets* (2 ed.). London: Springer.

Molchanov, I. and F. Molinari (2018). *Random Sets in Econometrics.* Econometric Society Monograph Series, Cambridge University Press, Cambridge UK.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton: Princeton University Press.

Schollmeyer, G. and T. Augustin (2015). 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.

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

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Vitale, R. A. (1985). l_p metrics for compact, convex sets. *Journal of Approximation Theory 45*, 280–287.