LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Georg Schollmeyer, Thomas Augustin

On Sharp Identification Regions for Regression Under Interval Data

Technical Report Number 143, 2013
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Abstract

The reliable analysis of interval data (coarsened data) is one of the most promising applications of imprecise probabilities in statistics. If one refrains from making untestable, and often materially unjustified, strong assumptions on the coarsening process, then the empirical distribution of the data is imprecise, and statistical models are, in Manski's terms, partially identified. We first elaborate some subtle differences between two natural ways of handling interval data in the dependent variable of regression models, distinguishing between two different types of identification regions, called *Sharp Marrow Region (SMR)* and *Sharp Collection Region (SCR)* here. Focusing on the case of linear regression analysis, we then derive some fundamental geometrical properties of SMR and SCR, allowing a comparison of the regions and providing some guidelines for their canonical construction.
Relying on the algebraic framework of adjunctions of two mappings between partially ordered sets, we characterize SMR as a right adjoint and as the monotone kernel of a criterion function based mapping, while SCR is indeed interpretable as the corresponding monotone hull. Finally we sketch some ideas on a compromise between SMR and SCR based on a set-domained loss function.
This paper is an extended version of a shorter paper with the same title, that is conditionally accepted for publication in the *Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*. In the present paper we added proofs and the seventh chapter with a small Monte-Carlo-Illustration, that would have made the original paper too long.

Keywords: partial identification, imprecise probabilities, interval data, sharp identification regions, coarse data, adjunctions, partially ordered sets, linear regression model, best linear predictor, set-domained loss function.

1 Introduction

The methodology of imprecise probabilities offers powerful methods for reliable handling of *coarse(ned) data*, see, e.g., the ISIPTA contributions by [18, 44, 40, 37, 41, 7, 20]. The term coarsened data, or epistemic data imprecision, is an umbrella term, comprising all situations where data are not observed in the resolution intended in the subject matter context. This means, there is a certain true precise value $y \in \mathcal{Y}$ of a generic variable Y of material interest, but we only observe a set $A \supseteq \{y\}$. An extreme special case of coarse data are *missing data*, where the missingness of value y_i of unit i can be interpreted as having observed the whole sample space $\mathcal{Y}$. In the case where A is an interval $[\underline{y}, \overline{y}]$ for $\underline{y}, \overline{y} \in \mathbb{R}$ coarse data are commonly called *interval data*.

Before turning to the formal framework, two issues with fundamental importance for practical applications shall be recalled.

First of all, it must be stressed that the term 'coarse' is a relative term. Whether data are coarse or not depends on the specified sample space, and therefore on the subject matter context to be investigated. If, for instance, the sample space is taken to consist of some a priori specified ranges for income data, and that is all what is needed, then data are not coarse, while if precise income values are of interest, the data are coarse.[1]

Secondly, it is important to emphasize that coarse data typically are not just the result of sloppy research, like an insufficient study design or improper data handling. On the contrary, coarse data are an integral part of data collection, in particular in social surveys. Interval data arise naturally from the use of categories in order to avoid refusals in the case of sensitive questions, and are a means to model roughly rounded responses (see, e.g., [24]). Coarsened categorical data are, for instance, produced by matching data sets with not fully overlapping categories, are the direct outcome of data protection by some anonymization techniques (see, e.g., [13]), or may be produced by the combination of spaces with given marginals by Frechèt bounds (see, e.g., [19]). Another prototypic setting is the case of systematically missing data, arising from treatment evaluations in non-randomized designs like observational studies.[2]

By confining themselves to precise probabilities, traditional statistical approaches to cope with coarse data are inevitably forced to try to escape the imprecision in the data eventually. An immediate way in the case of interval data $[\underline{y}_i, \overline{y}_i]$ for each unit $i = 1, \ldots, n$ in the sample is to replace each interval by the corresponding central value $\mathring{y}_i = (\underline{y}_i + \overline{y}_i)/2$, and then to proceed with a standard analysis based on that fictitious sample. More sophisticated approaches add complex, typically untestable assumptions, either to explicitly model the coarsening process by a precise model, or to characterize idealized situations where the coarsening can be included in standard likelihood and Bayesian inference without biasing the analysis systematically.[3]

In recent years, awareness in statistics and econometrics has grown that such strong assumptions quite often cannot be justified by substantive arguments, and thus the – too high – price for the seemingly precise result of the statistical analysis is the loss of credibility of the conclusions, and in the end consequen-

[1] Indeed, even unions of intervals may constitute precise observations, for instance as the response to the question 'When did you live in Munich?', measured in years. Then $\{[1986; 1991] \cup [1997; 2000]\}$ is a precise observation in the sample space of all finite unions of closed intervals $[a, b]$ with $a, b \in \mathbb{N}_+$. (See in particular the distinction between *conjunctive* and *disjunctive* random sets in [14, Section 1.4], from which also this example is adopted.)

[2] To evaluate effects of treatment or intervention A over treatment B, in principle, it would be necessary to have information from a parallel universe, so-to-say, i.e. to know in addition how the units treated with A *would* have reacted *if* they had been given the treatment B and vice versa.

This question has in particular attracted intensive attention in the partial identification literature in econometrics (see, for instance, the survey [39] or the instructive case study [36].

[3] Most prominent is here Little and Rubin's [22] classification, distinguishing situations of *missingness completely at random (MCAR)* or *missing at random (MAR)* from *missing not at random (MNAR)* settings, where a systematic bias has to be expected. This classification has been extended to coarsening by [16].

tially the practical relevance of the statistical analysis.[4] In the light of this, it is of particular importance to develop approaches that reflect the underlying imprecision in the data properly, resulting in potentially imprecise, but reliable results. The fascinating insight, corroborated by a variety of applications mainly in econometrics (see the exemplary references below), is that in many studies these results are still enough to answer important substantive science questions, and if not, the scientist is alerted that strong conclusions drawn from the data may be mere artefacts.

Related approaches, considering all possible data compatible with the observed set of values, have been developed almost independently in different settings, ranging from reliable computing and interval analysis in engineering (e.g., [29]) and extensions of generalized Bayesian inference [10, 45] to reliable descriptive statistics in social sciences ([32, Chapter 17f], [30]). This cautious way to proceed is closely related to set-based (profile-)likelihood approaches ([47, 7]) and to the methodology of partial identification, in particular propagated by Manski (e.g., [23]) in econometrics, and to systematic sensitivity analysis (e.g. [42]) in biometrics, where a general framework for imprecise data models, i.e. sets of observationally equivalent statistical models, has been developed. In these models instead of single valued parameters one obtains so-called **identification regions**, i.e. sets of all parameters compatible with the data. On the inferential side, there has been important progress in the development of appropriate confidence procedures (see, e.g., [5, 27, 6]), and computational techniques have maturated to the extent that routine use of basic procedures has become feasible (e.g., [8, 1, 38, 33]). As a result, applied contributions are now rather common and are particularly influential in econometrics and allied fields see, e.g., [28] for an analysis of income poverty measures based on coarsened survey data, [21] for a study of the German reform of unemployment compensation based on register data and [26] for an analysis of treatment effects in observational studies with an illustration based on the National Longitudinal Survey of Youth.

The paper is organized as follows. After some basic definitions (Section 2), we emphasize in Section 3 the distinction between different understandings and goals of regression models, leading to two different types of identification regions, called SMR and SCR here. Section 4 formulates some basic geometrical properties, while sections 5 applies an algebraic framework for investigating mappings between partially ordered sets. We recall the basic concepts needed here, and explain them exemplary in the context of Dempster-Shafer-Theory and by describing coherent lower previsions as hulls. Then SMR and SCR are characterized as the monotone kernel and monotone hull of a criterion function based mapping, respectively. Section 6 suggests another type of identification regions that is based on a strict set-valued perspective, relying on a loss function depending on sets of parameters. Section 7 illustrates all the three identification regions and the predictions made by these regions while Section 8 concludes.

[4] See Manski's Law of Decreasing Credibility [23, p. 1].

2 Basic Definitions

Let Θ be a parameter space and $P := \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ a corresponding statistical model on a measurable space $(\Omega, \mathcal{F})$ with the associated observable random variables $X, \underline{Y}, \overline{Y}$ and the unobserved random variable Y. We are interested in the relationship between X and Y, but we have no full information about Y, we only know that the unobserved variable Y is related to the observed $\underline{Y}$ and $\overline{Y}$ in the sense that Y fulfills a certain relation, for example $\mathbb{P}(\underline{Y} \leq Y \leq \overline{Y}) = 1$ or $\mathbb{E}(\underline{Y} \mid X) \leq \mathbb{E}(Y \mid X) \leq \mathbb{E}(\overline{Y} \mid X)$[5]. In the sequel, we assume the second condition with the additional assumption that $\mathbb{E}(\underline{Y} \mid x)$ and $\mathbb{E}(\overline{Y} \mid x)$ are continuous in x. With $\mathbb{P}$ we denote the unknown true model and with $\mathbb{E}$ the corresponding expectations. The expectations for a model $\mathbb{P}_\theta$ are denoted with $\mathbb{E}_\theta$. The joint distribution of the random variables $X, Y, \underline{Y}, \overline{Y}$ under a model P_θ is denoted with $F_\theta^{X,Y,\underline{Y},\overline{Y}}$ (or short F_θ) and the joint distribution under the true model $\mathbb{P}$ is denoted with $F^{X,Y,\underline{Y},\overline{Y}}$ (or short F). Analogously, the distribution of a subset of random variables, eg. $\{X, Y\}$ is denoted with $F_\theta^{X,Y}$ and $F^{X,Y}$ respectively. For arbitrary random variables like e.g. $X, Z, \underline{Y}, \overline{Y}$ we denote their joint distribution with $F^{X,Z,\underline{Y},\overline{Y}}$. Because Y is not observable, we have not the full information about Y, which generally leads to partially identified models, which we define in the sequel: Two parameters $\theta_1, \theta_2 \in \Theta$ are undistinguishable (i.e. $\theta_1 \sim \theta_2$) if the corresponding models $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are empirically undistinguishable, which means that the distributions of the observable variables are the same. A statistical model P is called **point-identified**, if any two different parameters θ_1 and θ_2 are empirically distinguishable. Otherwise it is called **partially identified**.

Example 1 *The simple linear model with interval outcomes:*

$$\Theta = B \times R$$

with $B = \mathbb{R}^2$ the actually interesting parameter space and $R = \mathbb{R}^\Omega \times \mathbb{R}^\Omega_{\geq 0} \times \mathbb{R}^\Omega_{\geq 0}$ describing the error-terms and the coarsening-process: For $\theta = (\beta, (\varepsilon, \delta_l, \delta_u)) \in \Theta$ the associated variables are defined as

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ \underline{Y} &= X\beta + \varepsilon - \delta_l \\ \overline{Y} &= X\beta + \varepsilon + \delta_u \end{aligned}$$

with $\varepsilon, \delta_l, \delta_u$ measurable and ε with existing conditional expectations $\mathbb{E}(\varepsilon \mid x) = 0$. The coarsening process is modeled by the random variables δ_l and δ_u that are nonnegative, which ensures $\underline{Y} \leq Y \leq \overline{Y}$. By abuse of notation we identify the random variable X with the matrix $(1, X)$ to use matrix notations like above, if useful. Furthermore, in the sequel we assume X as a fixed random variable with support $\mathbb{R}$ and therefore omit it in the parameter space Θ. It is clear that this model is only partially identified. For example $((\beta_0, \beta_1), (\varepsilon, 0, 1)) \sim ((\beta_0 + 1, \beta_1), (\varepsilon, 1, 0))$. Moreover, the quotient space $\Theta_{/\sim}$ is not of the form

[5] This means $\forall x : \mathbb{E}(\underline{Y} \mid x) \leq \mathbb{E}(Y \mid x) \leq \mathbb{E}(\overline{Y} \mid x)$.

$\Theta_{/\sim} = B_{/\sim_B} \times R_{/\sim_R}$ *for some relations* $\sim_B$ *and* $\sim_R$*, so we must factorize the whole space* Θ *and not only the interesting part* B *to make the model point-identified.*

3 Two Types of Identification Regions

There are two ideal type senses of what a statistical model is and what it should render. One can assume a statistical model as the exact true underlying probabilistic structure, from which one only has to know all details and then one knows the exact distributions of all involved random variables and can make inferences with this knowledge. In contrast one can see a statistical model not as a truth, but as a rough approximation of truth and use it as a parsimonious tool to predict for example future observations of some variables or to get a rough insight into the real underlying structure that is actually more complex. As examples for this differentiation one could see firstly the estimation of the intercept and the slope of a linear model and secondly the problem of finding the best linear predictor in the sense of [2], which makes predictions that are linear in the covariates, but the underlying model needs not to be linear. The main difference is here that in the first case we really assume a linear model and rely on it, whereas in the second case we use the linearity of the predictions only to have a parsimonious model for predictions or explanations, but we assume nothing about the true statistical model.
These views lead to different problem formulations, which we want to state now as we need it in our context. In order to efficiently tackle our goal, we leave the statistical perspective and join Manski ([23, p. 7]), who recommends that problems of identification become much clearer when one firstly separates non-identifiability from sample variation, and assumes all distributions to be known for the analytic treatment[6] (later on then sample counterparts may be constructed in the usual way). In particular, we also assume that the distribution of Y is known (and we have no variables $\underline{Y}$ and $\overline{Y}$) and later we generalize this to the case of an unobserved Y, which leads to different sharp identification regions that are then our objects of interest.

The first problem statement is: Given the distribution $F^{X,Y}$ of (X, Y), which is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, find all θ, such that $(X, Y) \sim F_\theta^{X,Y}$, which is equivalent to find all θ with $L(F_\theta^{X,Y}, F^{X,Y}) = 0$ for an arbitrary distance-function $L(\cdot, \cdot)$ or a similar function, which is zero if and only if both arguments are equal. Here we think of a kind of loss function and introduce this equivalent formulation to indicate the analogy to the second problem formulation:

Given the distribution $F^{X,Y}$, which is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, find all θ, such that $L(F_\theta^{X,Y}, F^{X,Y})$ is minimal. In contrast to the first

[6]The identification regions arising in this limit case are called *sharp* identification regions.

problem, this problem definition is also meaningful if $F^{X,Y} \notin \{F_\theta^{X,Y} \mid \theta \in \Theta\}$, i.e. if the model is not correctly specified. If the model is correctly specified, then both problems are often essentially the same in the sense that for example for a linear model, the BLUE-estimator and the best linear predictor (with a quadratic loss-function) are solving different tasks, but the parameter estimates are identical. The actual problem is now that $F^{X,Y}$ is unknown. One part of the problem is that also if we could observe Y, we could not know the exact distribution of Y and so we have to estimate it. In particular, we cannot decide with certainty, if $F^{X,Y}$ is an element of the class $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$, and so the two problem formulations are moving together a little bit. The other part of the problem is that the variable Y, we are actually interested in, is not observable. As argued above for the moment we only address this second part of the problem and assume that we know the exact distribution of all observable variables. Later in section 5.2 we also address the other part. If now Y is unobserved, we can generalize the two problems by applying them to all possible Y that are consistent with $(\underline{Y}, \overline{Y})$.This leads to different regions of parameters that were proposed in different papers: The region related to the first problem was introduced slightly differently in [9] and the other region was proposed as the sharp identification region for the best linear predictor in [2].

Definition 1 *Let $P = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ be a statistical model with the corresponding joint distributions $\{F_\theta^{X,Y} \mid \theta \in \Theta\}$ and $X, \underline{Y}, \overline{Y}$ given random variables.*

*(i) The **sharp marrow region (SMR)** is defined as:*

$$SMR := \{\theta \in \Theta \mid \mathbb{E}(\underline{Y} \mid X) \leq \mathbb{E}_\theta(Y \mid X) \leq \mathbb{E}(\overline{Y} \mid X)\}.$$

Note that the Y in the definition is the Y coming from the model $\mathbb{P}_\theta$, not the Y from the true model. If the model is correctly specified (or if at least $SMR \neq \emptyset$), this region can also be written as:[7]

$$SMR = \operatorname*{argmin}_{\theta \in \Theta} \left[\min_{Z \in \mathbb{E}([\underline{Y}, \overline{Y}] \mid X)} L\left(F_\theta^{X,Y}, F^{X,Z,}\right) \right]$$

with an arbitrary loss function L. Here with $\mathbb{E}([\underline{Y}, \overline{Y}] \mid X)$ we denote the set of all random variables Z fulfilling $\mathbb{E}(\underline{Y} \mid X) \leq \mathbb{E}(Z \mid X) \leq \mathbb{E}(\overline{Y} \mid X)$. This equivalent characterization of SMR is valid because a parameter $\theta \in \Theta$ is in SMR if and only if there exists a $Z \in \mathbb{E}([\underline{Y}, \overline{Y}] \mid X)$ with $F_\theta = F^{X,Z,\underline{Y},\overline{Y}}$ or equivalently $L(F_\theta, F^{X,Z,\underline{Y},\overline{Y}}) = 0$. From the above representation of SMR we can see that SMR can be written as the solution of a decision problem with a minimin decision rule.

*(ii) The **sharp collection region (SCR)** is defined as:*

$$SCR := \bigcup_{Z \in [\underline{Y}, \overline{Y}]} \operatorname*{argmin}_{\theta \in \Theta} L\left(F_\theta, F^{X,Z,\underline{Y},\overline{Y}}\right).$$

[7] Note that we use the set-valued definition of argmin.

With $[\underline{Y}, \overline{Y}]$ *we denote the set of all random variables* Y *that lie between* $\underline{Y}$ *and* $\overline{Y}$ *for all* $\omega \in \Omega$.

A first comparison of this two regions that emphasizes the case of misspecification and interpretational problems for the sharp marrow region in this case can be found in [31]: While the interpretation of the sharp collection region as the collection of all best linear predictors is clear, the interpretation of SMR seems to be not so useful under misspecification, especially if SMR is empty.[8] From an empty SMR we can conclude, that the model is misspecified, but not more. Furthermore in the above-mentioned paper the authors make clear that a tight SMR "cannot be viewed as an indicator that the underlying model contains a lot of information about the true but partially identified parameter."[9]. An illustration of these considerations, that we pick up in chapter 7 can also be found in [31].

4 Geometrical Properties of Identification Regions

From now on, we concentrate on the case of a linear model like in example 1 and the classical quadratic loss function. Since we are only interested in the components (β_0, β_1) of an element $\theta = ((\beta_0, \beta_1), (\varepsilon, \delta_l, \delta_u)) \in SMR$, by abuse of notation, we also denote the set $\{(\beta_0, \beta_1) \mid ((\beta_0, \beta_1), (\varepsilon, \delta_l, \delta_u)) \in SMR\}$ as the sharp marrow region (analogously for the sharp collection region). Then we have

$$\begin{aligned} SMR &= \left\{\beta \in B | \mathbb{E}(\underline{Y} \mid X) \leq X\beta \leq \mathbb{E}(\overline{Y} \mid X)\right\} \qquad \text{and} \\ SCR &= \{\operatorname*{argmin}_{\beta \in B} \ \mathbb{E}((X\beta - Y)^2) \mid Y \in [\underline{Y}, \overline{Y}]\}. \end{aligned}$$

Remark 4.1 *The sharp marrow region is always a subset of the sharp collection region, and this is the reason for calling it sharp marrow region, it is the marrow of all truly linear models that fit to some* $Z \in [\underline{Y}, \overline{Y}]$. *In contrast, the sharp collection region collects the best fitting parameters for every possible* $Z \in [\underline{Y}, \overline{Y}]$.

Proof: With $\beta \in SMR$ we have $\mathbb{E}(\underline{Y} \mid x) \leq x\beta \leq \mathbb{E}(\overline{Y} \mid x)$ for all x and therefore we can choose $Y(x) := \lambda(x)\,\underline{Y} + (1 - \lambda(x))\,\overline{Y}$ with $\lambda(x) \in [0, 1]$ such that $\mathbb{E}(Y \mid x) = x\beta$. It is clear that $Y \in [\underline{Y}, \overline{Y}]$ and with $\varepsilon := Y - X\beta$ we get $\mathbb{E}((X\beta - Y)^2) = \mathbb{E}((X\beta - X\beta + \varepsilon)^2) = \mathbb{E}(\varepsilon^2)$ and for all further $\tilde{\beta} \in \mathbb{R}^2$ we have $\mathbb{E}((X\tilde{\beta} - Y)^2) = \mathbb{E}((X\tilde{\beta} - X\beta + \varepsilon)^2) = \mathbb{E}((X\tilde{\beta} - X\beta)^2) + \mathbb{E}(\varepsilon)^2 + \mathbb{E}(2(X\tilde{\beta} - X\beta) \cdot \varepsilon) \geq \mathbb{E}(\varepsilon^2)$, since all conditional expectations of ε are zero. Thus we have shown $\beta \in$ SCR. ∎

It is easy to see that the sharp marrow region is convex and closed. Furthermore, all convex, compact sets can be represented as a sharp marrow region:

[8] But compare the remarks in the next to last paragraphs of chapters 5.1 and 5.2.
[9] [31, p. 202].

Proposition 4.1 *Let $I \subset \mathbb{R}^2$ be a compact convex set. Then there exist random variables $\underline{Y}, \overline{Y}$ such that $SMR(\underline{Y}, \overline{Y}) = I$, namely:*

$$\begin{aligned} \underline{Y} &= \min\{X\beta \mid \beta \in I\} \\ \overline{Y} &= \max\{X\beta \mid \beta \in I\}. \end{aligned}$$

Proof: From $\beta \in I$ it follows $\mathbb{E}(\underline{Y} \mid X) = \underline{Y} \leq X\beta \leq \overline{Y} = \mathbb{E}(\overline{Y} \mid X)$, which means that β is in $SMR(\underline{Y}, \overline{Y})$. If $\beta \notin I$ because of the separation lemma (see e.g. [17]), there exists a linear functional on $\mathbb{R}^2$, represented by a vector (x_0, x_1) with $x_0\beta_0 + x_1\beta_1 < \inf_{\beta \in I} x_0\beta_0 + x_1\beta_1$. If $x_0 > 0$ it follows $\beta_0 + \frac{x_1}{x_0}\beta_1 < \inf_{\beta \in I} \beta_0 + \frac{x_1}{x_0}\beta_1 = \underline{Y}(\frac{x_1}{x_0})$, which shows that $\beta \notin SMR(\underline{Y}, \overline{Y})$. For $x_0 = 0$ we have $x_1\beta_1 < \inf_{\beta \in I} x_1\beta_1$, which leads to $\beta_0 + nx_1\beta_1 < \inf_{\beta \in I} \beta_0 + \inf_{\beta \in I} nx_1\beta_1 \leq \inf_{\beta \in I} \beta_0 + nx_1\beta_1 = \underline{Y}(nx_1)$ if n is large enough. The case $x_0 < 0$ can be proved analogously to the case $x_0 > 0$. ∎

For the sharp collection region, the situation is more complicated. To analyze this, we need some definitions from geometry (cf. [46]):

Definition 2 *The Minkowski sum*

$$M = \bigoplus_{i=1}^{n} l_i = \left\{ \sum_{i=1}^{n} p_i \,\middle|\, p_i \in l_i \right\}$$

*of n line segments $l_i \subseteq \mathbb{R}^d$ is called a **zonotope**. A zonotope is a convex, compact and centrally symmetric polytope with finite many extreme points and centrally symmetric facets. A closed, centrally symmetric convex set $Z \subseteq \mathbb{R}^d$ is called a **zonoid** if it can be approximated arbitrarily closely by zonotopes (w.r.t. a metric, e.g. the Hausdorff distance). For $d = 2$ the zonoids are exactly the closed, centrally symmetric convex sets (see, e.g., [3]).*

Proposition 4.2 *Let $\mathbb{E}(\underline{Y}), \mathbb{E}(\overline{Y}), \mathbb{E}(\underline{Y} \cdot X), \mathbb{E}(\overline{Y} \cdot X)$ be finite and $\mathbb{V}ar(X) \neq 0$. Then the sharp collection region is a zonoid.*

Proof: The estimator $\hat{SCR} := \{(x'x)^{-1}x'y \mid y \in [\underline{y}, \overline{y}]\}$ for SCR proposed in [2] is, as the linear image of a cuboid, a zonotope (see [8, p. 36]). Since (, as shown in [2, p. 784]) $\hat{SCR}$ is a consistent estimator of SCR with respect to the Hausdorff distance, it is clear that every sharp collection region is a limit object of zonotopes, thus a zonoid. ∎

Now, the question arises, if every zonoid can be represented as a sharp collection region. At first glance this seems to be not the case. By looking at examples of (estimates of) sharp collection regions, like that in figure 1 one observes that this regions often have two points on its boundary at which the boundary is not smooth. Note that the situation for SMR is similar, if X has finite or compact support. If the separating functional of the proof of Proposition 4.1 has $x_0 = 0$, the argument nx_1 of $\underline{Y}(nx_1)$ could have such a high absolute value that this value only occurs with probability zero, which leads to similar unsmooth situations, see figure 1. Fortunately one can prove that every zonotope Z in

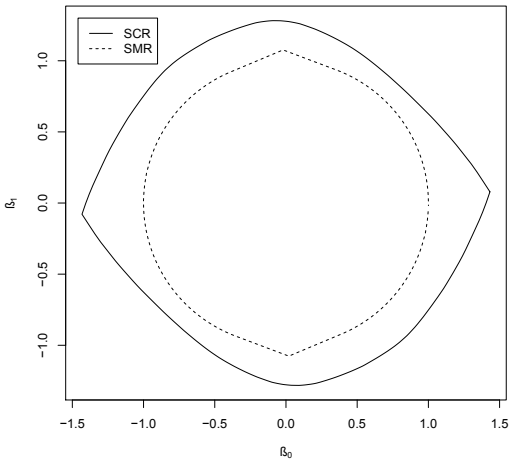

Figure 1: SCR and SMR with unsmooth boundary

general position (and, by looking on suitable limit-processes, also every zonoid) can be represented as a sharp collection region if we define the distribution of $X, \underline{Y}$ and $\overline{Y}$ in a certain way[10].

Proof: Without loss of generality, we have a zonotope Z generated by n line-segments that start in the origin of coordinates and have length d_i and slope s_i. With general position, we mean $s_i < \infty$. Now for $i = 1, \ldots, n$ take $x_i := s_i, \underline{y}_i = 0, \overline{y}_i = \frac{d_i}{\sqrt{1+x_i^2}}$ and add further $x_{n+1}, \ldots, x_m$ such that $(\frac{1}{m} x'x) = \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \bar{x^2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\underline{y}_{n+1} = \ldots = \underline{y}_m = \overline{y}_{n+1} = \ldots = \overline{y}_m = 0$. Then we have $\hat{SCR} = \{(x'x)^{-1}x'y \mid y \in [\underline{y}, \overline{y}]\} = \left\{ \frac{1}{m} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \mid y \in [0, \overline{y}] \right\}$ which is exactly the Minkowski addition of n line segments with slope $x_i = s_i$ and length $\frac{1}{m}\sqrt{y_i^2 + x_i^2 y_i^2} = \frac{yi}{m}\sqrt{1 + x_i^2} = d_i$. ∎

The last question is now, how independently from each other the regions SMR and SCR can be generated.

Proposition 4.3 *Let $I = SCR(\underline{Y}^*, \overline{Y}^*) \subseteq \mathbb{R}^2$ be a zonoid and $E \subseteq SMR(\underline{Y}^*, \overline{Y}^*)$ an arbitrary compact convex set. Then for every $\varepsilon > 0$ there exist random variables $\underline{Y}, \overline{Y}$ such that:*

$$\begin{aligned} d(SCR(\underline{Y}, \overline{Y}), I) &\leq \varepsilon \\ d(SMR(\underline{Y}, \overline{Y}), E) &\leq \varepsilon, \end{aligned}$$

[10]The main difference to SMR is that there we could construct regions for every arbitrary X with support $\mathbb{R}$.

where d is a metric on subsets of $\mathbb{R}^2$, e.g. the Hausdorff distance.

Proof: For $\varepsilon > 0$ define $S \subseteq \mathbb{R}$ such that the distance from any point x to S and the distance from x to S^C and $\mathbb{P}(X \in S)$ tends to zero as ε goes to zero. Then set $(\underline{Y}, \overline{Y})$ to $(\underline{Y}^*, \overline{Y}^*)$ if $x \in S^C$ and if $x \in S$ set $(\underline{Y}, \overline{Y})$ to the random variables $(\underline{Z}, \overline{Z})$ that generate E. Then $SMR((\underline{Y}, \overline{Y})) \longrightarrow SMR((\underline{Z}, \overline{Z})) = E$ and $SCR((\underline{Y}, \overline{Y})) \longrightarrow SCR((\underline{Y}^*, \overline{Y}^*)) = I$. ∎

5 An Algebraic View on Identification Regions

In the next section, we want to look at SMR and SCR as mappings. To analyze the algebraic structure of these mappings, we need some facts about adjunctions. Adjunctions arise in many contexts and often make life a bit easier, see the next examples. For an introduction to partially ordered sets and adjunctions see, e.g., [11, 15].

Definition 3 *Let $(P, \leq)$ and $(Q, \sqsubseteq)$ be partially ordered sets. A pair (f, g) of mappings $f : P \longrightarrow Q$ and $g : Q \longrightarrow P$ is called **adjunction**, if:*

$$\forall p \in P \forall q \in Q : \quad p \leq g(q) \iff f(p) \sqsubseteq q.$$

*In this case, f is called **left adjoint** and g is called **right adjoint**.*

Lemma 5.1 *Let (f, g) be an adjunction. Then the following holds:*

- *A1 $g \circ f$ is extensive and $f \circ g$ is intensive, i.e.:*
  $\forall p \in P, q \in Q : g(f(p)) \geq p \quad \& \quad f(g(q)) \sqsubseteq q.$
- *A2 f and g are order-preserving (monotone).*
- *A3 $f \circ g \circ f = f$ and $g \circ f \circ g = g$ and thus $f \circ g$ and $g \circ f$ are idempotent.*
- *A4 From A1 - A3 it follows that $g \circ f$ is a closure operator and $f \circ g$ is a kernel operator.*[11]
- *A5 The adjoints f and g are determining each other unambiguously.*
- *A6 f preserves existing joins and g preserves existing meets.*

To illustrate the concept of adjunctions, we apply it to two areas of the theory of imprecise probability.

Example 2 *Dempster-Shafer-Theory*[12]*:*
In [12] we have the multivalued mapping $\Gamma : X \longrightarrow 2^S$ with which we can associate a set-domained version

$$\tilde{\Gamma} : (2^X, \subseteq) \longrightarrow (2^S, \subseteq) : A \mapsto \bigcup_{a \in A} \Gamma(a).$$

[11] A closure operator is a monotone, extensive and idempotent mapping and a kernel operator is a monotone, intensive and idempotent one.

[12] For an introduction, see [12] and [34].

Furthermore we have the operator

$$\begin{array}{rcl} {}_* : (2^S, \subseteq) & \longrightarrow & (2^X, \subseteq) : \\ T & \mapsto & T_* := \{x \in X \mid \Gamma(x) \subseteq T\}. \end{array}$$

Then it is obvious that the pair $(\tilde{\Gamma}, {}_*)$ *is an adjunction because both* $A \subseteq T_*$ *and* $\tilde{\Gamma}(A) \subseteq T$ *are meaning exactly that all* $a \in A$ *are mapped to subsets of* T*. From this, the* ∞*-monotonicity of a belief function* $B = P \circ {}_*$ *with* P *a probability-measure follows immediately, since* P *is* ∞*-monotone and* ${}_*$ *is meet-preserving:*

$$\begin{aligned} & B(\bigcup_{i=1}^{k} T_i) = P((\bigcup_{i=1}^{k} T_i)_*) \geq P(\bigcup_{i=1}^{k} (T_i)_*) \\ & \geq \sum_{J \neq \emptyset} (-1)^{|J|+1} P(\bigcap_{i \in J} (T_i)_*) = \sum_{J \neq \emptyset} (-1)^{|J|+1} P((\bigcap_{i \in J} T_i)_*) \\ & = \sum_{J \neq \emptyset} (-1)^{|J|+1} B(\bigcap_{i \in J} T_i). \end{aligned}$$

Furthermore, it is clear that also the composition of a belief function and ${}_*$ *or another meet-preserving mapping is* ∞*-monotone.*

Example 3 *Lower Coherent Previsions*[13]*:*
With $(\mathbb{R}^{\mathcal{L}(\Omega)}, \leq)$ *the set of all previsions that are defined on all gambles and avoid sure loss, equipped with the dominance relation* $\underline{P}_1 \leq \underline{P}_2 : \Longleftrightarrow \ \forall X \in \mathcal{L}(\Omega) : \underline{P}_1(X) \leq \underline{P}_2(X)$ *and* $(2^{\mathscr{P}(\Omega)}, \supseteq)$ *the set of all nonempty sets of finitely additive probability-measures on* Ω *with the ordinary superset relation, we can construct the following adjunction:*

$$\begin{array}{rcl} f : (\mathbb{R}^{\mathcal{L}(\Omega)}, \leq) & \longrightarrow & (2^{\mathscr{P}(\Omega)}, \supseteq) : \underline{P} \mapsto \mathcal{M}(\underline{P}) \\ g : (2^{\mathscr{P}(\Omega)}, \supseteq) & \longrightarrow & (\mathbb{R}^{\mathcal{L}(\Omega)}, \leq) : M \mapsto \underline{P}_M \end{array}$$

with $\mathcal{M}(\underline{P}) = \{p \in \mathscr{P}(\Omega) \mid \forall X \in \mathcal{L}(\Omega) : p(X) \geq \underline{P}(X)\}$*, where* $\mathscr{P}(\Omega)$ *is the set of all finitely additive probability-measures and* $\underline{P}_M : \mathcal{L}(\Omega) \longrightarrow \mathbb{R} : X \mapsto \inf_{p \in M} p(X)$*. In this language, because of the lower envelope theorem*[14] *coherent lower previsions are exactly the hulls*[15] *of the closure operator* $g \circ f$*, which maps a lower prevision that avoids sure loss to its natural extension. It is now easy to see that the natural extension of a prevision* $\underline{P}$ *is the lowest coherent prevision that dominates* $\underline{P}$*: If* $\underline{P}_2 \geq \underline{P}$ *is another coherent prevision that dominates* $\underline{P}$*, then it is a hull* $(g \circ f)(Q)$ *for some* Q *and with the idempotence and the monotonicity of* $g \circ f$ *we have* $\underline{P}_2 = (g \circ f)(Q) = (g \circ f \circ g \circ f)(Q) \geq (g \circ f)(\underline{P})$*, where the right hand side is the natural extension of* $\underline{P}$*.*

[13] For an introduction, see [43].
[14] See [43, p. 134].
[15] Hulls are the images of a closure operator and similarly kernels are the images of a kernel operator.

5.1 SMR as a Right Adjoint

Proposition 5.2 *Let $(\mathscr{Y}, \leq)$ be the set of pairs of numeric random variables $\mathcal{Y} = (\underline{Y}, \overline{Y})$, equipped with the relation $\leq$ defined by*

$$\begin{aligned} \mathcal{Y}_1 \leq \mathcal{Y}_2 : \quad \Longleftrightarrow \quad & \mathbb{E}(\overline{Y}_1 \mid X) \leq \mathbb{E}(\overline{Y}_2 \mid X) \quad \& \\ & \mathbb{E}(\underline{Y}_1 \mid X) \geq \mathbb{E}(\underline{Y}_2 \mid X). \end{aligned}$$

This means that if $\mathcal{Y}_1 \leq \mathcal{Y}_2$, the observable variables $(\underline{Y}_1, \overline{Y}_1)$ are more informative than $(\underline{Y}_2, \overline{Y}_2)$ or equally informative, because from $(\underline{Y}_1, \overline{Y}_1)$ we can learn more or the same about the conditional expectations of the unobserved variable Y, we are actually interested in. The mapping

$$\begin{aligned} SMR : (\mathscr{Y}, \leq) \; &\longrightarrow (2^B, \subseteq) : \\ (\underline{Y}, \overline{Y}) \; &\mapsto \{\beta \mid \mathbb{E}(\underline{Y} \,|X) \leq X\beta \leq \mathbb{E}(\overline{Y} \,|X)\} \end{aligned}$$

is a right adjoint. The corresponding left adjoint is the prediction-operator[16]*:*

$$\begin{aligned} PR : (2^B, \subseteq) \quad &\longrightarrow \quad (\mathscr{Y}, \leq) : \\ \Gamma \quad &\mapsto \quad \left(\inf_{\beta \in \Gamma} X\beta, \sup_{\beta \in \Gamma} X\beta \right). \end{aligned}$$

Proof: We show $\Gamma \subseteq SMR(\mathcal{Y}) \iff PR(\Gamma) \subseteq \mathcal{Y}$: $\Rightarrow$: Let $Z = (\underline{Z}, \overline{Z}) := PR(\Gamma)$ and $\mathcal{Y} = (\underline{Y}, \overline{Y})$). Then we have $\underline{Z}(x) = \inf_{\beta \in \Gamma} x\beta \geq \inf_{\beta \in SMR(\mathcal{Y})} x\beta \geq \mathbb{E}(\underline{Y} \mid x)$ and $\overline{Z}(x) = \sup_{\beta \in \Gamma} x\beta \leq \sup_{\beta \in SMR(\mathcal{Y})} x\beta \leq \mathbb{E}(\overline{Y} \mid x)$.
$\Leftarrow$: Let $\beta \in \Gamma$. We have to show $\beta \in SMR(\mathcal{Y})$ or equivalently $\forall x : \mathbb{E}(\underline{Y} \mid x) \leq x\beta \leq \mathbb{E}(\overline{Y} \mid x)$:
$x\beta \leq \sup_{\beta \in \Gamma} x\beta \leq \mathbb{E}(\overline{Y} \mid x)$ and $x\beta \geq \inf_{\beta \in \Gamma} x\beta \geq \mathbb{E}(\underline{Y} \mid x)$. ∎

Because SMR is a right adjoint, it has the properties $A1 - A6$. The monotonicity $A2$ means that $SMR(\mathcal{Y})$ is more informative if $\mathcal{Y}$ is more informative. The idempotence $A3$ means that if we estimate, predict and then estimate again, we get the same information as if we had only estimated one time. Analogously if we predict, estimate and then predict once more, we get the same prediction as we would get, if we predicted only once. This property is often satisfied by classical estimators, for example the classical least squares estimator has an idempotent prediction matrix. Because $PR \circ SMR$ is a kernel operator, we can now give a clear interpretation of SMR, which is also valid in the misspecified case: The sharp marrow region is the largest region for which the corresponding predictions constitutes the largest inner approximation of the conditional expectations[17]. This interpretation may be not so useful in the misspecified situation, but it is clearly stated. The monotonicity is also shared by SCR,

[16] Here, the empty infimum is defined as ∞ and the empty supremum is defined as $-\infty$.

[17] An empty SMR means, that there is no inner approximation induced by the prediction of a set of parameters.

but SCR is no right adjoint, since it is not meet-preserving, because the intersection of two zonoids is generally not a zonoid. Furthermore, generally only $(SCR \circ PR \circ SCR)(\mathcal{Y}) \supset SCR(\mathcal{Y})$ holds, which means that we generally loose information if we predict and estimate once more.

5.2 SMR and SCR as a Kernel and a Hull

In [9] a criterion function based identification region is proposed. The criterion function (see Prop. 5.3) is based on a generalization of the expected squared errors to the expected squared minimal errors. The proposed sharp identification region is the argmin of this criterion function and it is very similar to SMR, but it is not monotone. It shows up that SMR is the highest lower and SCR is the lowest upper monotone approximation of this region.

Definition 4 *Let $E : (P, \leq) \longrightarrow (Q, \sqsubseteq)$ be a mapping. The monotone hull of E is defined as:*

$$H(E) \quad : \quad (P, \leq) \longrightarrow (Q, \sqsubseteq) : X \mapsto \bigvee_{Y \leq X} E(Y).$$

The monotone kernel of E is defined as:

$$K(E) \quad : \quad (P, \leq) \longrightarrow (Q, \sqsubseteq) : X \mapsto \bigwedge_{Y \geq X} E(Y).$$

These set-valued mappings are both order-preserving. Furthermore, the mapping $E \mapsto H(E)$ is a closure operator and the mapping $E \mapsto K(E)$ is a kernel operator, thus indeed $H(E)$ is a hull and $K(E)$ is a kernel. In particular, $H(E)$ is the lowest order-preserving mapping that is higher than E. Analogously, $K(E)$ is the highest order-preserving mapping that is lower than E.

Proposition 5.3 *Let the criterion function*
$Q : B \to \mathbb{R}$ be defined as

$$\begin{aligned} Q(\beta) &= \int \left(\mathbb{E}(\underline{Y}\,|x) - x\beta\right)_+^2 + \left(\mathbb{E}(\overline{Y}\,|x) - x\beta\right)_-^2 d\mathbb{P}(x) \\ &= \int \min_{Y \in [\underline{Y}, \overline{Y}]} (\mathbb{E}(Y \mid x) - x\beta)^2 d\mathbb{P}(x). \end{aligned}$$

Then the criterion function based mapping

$$\begin{array}{rcl} E_Q : (\mathscr{Y}, \leq) & \longrightarrow & (2^B, \subseteq) : \\ (\underline{Y}, \overline{Y}) & \mapsto & \operatorname*{argmin}_{\beta \in B} Q(\beta) \end{array}$$

is a source of SMR and SCR:

$$SMR = K(E_Q) \qquad \text{and} \qquad SCR = H(E_Q).$$

Proof: First, we mention that SCR is the monotone hull of the ordinary least squares estimator defined as $OLS : (\mathscr{Y}, \leq) \longrightarrow (2^B, \subseteq)$:

$$(\underline{Y}, \overline{Y}) \mapsto \begin{cases} \underset{\beta \in B}{\operatorname{argmin}} \mathbb{E}((X\beta - \underline{Y})^2) & \text{if } \underline{Y} = \overline{Y} \\ \emptyset & \text{else .} \end{cases}$$

It is clear that OLS is lower than E_Q, because, for precise $Y = \underline{Y} = \overline{Y}$ they are the same and otherwise $OLS(\mathcal{Y})$ is empty. with the monotonicity of the mapping $E \mapsto H(E)$ we get $SCR = H(OLS) \subseteq H(E_Q)$.[18] Furthermore, if we have a β that minimizes Q, then we can look on that variable $Y \in [\underline{Y}, \overline{Y}]$ with minimal distance to $X\beta$ and see that for this Y, β also minimizes $\mathbb{E}((X\beta - Y)^2)$. So we have $E_Q \subseteq SCR$. Since SCR is monotone and $H(E_Q)$ is the lowest monotone mapping that is greater than E_Q, we have $H(E_Q) \subseteq SCR$, which completes the proof of the first statement $SCR = H(E_Q)$. To show $SMR(\mathcal{Y}) = (K(E_Q))(\mathcal{Y})$ for all $\mathcal{Y} \in \mathscr{Y}$, we distinguish two cases:

a) $\min\limits_{\beta \in B} Q(\beta) = 0$ or equivalently $SMR(\mathcal{Y})$ is not empty. Then we have $\beta \in E_Q(\mathcal{Y}) \iff \mathbb{E}(\underline{Y} \mid X) \leq X\beta \leq \mathbb{E}\,\overline{Y} \mid X)$, which shows $SMR(\mathcal{Y}) = E_Q(\mathcal{Y})$ and furthermore, the implication $\mathcal{Z} \geq \mathcal{Y} \Longrightarrow E_Q(\mathcal{Z}) \supseteq E_Q(\mathcal{Y})$ shows that in this case also $SMR(\mathcal{Y}) = (K(E_Q))(\mathcal{Y})$.

b) $\min\limits_{\beta \in B} Q(\beta) > 0$ or equivalently $SMR(\mathcal{Y})$ is empty. Because of the strict convexity of Q, the argmin is unique and we denote it with β^*. Because the minimum of Q is not zero, there exists a set $S \subseteq \Omega$ with $\mathbb{P}(S) > 0$ and i): $\mathbb{E}(\overline{Y} \mid x(s)) + \delta < x(s)\beta^*$ or ii): $\mathbb{E}(\underline{Y} \mid x(s)) - \delta > x(s)\beta^*$ for some $\delta > 0$ and all $s \in S$. In case i), with $\mathcal{Z} = (\underline{Z}, \overline{Z})$ defined by

$$\underline{Z} = \underline{Y}, \qquad \overline{Z}(\omega) = \begin{cases} \overline{Y}(\omega) + \frac{\delta}{2} & \text{if } \omega \in S \\ \overline{Y}(\omega) & \text{else} \end{cases}$$

we get $\mathcal{Y} \leq \mathcal{Z}$. Because for this $\mathcal{Z}$, the corresponding $\min\limits_{\beta \in B} Q(\beta)$ is also greater than zero and the therefore unique argmin is different from β^*. From this it follows $(K(E_Q))(\mathcal{Y}) \subseteq E_Q(\mathcal{Y}) \cap E_Q(\mathcal{Z}) = \emptyset = SMR(\mathcal{Y})$. The case ii) can be shown in an analogous way.

■

From all above, the region SMR seems to be (at least in algebraic terms) a more satisfying region, but note that this region assumes that the model is in fact linear, which is generally untestable in this context. But the linearity assumption could be understood differently, firstly as an assumption on the true model and secondly as something like a regularization or simplification method to avoid overfitting or to have a parsimonious model. The first case points to the sharp marrow region and the second seemingly to the sharp collection region, but the parsimoniousness is decreasing if we allow for sets of paramters β instead of a single parameter and it is not a matter of course, if the SCR, constructed as the union of all reasonable best linear predictors, is still a useful

[18] Note that the relation $\subseteq$ is defined pointwise: $E_1 \subseteq E_2 : \iff \forall \mathcal{Y} : E_1(\mathcal{Y}) \subseteq E_2(\mathcal{Y})$.

model of the data.[19]
The region SCR can be estimated from samples in a consistent, monotone and nonpartial way. With nonpartial we mean that no pair $\underline{y} \leq \overline{y}$ of data would lead to the empty set as the estimate for SCR. One possibility is the estimator proposed in [2]. In contrast, also a nonempty SMR cannot be estimated in such a way[20]. To see this, take a sample $(\underline{y}, \overline{y}) = (e^{-x^2}, e^{-x^2})$, $(\underline{z}_1, \overline{z}_1) = (0, \overline{y}) \geq (\underline{y}, \overline{y})$ and $(\underline{z}_2, \overline{z}_2) = (\underline{y}, 1) \geq (\underline{y}, \overline{y})$. If an estimator $\hat{SMR}$ is consistent and monotone then for n large enough it should satisfy $\hat{SMR}((\underline{y}, \overline{y})) \subseteq \hat{SMR}((\underline{z}_1, \overline{z}_1)) \cap \hat{SMR}((\underline{z}_2, \overline{z}_2)) \approx \{(0,0)\} \cap \{(1,0)\} = \emptyset$. Furthermore SMR could not be estimated robustly in the sense that if one has a mixture in the sense of the proof of Proposition 4.3 then for ε small enough it is not clear what should be the estimated SMR, because that part of the data from the smaller region could be outliers or not, which would lead to different regions.

6 An Identification Region Based on a Set-Domained Loss Function

Now we try to establish a region, which could be understood as a compromise between SMR and SCR. The idea here is that we look on loss functions that are dependent on sets of parameters instead of single parameters. So in a sense we take the fact seriously that the region is a whole set that constitutes an imprecise probability structure. We do not look explicitly at every point of the set and then temporarily forget that the envisaged point is only one point of the set and maltreat it with a classical method. Instead, we see the set as a whole and do not look into it too deeply. We will construct a distance function between the set of conditional expectations of Y that cannot be refuted and the set of conditional expectations that are predicted by a set Γ of parameters. Here we do not assume that the true model is a linear one (if we would make this assumption, then we would get the region SMR again). Since we have to measure the distance between the two sets $\mathcal{A}(\mathcal{Y}) := \{(x, \mathbb{E}(Y \mid x)) \mid Y \in [\underline{Y}, \overline{Y}], x \in \mathbb{R}\}$ and $\mathcal{B}(\Gamma) := \{(x, x\beta) \mid \beta \in \Gamma, x \in \mathbb{R}\}$, we could use for example the Hausdorff distance

$$d_H(\mathcal{A}, \mathcal{B}) = \max \left\{ \sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a,b), \quad \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} d(a,b) \right\}$$

with some metric d of $\mathbb{R}^2$, which possibly takes the distribution of X into account and weights the distance according to the density $f(x)$. For a fixed x we have the possible conditional expectations of Y and the conditional expectations that are predicted by the parameter set Γ. Thus, both point sets are matched in a sense. Because the Hausdorff distance does not match the points of the two sets but compares all points of the two sets to each other, this distance seems

[19] In terms of parsimoniousness SMR is comparable to SCR and in fact SMR sometimes describes the data better, e.g., if $\mathcal{Y} = PR(\Gamma)$ for some Γ because then we have $PR(SMR(\mathcal{Y})) = \mathcal{Y}$ but generally only $PR(SCR(\mathcal{Y})) > \mathcal{Y}$.

[20] Note that the estimator proposed in [9] assumes a finite support of X and is not monotone.

to be a little bit counterintuitive. Thus, we propose a slightly different matched distance:

$$d_M(\mathcal{A},\mathcal{B}) := \int \left(\sup_{(x,a_2)\in\mathcal{A}} a_2 - \sup_{(x,b_2)\in\mathcal{B}} b_2 \right)^2 + \left(\inf_{(x,a_2)\in\mathcal{A}} a_2 - \inf_{(x,b_2)\in\mathcal{B}} b_2 \right)^2 d\mathbb{P}(x).$$

Now we can define a set-domained loss function as

$$L_S(\mathcal{Y},\Gamma) \quad = \quad d_M\left(\mathcal{A}(\mathcal{Y}), \mathcal{B}(\Gamma)\right)$$

and construct the sharp identification region for the minimizers of the set-domained loss (short: sharp setloss region) $SSR := \bigcup_{\Gamma\subseteq B} \operatorname{argmin} L_S(\mathcal{Y},\Gamma)$. Note that the argmin is not always unique, so that we have to take the union of all sets that minimizes L_S. To compute SSR one can look at the space $\mathscr{K} = \{PR(\Gamma) \mid \Gamma \subseteq B\}$ of all pairs of random variables $(\underline{Z}, \overline{Z})$ that are predicted by some set Γ. Since the predicted variables are only dependent on x, we treat them as functions from $\mathbb{R}$ to $\mathbb{R}$. The set $\mathscr{K}$ is then exactly the set of all $(\underline{Z}, \overline{Z})$ satisfying $\forall x_3 \notin [x_1, x_2]$:

$$\begin{aligned} &\overline{Z}(x_1) + (x_3 - x_1) \cdot \frac{\overline{Z}(x_2) - \overline{Z}(x_1)}{x_2 - x_1} \in [\underline{Z}(x_3), \overline{Z}(x_3)] \ \ \& \\ &\underline{Z}(x_1) + (x_3 - x_1) \cdot \frac{\underline{Z}(x_2) - \underline{Z}(x_1)}{x_2 - x_1} \in [\underline{Z}(x_3), \overline{Z}(x_3)]. \end{aligned} \tag{1}$$

That implies particularly that $\overline{Z}$ is convex and $\underline{Z}$ is concave. The task is now to find a pair $(\underline{Z}^*, \overline{Z}^*) \in \mathscr{K}$ that minimizes

$$\int (\underline{Z}(x) - \underline{Y}(x))^2 + (\overline{Z}(x) - \overline{Y}(x))^2 d\mathbb{P}(x).$$

This problem is nothing else than the problem of finding the projection of $(\underline{Y}, \overline{Y})$ on $\mathscr{K}$ and since $\mathscr{Y}$ is a Hilbert space and $\mathscr{K}$ is a closed convex set, this projection is unique. The candidate for the sharp setloss region is then $SMR((\underline{Z}^*, \overline{Z}^*))$. Because of $(\underline{Z}^*, \overline{Z}^*) = PR(\Gamma)$ for some Γ, we have

$$PR(SMR((\underline{Z}^*, \overline{Z}^*)) = PR(SMR(PR(\Gamma))) = PR(\Gamma) = (\underline{Z}^*, \overline{Z}^*),$$

which means that our region predicts exactly $(\underline{Z}^*, \overline{Z}^*)$. Furthermore, every other set that also predicts $(\underline{Z}^*, \overline{Z}^*)$ has to be a subset of our region and thus we have $SSR = SMR((\underline{Z}^*, \overline{Z}^*))$. From the construction of SSR it is also clear that the compositions $PR \circ SSR$ and $SSR \circ PR$ are also idempotent. To estimate the region SSR from a sample, we can analogously project the pair of vectors $(\underline{y}, \overline{y})$ on the set of pairs of vectors $(\underline{z}, \overline{z})$ satisfying

$$\begin{aligned} \forall x_k \notin [x_i, x_j] : \qquad &\overline{z}_i + (x_k - x_i)\frac{\overline{z}_j - \overline{z}_i}{x_j - x_i} \in [\underline{z}_k, \overline{z}_k] \quad \& \\ &\underline{z}_i + (x_k - x_i)\frac{\underline{z}_j - \underline{z}_i}{x_j - x_i} \in [\underline{z}_k, \overline{z}_k]. \end{aligned}$$

With $\theta = (\underline{z}_1, \ldots, \underline{z}_n, \overline{z}_1, \ldots \overline{z}_n)$ this problem can be written as the minimization of $\theta' Q \theta + c'\theta$ subject to $A\theta \geq 0$ for a (positive definite) matrix Q, a matrix A and a vector c. To compute the solution, one can use for example the algorithm proposed in [25]. To compute the final set $SMR((\underline{z}^*, \overline{z}^*))$, one can use standard linear programming techniques. The method can be robustified by modifying the loss function, but then, the solution may be not unique anymore. The minimization problem would get nonlinear, but the dimension of the problem would be n, which is maybe still acceptable[21]. Another idea is to allow only special sets of parameters. Here especially sets of sets of parameters that are closed under Minkowski convex-combinations are interesting, because this would ensure the uniqueness of the solution, because then the set of predictions made by such sets is convex. Such sets of sets are e.g. the set of all zonoids or the set of all zonotopes that are generated by line-segments that have a special angle. The minimization of L_S is then still tractable if the set of sets is parametrizable with a not too high number of parameters. An advantage of using special sets is that these sets are possibly better interpretable, especially if one has a higher number of covariates. For example an arbitrary high dimensional convex point set represented by all its extreme points is harder to figure out than a high dimensional ellipsoid represented by its location and the direction and spread of all main axes.

7 A small Monte Carlo Illustration

We now ran a small Monte Carlo experiment to illustrate the three regions. We use the same two simulation settings as in [31]. That is first a dependent variable Y with $\mathbb{E}(Y \mid x) = 5 + x$ with corresponding $\underline{Y} = Y - 0.5 - 0.2 \cdot X^2$ and $\overline{Y} = Y + 2.5 + 0.5 \cdot X^2$ where X is uniformly distributet on $[0, 5]$. This first setting represents the correctly specified case. For the second situation, we change only the formula for Y to $\mathbb{E}(Y \mid x) = 5 + (x-2)^3 - x^2$ to have a misspecified case. The conditional Expectations are illustrated in figure 2. Figure 3 shows the corresponding identification regions. In the misspecified case the sharp marrow region is empty. Since SSR is a superset of SMR it also contains the true parameter in the correctly specified case. In the misspecified case, the parameters are meaningless in the first place, but we can compare the predictions made by the different identification regions, which are illustrated in figure 4.

[21] Note that the naive robustification of SCR seems to be not so easy, because one has to look at the robust estimates for all $y \in [\underline{y}, \overline{y}]$ and this is not as easy as the computation of the image of $[\underline{y}, \overline{y}]$ under a linear mapping.

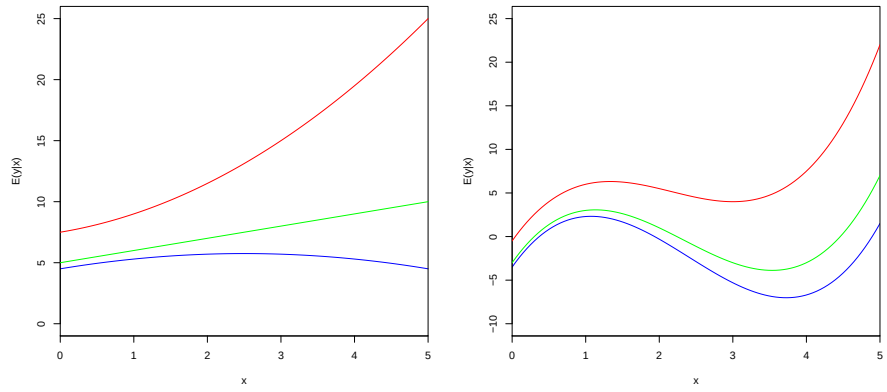

Figure 2: The conditional expectations $\mathbb{E}(Y \mid x)$ (green), $\mathbb{E}(\underline{Y} \mid x)$ (blue) and $\mathbb{E}(\overline{Y} \mid x)$ (red) for a well-specified (left) and a misspecified (right) situation.

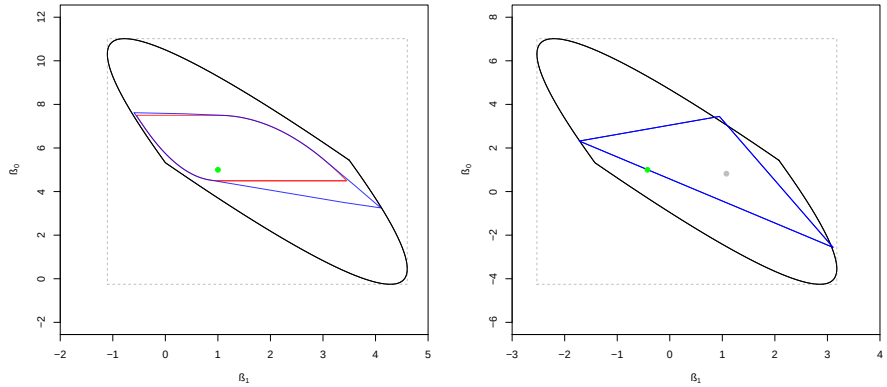

Figure 3: The identification regions SMR (red), SCR (black) and SSR (blue) for a well-specified (left) and a misspecified (right) situation. The true Parameter is dotted green (for the misspecified situation the green point is the best linear predictor for y). In the misspecified case SMR is empty and the grey point indicates the unique argmin of the criterion function Q.

If we really want to predict the value y for a next observation with the covariate-value x we can generally only predict an interval $[\hat{\underline{y}}, \hat{\overline{y}}]$. Here the question arises, if the collection of all possible best linear predictions induced by all possible $Y \in [\underline{Y}, \overline{Y}]$ is a useful tool for this kind of prediction. Generally, the predictions are too rough and are getting rougher and rougher if we estimate

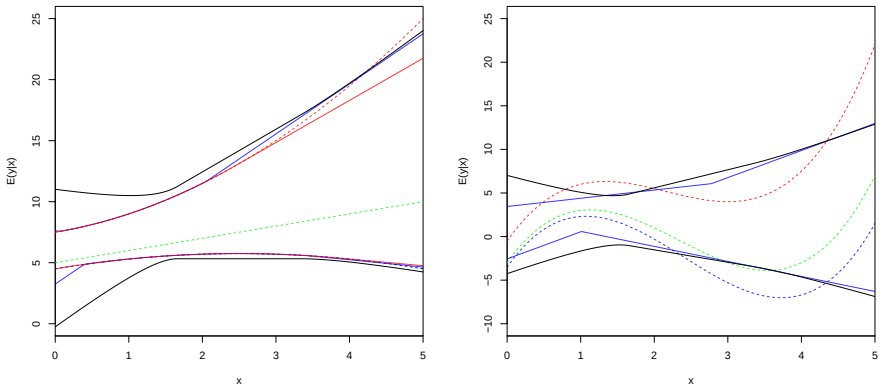

Figure 4: The predictions made by the regions SMR (red), SCR (black) and SSR (blue). The original conditional expectations are dashed ($\mathbb{E}(Y \mid x)$ green, $\mathbb{E}(\underline{Y} \mid x)$ blue and $\mathbb{E}(\overline{Y} \mid x)$ red).

and predict again and again, which is illustrated in figure 5.

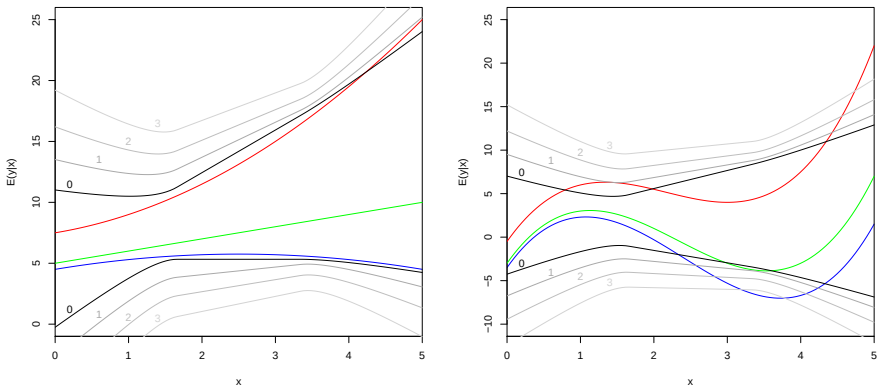

Figure 5: Different powers (0-3) of the operator $PR \circ SCR$ applied to $SCR((\underline{Y}, \overline{Y}))$ for the correctly specified (left) and the misspecified (right) case: the predictions are getting rougher and rougher if we estimate and predict again and again.

Of course, if one is interested in the set of all possible best linear predictors, one can get it with SCR, but what to do with this region? If one only needs possible moments of X, Y and XY then SCR would be helpful, but if we are

forced to make interval-valued predictions on outcomes y, we possibly use better SMR if we can rely on the linearity assumption, or SSR if we cannot do this, because it describes at least the bounds $\mathbb{E}(\underline{Y} \mid x)$ and $\mathbb{E}(\overline{Y} \mid x)$ better than SCR in the sense that the loss function L_S is smaller or equal. One can now argue, that SCR is, as a zonoid, a more regular or parsimonious region and therefore prefer SCR, but note that also if we restrict the problem to cases, were SMR is already a zonoid, then also in this restricted situation SCR is bigger than SMR. This is a main difference to e.g. the decision between a linear and a quadratic model: if we fit a quadratic model to a linear setting then we would get the same model as if we had chosen a linear model.

8 Concluding Remarks

We have worked out some differences between two types of identification regions in regression analysis under interval data, and discussed some of their properties. Indeed, SMR, relying so-to-say on the marrow of the regression model, and SCR, taking in a collection procedure all potential combinations of data points equally seriously, can be characterized as the monotone kernel and the monotone hull of a criterion function based mapping.
Furthermore, we sketched an appealing, rigorously set-based compromise, whose properties have still to be investigated in more detail. Other topics of further research include the additional inclusion of coarse covariates and an extension to generalized linear models. For generalized linear predictors in [35] a characterization of the sharp collection region is already given. If also covariates are interval-valued, the description of SCR becomes more complicated and a reformulation relying on roots of likelihood-based score-functions seems promising.[22] For the sharp marrow region the crucial role the conditional expectation $\mathbb{E}(Y|X)$ plays in the definition of SMR provides an immediate, promising link. Another direction of future research might be the analysis of models with instrumental variables. For this case a sharp characterization of SCR in terms of the support function of the identified set as well as some asymptotics of corresponding estimates can be found in [4].

References

[1] Beresteanu, A., Molchanov, I., & Molinari, F. (2011): Sharp identification regions in models with convex moment predictions, *Econometrica*, 79, 1785–1821.

[2] Beresteanu, A., & Molinari, F. (2008): Asymptotic properties for a class of partially identified models, Econometrica, 76, 763–814.

[22] See [33], who also developed algorithms for calculating SCR's like regions in a GLM based on an exponential model.

[3] Bolker, E.D. (1971): The zonoid problem, *The American Mathematical Monthly*, 78, 529–531.

[4] Bontemps, C., Magnac, T., & Maurin, E. (2012): Set identified linear Models, *Econometrica*, 80, 1129–1155.

[5] Bugni, F. A. (2010): Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set, *Econometrica*, 78, 735–753.

[6] Canay, I. A. (2010): El inference for partially identified models: Large deviations optimality and bootstrap validity, *Journal of Econometrics*, 156, 408–425.

[7] Cattaneo, M. E. G. V. & Wiencierz, A. (2012): Likelihood-based imprecise regression, *International Journal of Approximate Reasoning*, 53, 1137–1154, extended version of the 2011 ISIPTA paper. see F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 119–128, Innsbruck.

[8] Černý, M. & Rada M. (2011): On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurement Science Review*, 11, 34–40.

[9] Chernozhukov, V., Hong, H., & Tamer, E. (2007): Estimation and confidence regions for parameter sets in econometric models, *Econometrica*, 75, 1243–1284.

[10] de Cooman, G. & Zaffalon, M. (2004): Updating beliefs with incomplete observations. *Artificial Intelligence*, 159, 75–125.

[11] Davey, B.A. & Priestley, H.A. (2002): *Introduction to Lattices and Order*, 2nd edition. Cambridge UP.

[12] Dempster, A.P. (1967): Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38, 325–339.

[13] Dobra, A.& Fienberg, S. (2000): Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 11885–11892.

[14] Dubois, D. (1986): Belief structure, possibility theory and decomposable confidence measures on finite sets. *Computers and Artificial Intelligence*, 5, 403–416.

[15] Heijmans, H.J.A.M., (1994): *Morphological image operators*, Academic Press, Boston.

[16] Heitjan, D. & Rubin, D. (1991): Ignorability and coarse data. *The Annals of Statistics*, 19, 2244–2253.

[17] Holmes, R. N. (1975): *Geometric Functional Analysis and its Applications*. Springer, New York.

[18] Horowitz, J. L. & Manski, C. F. (2001): Imprecise identification from incomplete data. In: G. de Cooman, T. Fine, & T. Seidenfeld (eds.): *ISIPTA'01: Proc. Second Int. Symp. on Imprecise Probabilities and Their Applications*, 213–218, Maastricht, Shaker.

[19] Kwerel, S. (1983): Fréchet Bounds, In: S. Kotz & N. Johnson (eds.): *Encyclopedia of Statistical Sciences: Volume 3*, 203–209. Wiley, New York.

[20] Küchenhoff, H., Augustin, T. & Kunz, A. (2012): Partially identified prevalence estimation under misclassification using the kappa coefficient. *International Journal of Approximate Reasoning*, 53, 1168–1182, extended version of the 2011 ISIPTA paper. see F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 237–246, Innsbruck.

[21] Lee, S. & Wilke, R. A. (2009): Reform of unemployment compensation in Germany: A nonparametric bounds analysis using register data, *Journal of Business & Economic Statistics*, 27, 193–205.

[22] Little, R. & Rubin, D. (1987): *Statistical Analysis with Missing Data*. Wiley, New York.

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

[24] Manski, C. F. & Molinari, F. (2010): Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics*, 28, 219–231.

[25] Meyer, M. C. (2013): A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics*, 42, 1126–1139.

[26] Molinari, F. (2010): Missing treatments, *Journal of Business & Economic Statistics*, 28, 82–95.

[27] Moon, H. R. & Schorfheide, F. (2012): Bayesian and frequentist inference in partially identified models, *Econometrica*, 80, 755–782.

[28] Nicoletti, C., Peracchi, F., & Foliano, F. (2011): Estimating income poverty in the presence of missing data and measurement error, *Journal of Business & Economic Statistics*, 29, 61–72.

[29] Nguyen, H., Kreinovich, V., Wu, B., & Xiang, G. (2011): *Computing Statistics under Interval and Fuzzy Uncertainty: Applications to Computer Science and Engineering*. Springer, Berlin/Heidelberg.

[30] Pötter, U. (2008): *Statistical Models of Incomplete Data and Their Use in Social Sciences*, Ruhr-Universität Bochum (Habilitation Thesis).

[31] Ponomareva, M., & Tamer, E. (2011): Misspecification in moment inequality models: back to moment equalities?. *Econometrics Journal*, 14, 186-203.

[32] Rohwer, G. & Pötter, U. (2001): *Grundzüge der sozialwissenschaftlichen Statistik*. Juventa, Weinheim.

[33] Seitz, M. (2012): Estimating partially identified parameters in generalized linear models under interval data (in German), Master Thesis, Department of Statistics, LMU Munich.

[34] Shafer, G. (1976): *A Mathematical Theory of Evidence.* Princeton U. P.

[35] Stoye, J. (2007): Bounds on generalized ginear predictors with incomplete outcome data. *Reliable Computing*, 13, 293–302.

[36] Stoye, J. (2009a): Partial identification and robust treatment choice: An application to young offenders. *Journal of Statistical Theory and Practice*, 3, 239–254.

[37] Stoye, J. (2009b): Statistical inference for interval identified parameters. In: T. Augustin, F. Coolen, S. Moral, & M. Troffaes (eds.): *ISIPTA '09: Proc. Sixth Int. Symp. on Imprecise Probability: Theories and Applications*, 395–404, Durham, UK, SIPTA.

[38] Stoye, J. (2010): Partial identification of spread parameters. *Quantitative Economics*, 1, 323–357.

[39] Tamer, E. (2010): Partial identification in econometrics. *Annual Review of Economics*, 2, 167–195.

[40] Utkin, L. V. & Augustin, T. (2007): Decision making under imperfect measurement using the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 44, 322–338; extended version of the 2005 ISIPTA paper. see F. Cozman, R. Nau, & T. Seidenfeld (eds.): *ISIPTA '05: Proc. Fourth Int. Symp. on Imprecise Probabilities and Their Applications*, Manno.

[41] Utkin, L. V. & Coolen, F.P.A. (2011): Interval-valued regression and classification models in the framework of machine learning. In: F. P. A. Coolen, G. de Cooman, T. Fetz, & M. Oberguggenberger (eds.): *ISIPTA '11: Proc. Seventh Int. Symp. on Imprecise Probability: Theories and Applications*, 371–380, Innsbruck.

[42] Vansteelandt, S., Goetghebeur, E., Kenward, M., & Molenberghs, G. (2006): Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16, 953–979.

[43] Walley, P. (1991): *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

[44] Zaffalon, M. (2005): Conservative rules for predictive inference with incomplete data. In: F. Cozman, R. Nau, & T. Seidenfeld (eds.): *ISIPTA '05: Proc. Fourth Int. Symp. on Imprecise Probabilities and Their Applications*, 406–415, Manno.

[45] Zaffalon, M. & Miranda, E. (2009): Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34, 757–821.

[46] Ziegler, G.M. (2008): *Lectures on Polytopes.* Graduate Texts in Mathematics, vol. 152, Springer, New York.

[47] Zhang, Z. (2010): Profile likelihood and incomplete data. *International Statistical Review*, 78, 102-116.