Regression analysis with imprecise data

Andrea Wiencierz

Dissertation an der
Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

30. Oktober 2013

Erster Berichterstatter:
Prof. Dr. Thomas Augustin (LMU München)

Zweiter Berichterstatter:
Prof. Dr. Lev V. Utkin (FTU St. Petersburg)

Disputation: 13. Dezember 2013

Für meine Eltern

Susanne und Paul Wiencierz

Zusammenfassung

Methoden der statistischen Datenanalyse setzen in der Regel voraus, dass die vorhandenen Daten präzise und korrekte Beobachtungen der untersuchten Größen sind. Häufig können aber bei praktischen Studien die interessierenden Werte nur unvollständig oder unscharf beobachtet werden. Die vorliegende Arbeit beschäftigt sich mit der Fragestellung, wie Regressionsanalysen bei unscharfen Daten sinnvoll durchgeführt werden können.

Zunächst werden verschiedene Ansätze zum Umgang mit unscharf beobachteten Variablen diskutiert, bevor eine neue Likelihood-basierte Methodologie für Regression mit unscharfen Daten eingeführt wird. Als Ergebnis der Regressionsanalyse wird bei diesem Ansatz keine einzelne Regressionsfunktion angestrebt, sondern die gesamte Menge aller anhand der Daten plausiblen Regressionsfunktionen betrachtet, welche als Konfidenzbereich für den untersuchten Zusammenhang interpretiert werden kann. Im darauffolgenden Kapitel wird im Rahmen dieser Methodologie eine Regressionsmethode entwickelt, die sehr allgemein bezüglich der Form der unscharfen Beobachtungen, der möglichen Verteilungen der Zufallsgrößen sowie der Form des funktionalen Zusammenhangs zwischen den untersuchten Variablen ist. Zudem werden ein exakter Algorithmus für den Spezialfall der linearen Einfachregression mit Intervalldaten entwickelt und einige statistische Eigenschaften der Methode näher untersucht. Dabei stellt sich heraus, dass die entwickelte Regressionsmethode sowohl robust im Sinne eines hohen Bruchpunktes ist, als auch sehr verlässliche Erkenntnisse hervorbringt, was sich in einer hohen Überdeckungswahrscheinlichkeit der Ergebnismenge äußert. Darüber hinaus wird in einem weiteren Kapitel ein in der Literatur vorgeschlagener Alternativansatz ausführlich diskutiert, der auf Support Vector Regression aufbaut. Dieser wird durch Einbettung in den methodologischen Rahmen des vorher eingeführten Likelihood-basierten Ansatzes weiter verallgemeinert. Abschließend werden die behandelten Regressionsmethoden auf zwei praktische Probleme angewandt.

Regression analysis with imprecise data

Andrea Wiencierz

Dissertation submitted to
the Faculty of Mathematics, Informatics, and Statistics
of the LMU Munich

for the academic degree of
Doctor rerum naturalium (Dr. rer. nat.)

October 30, 2013

First referee: Prof. Dr. Thomas Augustin (LMU Munich)
Second referee: Prof. Dr. Lev V. Utkin (FTU St. Petersburg)
Defense: December 13, 2013

For my parents

Susanne and Paul Wiencierz

Abstract

Statistical methods usually require that the analyzed data are correct and precise observations of the variables of interest. In practice, however, often only incomplete or uncertain information about the quantities of interest is available. The question studied in the present thesis is, how a regression analysis can reasonably be performed when the variables are only imprecisely observed.

At first, different approaches to analyzing imprecisely observed variables that were proposed in the Statistics literature are discussed. Then, a new likelihood-based methodology for regression analysis with imprecise data called Likelihood-based Imprecise Regression is introduced. The corresponding methodological framework is very broad and permits accounting for coarsening errors, in contrast to most alternative approaches to analyzing imprecise data. The methodology suggests considering as the result of a regression analysis the entire set of all regression functions that cannot be excluded in the light of the data, which can be interpreted as a confidence set. In the subsequent chapter, a very general regression method is derived from the likelihood-based methodology. This regression method does not impose restrictive assumptions about the form of the imprecise observations, about the underlying probability distribution, and about the shape of the relationship between the variables. Moreover, an exact algorithm is developed for the special case of simple linear regression with interval data and selected statistical properties of this regression method are studied. The proposed regression method turns out to be robust in terms of a high breakdown point and to provide very reliable insights in the sense of a set-valued result with a high coverage probability. In addition, an alternative approach proposed in the literature based on Support Vector Regression is studied in detail and generalized by embedding it into the framework of the formerly introduced likelihood-based methodology. In the end, the discussed regression methods are applied to two practical questions.

Contents

Chapter 1

Introduction

The present thesis deals with the statistical problem of analyzing the relationship between two or more real-valued variables when these quantities are only imprecisely observed.

1.1 Motivation

The term regression refers to the most popular and commonly employed methods of statistical data analysis. The goal of a regression analysis is to obtain a quantitative description of the relationship between one or more explanatory variables and a response variable. For example, regression methods can be used to analyze the relationship between the income, the age and further sociodemographic characteristics of an individual and the overall life satisfaction (Wunder et al., 2013) or to investigate how the number of earthquakes per day in a seismically active region can be explained by the amount of preceding rainfall together with the air temperature and further quantities describing the impact of the earth tide (Svejdar et al., 2011). There is a large variety of regression methods for many different situations, which is regularly complemented by new suggestions. Overviews of numerous established regression methods with many references for interested readers can be found, for example, in the textbooks Fahrmeir et al. (2013) and Hastie et al. (2009).

Like most statistical tools, regression methods are usually based on the assumption that the analyzed data are precise and correct observations of the variables of interest. In statistical practice, however, often only incomplete or uncertain information about the data values is available. For example, consider the personal income, which is a key variable for many socioeconomic questions. Data on personal income are usually collected by surveys, because privacy laws prevent other ways. Faced with the question about the exact figure of the total or net income received during the last month or year, a respondent is very likely to give a value that is rounded to a multiple of hundreds or thousands. Hence, the recorded value often contains only the information that the exact figure lies in some interval around the given value. Moreover, as it is a very delicate question, many respondents are not willing answer at all. For those cases, the data set contains missing values, providing only the information that the income figures are numbers in the observation space of this variable. A common practice to obtain more informative answers is to ask those who refuse to give a precise value in a second step to indicate in which of the categories of a partitioned income range their income lies. As revealing a coarse income category is less informative than revealing the exact value, people are more likely to give this information. For those answering only this categorized question, the data provide the information that the income belongs to one of the intervals that constitute the partition of the income range. In all these cases, there is uncertainty about the exact data values. In fact, continuous variables are always observed only with limited precision, because the recorded number of digits is always finite. Further common examples of imperfect observations include censored survival times, variables that are observed on different aggregation levels, or missing values.

In all of these cases, the incomplete or uncertain information about the precise values of interest can be expressed by subsets of the observation space. For example, interval-censored and rounded data can be represented by intervals. Furthermore, if a value is precisely observed the observed set is a singleton and if a value is missing it is represented by the entire observation space of the corresponding variable. As the representation by

subsets allows considering many different forms of uncertainty about data within the same framework, this representation is adopted in this thesis. In the following, set-valued observations of real-valued variables are simply called imprecise data.

It is important to note that the notion imprecise data is sometimes used with a different meaning, for example, in the context of Fuzzy Statistics. In this and other frameworks, the quantities of interest are supposed to be inherently imprecise and therefore modeled as (fuzzy) sets. Consequently, in this context, the (fuzzy) set-valued data constitute exact observations of imprecise variables, and in a regression analysis, the relationship between imprecise quantities is investigated. How to approach this statistical problem, was studied, for example, in Blanco-Fernández et al. (2011); Domingues et al. (2010); Ferraro et al. (2010); Lima Neto and de Carvalho (2008); Coppi et al. (2006); Körner and Näther (1998); Diamond (1990). By contrast, this thesis is about analyzing the relationship between some precise variables in the situation in which only set-valued data on these quantities are available, because this problem appears to be more relevant for statistical practice, given the many different examples mentioned above. So far, there is no standard methodology for analyzing data that are imprecise in this sense.

The aim of the present thesis is to find a regression method that provides reliable insights about the relationship of interest, even if the variables are only imprecisely observed. Furthermore, the regression method should be general in the sense that it does not impose restrictive assumptions about the form of the imprecise observations, about the underlying probability distribution, and about the shape of the relationship between the variables.

1.2 Outline of the thesis

The core of this thesis starts with a review of different approaches to analyzing imprecise data that were proposed in the literature. Then, in Chapter 3, the formal framework of a new general likelihood-based approach to

regression with imprecisely observed variables is presented. Chapter 4 is devoted to a robust regression method derived from this general framework, and in Chapter 5, an alternative regression method for imprecisely observed responses is studied. Finally, two applications are presented in Chapter 6, before some general comments and a short outlook in Chapter 7 conclude this thesis.

A certain part of the ideas and results presented in this thesis were already published in a total of four publications. The following list indicates what sections of the present thesis are concerned and in which way the four publications contribute to these sections.

- Chapter 3 is based on Cattaneo and Wiencierz (2012, Sections 2 and 3) and contains many additional remarks and explanations.
- Section 4.1 is based on Wiencierz and Cattaneo (2012, Section 2), on Cattaneo and Wiencierz (2012, Section 3), and on Cattaneo and Wiencierz (2011, Sections 2 and 3) and contains many additional remarks and explanations.
- Section 4.2 is in part taken from Cattaneo and Wiencierz (2011, Section 4) together with additional exemplifications and remarks.
- Section 4.3 is based on Cattaneo and Wiencierz (2013, Section 3) and on Wiencierz and Cattaneo (2012, Section 3) and contains many additional remarks and explanations.
- Section 6.1 is for the most part taken from Cattaneo and Wiencierz (2013, Section 4).

Chapter 2

Analyzing imprecise data

In this chapter, different approaches to analyzing imprecise data are discussed with a focus on regression. As explained in Chapter 1, the term imprecise data stands for set-valued observations of precise variables, which covers amongst others actually precise observations and completely missing values as special cases.

2.1 Approaches aiming at a precise result

A simple ad hoc approach to dealing with imprecise data could be to reduce the observed sets each to a single value and to apply a standard method to the thus obtained precise data set. For example, if we want to perform a regression analysis and some of the analyzed variables are observed as intervals representing rounded values, the intervals could be replaced by their midpoints and a standard regression method could be applied to the midpoint data set, which yields a single estimated regression function. However, proceeding in this way in general does not provide correct estimates, as it was discussed already more than a century ago by Sheppard (1898) and for the example of linear regression with rounded data, e.g., by Dempster and Rubin (1983) and Beaton et al. (1976). Nevertheless, by imposing assumptions about the (random) behavior of the rounding error, the estimates may be corrected for the error in several sit-

uations. Many correction methods were developed for various statistical methods and different kinds of measurement errors. An extensive overview of modern measurement error models is given, for instance, in Carroll et al. (2006).

In the literature, further approaches explicitly modeling the mechanism that leads to the imprecise observations were suggested for various special cases of imprecise data and particular statistical methods. For example, in the case of completely missing values, the missingness mechanism can be described by a random quantity indicating whether the value is observed or not. Using this description of the uncertainty in the data set, Rubin (1976) defined the condition of Missing At Random (MAR), requiring that the fact that a value is missing must not depend on the unobserved value itself. Provided the missingness mechanism is uninformative in this sense, Rubin (1976) showed that valid likelihood-based inferences may be obtained ignoring the mechanism. Based on this result, many other suggestions to dealing with missing data were made, including sophisticated imputation methods. For more details on these methods, see, for instance, Little and Rubin (2002). Another common type of imprecise data constitute censored event or life times. These occur, for example, in data on incidence times of patients suffering from a certain disease or on the age at failure of technical devices, for which only lower or upper bounds (or both) are known, because some patients were sick for a certain time before the disease was diagnosed or because some devices were replaced before they failed. Here, the censoring mechanism can be modeled by a quantity indicating for each observation whether the actual time of interest was observed or whether a censoring time was observed as upper or lower bound to the actual time. Different statistical methods for analyzing data sets containing censored data were proposed, e.g., by Salibian-Barrera and Yohai (2008); Gómez et al. (2003); Lindsey (1998); Heitjan and Rubin (1990). Since there are many other practical settings where uncertain or partial information about some data values is available, Heitjan and Rubin (1991) generalize the MAR concept to other kinds of imprecise data and deduce a similar ignorability result. As Heitjan and Rubin (1991)

refer to imprecise observations like rounded values, censored life times, or general subsets of the observation space as coarse data, the corresponding condition is called Coarsening At Random (CAR). Generalizing the MAR condition, CAR requires that the coarsening mechanism is independent of the underlying precise value. Moreover, Heitjan and Rubin (1991) showed that likelihood-based inferences can be obtained from coarse data without explicitly accounting for the coarsening mechanism, if the latter respects the CAR condition. In practice, however, it is generally impossible to check whether this condition (or its special case MAR) is fulfilled or not. For a detailed discussion about the CAR condition, see, for example, Pötter (2008, Chapter 2).

All approaches to analyzing imprecise data discussed so far follow the idea that the mechanism leading to the imperfect observations is explicitly considered in the probability model underlying the statistical analysis and assumptions about the (random) behavior of this mechanism ensure to obtain a precise result. Yet, these assumptions are in some cases very restrictive and can never be verified in a practical setting. Another drawback of these approaches is that usually only one special type of uncertain information in the data can be considered at a time. As many different kinds of imprecision in data can be expressed by subsets of the observation space, a general methodology for the analysis of imprecise data should directly start with the observed sets.

A different methodology for regression analysis with interval data was proposed in Utkin and Coolen (2011). The proposed methodology yields a precise results by adopting either a minimin or a minimax strategy. It is extensively discussed in Chapter 5.

2.2 Approaches admitting an imprecise result

A simple approach could be not to aim at a precise result but to consider as the imprecise result of the statistical analysis the whole set of all precise estimates resulting from precise data sets that are compatible with the imprecise data. This approach can be generally applied to set-valued data,

no matter what the source of the imprecision is, and implies no assumption like CAR. It was proposed, for example, by Ferson et al. (2007); Gioia and Lauro (2005); Marino and Palumbo (2002). The set-valued result collects all precise estimates that would be obtained if the data were precisely observed at locations within the observed sets. However, it is not clear what inferences can be deduced from the imprecise result of this ad hoc approach, because it is not based on a statistical model for inference with imprecise data.

Another approach that allows an imprecise result and that can provide a foundation for the ad hoc approach is known as Partial Identification. This approach emerged during the past 25 years mainly in Econometrics and Biostatistics. Partial Identification is based on the idea that, if the analyzed variables are only imprecisely observed, only partial knowledge about the characteristics of interest can be obtained, avoiding strong assumptions about the coarsening mechanism like CAR. Hence, bounds for the value of a characteristic of interest are derived, in considering all probability measures with support on the imprecise data as possible probability distributions of the precise values given the imprecise observations. The resulting set is called identification or ignorance region for the characteristic and can be reliably estimated from the imprecise data. However, care has to be taken when evaluating a characteristic as partially identified in some setting. Without specifying in detail what imprecise data can be observed, this only means that the quantity of interest is in general not completely identified, but in some special cases the estimated identification region may actually be a singleton, e.g., if the variables are precisely observed with probability one, or may become a point as more and more data are observed. A thorough presentation of the main concepts of Partial Identification together with an overview of applications of this approach is provided by Manski (2003), while Manski and Tamer (2002) study the special case of regression with interval data in detail, and Horowitz and Manski (1995) discuss the distinction of the Partial Identification approach from Robust Statistics. In a practical analysis, it was initially suggested to estimate the probability distribution of the imprecise data by their empir-

ical distribution and to determine the identification region associated with this probability distribution, which corresponds, in fact, to the result of the ad hoc approach mentioned above. Once the estimate of the identification region is determined, the power of additional assumptions to reduce the size of the set-valued result can be investigated. Hence, Partial Identification methods are generally not intended to completely refrain from further assumptions about the unobserved data like CAR, but to become aware of their strength when imposing them in a particular analysis. In recent years, many other statistical methods were developed in the framework of Partial Identification, in particular, methods that allow taking also the estimation uncertainty into account. Different confidence regions for real-valued distribution characteristics or for regression parameters were proposed, for example, by Schollmeyer and Augustin (2013); Beresteanu et al. (2012); Beresteanu and Molinari (2008); Vansteelandt et al. (2006).

Apart from the Partial Identification approach, likelihood inference provides a very general and flexible framework for analyzing imprecise data. It directly allows accounting for the imprecision of the data as well as for the statistical uncertainty associated with the estimation on the basis of a finite number of observations. Considering a joint probability model for the precise variables and the imprecise observables, the imprecise data induce a (nonparametric or parametric) likelihood function on the set of considered probability measures, from which a profile likelihood function for some characteristic of the probability distribution of the variables of interest can be derived. Based on this profile likelihood function, confidence regions for the characteristic can be easily obtained by cutting the graph of the likelihood function at a chosen height determining the coverage level. This methodology is very general, because it can be applied to data sets containing at the same time precise and set-valued observations representing different kinds of data imprecision. Furthermore, no assumptions like CAR are necessary, however, additional assumptions about the coarsening mechanism can be considered by choosing a corresponding set of joint probability models of the analyzed situation. As the likelihood framework is very flexible, we used this inference framework in

combination with results about likelihood-based decisions from Cattaneo (2007) to develop a regression methodology for imprecise data in Cattaneo and Wiencierz (2012). The general methodology for likelihood inference with imprecise data was also proposed by Zhang (2010, 2009), but not yet considered in the context of regression analysis. How we employ this framework to develop a general methodology for regression analysis with imprecise data, is described in detail in the following chapter.

Chapter 3

Likelihood-based Imprecise Regression

In this chapter, the methodology for regression with imprecise data developed in Cattaneo and Wiencierz (2012) is presented in detail. As the approach is based on likelihood inference and usually yields a set-valued result, it is called Likelihood-based Imprecise Regression (LIR). The LIR approach is based on a general methodology for likelihood inference with imprecise data and is derived within the framework for likelihood-based decisions developed in Cattaneo (2013, 2007). The regression problem is thus formalized as a decision problem about which regression function best describes the relationship of interest in the light of the (possibly) imprecise observations. In the considered data situation, it is difficult to obtain a precise evaluation of each of the considered functions without imposing strong assumptions about the coarsening mechanism. To avoid such restrictions and to additionally take the statistical uncertainty into account, confidence regions for the loss associated with each regression function are considered, which can reliably be learned from the imprecise data. Thus, the aim of a LIR analysis is not to obtain a single estimated regression function at any price, but rather to describe the whole uncertainty involved in the regression problem with imprecise data.

3.1 The LIR methodology for precise data

In regression analysis, the relationship between some explanatory variables $X \in \mathcal{X} \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}$, and a response variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is investigated. Typically, it is supposed that $\mathcal{X} \times \mathcal{Y}$ is the Cartesian product of $d+1$ possibly unbounded intervals. The relationship of interest can be formalized by a function

$$f : \mathcal{X} \to \mathbb{R}.$$

The quantities X and Y are regarded as random variables, the joint random object is denoted by $V = (X, Y)$ with observation space $\mathcal{V} = \mathcal{X} \times \mathcal{Y}$.

Usually a sample of possible realizations $V_1, \ldots, V_n$, with $n \in \mathbb{N}$, is considered, from which it shall be inferred which of the functions in a certain predefined set $\mathcal{F}$ best describes the relationship between the variables of interest, X and Y. As commonly done, we assume that $V_1, \ldots, V_n$ are independent and identically distributed (i.i.d.) according to some probability measure P_V on $\mathcal{V}$.

The task of identifying the function $f \in \mathcal{F}$ that best describes the relationship of interest can be formulated as a decision problem with $\mathcal{F}$ being the set of possible decisions, $\mathcal{P}_V$ the set of probability distributions for V that are considered as possible models of the analyzed situation, and L the associated loss function on $\mathcal{F} \times \mathcal{P}_V$. The closer $L(f, P_V)$ is to zero for some $P_V \in \mathcal{P}_V$, the better the function f describes the relationship between X and Y, provided that P_V is the true model.

Most of the common loss functions in the regression context are expressed by means of the (absolute) residual, defined for each $f \in \mathcal{F}$ by

$$R_f = |Y - f(X)|.$$

If we consider a sample $V_1, \ldots, V_n$, with $V_i \sim P_V$, for all $i \in \{1, \ldots, n\}$, the corresponding residuals $R_{f,1}, \ldots, R_{f,n}$ are also i.i.d. random quantities

with probability distribution P_{R_f} given by

$$P_{R_f}(R_f \leq r) = \int_{\mathcal{V}} \mathbb{I}_{\{(x',y')=v' \in \mathcal{V} \,:\, |y'-f(x')| \leq r\}}(v)\, dP_V(v),$$

for all $r \in \mathbb{R}_{\geq 0}$, where $\mathbb{I}_{\mathcal{S}}$ denotes the indicator function of a set $\mathcal{S}$, defined on a suitable space. To avoid notational overload, we write throughout this thesis $P_V(R_f \leq r) = P_V(V \in \{(x',y') = v' \in \mathcal{V} : |y' - f(x')| \leq r\})$ instead of $P_{R_f}(R_f \leq r)$. As loss function, usually, some characteristic of the residuals' distribution is considered, for instance, a moment or a quantile. A famous example is the loss function given by $L(f, P_V) = \mathbb{E}(R_f^2)$, that is, by the second moment of the distribution of the residuals (under P_V), whose minimization corresponds to the regression method of Least Squares (LS). The LS solution is given by the regression function satisfying $f(x) = \mathbb{E}(Y|x)$ for all $x \in \mathcal{X}$, where $\mathbb{E}(Y|x)$ is the conditional expectation (under P_V) of Y given $X = x$.

Given the true probability measure P_V, the best description of the relationship between X and Y is the function minimizing $L(\cdot, P_V)$. However, usually P_V is unknown. Many regression methods overcome this problem by substituting P_V with the empirical distribution, after having obtained the data $V_1 = v_1, \ldots, V_n = v_n$. The empirical distribution denoted by $\hat{P}_V$ is the discrete distribution over $\mathcal{V}$ with probability mass $1/n$ at each observed point $v_1, \ldots, v_n$. In the fully nonparametric case where $\mathcal{P}_V$ corresponds to the set of all probability measures on $\mathcal{V}$, the empirical distribution is the maximum likelihood (ML) estimate of P_V. As pointed out in Cattaneo (2007, Section 1.3), the minimization of $L(\cdot, \hat{P}_V)$ leads to the ML estimate of f, if some weak regularity conditions are fulfilled. Therefore, it is reasonable to proceed in this way if one aims at a precise evaluation ignoring the involved uncertainties. Yet, we follow a more general approach to likelihood inference, where we make use of the information of the entire likelihood function instead of focusing only on its maximum. Moreover, we define the likelihood function generally as a function of the probability distribution, which allows considering also nonparametric models. This general approach to likelihood inference was also

adopted in Cattaneo (2007) and had formerly been suggested, for example, by Owen (1988).

Given the observations $V_1 = v_1, \ldots, V_n = v_n$, we define the induced (normalized) likelihood function $lik_V : \mathcal{P}_V \to [0,1]$ by

$$lik_V(P_V) = \frac{P_V(V_1 = v_1, \ldots, V_n = v_n)}{\sup_{P'_V \in \mathcal{P}_V} P'_V(V_1 = v_1, \ldots, V_n = v_n)}.$$

If P_V is a continuous probability measure, $v_1, \ldots, v_n$ can be replaced by small intervals around the observed values to ensure that $lik_V(P_V)$ is well defined. This is justified because continuous quantities can always be measured only with finite precision. Alternatively, in case that all $P_V \in \mathcal{P}_V$ are continuous, $lik_V(P_V)$ can be approximated by the ratio of the corresponding densities, as it is commonly done in Statistics.

Hence, the (normalized) likelihood function is given by the probabilities with which each $P_V \in \mathcal{P}_V$ would have predicted the observations relative to the probability assigned by the best-predicting probability model. Therefore, lik_V provides detailed information about which probability models in $\mathcal{P}_V$ are more plausible than others in the light of the available data. For any $\beta \in (0,1)$, $\mathcal{P}_V$ can be reduced to the set

$$\mathcal{P}_{V,>\beta} = \{P_V \in \mathcal{P}_V : lik_V(P_V) > \beta\}$$

of all probability distributions whose (normalized) likelihood exceeds the threshold β, i.e., which assign at least a certain probability to the observed values.

The set $\mathcal{P}_{V,>\beta}$ allows deriving likelihood-based confidence regions for some characteristic of the probability models considered. Define the characteristic g as a (possibly) set-valued function from $\mathcal{P}_V$ to a set $\mathcal{G} \subseteq \mathbb{R}$, that is, formally $g : \mathcal{P}_V \to 2^{\mathcal{G}} \backslash \{\varnothing\}$, where $2^{\mathcal{S}}$ denotes the power set of a set $\mathcal{S}$. For instance, g can be the function assigning to each probability measure P_V the corresponding value (or interval) $g(P_V)$ of a certain quantile of the distribution of the residuals associated with some $f \in \mathcal{F}$. For each $p \in (0,1)$, this p-quantile can be defined as any value $q \in \mathbb{R}$ such that

$P_V(R_f < q) \leq p \leq P_V(R_f \leq q)$, which is only unique if the corresponding cumulative distribution function is strictly increasing, i.e., if P_V is a continuous probability measure. As the LIR methodology is not restricted to this case, g is generally defined as a multi-valued mapping, where in case of the p-quantile $g(P_V) = \{q \in \mathbb{R} : P_V(R_f < q) \leq p \leq P_V(R_f \leq q)\}$. Then, for each $\beta \in (0,1)$, the set

$$\mathcal{G}_{>\beta} = \bigcup_{P_V \in \mathcal{P}_{V,>\beta}} g(P_V)$$

defines a likelihood-based confidence region with cutoff point β for the characteristic g. This set can alternatively be represented as

$$\mathcal{G}_{>\beta} = \{\gamma \in \mathcal{G} : lik_g(\gamma) > \beta\},$$

where $lik_g : \mathcal{G} \to [0,1]$ is the (normalized) profile likelihood function for g defined by

$$lik_g(\gamma) = \sup_{P_V \in \mathcal{P}_V : \gamma \in g(P_V)} lik_V(P_V)$$

(see also Cattaneo and Wiencierz, 2012, Lemma 1).

When $\mathcal{P}_V$ is a family of parametric probability distributions and g a corresponding parameter, the confidence region $\mathcal{G}_{>\beta}$ corresponds to all values γ of the characteristic of interest that would not be rejected in a likelihood ratio test of the simple hypothesis $H_0 : g = \gamma$ versus the alternative $H_1 : g \neq \gamma$. Under suitable regularity conditions, the likelihood ratio test statistic $-2\log(lik_g(\gamma))$ has an asymptotic χ^2-distribution with one degree of freedom, as shown by Wilks (1938). For the nonparametric case, where $\mathcal{P}_V$ is the set of all probability measures on $\mathcal{V}$ and g is some characteristic of these distributions, Owen (1988) derived the same asymptotic distribution of this test statistic, provided some regularity conditions are fulfilled. Hence, the asymptotic confidence level of $\mathcal{G}_{>\beta}$ is directly determined by β. For any $\beta \in (0,1)$, the asymptotic level of the likelihood-based confidence region $\mathcal{G}_{>\beta}$ is given by $F_{\chi^2_1}(-2\log(\beta))$, where $F_{\chi^2_1}$ is the cumulative

distribution function of the χ^2-distribution with one degree of freedom. The lower the cutoff point β is chosen, the higher the confidence level of $\mathcal{G}_{>\beta}$, for example, the choice of $\beta = 0.15$ corresponds to an asymptotic confidence level of approximately 95%, while $\beta = 0.5$ implies a level of about 76%.

In the context of the regression problem, the characteristic of interest is the loss associated with each regression function $f \in \mathcal{F}$. Hence, as g we consider the function-specific loss function L_f defined for all $f \in \mathcal{F}$ by $L_f(P_V) = L(f, P_V)$. That is, L is also considered to be a possibly multi-valued mapping here. Given some $\beta \in (0, 1)$, we obtain for each function $f \in \mathcal{F}$ a confidence region $\mathcal{C}_{f,>\beta}$ for the associated loss. In the example case of the loss function assigning to each pair (f, P_V) the p-quantile of the residuals' distribution, we have that $\mathcal{C}_{f,>\beta}$ is always an interval, but this is not necessarily true for other loss functions (see Owen, 1988). In the LIR methodology, we use the confidence region $\mathcal{C}_{f,>\beta}$ as the decision criterion for the regression problem. Since the cutoff point is the same for all $f \in \mathcal{F}$ in the same LIR analysis, we suppress β in the notation of the confidence regions in the following. Being a set-valued decision criterion, the confidence region for the loss induces only a partial order on $\mathcal{F}$, and therefore, cannot simply be minimized. Yet, it is possible to apply generalized decision rules or weak decision principles to obtain a (possibly set-valued) solution. For example, all regression functions that are not strictly dominated by another function can be considered as the imprecise result of the regression analysis. A function f strictly dominates another function f' if

$$\sup_{P_V \in \mathcal{P}_{V,>\beta}} L_f(P_V) < \inf_{P_V \in \mathcal{P}_{V,>\beta}} L_{f'}(P_V) \quad \Longleftrightarrow \quad \sup \mathcal{C}_f < \inf \mathcal{C}_{f'}.$$

The obtained set of functions can be interpreted as a confidence set for the true function describing the relationship between X and Y, thus, its extent reflects the amount of statistical uncertainty regarded in the analysis according to the choice of β.

3.2 The LIR methodology for imprecise data

Now, let us consider the situation in which it is impossible to observe the variables precisely, instead only partial information about $V_1, \ldots, V_n$ is available. The corresponding imprecise data are represented by the random sets $V_1^*, \ldots, V_n^*$ taking values in the set $\mathcal{V}^* \subseteq 2^{\mathcal{V}}$. The set-valued observations can be arbitrary subsets of $\mathcal{V}$, including as extreme cases actually precise observations (when $V_i^* = \{V_i\}$) and completely missing data (when $V_i^* = \mathcal{V}$). We assume that the joint random objects $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ are i.i.d. according to some probability measure $P \in \mathcal{P}$, where $\mathcal{P}$ is a subset of the set $\mathcal{P}_\varepsilon$ of all probability models satisfying

$$P(V \in V^*) \geq 1 - \varepsilon, \tag{3.1}$$

for some $\varepsilon \in [0, 1/2)$. This is a very general model for the considered data situation, according to which for each realization of the random variables of interest, X and Y, there is an unobservable precise version $V_i = v_i \in \mathcal{V}$ and an observable imprecise version $V_i^* = A_i \in \mathcal{V}^*$. In Figure 3.1, this idea is illustrated by means of an artificial data example. How the two versions are related is mainly determined by the model parameter ε corresponding to the upper bound to the probability of a wrong coarsening. The event $V_i \notin V_i^*$ might occur due to, for example, data processing errors or bad memory of respondents in a survey. Requiring $\varepsilon = 0$ in (3.1), as other approaches to the analysis of imprecise observations usually do, corresponds to assuming that the (imprecise) data were perfectly recorded. However, in many practical settings, such an assumption is not reasonable, hence, the general model for the imprecise data of the LIR approach is more flexible and allows accounting for measurement errors. In the fully nonparametric setting where $\mathcal{P} = \mathcal{P}_\varepsilon$, Assumption (3.1) does not even exclude informative coarsening. Stronger assumptions about the coarsening mechanism may also be included by the choice of an appropriate set $\mathcal{P}$.

On the basis of the model for the (unobserved) precise and (observed) imprecise data, we can, completely analogously to the case above, derive

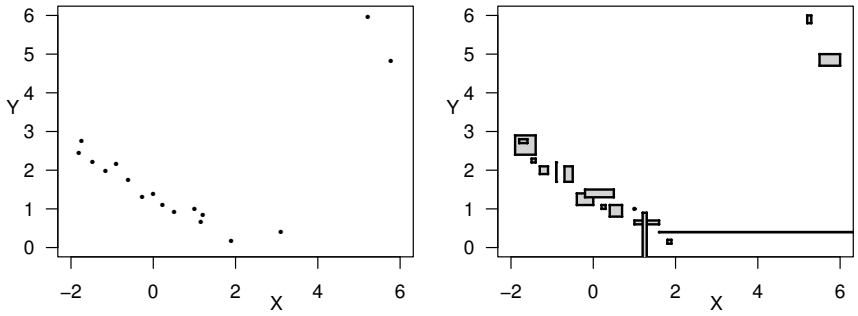

Figure 3.1: Precise (left) and imprecise (right) versions of a two-dimensional synthetic data set with $n = 17$. The imprecise data have varying amounts of imprecision: there is one actually precisely observed data point with $V_i^* = [1,1] \times [1,1] = \{(1,1)\}$, there are two line segments (one of which is unbounded towards $+\infty$ in the X-dimension), and finally, there are 14 rectangles of different sizes and shapes (one of which is unbounded towards $-\infty$ in the Y-dimension).

likelihood-based confidence regions for the characteristic of the marginal distribution of the precise data P_V that is regarded as loss function of the regression problem. Note that the aim of the regression analysis remains unchanged, that is, we still want to analyze the relationship between the (precise) quantities X and Y, only the quality of the available data is different now.

Consider that some imprecise observations $V_1^* = A_1, \ldots, V_n^* = A_n$ were made. The (normalized) likelihood function lik on $\mathcal{P}$ induced by these observations is defined by

$$\begin{aligned} lik(P) &= \frac{P(V_1^* = A_1, \ldots, V_n^* = A_n)}{\sup_{P' \in \mathcal{P}} P'(V_1^* = A_1, \ldots, V_n^* = A_n)} \\ &= \frac{\prod_{i=1}^n P_{V^*}(V_i^* = A_i)}{\sup_{P' \in \mathcal{P}} \prod_{i=1}^n P'_{V^*}(V_i^* = A_i)}, \end{aligned} \tag{3.2}$$

where P_{V^*} denotes the marginal distribution of the imprecise data associated with a probability model $P \in \mathcal{P}$. As the value of the likelihood function for each P is given (up to a multiplicative constant) by the probability with which this probability model had predicted the data at hand and since we only observed the imprecise data, the value of $lik(P)$ only

depends on the marginal distribution P_{V^*}. As in the precise data case, lik can be used to reduce $\mathcal{P}$ to the set

$$\mathcal{P}_{>\beta} = \{P \in \mathcal{P} : lik(P) > \beta\}$$

of all plausible probability models given the data, for some $\beta \in (0, 1)$.

Furthermore, likelihood-based confidence regions $\mathcal{G}_{>\beta}$ for some characteristic g of the probability distributions $P \in \mathcal{P}$ can be determined analogously to the precise data case, i.e., as $\mathcal{G}_{>\beta} = \{\gamma \in \mathcal{G} : lik_g(\gamma) > \beta\}$. However, it is more complicated to derive the (normalized) profile likelihood function for g here. This is because the characteristics considered as loss functions in the regression problem are usually characteristics of the distribution of the (unobservable precise) residuals $R_{f,i}$, with $i \in \{1, \ldots, n\}$, and thus, only depend on the marginal distribution P_V of the precise data, while the likelihood function is entirely determined by the marginal distribution P_{V^*} of the imprecise data. Hence, the uncertainty about the probability distribution P_V of the quantities of interest is more complex here. It is composed of two parts: on the one hand, there is the statistical uncertainty about the correct distribution P_{V^*} of the imprecise data, and on the other hand, there is the indetermination regarding which marginal probability distribution P_V of the precise data that is compatible with P_{V^*} is the correct one, which is due to the fact that the data are only imprecisely obtained. In general, the statistical uncertainty decreases as more data are observed, while the indetermination remains.

To deduce an expression for lik_g, we denote $g(P)$ by $g'(P_V)$ for all $P \in \mathcal{P}$ and we define an imprecise version g^* on $\mathcal{P}_{V^*}$ of the multi-valued mapping describing the characteristic of interest for all $P_{V^*} \in \mathcal{P}_{V^*}$ by

$$g^*(P_{V^*}) = \bigcup_{P_V \in [P_{V^*}]} g'(P_V), \tag{3.3}$$

where $[P_{V^*}]$ is the set of all probability distributions P'_V of the precise data corresponding to models $P' \in \mathcal{P}$ with marginal distribution $P'_{V^*} = P_{V^*}$ for the imprecise data. If we consider, for instance, the fully nonparametric

assumption $\mathcal{P} = \mathcal{P}_\varepsilon$ with $\varepsilon = 0$, for a fixed P_{V^*}, the set $[P_{V^*}]$ is composed of all marginal distributions P_V of the precise data satisfying for all measurable events $A \subseteq \mathcal{V}$

$$\begin{aligned} P_V(V \in A) &\geq \int_{\mathcal{V}^*} \mathbb{I}_{\{A' \in \mathcal{V}^* \,:\, A' \subseteq A\}}(\tilde{A})\, dP_{V^*}(\tilde{A}) \quad \text{and} \\ P_V(V \in A) &\leq \int_{\mathcal{V}^*} \mathbb{I}_{\{A' \in \mathcal{V}^* \,:\, A' \cap A \neq \varnothing\}}(\tilde{A})\, dP_{V^*}(\tilde{A}). \end{aligned} \tag{3.4}$$

The expressions on the right-hand side of these inequalities are often referred to as lower and upper probabilities or as belief and plausibility function, respectively. For a closer look at these concepts, see, for example, Destercke et al. (2008); Nguyen and Wu (2006); Smets (2005); Dempster (1968).

As the likelihood function lik only depends on P_{V^*}, we furthermore define by $lik^*(P_{V^*}) = lik(P)$ the (normalized) likelihood function lik^* on the set $\mathcal{P}_{V^*}$ of all marginal distributions of the imprecise data associated with the considered probability measures $P \in \mathcal{P}$. This definition permits expressing $\mathcal{G}_{>\beta}$ as

$$\mathcal{G}_{>\beta} = \bigcup_{P_{V^*} \in \mathcal{P}_{V^*} \,:\, lik^*(P_{V^*}) > \beta} g^*(P_{V^*}),$$

and deriving the (normalized) profile likelihood function $lik^*_{g^*}$ on $\mathcal{G}$ associated with g^* as

$$lik^*_{g^*}(\gamma) = \sup_{P_{V^*} \in \mathcal{P}_{V^*} \,:\, \gamma \in g^*(P_{V^*})} lik^*(P_{V^*}). \tag{3.5}$$

Now, it is straightforward to conclude that for all $\gamma \in \mathcal{G}$

$$lik_g(\gamma) = lik^*_{g^*}(\gamma) \tag{3.6}$$

(see also Cattaneo and Wiencierz, 2012, Lemma 2).

Hence, as in the precise data case, each possible regression function $f \in \mathcal{F}$ is evaluated by a set-valued decision criterion $\mathcal{C}_f$, corresponding

to a confidence region for some characteristic of the residuals' distribution associated with f. The underlying general methodology for likelihood inference with imprecise data via likelihood-based confidence regions was also proposed by Zhang (2010, 2009). As with the imprecision of the observations the confidence regions get larger, the asymptotic confidence level $F_{\chi^2_1}(-2\log(\beta))$ provides the lower bound to the actual asymptotic coverage probability of $\mathcal{C}_f$ here. With $\beta = 0.15$, for example, $\mathcal{C}_f$ is asymptotically a conservative 95% confidence set.

To solve the decision problem of the regression analysis, generalized decision rules or weak decision principles can be applied. Although decisions rules like, for example, the Likelihood-based Region Minimax (LRM) developed in Cattaneo (2007, Section 1.3) may allow singling out one optimal function on the basis of the confidence regions, we find that a precise solution is not appropriate in the context of the statistical analysis of imprecise data. On the contrary, the aim should be to describe the whole uncertainty about which regression function best describes the relationship of interest in the light of the (possibly) imprecisely observed data. Therefore, we suggest applying the dominance principle. Thus, we consider all regression functions that are not strictly dominated by another function as the imprecise result of the regression analysis. The resulting set of regression functions consists of all functions that are plausible descriptions of the relationship between X and Y, i.e., of all functions that cannot be excluded by the likelihood inference. As the width of the sets $\mathcal{C}_f$ is determined by the choice of the confidence level through β as well as by the degree of imprecision of the observations, also the extent of the set of plausible regression functions reflects not only the amount of statistical uncertainty according to the choice of β but also the indetermination due to the fact that the variables are only imprecisely observed.

To summarize, the LIR approach provides a very general framework for regression analysis with imprecise data. The imprecise data can be any subsets of the observation space of the variables of interest, including as special cases actually precise data and completely missing data. The LIR methodology consists in using likelihood-based confidence regions for some

characteristic of the residuals' distribution as set-valued decision criterion in the regression problem and in applying the dominance principle to these in order to extract all regression functions that are plausible in the light of the data. The confidence regions can be derived on the basis of a very general model connecting the precise data with the (possibly) imprecise observations. In the framework of LIR, the result of the regression analysis is in general set-valued, even in the special case where the data are in fact precisely observed. It consists of all descriptions of the relationship of interest that are not eliminated by the likelihood inference, and thus, the obtained result can be regarded as a confidence set for the true regression function. Thus, the idea is to directly obtain a result representing the uncertainties involved in the regression problem with imprecise data, instead of a single regression function one cannot be certain about.

However, if it is actually possible to obtain informative confidence regions, of course, depends on the concrete choices of $\mathcal{P}$ and L_f. In the following chapter, a regression method within the LIR framework is proposed, which is based on the fully nonparametric probability assumption and where a quantile of the distribution of the residuals is considered as loss function.

Chapter 4

A robust regression method within the LIR framework

In this chapter, the mathematical framework of a robust regression method derived from the general LIR approach is presented in detail. Then, some features of the robust LIR method are discussed with the help of an illustrative example, before its statistical properties are thoroughly investigated. Furthermore, the implementation of the robust LIR method is extensively discussed and its realization as a package for the statistical software environment R (R Core Team, 2013) is presented.

4.1 The robust LIR method

On the basis of the general LIR methodology described in the previous chapter, we developed in Cattaneo and Wiencierz (2012, Section 3) a robust regression method for imprecise data. In the robust LIR method, the p-quantile, for some $p \in (0, 1)$, of the residuals' distribution is considered as evaluation L_f of each regression function $f \in \mathcal{F}$ and the nonparametric distributional assumption $\mathcal{P} = \mathcal{P}_\varepsilon$, for some $\varepsilon \in [0, 1/2)$, is adopted.

With this general nonparametric assumption, where the data can be generated by any distribution satisfying Condition (3.1), it is generally impossible to obtain informative confidence regions for moments of the

residuals' distribution, because in this case the profile likelihood function is constant, and therefore, for each $\beta \in (0, 1)$, the confidence region is the entire set of possible values. This is due to the fact that moments are very sensitive to small contaminations. For example, if we consider for simplicity the situation in which only precise data are observed and where the loss L_f is given by the residual's expectation, it is easy to see that the profile likelihood function is constant equal one over the entire domain of R_f. Every value on the domain can be obtained as the expected value of a mixture distribution between the empirical distribution of the data and another distribution. As the latter can be an arbitrary distribution with very large or even infinite expected value, also the expectation of the mixture distribution may be arbitrarily high because it is given by the convex combination of the expectations of both involved probability measures. Furthermore, the mixture distribution can be arbitrarily similar to the empirical distribution when the weight of the contamination is small enough, and thus, assign practically the same probability to the observed data. Therefore, all possible values of the expectation are equally plausible and the profile likelihood function takes the value one for all of them.

In contrast to moments, quantiles are robust distribution characteristics, which are resistant to small changes in the probability distribution. For instance, the median being the $1/2$-quantile is the robust counterpart to the expectation being the first moment. Thus, informative confidence regions for quantiles can also be obtained in the fully nonparametric setting and these likelihood-based confidence regions are generally intervals (see, e.g., Owen, 2001, Section 3.6). That is why the robust LIR method combines the general nonparametric probability model with the loss function assigning to each pair (f, P) the p-quantile of the distribution under P of the residuals associated with the function f. In the same way the minimization of $\mathbb{E}(R_f^2)$ is associated with the LS regression method, the idea of minimizing the p-quantile is the rationale behind the Least Quantile of Squares (LQS, Rousseeuw and Leroy, 1987, Section 3.4) regression method, which is known to be very robust. Hence, the proposed LIR method can be regarded as a twofold generalization of LQS regression, on

the one hand, to imprecise data, and on the other hand, to directly accounting for statistical uncertainty in the result of the regression analysis.

Given the particular choices of $\mathcal{P}$ and L_f corresponding to the robust LIR method, we deduce an explicit formula for the profile likelihood function for the p-quantile, before we derive a simpler expression that permits determining directly the confidence regions $\mathcal{C}_f$, which are finally used to obtain the imprecise result of the regression analysis.

4.1.1 Profile likelihood for the p-quantile of the residuals' distribution

For each function $f \in \mathcal{F}$, let Q_f denote the function-specific loss function assigning to each probability measure $P \in \mathcal{P}$ the p-quantile, with $p \in (0,1)$, of the distribution of R_f under P and let $\mathcal{Q}_f \subseteq \mathbb{R}_{\geq 0}$ be the (possibly unbounded) interval of all possible values of this p-quantile. To derive the corresponding profile likelihood function $lik_{Q_f} : \mathcal{Q}_f \to [0,1]$ induced by some imprecise observations $V_1^* = A_1, \ldots, V_n^* = A_n$, we consider the vertical bands around f defined for each $q \in \mathbb{R}_{\geq 0}$ by

$$\overline{B}_{f,q} = \{(x,y) \in \mathcal{V} : |y - f(x)| \leq q\} \quad \text{and}$$

$$\underline{B}_{f,q} = \{(x,y) \in \mathcal{V} : |y - f(x)| < q\}.$$

A graphical illustration of the defined bands is given in Figure 4.1. To show why these bands provide a good starting point for finding lik_{Q_f}, we consider the case $p = 1/2$. For simplicity, we furthermore assume that the observations are actually precise, i.e., $A_i = \{v_i\}$ for all $i \in \{1, \ldots, n\}$, where n is an odd number, and $\varepsilon = 0$. Then, for any $f \in \mathcal{F}$, the nonparametric ML estimator of Q_f is the median of the empirical distribution of the observed residuals $r_{f,1}, \ldots, r_{f,n}$, which is given by $r_{f,(n+1/2)}$, where $r_{f,(i)}$ denotes the i-th smallest residual. Hence, for $q = r_{f,(n+1/2)}$, we know that $lik_{Q_f}(q) = 1$, and moreover, the band $\overline{B}_{f,q}$ around f can be characterized by the fact that it contains at least $n+1/2$ data. This characterization will be also useful later. Furthermore, recall the general definition of the

p-quantile of the residuals' distribution, which is any value $q \in \mathbb{R}_{\geq 0}$ such that $P_V(R_f < q) \leq p \leq P_V(R_f \leq q)$, where P_V is the marginal distribution of the precise data, which is also the marginal distribution of the imprecise data in the special case considered here. With the above definitions of the vertical bands, we can write this defining property as $P_V(V \in \underline{B}_{f,q}) \leq p \leq P_V(V \in \overline{B}_{f,q})$. Then, in the simple situation we consider here, determining $lik_{Q_f}(q)$ for a fixed $q \in \mathcal{Q}_f$ becomes simply counting the data inside the band $\overline{B}_{f,q}$ and those outside, because the probability measure attaining the highest likelihood value, is always the discrete probability measure distributing in equal parts probability mass p among the observations inside $\overline{B}_{f,q}$ and probability mass $1-p$ among those outside the band. By a similar reasoning, it is possible to derive the entire profile likelihood function for Q_f, also in the general case of imprecise observations. But before we put down a formal expression for lik_{Q_f}, we introduce some further definitions.

For each $f \in \mathcal{F}$, the corresponding functions $\overline{k}_f$ and $\underline{k}_f$ are defined, whose values are for all $q \in \mathbb{R}_{\geq 0}$ given by

$$\overline{k}_f(q) = |\{i \in \{1, \ldots, n\} \,:\, A_i \cap \overline{B}_{f,q} \neq \varnothing\}| \quad \text{and}$$

$$\underline{k}_f(q) = |\{i \in \{1, \ldots, n\} \,:\, A_i \subseteq \underline{B}_{f,q}\}|\,,$$

where $|\mathcal{S}|$ is the number of elements of a set $\mathcal{S}$. Hence, $\overline{k}_f(q)$ is the number of imprecise data intersecting the closed band $\overline{B}_{f,q}$ of vertical bandwidth $2\,q$ around the function f, while $\underline{k}_f(q)$ corresponds to the number of imprecise data completely included in the open band $\underline{B}_{f,q}$. From the definition follows that $\overline{k}_f$ and $\underline{k}_f$ are monotonically increasing functions of q, and $\underline{k}_f(q) \leq \overline{k}_f(q)$ for all $q \in \mathbb{R}_{\geq 0}$. Finally, the function $h : [0,1] \times (0,1) \to (0,1]$ is defined with

$$h(s,t) = \begin{cases} 1-t & \text{if} \quad s = 0, \\ \left(\dfrac{t}{s}\right)^s \left(\dfrac{1-t}{1-s}\right)^{1-s} & \text{if} \quad 0 < s < 1, \\ t & \text{if} \quad s = 1, \end{cases}$$

for all $s \in [0,1]$ and all $t \in (0,1)$. Before restating Theorem 1 of Cattaneo and Wiencierz (2012), which provides a precise expression for lik_{Q_f}, we recapitulate the setting considered in the robust LIR method.

- The aim is to find those of the functions $f \in \mathcal{F}$ that are plausible formalizations of the relationship between the variables X and Y.
- The random vector $V = (X, Y)$ summarizes the (precise) variables of interest, while V^* is the random set representing the imprecise observation of V.
- We assume that $(V, V^*) \sim P \in \mathcal{P} = \mathcal{P}_\varepsilon$, where $\mathcal{P}_\varepsilon$ is the set of all probability measures on $\mathcal{V} \times \mathcal{V}^*$ satisfying (3.1), for some $\varepsilon \in [0, 1/2)$.
- Then, there is a sample of n i.i.d. realizations of these random objects, with $V_i = v_i$ and $V_i^* = A_i$ for all $i \in \{1, \ldots, n\}$, but only $A_1, \ldots, A_n$ are observed.
- As loss function of the regression problem we consider the p-quantile, with $p \in (0,1)$, of the residuals' distribution. That is, for each function $f \in \mathcal{F}$ and some $P \in \mathcal{P}$, the loss associated with f is given by the p-quantile Q_f of the distribution (under P) of the (unobservable precise) residuals $R_{f,1}, \ldots, R_{f,n}$.
- According to the general LIR methodology explained in Chapter 3, likelihood-based confidence regions $\mathcal{C}_f$ for Q_f can be obtained from the imprecise data $A_1, \ldots, A_n$ and those are used as decision criterion of the regression problem, i.e., to finally identify the set of all plausible regression functions.

To obtain these confidence regions, the profile likelihood function lik_{Q_f} has to be determined, which only depends on the marginal distributions P_{V^*} of the imprecise data corresponding to the probability measures $P \in \mathcal{P}$.

Theorem 1. *For each $f \in \mathcal{F}$, the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of the residuals $R_{f,i}$ can be expressed as*

follows, for all $q \in \mathcal{Q}_f$*:*

$$lik_{Q_f}(q) = \begin{cases} h\left(\dfrac{\overline{k}_f(q)}{n},\, p-\varepsilon\right)^n & \text{if} \quad \overline{k}_f(q) < (p-\varepsilon)\, n, \\ 1 & \text{if} \quad [\underline{k}_f(q),\, \overline{k}_f(q)] \cap [(p-\varepsilon)\, n,\, (p+\varepsilon)\, n] \neq \varnothing, \\ h\left(\dfrac{\underline{k}_f(q)}{n},\, p+\varepsilon\right)^n & \text{if} \quad \underline{k}_f(q) > (p+\varepsilon)\, n. \end{cases}$$

Proof. Here, we only give the idea how this expression can be deduced, the complete and detailed proof can be found in Cattaneo and Wiencierz (2012, Section 3).

By Equations (3.2), (3.3), (3.5), and (3.6), we know that for each $q \in \mathcal{Q}_f$ the value of $lik_{Q_f}(q)$ is given by

$$lik_{Q_f}(q) = \sup_{\substack{P_{V^*} \in \mathcal{P}_{V^*}\,: \\ q \in \bigcup_{P_V \in [P_{V^*}]} Q_f(P_V)}} \frac{\prod_{i=1}^n P_{V^*}(V_i^* = A_i)}{\sup_{P'_{V^*} \in \mathcal{P}_{V^*}} \prod_{i=1}^n P'_{V^*}(V_i^* = A_i)},$$

that is, by the supremum of the likelihoods $lik^*(P_{V^*})$ of all marginal distributions of the imprecise data such that there is a compatible marginal distribution P_V of the precise data whose p-quantile of the corresponding distribution of the residuals covers the value q. Therefore, we at first look for the P_{V^*} assigning the highest possible probability to the observations at hand. This distribution is usually a discrete distribution with probability masses larger than zero only at the observed imprecise data. However, note that all other probability measures in $\mathcal{P}$ can be thought of as points (or line segments) under the curve of lik_{Q_f}, thus, the confidence regions $\mathcal{C}_f$ based on lik_{Q_f} cover the p-quantiles corresponding to all $P \in \mathcal{P}$ with $lik(P) > \beta$. For a fixed $q \in \mathcal{Q}_f$, the distribution attaining the highest likelihood can be obtained by looking for the allocation of probability mass on the observations at hand implying the highest possible likelihood, while respecting the restrictions imposed by the definition of the p-quantile of the residuals' distributions and the additional flexibility given by ε.

For all $q \in \mathcal{Q}_f$ such that $[\underline{k}_f(q),\, \overline{k}_f(q)] \cap [(p-\varepsilon)\, n,\, (p+\varepsilon)\, n] \neq \varnothing$, there is a precise data distribution compatible with the empirical distribution $\hat{P}_{V^*}$ of the imprecise data and covering q in its p-quantile, thus, for these q we get $lik_{Q_f}(q) = lik^*(\hat{P}_{V^*}) = 1$.

In the case of all $q \in \mathcal{Q}_f$ such that $\overline{k}_f(q) < (p-\varepsilon)\, n$, the probability distribution for the imprecise data attaining the highest likelihood value is the distribution that is as similar to the empirical distribution $\hat{P}_{V^*}$ as possible, given the restrictions imposed by the definition of the p-quantile of the residuals' distributions. To obtain this probability measure, the mass $p-\varepsilon$ is equally distributed among the $\overline{k}_f(q)$ imprecise data intersecting the closed band $\overline{B}_{f,q}$ and the remaining probability mass $1-p+\varepsilon$ is assigned to the $n - \overline{k}_f(q)$ imprecise data not intersecting $\overline{B}_{f,q}$. Then, computing the corresponding probability of the observed sample and dividing it by the highest possible value $(1/n)^n$ leads to the above expression.

In the case of all $q \in \mathcal{Q}_f$ such that $\underline{k}_f(q) > (p+\varepsilon)\, n$, the expression for $lik_{Q_f}(q)$ can be derived by analogous reasoning. □

These explanations suggest that, if, for some interval $[\underline{q}, \overline{q}] \subseteq \mathcal{Q}_f$, the borders of all bands $\overline{B}_{f,q}$ with $q \in [\underline{q}, \overline{q}]$ do not intersect any observation, then $lik_{Q_f}(q)$ is the same for all quantile values in this interval. This intuition can be formalized, leading to a simpler expression for the profile likelihood function of a p-quantile of the distribution of R_f. Figure 4.1 shows an example of the profile likelihood function on the basis of the imprecise data set introduced in Section 3.2.

For each function $f \in \mathcal{F}$, we consider lower and upper residuals denoted by $\underline{r}_{f,i}$ and $\overline{r}_{f,i}$, respectively, which are defined for all imprecise observations A_i, with $i \in \{1, \ldots, n\}$, by

$$\underline{r}_{f,i} = \inf_{(x,y)\in A_i} |y - f(x)| \quad \text{and} \quad \overline{r}_{f,i} = \sup_{(x,y)\in A_i} |y - f(x)|\,.$$

The intervals $[\underline{r}_{f,i}, \overline{r}_{f,i}]$ correspond to the imprecise observations of the residuals $R_{f,i}$, for all $i \in \{1, \ldots, n\}$. When we consider the ordered lower residuals $\underline{r}_{f,(1)} \leq \ldots \leq \underline{r}_{f,(n)}$ and the ordered upper residuals $\overline{r}_{f,(1)} \leq \ldots \leq \overline{r}_{f,(n)}$ and we define $\underline{r}_{f,(0)}$ and $\overline{r}_{f,(0)}$ as $\inf \mathcal{Q}_f$ and similarly $\underline{r}_{f,(n+1)}$

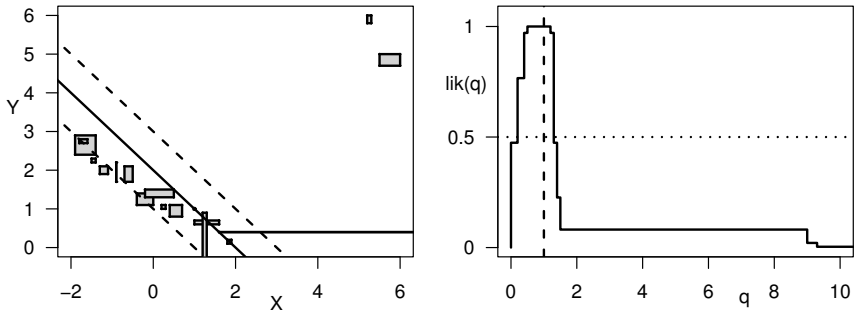

Figure 4.1: Linear regression function given by $f(x) = 2 - x$ for all $x \in \mathbb{R}$ (solid line) with band $\overline{B}_{f,q}$ (dashed lines) for $q = 1$ (left) and corresponding function lik_{Q_f} for the median with $\varepsilon = 0$ (right). For $q = 1$, $\overline{B}_{f,1}$ intersects $\overline{k}_f(1) = 14$ imprecise data, while $\underline{B}_{f,1}$ contains $\underline{k}_f(1) = 6$ imprecise data. As we have $[(p-\varepsilon)\,n,\,(p+\varepsilon)\,n] = [8.5, 8.5]$ and $[6, 14] \cap [8.5, 8.5] \neq \varnothing$, we obtain $lik_{Q_f}(1) = 1$ (right, dashed line). If the graph of lik_{Q_f} is cut at 0.5 (right, dotted line), the coordinates of the intersection points on the q-axis give the endpoints of the confidence interval $\mathcal{C}_f$ for $\beta = 0.5$.

and $\overline{r}_{f,(n+1)}$ as $\sup \mathcal{Q}_f$, we obtain $\underline{r}_{f,(0)} \leq \ldots \leq \underline{r}_{f,(n+1)}$ and $\overline{r}_{f,(0)} \leq \ldots \leq \overline{r}_{f,(n+1)}$. Finally, we define the integers $\underline{i}$ and $\overline{i}$ as $\underline{i} = \max\left(\lceil (p-\varepsilon)\,n \rceil,\, 0\right)$ and $\overline{i} = \min\left(\lfloor (p+\varepsilon)\,n \rfloor,\, n\right) + 1$, respectively, where $\underline{i} \in \{0, \ldots, n\}$ and $\overline{i} \in \{1, \ldots, n+1\}$ with $\underline{i} \leq \overline{i}$.

These definitions allow us to express the points of discontinuity of the functions $\underline{k}_f$ and $\overline{k}_f$ (restricted to the set $\mathcal{Q}_f$) as the ordered upper and lower residuals, respectively. Moreover, it is easy to see that for all $q \notin \{\overline{r}_{f,(0)}, \ldots, \overline{r}_{f,(n+1)}\}$ the function $\underline{k}_f$ is given by $\underline{k}_f(q) = i$ if $\overline{r}_{f,(i)} < q < \overline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n+1\}$, while for all $q \notin \{\underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(n+1)}\}$ the function $\overline{k}_f$ is given by $\overline{k}_f(q) = i$ if $\underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n+1\}$ (see also Cattaneo and Wiencierz, 2012, Lemma 3). Now, we can restate Corollary 1 of Cattaneo and Wiencierz (2012), providing the simpler expression for lik_{Q_f}.

Corollary 1. *For each $f \in \mathcal{F}$, the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of the residuals $R_{f,i}$ is a piecewise constant function, which can take at most $n+2$ different values.*

The points of discontinuity of lik_{Q_f}, including the endpoints of $\mathcal{Q}_f$, are (in ascending order, with possible repetitions)

$$\underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\underline{i})}, \overline{r}_{f,(\bar{i})}, \ldots, \overline{r}_{f,(n+1)},$$

and for all other values of $q \in \mathcal{Q}_f$,

$$lik_{Q_f}(q) =$$

$$\begin{cases} h\left(\dfrac{i}{n}, p-\varepsilon\right)^n & \text{if} \quad \underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)} \ \text{ with } \ i \in \{0, \ldots, \underline{i}-1\} \\ & \text{(when } \underline{i} \geq 1\text{)}, \\ 1 & \text{if} \quad \underline{r}_{f,(\underline{i})} < q < \overline{r}_{f,(\bar{i})}, \\ h\left(\dfrac{i}{n}, p+\varepsilon\right)^n & \text{if} \quad \overline{r}_{f,(i)} < q < \overline{r}_{f,(i+1)} \ \text{ and } \ i \in \{\bar{i}, \ldots, n\} \\ & \text{(when } \bar{i} \leq n\text{)}. \end{cases}$$

Proof. The above expression can easily be proved employing the formerly introduced definitions and Theorem 1. The complete and detailed proof can be found in Cattaneo and Wiencierz (2012, Section 3). □

4.1.2 Likelihood-based confidence regions for the p-quantile of the residuals' distribution

Furthermore, the following result can be derived, which was formulated as Corollary 2 in Cattaneo and Wiencierz (2012). It provides a method to determine for each cutoff point $\beta \in (0,1)$ the likelihood-based confidence regions $\mathcal{C}_f$ for the quantiles of the residuals' distribution used to evaluate the considered regression functions $f \in \mathcal{F}$.

Corollary 2. *If ε is sufficiently small and n is sufficiently large so that*

$$(\max\{p, 1-p\} + \varepsilon)^n \leq \beta \tag{4.1}$$

holds, then the integers

$$\underline{k} = \max\left\{k \in \{0, \ldots, \underline{i}-1\} \,:\, h\left(\frac{k}{n}, p-\varepsilon\right) \leq \sqrt[n]{\beta}\right\} \quad \textit{and}$$

$$\overline{k} = \min\left\{k \in \{\overline{i}, \ldots, n\} \,:\, h\left(\frac{k}{n}, p+\varepsilon\right) \leq \sqrt[n]{\beta}\right\}$$

are well-defined and satisfy

$$0 \leq \underline{k} < (p-\varepsilon)\, n \leq p\, n \leq (p+\varepsilon)\, n < \overline{k} \leq n,$$

and for each $f \in \mathcal{F}$, the likelihood-based confidence region with cutoff point β for the p-quantile of the distribution of the residuals $R_{f,i}$ is the nonempty interval

$$\mathcal{C}_f = \{q \in \mathbb{R}_{\geq 0} \,:\, [\underline{k}_f(q), \overline{k}_f(q)] \cap (\underline{k}, \overline{k}) \neq \varnothing\},$$

whose lower and upper endpoints are $\underline{r}_{f,(\underline{k}+1)}$ and $\overline{r}_{f,(\overline{k})}$, respectively.

Proof. Again, we here explain only the idea how this result can be obtained, the complete and detailed proof can be found in Cattaneo and Wiencierz (2012, Section 3).

For a given proportion $p \in (0,1)$, a fixed $\varepsilon \in [0, 1/2)$, and a chosen cutoff point $\beta \in [(\max\{p, 1-p\}+\varepsilon)^n, 1)$, we can use the first case of the expression of lik_{Q_f} in Theorem 1 to identify $\underline{k}$, the maximum number smaller than $(p-\varepsilon)\, n$ of observations that may intersect the closed band $\overline{B}_{f,q}$ when $lik_{Q_f}(q) \leq \beta$ holds. Then, according to the corresponding expression in Corollary 1, the index of the lower residual at which the profile-likelihood function at first exceeds the threshold β is $\underline{k}+1$. Likewise, we use the third case of the two expressions for lik_{Q_f} to obtain $\overline{k}$, the minimum number larger than $(p+\varepsilon)\, n$ of observations that must be included in the open band $\underline{B}_{f,q}$ when $lik_{Q_f}(q) \leq \beta$ holds, and thereby, we obtain also the index of the upper residual after which the function lik_{Q_f} jumps down below the threshold β again, which is $\overline{k}$. Thus, $\underline{r}_{f,(\underline{k}+1)}$ and $\overline{r}_{f,(\overline{k})}$ are the lower and upper endpoints of the interval of quantile values $q \in \mathbb{R}_{\geq 0}$ with

$lik_{Q_f}(q) > \beta$, which is the likelihood-based confidence region $\mathcal{C}_f$ for the p-quantile of the residuals' distribution. $\square$

The thus obtained intervals $\mathcal{C}_f$, for all functions $f \in \mathcal{F}$, contain all values $q \in \mathbb{R}_{\geq 0}$ for which the open band $\underline{B}_{f,q}$ contains at most $\overline{k} - 1$ imprecise data, while the closed band $\overline{B}_{f,q}$ intersects at least $\underline{k} + 1$ data. The width of the confidence intervals is determined on the one hand by the statistical uncertainty accounted for, which is reflected by the difference between $\underline{k} + 1$ and $\overline{k}$, and on the other hand by the degree of coarseness of the data, which is represented by the distinction between containing and intersecting imprecise data. If β is sufficiently close to one such that $\underline{k} = \underline{i} - 1 = \max\left(\lceil (p-\varepsilon)\, n \rceil\,,\, 0\right) - 1$ (when $\underline{i} \geq 1$) and $\overline{k} = \overline{i} = \min\left(\lfloor (p+\varepsilon)\, n \rfloor\,,\, n\right) + 1$ (when $\overline{i} \leq n$), the intervals $\mathcal{C}_f$ consist of all $q \in \mathbb{R}_{\geq 0}$ that are ML estimates of the p-quantile of the distribution of the residuals $R_{f,i}$. Even in this case and even if the data are in fact precisely observed when supposing that $\varepsilon = 0$ in Assumption (3.1), the confidence regions are usually proper intervals, because quantiles are in general not unique.

As mentioned already in the previous section, the intervals $\mathcal{C}_f$ are likelihood-based confidence regions whose asymptotic confidence level is bounded from below by $F_{\chi^2_1}(-2\log(\beta))$. Whether the endpoints $\underline{r}_{f,(\underline{k}+1)}$ and $\overline{r}_{f,(\overline{k})}$ are included in the confidence regions or not, depends on the observations at hand. For example, if all $A_1, \ldots, A_n$ are closed intervals, the confidence intervals $\mathcal{C}_f$ are closed, too.

4.1.3 Imprecise result of the robust LIR method

Given the confidence regions as interval-valued evaluations of each considered regression function, the final result of the regression problem can be determined. As explained in Chapter 3, the aim of the LIR analysis is not to obtain one single regression estimate, but to obtain a result that reflects the whole uncertainty about which of the considered functions best describes the relationship between the variables of interest. Therefore, the LIR result consists of all regression functions that are not strictly domi-

nated by another regression function with respect to their likelihood-based confidence intervals for some quantile of the residuals' distribution. According to the definition at the end of Section 3.2, a function $f \in \mathcal{F}$ is undominated if $\inf \mathcal{C}_f \leq \inf_{f' \in \mathcal{F}} \sup \mathcal{C}_{f'}$. The functions $f \in \mathcal{F}$ such that $\sup \mathcal{C}_f = \inf_{f' \in \mathcal{F}} \sup \mathcal{C}_{f'}$ are optimal according to the LRM rule mentioned in the previous subsection, and therefore, called LRM functions. If there is a unique LRM function, it is denoted by f_{LRM}. Furthermore, we define

$$\overline{q}_{LRM} = \inf_{f \in \mathcal{F}} \sup \mathcal{C}_f, \tag{4.2}$$

which corresponds to the upper endpoint of the confidence intervals of the LRM functions. If there is a unique LRM function and the corresponding confidence interval $\mathcal{C}_{f_{LRM}}$ is right-closed, that is, if $\overline{q}_{LRM} \in \mathcal{C}_{f_{LRM}}$, then the function f_{LRM} is characterized by the fact that the closed band $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q}$ that contains at least $\overline{k}$ imprecise data. This characterization can be extended to all LRM functions for which the corresponding confidence intervals are right-closed. Moreover, if all data are in fact precisely observed, we assume $\varepsilon = 0$ in (3.1), and there is a unique LRM function, then f_{LRM} corresponds to the LQS regression function for the $\overline{k}/n$-quantile.

Finally, provided Condition (4.1) holds, the set $\mathcal{U} \subseteq \mathcal{F}$ of all undominated regression functions can be defined as

$$\mathcal{U} = \{f \in \mathcal{F} : \inf \mathcal{C}_f \leq \overline{q}_{LRM}\} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}. \tag{4.3}$$

The whole set $\mathcal{U}$ constitutes the result of the robust LIR analysis. The undominated functions $f \in \mathcal{U}$ are geometrically characterized by the fact that the corresponding closed bands $\overline{B}_{f, \overline{q}_{LRM}}$ of width $2\,\overline{q}_{LRM}$ intersect at least $\underline{k} + 1$ imprecise data. Furthermore, the set $\mathcal{U}$ always contains the set $\mathcal{T}$ of all LQS regression functions for the $\overline{k}/n$-quantile obtained from precise data sets that are compatible with the imprecise data $A_1, \ldots, A_n$. To explain this, we take a closer look at the set $\mathcal{T}$, which can be defined as $\mathcal{T} = \{f \in \mathcal{F} : \exists\, v_1, \ldots, v_n$ with $v_i \in A_i\ \forall i \in \{1, \ldots, n\}$ and $r_{f,(\overline{k})} = \inf_{f' \in \mathcal{F}} r_{f',(\overline{k})}\}$. For each $f \in \mathcal{T}$, there is a compatible precise data set

implying the residuals $r_{f,1}, \ldots, r_{f,n}$ with $r_{f,(\overline{k})} = \inf_{f' \in \mathcal{F}} r_{f',(\overline{k})}$. Obviously, for all $i \in \{1, \ldots, n\}$ we have $\underline{r}_{f,i} \leq r_{f,i}$ and $r_{f,i} \leq \overline{r}_{f,i}$, hence, $\underline{r}_{f,(\overline{k})} \leq r_{f,(\overline{k})} \leq \overline{r}_{f,(\overline{k})}$. Moreover, by the definitions of $\underline{k}$ and $\overline{k}$ we know that $\underline{r}_{f,(\underline{k}+1)} \leq \underline{r}_{f,(\overline{k})}$. Altogether we obtain that $\underline{r}_{f,(\underline{k}+1)} \leq r_{f,(\overline{k})} = \inf_{f' \in \mathcal{F}} r_{f',(\overline{k})} \leq \inf_{f' \in \mathcal{F}} \overline{r}_{f',(\overline{k})}$, and thus, $f \in \mathcal{U}$. Note that the extent of the set $\mathcal{T}$ is only determined by the imprecision of the data. The result of the robust LIR method is usually larger than this, because the LIR analysis also accounts for the statistical uncertainty according to the chosen β (and because the LIR method does not use an interpolation scheme to obtain precise ML estimates for the quantiles, when these are not unique).

4.2 Illustration of the robust LIR method

In this section, some features of the robust LIR method are studied and illustrated. For this purpose, we consider again the artificial data set introduced in Chapter 3, consisting of 17 imprecise observations $V_1^* = A_1, \ldots, V_{17}^* = A_{17}$ of two variables $(X, Y) = V$ with $V \in \mathcal{V} = \mathbb{R}^2$. In this example, both variables are available as (possibly unbounded) intervals, that is, the imprecise data are of the type $V^* = [\underline{X}, \overline{X}] \times [\underline{Y}, \overline{Y}]$ with $\underline{X}, \overline{X}, \underline{Y}, \overline{Y} \in \mathbb{R} \cup \{-\infty, +\infty\}$. This kind of imprecise data represents the most relevant special case for statistical practice. It is sometimes referred to as interval-censored data, but we call it simply interval data. The nonparametric probability model underlying the robust LIR method implies that $P(\underline{X} \leq X \leq \overline{X} \text{ and } \underline{Y} \leq Y \leq \overline{Y}) \geq 1 - \varepsilon$, for some $\varepsilon \in [0, 1/2)$. As the majority of the data indicate a (decreasing) linear relationship, we here consider as possible regression functions all functions in the set $\mathcal{F} = \{f_{a,b} : (a, b) \in \mathbb{R}^2\}$ of linear functions $f_{a,b} : \mathbb{R} \to \mathbb{R}$, defined by $f_{a,b}(x) = a + b\,x$ for all $x \in \mathcal{X}$. However, note that the robust LIR method is not restricted to linear regression, on the contrary, the theoretical framework allows considering arbitrary functions to describe the relationship between the analyzed variables. Furthermore, we here focus on the regression method with $p = 1/2$, where the median of the probability distribution of R_f is minimized.

With the particular configuration of the robust LIR method and the interval data considered here, given a fixed $\varepsilon \in [0, 1/2)$ and a chosen $\beta \in (0, 1)$ that satisfies Condition (4.1), the likelihood-based confidence regions $\mathcal{C}_f$, for all $f \in \mathcal{F}$, are given by the closed intervals $[\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$ (see Corollary 2). According to Equation (4.3), the set $\mathcal{U}$ of undominated functions is the set of all functions f satisfying $\underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}$. As the considered regression functions in the set $\mathcal{F}$ are linear functions indexed by two parameters $(a, b) \in \mathbb{R}^2$, the set $\mathcal{U}$ can be alternatively represented by the corresponding subset of the parameter space $\mathbb{R}^2$. Hence, we furthermore define the set

$$\mathcal{U}' = \left\{(a, b) \in \mathbb{R}^2 \,:\, f_{a,b} \in \mathcal{U}\right\}, \tag{4.4}$$

which contains all parameter combinations associated with the undominated functions. This set can be described as the union of finitely many (possibly unbounded) polygons as it is explained in Section 4.3 (see also Cattaneo and Wiencierz, 2013, Section 3).

For the case of simple linear regression with interval data that is considered here, the robust LIR method is implemented in the package `linLIR` (Wiencierz, 2013) for the statistical software environment R (R Core Team, 2013) and it is presented in detail in Section 4.3. Hence, all results and graphs in this section are obtained by using the `linLIR` package. Figure 4.2 shows the result of the regression analysis of the example data set for the choice $\beta = 0.5$ and under the assumption $\varepsilon = 0$. Most of the 500 undominated regression lines plotted in the left graph are decreasing functions, but slightly increasing lines are also present, which is confirmed by the corresponding subset of parameter combinations in the right graph. Furthermore, we find a unique function $f \in \mathcal{F}$ such that $\overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}$, which is indicated by a black line or point. Next, we examine how different choices of β and different assumptions about ε in (3.1) affect the regression's result.

Different choices of $\beta \in [(1/2 + \varepsilon)^{17}, 1)$, i.e., satisfying Condition (4.1), imply different confidence levels of the interval estimates $\mathcal{C}_f$. For example,

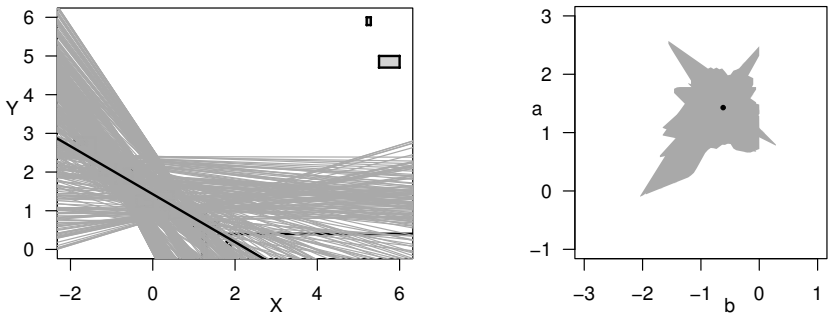

Figure 4.2: Draft of the set of undominated regression functions (left, 500 randomly chosen functions) and set of corresponding parameter combinations (right), for $\beta = 0.5$ and $\varepsilon = 0$. The black line or point indicates the LRM function.

under the assumption $\varepsilon = 0$, a likelihood-based confidence interval with cutoff point $\beta = 0.5$ would be a conservative 76% confidence interval for the median of the residuals' distribution. A high confidence level of the interval estimates of the median of the absolute residuals requires a small choice of β and vice versa. Thus, the higher β, the lower the confidence level and consequently the narrower the set of undominated regression functions. In a practical setting, confidence in and inferential strength of the result have to be balanced in light of the analyzed question and the purpose of the analysis in order to choose an appropriate value for β. In Figure 4.3, different results of LIR analyses with other choices of β are displayed. For a low cutoff point such as $\beta = 0.15$, the regression's result is very imprecise, admitting different directions of the relationship between the analyzed variables. In contrast to that, a high cutoff point such as $\beta = 0.8$ leads to a less imprecise result containing no increasing lines.

In fact, the sets of undominated regression functions for different values of the cutoff point β are nested. To explain this, we consider $\beta_1 > \beta_2$ and the corresponding resulting sets $\mathcal{U}_{\beta_1}$ and $\mathcal{U}_{\beta_2}$. For $\beta_1 > \beta_2$, we have that $\underline{k}_1 \geq \underline{k}_2$ and $\overline{k}_1 \leq \overline{k}_2$. The latter implies that $\overline{q}_{LRM,1} \leq \overline{q}_{LRM,2}$, which means that for each $f \in \mathcal{F}$ we have $\overline{B}_{f,\overline{q}_{LRM,2}} \supseteq \overline{B}_{f,\overline{q}_{LRM,1}}$. As for each $f \in \mathcal{U}_{\beta_1}$ the closed band $\overline{B}_{f,\overline{q}_{LRM,1}}$ intersects at least $\underline{k}_1 + 1$ imprecise data

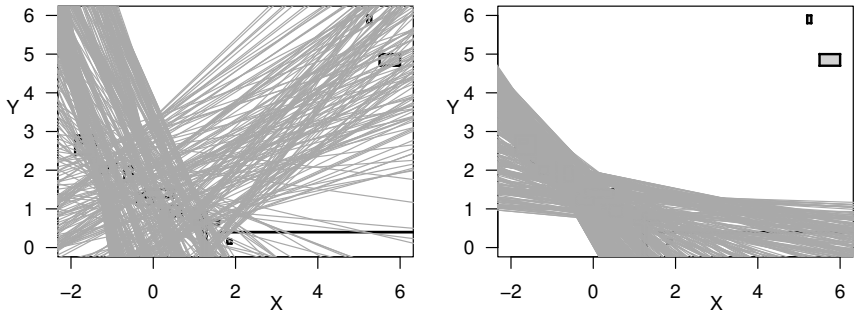

Figure 4.3: Drafts of the sets of undominated regression functions (500 randomly chosen functions) for $\varepsilon = 0$ and two different choices of β, namely $\beta = 0.15$ (left) and $\beta = 0.8$ (right), corresponding to the confidence levels 95% and 50%, respectively.

and $\underline{k}_1 \geq \underline{k}_2$, the band $\overline{B}_{f,\overline{q}_{LRM,2}}$ also intersects at least $\underline{k}_2 + 1$ imprecise data, and therefore, $f \in \mathcal{U}_{\beta_2}$.

In addition to the choice of the cutoff point of the likelihood, also $\varepsilon \in [0, 1/2)$ has to be set a priori. According to (3.1), ε is the upper bound to the probability that an imprecise observation does not contain the correct precise value in the (nonparametric) probability model underlying the presented LIR method. In most approaches to analyzing imprecise data, ε is assumed to be zero, but there may be situations in which the analyst has concerns about the correctness of the imprecise data. For example, in the case of survey data, there are many different sources for biases that should be accounted for in the analysis of such data, at least with a small probability. Hence, the consideration of an $\varepsilon > 0$ means to account for some more uncertainty about the data in addition to the indetermination issuing from the coarseness of the data.

It follows directly from definitions of $\underline{k}$ and $\overline{k}$ in Corollary 2 that increasing ε has the same effect on the width of the confidence intervals as decreasing β, since in both cases, $\underline{k}$ decreases and $\overline{k}$ increases. Thus, the worse the assumed data quality, the more imprecise the result. If we choose $\beta = 0.5$ in our example and assume $\varepsilon = 1/10$, we obtain the same result as for $\beta = 0.15$ assuming $\varepsilon = 0$, shown in Figure 4.3, because in

both cases $\underline{k} = 4$ and $\overline{k} = 13$. However, the interpretation is different. While in the case of increasing β for a fixed ε the amount of statistical uncertainty reflected by the result is reduced, in the case of increasing ε for a fixed β the assumptions of the underlying nonparametric probability model are weakened.

We showed how different choices of the confidence level and different assumptions about the correctness of the (imprecise) data are reflected in the regression's result. In order to illustrate how varying degrees of imprecision of the data are represented in the result of a LIR analysis, we compare the above results with those obtained from an actually precise data set compatible with the imprecise data of the example data set and from a compatible data set where only Y is imprecisely observed. Both data sets are displayed in Figure 4.4, where the left data set is the same as in Figure 3.1 and the right one is obtained by combining the precise values for X of this data set with the imprecise values for Y.

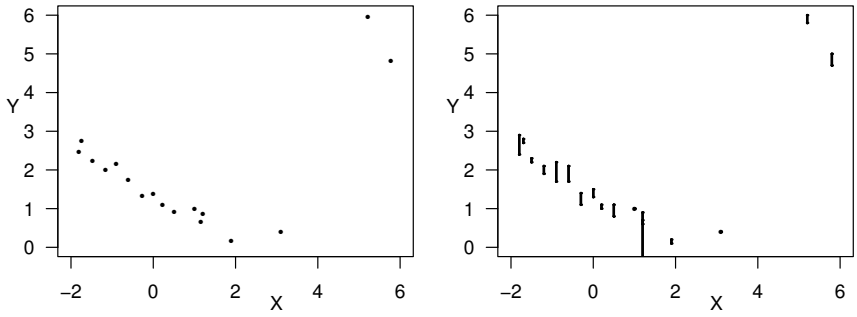

Figure 4.4: Precise data set compatible with the example data (left) and compatible data set where only Y is imprecisely observed (right).

For the LIR analyses of these less imprecise data sets, we assume that $\varepsilon = 0$ and we choose $\beta = 0.15$ to obtain a reliable result. The resulting sets of undominated functions are shown in Figure 4.5. Under the here adopted assumption that all observations are correct, the analysis of the actually precise data set leads to the most determined result, admitting only decreasing lines. In the case where Y is imprecisely observed, the set of undominated functions admits different directions of the relationship,

although the extent of $\mathcal{U}$ is visibly smaller than in the case where also X is observed as intervals, which was displayed in Figure 4.3 (left). If β was chosen close to one in order to obtain the ML estimate, the extent of the set $\mathcal{U}$ would only reflect the imprecision of the observations. However, even in the case of the actually precise data set together with the assumption $\varepsilon = 0$, we obtain in general an imprecise result in the present situation. That is because, given $p = 1/2$ and $n = 17$, for β close to one we have $\underline{k} = \lceil p\,n \rceil - 1 = 8$ and $\overline{k} = \lfloor p\,n \rfloor + 1 = 9$, implying that $\mathcal{C}_f = [\underline{r}_{f,(8+1)}, \overline{r}_{f,(9)}]$ for all $f \in \mathcal{F}$. Furthermore, when the variables are in fact precisely observed, $\underline{r}_{f,(i)} = \overline{r}_{f,(i)} = r_{f,(i)}$ for all $i \in \{1, \ldots, n\}$, and thus, $\mathcal{C}_f = \{r_{f,(9)}\}$ for all $f \in \mathcal{F}$. Hence, $\mathcal{U}$ consists of all functions f with $r_{f,(9)} = \inf_{f' \in \mathcal{F}} r_{f',(9)} = \overline{q}_{LRM}$, and in general, there can be more than one function attaining $\overline{q}_{LRM}$.

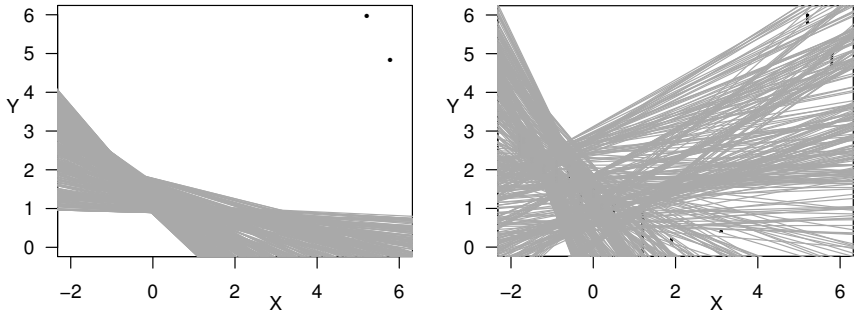

Figure 4.5: Sets of undominated regression functions (500 randomly chosen functions) for the compatible data sets where X and Y are in fact precisely observed (left) and where only X is precisely observed (right), for $\beta = 0.15$ and $\varepsilon = 0$.

In this section, we investigated the impact of different assumptions about the data generating process, amounts of statistical uncertainty selected, and degrees of coarseness of the data on the result of the robust LIR analysis. Both aspects of the uncertainty of a statistical analysis of imprecise data, statistical uncertainty and indetermination, are crucial and should be reflected in the result. Within the LIR framework both parts of the uncertainty are expressed in the same way, that is, they determine the extent of the generally imprecise result of the LIR analysis. Further properties of the robust LIR method are investigated in Section 4.4.

4.3 Implementation of the robust LIR method for linear regression with interval data

This section deals with the implementation of the robust LIR method. To determine the set-valued result of the robust LIR analysis, is a demanding problem. Even in the very particular case in which the robust LIR method reduces to the standard LQS regression method with quantile $\overline{k}/n$, it is a challenging task, as the objective function of the minimization problem is neither differentiable nor convex (see, e.g., Watson, 1998). Thus, common optimization techniques cannot be applied. However, for the LQS regression method, an exact algorithm was developed (Rousseeuw and Leroy, 1987, Chapter 5), which was further studied and improved, for example, in Watson (1998); Stromberg (1993); Steele and Steiger (1986).

For the robust LIR method, the implementation task is even more challenging, because the objective function and the solution are in general set-valued. Therefore, we suggested a first implementation based on a grid search over the space of parameters identifying the considered regression functions in Cattaneo and Wiencierz (2012) and we applied a random search to determine the result of the robust LIR analysis in Cattaneo and Wiencierz (2011). Then, for the special case of simple linear regression with interval data, we derived an exact algorithm in Cattaneo and Wiencierz (2013), based on the ideas set out in Wiencierz and Cattaneo (2012). This algorithm consists of two parts: at first, the smallest upper endpoint $\overline{q}_{LRM}$ of all confidence regions for the p-quantile of the residuals' distribution is determined, and then, it is used in the second part to identify the set of all undominated regression functions. The first part of the algorithm generalizes the initial algorithm for LQS regression and the entire algorithm has the same computational complexity of $O(n^3 \log n)$ (see, e.g., Steele and Steiger, 1986). Moreover, as it was done with the initial LQS algorithm, it is possible to adapt the algorithm for robust simple linear LIR to multiple linear regression and also to other kinds of imprecise data than intervals. In the following subsections, the exact algorithm is deduced in detail and its realization in the R package `linLIR` (Wiencierz, 2013) is presented.

4.3.1 An exact algorithm for simple linear regression

Here, we consider a similar setting as in Section 4.2, where the presumably linear relationship between two real-valued variables, X and Y, is analyzed on the basis of imprecise observations, which are of the form of (possibly unbounded) intervals for each of the variables. To determine the set-valued result of the robust LIR analysis in this situation, we developed an exact algorithm by exploiting the geometrical characterizations of the LRM functions and of the undominated functions. Before we derive the exact algorithm, we recapitulate the core elements of the robust LIR method in this special case.

- The relationship between the variables $X \in \mathcal{X} = \mathbb{R}$ and $Y \in \mathcal{Y} = \mathbb{R}$ is investigated.

- As possible descriptions of the relationship between X and Y we consider the set $\mathcal{F}$ consisting of all linear functions $f_{a,b} : \mathcal{X} \to \mathbb{R}$, with $f_{a,b}(x) = a + b\,x$, for all $x \in \mathcal{X}$, where $(a, b) \in \mathbb{R}^2$.

- In the present case, the random set V^* representing the imprecise observation of $V = (X, Y)$ can take as values only rectangles formed by closed (possibly unbounded) intervals, i.e., $V^* = [\underline{X}, \overline{X}] \times [\underline{Y}, \overline{Y}]$, with $\underline{X}, \overline{X}, \underline{Y}, \overline{Y} \in \mathbb{R} \cup \{-\infty, +\infty\}$.

- We assume that $(V, V^*) \sim P \in \mathcal{P} = \mathcal{P}_\varepsilon$, where $\mathcal{P}_\varepsilon$ is the set of all probability measures satisfying (3.1), that is, all distributions P' with $P'(\underline{X} \leq X \leq \overline{X} \text{ and } \underline{Y} \leq Y \leq \overline{Y}) \geq 1 - \varepsilon$, for some $\varepsilon \in [0, 1/2)$.

- Then, there is a sample of n i.i.d. realizations of the random objects V and V^*, with $V_i = v_i$ and $V_i^* = [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ for all $i \in \{1, \ldots, n\}$, but only the imprecise realizations $[\underline{x}_1, \overline{x}_1] \times [\underline{y}_1, \overline{y}_1], \ldots, [\underline{x}_n, \overline{x}_n] \times [\underline{y}_n, \overline{y}_n]$ are observed.

- As stated in Section 4.1, in the robust LIR method, we consider as the loss associated with a possible regression function $f \in \mathcal{F}$ the p-quantile Q_f, with $p \in (0, 1)$, of the distribution of the (unobservable precise) residuals $R_{f,1}, \ldots, R_{f,n}$.

- On the basis of the imprecise data, likelihood-based confidence regions $\mathcal{C}_f$ for Q_f can be obtained as described in Section 4.1, applying the general LIR methodology explained in Chapter 3.

- In the present case of interval data, given some choice of β satisfying Condition (4.1), for each $f \in \mathcal{F}$, the confidence region $\mathcal{C}_f$ is the interval $[\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$, where the integers $\underline{k}$ and $\overline{k}$ depend on p, n, ε, and β (see Corollary 2).

- Finally, according to (4.3), the imprecise result of the robust LIR analysis is given by the set $\mathcal{U} = \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$, where $\overline{q}_{LRM} = \inf_{f \in \mathcal{F}} \overline{r}_{f,(\overline{k})}$ by its definition in (4.2).

Hence, to determine the set of all undominated functions, we have to find the smallest upper endpoint $\overline{q}_{LRM}$ of all confidence intervals $\mathcal{C}_f$, with $f \in \mathcal{F}$, before all functions f' with $\underline{r}_{f',(\underline{k}+1)} \leq \overline{q}_{LRM}$ can be identified.

Part 1: Determining $\overline{q}_{LRM}$

According to the explanations in Subsection 4.1.3, the LRM functions are characterized by the fact that the closed bands $\overline{B}_{f,\overline{q}_{LRM}}$ around them have the thinnest bandwidth of all bands $\overline{B}_{f',q}$, with $f' \in \mathcal{F}$ and $q = \overline{r}_{f',(\overline{k})}$, that is, containing at least $\overline{k}$ imprecise data. Thus, in order to obtain $\overline{q}_{LRM}$, we need to find the linear functions minimizing this bandwidth. For a given slope $b \in \mathbb{R}$, the corresponding intercept value $a \in \mathbb{R}$ such that $\overline{r}_{f_{a,b},(\overline{k})}$ becomes minimal can easily be found. Moreover, similar to the results obtained by Stromberg (1993) and by Steele and Steiger (1986) for the LQS method, it can be deduced that, for an LRM function f, some of the $\overline{k}$ imprecise data included in $\overline{B}_{f,\overline{q}_{LRM}}$ touch the boundaries of the closed band in at least two different points. As the boundaries of the bands are parallel to the central lines, the slopes of the LRM functions are either zero or determined by the corresponding vertices of two bounded imprecise observations included in the band. Hence, to identify $\overline{q}_{LRM}$, it suffices to consider as possible LRM functions all functions with slopes given by the four slopes between the corresponding vertices of each pair of (nonidentical) bounded imprecise observations or zero and corresponding optimal

intercepts. In this way, the general minimization problem is reduced to a discrete problem with at most $4\binom{n}{2}+1$ possible solutions.

Of course, the result also depends on the degree of coarseness of the data. If there are less than $\overline{k}$ imprecise data that are bounded with respect to Y, we obtain $\overline{r}_{f,(\overline{k})} = +\infty$ for all $f \in \mathcal{F}$. Therefore, in this case $\overline{q}_{LRM} = +\infty$ and $\mathcal{U} = \mathcal{F}$. If all observations are bounded with respect to Y, but there are less than $\overline{k}$ imprecise data that are bounded with respect to X, the only candidate slope is zero, because in this situation only bands around a horizontal line can include at least $\overline{k}$ imprecise data. More generally, if there are less than $\overline{k}$ imprecise observations such that the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is bounded, either the only possible slope is zero or we obtain $\overline{q}_{LRM} = +\infty$.

To formalize this, let $\mathcal{D} \subseteq \{1, \ldots, n\}$ be the set of indices of those imprecise observations for which $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is bounded. If $|\mathcal{D}| < \overline{k}$, the set $\mathcal{B}$ of all possible slopes of the LRM functions is the singleton $\{0\}$, else (i.e., if $|\mathcal{D}| \geq \overline{k}$) the set $\mathcal{B}$ is given by

$$\begin{aligned}
\mathcal{B} = &\left\{ \frac{\overline{y}_i - \overline{y}_j}{\underline{x}_i - \underline{x}_j} \,:\, (i,j) \in \mathcal{D}^2 \text{ and } \underline{x}_i > \underline{x}_j \text{ and } \overline{y}_i > \overline{y}_j \right\} \\
&\cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\underline{x}_i - \underline{x}_j} \,:\, (i,j) \in \mathcal{D}^2 \text{ and } \underline{x}_i > \underline{x}_j \text{ and } \underline{y}_i < \underline{y}_j \right\} \\
&\cup \left\{ \frac{\overline{y}_i - \overline{y}_j}{\overline{x}_i - \overline{x}_j} \,:\, (i,j) \in \mathcal{D}^2 \text{ and } \overline{x}_i > \overline{x}_j \text{ and } \overline{y}_i < \overline{y}_j \right\} \\
&\cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\overline{x}_i - \overline{x}_j} \,:\, (i,j) \in \mathcal{D}^2 \text{ and } \overline{x}_i > \overline{x}_j \text{ and } \underline{y}_i > \underline{y}_j \right\} \\
&\cup \{0\}\,.
\end{aligned}$$

Furthermore, we define for each $b \in \mathbb{R}$ and each $i \in \{1, \ldots, n\}$ the transformed data $[\underline{z}_{b,i}, \overline{z}_{b,i}]$ whose endpoints are given by

$$\underline{z}_{b,i} = \begin{cases} \underline{y}_i - b\,\underline{x}_i & \text{if} \quad b < 0, \\ \underline{y}_i & \text{if} \quad b = 0, \\ \underline{y}_i - b\,\overline{x}_i & \text{if} \quad b > 0, \end{cases} \quad \text{and} \quad \overline{z}_{b,i} = \begin{cases} \overline{y}_i - b\,\overline{x}_i & \text{if} \quad b < 0, \\ \overline{y}_i & \text{if} \quad b = 0, \\ \overline{y}_i - b\,\underline{x}_i & \text{if} \quad b > 0. \end{cases}$$

As usual, $\underline{z}_{b,(1)}, \ldots, \underline{z}_{b,(n)}$ and $\overline{z}_{b,(1)}, \ldots, \overline{z}_{b,(n)}$ denote the ordered lower and upper endpoints, respectively. Furthermore, for each $b \in \mathbb{R}$ and each $j \in \{1, \ldots, n - \overline{k} + 1\}$, we denote by $\overline{z}_{b,[j]}$ the $\overline{k}$-th smallest value among those $\overline{z}_{b,i}$ for which $\underline{z}_{b,i} \geq \underline{z}_{b,(j)}$. By means of these definitions, the above explanations can be summarized in the following theorem, which constitutes Theorem 1 of Cattaneo and Wiencierz (2013).

Theorem 2. *If there are less than $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded, then*

$$\overline{q}_{LRM} = +\infty,$$
$$\{f \in \mathcal{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} = \mathcal{F}.$$

Otherwise (i.e., when there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded),

$$\overline{q}_{LRM} = \frac{1}{2} \min_{(b,j) \in \mathcal{B} \times \{1, \ldots, n - \overline{k} + 1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}),$$

$$\{f \in \mathcal{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} \supseteq$$
$$\left\{ f_{a',b'} : (b', j') \in \operatorname*{arg\,min}_{(b,j) \in \mathcal{B} \times \{1, \ldots, n - \overline{k} + 1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}) \right.$$
$$\left. \text{and} \quad a' = \frac{1}{2} \left(\underline{z}_{b',(j')} + \overline{z}_{b',[j']} \right) \right\},$$

where the set on the left-hand side is infinite when the inclusion is strict. However, the inclusion is certainly an equality when the following condition is satisfied: if there is a pair $(i, j) \in \mathcal{D}^2$ such that $\underline{x}_i = \overline{x}_j$ and $\max\{\overline{y}_i, \overline{y}_j\} - \min\{\underline{y}_i, \underline{y}_j\} = 2\,\overline{q}_{LRM}$, then $i \neq j$ and the two intervals $[\underline{y}_i, \overline{y}_i]$ and $[\underline{y}_j, \overline{y}_j]$ are nested (i.e., either $[\underline{y}_i, \overline{y}_i] \subseteq [\underline{y}_j, \overline{y}_j]$, or $[\underline{y}_j, \overline{y}_j] \subseteq [\underline{y}_i, \overline{y}_i]$).

Proof. The first part of this theorem can easily be proved by the argumentation above, while the proof of the second part requires further arguments, some of which are briefly sketched here. The complete and detailed proof can be found in Cattaneo and Wiencierz (2013, Subsection A.1).

The second part of Theorem 2 regards the situation in which there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded. In this case, $\overline{q}_{LRM} < +\infty$, because $\overline{r}_{f,(\overline{k})} < +\infty$ at least for the constant functions in $\mathcal{F}$. To obtain $\overline{q}_{LRM}$, we have to minimize $\overline{r}_{f_{a,b},(\overline{k})}$ over all linear functions $f_{a,b} \in \mathcal{F}$ with $(a, b) \in \mathbb{R}^2$, which corresponds to finding the thinnest bands of parallel lines containing at least $\overline{k}$ imprecise data. With the above definitions, for a fixed $b \in \mathbb{R}$, finding the thinnest band including $\overline{k}$ or more imprecise data is equivalent to identifying the shortest interval containing at least $\overline{k}$ transformed data $[\underline{z}_{b,1}, \overline{z}_{b,1}], \ldots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. For each $b \in \mathbb{R}$, the intervals including at least $\overline{k}$ transformed data can be written as $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ with $j \in \{1, \ldots, n-\overline{k}+1\}$, where an interval $[\underline{c}, \overline{c}]$ is empty if $\underline{c} > \overline{c}$. Thus, we obtain a general expression for $\overline{q}_{LRM}$ like the one in the theorem, but where we have $\mathbb{R}$ instead of $\mathcal{B}$. Moreover, as for each $b \in \mathbb{R}$ the corresponding intercept a minimizing the bandwidth of the closed band around the function $f_{a,b}$ including at least $\overline{k}$ imprecise data is given by the center of the shortest of the intervals $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ with $j \in \{1, \ldots, n - \overline{k} + 1\}$, we also get the analogous expression for the set of the LRM functions.

Finally, some effort is needed to prove that the slopes of all functions associated with $\overline{q}_{LRM}$ are indeed elements of $\mathcal{B}$ when the condition at the end of the theorem is satisfied. Following Stromberg (1993), this can be achieved by formulating part of the optimization problem as a so-called Chebyshev approximation problem and applying general results about the solutions of these problems from Cheney (1982, Chapters 1 and 2). The condition at the end of Theorem 2 provides a sufficient but not necessary condition for the set of LRM functions to be the finite set given by the expression on the right side of the equation above. The condition excludes situations in which the set of LRM functions can be an infinite proper subset of $\mathcal{F}$. However, in principle, it is possible to give the precise expression for the set of LRM functions also if this condition is not satisfied, which requires many case distinctions. As our main interest lies in determining $\overline{q}_{LRM}$ here, these details about the set of LRM functions are neglected. □

Hence, Theorem 2 constitutes the basis for the first part of the exact algorithm. It provides a way to determine $\overline{q}_{LRM}$ as the solution of a discrete optimization problem and to identify all of them, if there are finitely many associated LRM functions. At first, the set $\mathcal{B}$ is determined, then, for each $b \in \mathcal{B}$, the shortest interval containing $\overline{k}$ of the transformed data $[\underline{z}_{b,1}, \overline{z}_{b,1}], \ldots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$ is identified, providing the corresponding optimal intercept $a \in \mathbb{R}$ by its center and the associated $\overline{r}_{f_{a,b},(\overline{k})}$ by half of its length. Finally, $\overline{q}_{LRM}$ is obtained as the minimum of these $\overline{r}_{f_{a,b},(\overline{k})}$ over all $f_{a,b}$ with $b \in \mathcal{B}$ and optimal intercept a, and the LRM functions are given by the functions corresponding to this minimal upper endpoint.

Part 2: Identifying the set $\mathcal{U}$

On the basis of $\overline{q}_{LRM}$, the set $\mathcal{U}$ of all undominated regression functions is to be determined. For each $f \in \mathcal{U}$, we know that the closed band $\overline{B}_{f,\overline{q}_{LRM}}$ of width $2\,\overline{q}_{LRM}$ intersects at least $\underline{k}+1$ imprecise data. The second part of the algorithm is derived by exploiting this geometrical characterization of the undominated functions. For some fixed $b \in \mathcal{B}$, to identify the undominated functions with slope b, we have to find all intercept values $a \in \mathbb{R}$ for which the corresponding bands $\overline{B}_{f_{a,b},\overline{q}_{LRM}}$ intersect at least $\underline{k}+1$ imprecise data. This is equivalent to finding the centers $a \in \mathbb{R}$ of all intervals $[a-\overline{q}_{LRM}, a+\overline{q}_{LRM}]$ that intersect at least $\underline{k}+1$ of the transformed imprecise data $[\underline{z}_{b,1}, \overline{z}_{b,1}], \ldots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. For each subset of the transformed data $\{[\underline{z}_{b,i}, \overline{z}_{b,i}] : i \in \mathcal{I} \subseteq \{1,\ldots,n\}\}$ of size $|\mathcal{I}| = \underline{k}+1$, the interval $[a-\overline{q}_{LRM}, a+\overline{q}_{LRM}]$ intersects all of these $\underline{k}+1$ transformed data if $a \in [\max_{i\in\mathcal{I}} \underline{z}_{b,i} - \overline{q}_{LRM}, \min_{i\in\mathcal{I}} \overline{z}_{b,i} + \overline{q}_{LRM}]$. Thus, for each $b \in \mathbb{R}$, the (possibly empty) set $\mathcal{A}_b$ of all intercept values $a \in \mathbb{R}$ for which the corresponding bands $\overline{B}_{f_{a,b},\overline{q}_{LRM}}$ intersect $\underline{k}+1$ or more imprecise data is

$$\mathcal{A}_b = \bigcup_{\mathcal{I}\subseteq\{1,\ldots,n\}\,:\,|\mathcal{I}|=\underline{k}+1} \Big[\max_{i\in\mathcal{I}} \underline{z}_{b,i} - \overline{q}_{LRM}, \min_{i\in\mathcal{I}} \overline{z}_{b,i} + \overline{q}_{LRM} \Big].$$

With this definition, the set of all undominated functions can be formulated as the set of all linear functions with slopes $b \in \mathbb{R}$ and corresponding

intercepts $a \in \mathcal{A}_b$, that is,

$$\mathcal{U} = \{f_{a,b} \,:\, b \in \mathbb{R} \text{ and } a \in \mathcal{A}_b\}\,.$$

Thanks to Theorem 2 of Cattaneo and Wiencierz (2013) restated in the following, we obtain a simpler expression for the sets $\mathcal{A}_b$ for all $b \in \mathbb{R}$, and thus, also for $\mathcal{U}$.

Theorem 3.

$$\mathcal{U} = \left\{ f_{a,b} \,:\, b \in \mathbb{R} \;\; \textit{and} \;\; a \in \bigcup_{j=1}^{n-\underline{k}} \left[\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM},\, \overline{z}_{b,(j)} + \overline{q}_{LRM} \right] \right\}.$$

Proof. The expression for $\mathcal{U}$ can be proved with the above explanations and by means of the technical result stated as Lemma 1 in Cattaneo and Wiencierz (2013, Section A). The complete and detailed proof can be found in Cattaneo and Wiencierz (2013, Subsection A.2). □

Theorem 3 makes it possible to determine in a simple way for each $b \in \mathbb{R}$ the corresponding set $\mathcal{A}_b$, containing the intercept values of all undominated functions with slope b. For each $b \in \mathbb{R}$, this set is can be obtained as the union of the $n-\underline{k}$ intervals $[\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{b,(j)} + \overline{q}_{LRM}]$. For most purposes, this would already be sufficient, as it is always possible to use a fine grid over a suitable range of slope values to obtain an acceptable approximation of the set $\mathcal{U}$. Yet, the result of Theorem 3 additionally implies a precise description of the set $\mathcal{U}' \subseteq \mathbb{R}^2$ of parameter values associated with the undominated functions, namely as the union of finitely many (possibly unbounded) polygons.

Precise description of $\mathcal{U}'$

According to (4.4), the set $\mathcal{U}'$ of parameter combinations corresponding to the undominated functions is given by $\mathcal{U}' = \{(a,b) \,:\, b \in \mathbb{R} \text{ and } a \in \mathcal{A}_b\}$. For each $b \in \mathbb{R}$, the set $\mathcal{A}_b$ can be determined as the union of the $n-\underline{k}$ (possibly empty or unbounded) intervals $[\underline{z}_{b,(\underline{k}+1)} - \overline{q}_{LRM}, \overline{z}_{b,(1)} + \overline{q}_{LRM}], \ldots,$ $[\underline{z}_{b,(n)} - \overline{q}_{LRM}, \overline{z}_{b,(n-\underline{k})} + \overline{q}_{LRM}]$. If we consider for each $j \in \{1, \ldots, n-\underline{k}\}$

the interval endpoints $\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $\overline{z}_{b,(j)} + \overline{q}_{LRM}$ as functions of $b \in \mathbb{R}$, we find that these functions are either piecewise linear or constant equal $\pm\infty$ (with a possible discontinuity at $b = 0$). As a polygon can be defined as a subset of $\mathbb{R}^2$ bounded by finitely many line segments and half-lines (see, e.g., Alexandrov, 2005, Subsection 1.1.1), for each $j \in \{1, \ldots, n-\underline{k}\}$, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ determine a polygon. Hence, $\mathcal{U}'$ can be represented as the union of these $n - \underline{k}$ (possibly unbounded) polygons. However, in general, $\mathcal{U}'$ is neither closed nor convex nor connected. An illustration of the complex shape of $\mathcal{U}'$ was given in Figure 4.2 (right).

More precisely, when we consider the case in which all imprecise data are bounded, i.e., when $|\mathcal{D}| = n$, the definitions of $\underline{z}_{b,i}$ and $\overline{z}_{b,i}$ imply that, for all $i \in \{1, \ldots, n\}$, the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ are continuous and piecewise linear, each consisting of two half-lines joined at $b = 0$. Therefore, for each $j \in \{1, \ldots, n - \underline{k}\}$, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ consist of segments of some of the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$, connecting the points at which the indices of the $(\underline{k} + j)$-th smallest $\underline{z}_{b,i}$ and the j-th smallest $\overline{z}_{b,i}$, respectively, change, that is, where the corresponding functions cross each other. Hence, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ are also continuous and piecewise linear for all $j \in \{1, \ldots, n - \underline{k}\}$. If these functions moreover intersect on $\mathbb{R}_{<0}$ and on $\mathbb{R}_{>0}$ for all $j \in \{1, \ldots, n - \underline{k}\}$, the set $\mathcal{U}'$ is a closed and bounded subset of $\mathbb{R}^2$, since it is the union of the polygons determined by these functions. In case that the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ do not intersect twice on $\mathbb{R}_{<0}$ and on $\mathbb{R}_{>0}$ for some $j \in \{1, \ldots, n - \underline{k}\}$, the set $\mathcal{U}'$ is unbounded, but still closed in $\mathbb{R}^2$.

When we consider more general data situations, two cases have to be distinguished. At first, we consider the situation in which there is no imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{x}_i, \overline{x}_i]$ is unbounded and $[\underline{y}_i, \overline{y}_i] \neq [-\infty, +\infty]$. In this case, for all $i \in \{1, \ldots, n\}$, the function $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ is either continuous and piecewise linear (when $\underline{z}_{b,i} > -\infty$) like in the situation where all data are bounded or it is constant

equal $-\infty$. Analogously, the function $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ is either continuous and piecewise linear or constant equal $+\infty$. Therefore, also the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ are each either continuous and piecewise linear or constant equal $\pm\infty$. Hence, the polygons determined by these functions are not necessarily bounded, and thus, also in this data situation $\mathcal{U}'$ is a closed but possibly unbounded subset of $\mathbb{R}^2$.

Finally, we consider the situation in which there is an imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{x}_i, \overline{x}_i]$ is unbounded and $[\underline{y}_i, \overline{y}_i] \neq [-\infty, +\infty]$. Here, at least one of the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ associated with this observation has a discontinuity at $b = 0$. Therefore, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ can be discontinuous at $b = 0$. By consequence, the resulting set $\mathcal{U}'$ is not necessarily closed in this situation. However, if $\mathcal{U}'$ is not closed in $\mathbb{R}^2$, the two parts $\mathcal{U}' \cap (\mathbb{R} \times \{0\})$ and $\mathcal{U}' \cap (\mathbb{R} \times \mathbb{R}_{\neq 0})$ are still closed in their corresponding subspaces $\mathbb{R} \times \{0\}$ and $\mathbb{R} \times \mathbb{R}_{\neq 0}$, respectively.

When the exact shape of $\mathcal{U}'$ is to be determined, for example, in order to visualize the set of undominated parameters, it suffices to consider the finite set of all slopes at which any two of the $2\,n$ functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ cross each other together with the slope zero, because these values constitute all possible locations of the vertices of $\mathcal{U}'$. In addition, two values closely above and below zero as well as one very small and one very large value should be considered to capture the limits. Over the interval defined by the smallest and the largest of the thus determined slopes b', for each $j \in \{1, \ldots, n - \underline{k}\}$, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ correspond to the paths connecting the points at these slope values. Hence, if the set of undominated parameters is bounded, it can be precisely drawn in a graph by connecting, for each $j \in \{1, \ldots, n - \underline{k}\}$, the points $(b', \underline{z}_{b',(\underline{k}+j)} - \overline{q}_{LRM})$ and $(b', \overline{z}_{b',(j)} + \overline{q}_{LRM})$, respectively, located at those of the relevant slopes where $\underline{z}_{b',(\underline{k}+j)} - \overline{q}_{LRM} \leq \overline{z}_{b',(j)} + \overline{q}_{LRM}$. Otherwise the exact shape of $\mathcal{U}'$ can be depicted over a predefined range of slopes.

To identify the set of all relevant slopes, we at first define the set $\tilde{\mathcal{B}}$ of all $b \in \mathbb{R}$ at which either two functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto$

$\underline{z}_{b,j} - \overline{q}_{LRM}$ intersect, for some $(i,j) \in \{1,\dots,n\}^2$ with $i \neq j$, or two functions $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,j} + \overline{q}_{LRM}$ cross each other, for some $(i,j) \in \{1,\dots,n\}^2$ with $i \neq j$. With the definition $^{+\infty}/_{+\infty} = 0$, this set can be written as

$$\begin{aligned}\tilde{\mathcal{B}} = \Bigg(& \left\{ \frac{\overline{y}_i - \overline{y}_j}{\underline{x}_i - \underline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \ \underline{x}_i > \underline{x}_j \ \text{ and } \ \overline{y}_i > \overline{y}_j \right\} \\ & \cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\underline{x}_i - \underline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \ \underline{x}_i > \underline{x}_j \ \text{ and } \ \underline{y}_i < \underline{y}_j \right\} \\ & \cup \left\{ \frac{\overline{y}_i - \overline{y}_j}{\overline{x}_i - \overline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \ \overline{x}_i > \overline{x}_j \ \text{ and } \ \overline{y}_i < \overline{y}_j \right\} \\ & \cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\overline{x}_i - \overline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \ \overline{x}_i > \overline{x}_j \ \text{ and } \ \underline{y}_i > \underline{y}_j \right\} \Bigg) \\ & \cap \ \mathbb{R}.\end{aligned}$$

In fact, the set $\tilde{\mathcal{B}}$ is very similar to the set $\mathcal{B}$ considered in the first part of the algorithm, which consists of the slopes determined by the vertices of all bounded data $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ with $i \in \mathcal{D}$ and zero. Since two functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \underline{z}_{b,j} - \overline{q}_{LRM}$ or two functions $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,j} + \overline{q}_{LRM}$ intersect when b is the slope determined by the corresponding vertices of the imprecise data A_i and A_j, we have that $(\tilde{\mathcal{B}} \cup \{0\})$ is a superset of $\mathcal{B}$. Furthermore, we define the set $\check{\mathcal{B}}$ of all slopes $b \in \mathbb{R}$ at which the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,j} + \overline{q}_{LRM}$ cross each other, for all $(i,j) \in \{1,\dots,n\}^2$. The set $\check{\mathcal{B}}$ is given by

$$\begin{aligned}\check{\mathcal{B}} = \Bigg(& \left\{ \frac{(\overline{y}_i + 2\,\overline{q}_{LRM}) - \underline{y}_j}{\underline{x}_i - \overline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \right. \\ & \qquad \left. \underline{x}_i > \overline{x}_j \ \text{ and } \ (\overline{y}_i + 2\,\overline{q}_{LRM}) > \underline{y}_j \right\} \\ & \cup \left\{ \frac{\underline{y}_i - (\overline{y}_j + 2\,\overline{q}_{LRM})}{\underline{x}_i - \overline{x}_j} \ : \ (i,j) \in \{1,\dots,n\}^2 \ \text{ and } \right. \\ & \qquad \left. \underline{x}_i > \overline{x}_j \ \text{ and } \ \underline{y}_i < (\overline{y}_j + 2\,\overline{q}_{LRM}) \right\}\end{aligned}$$

$$
\begin{aligned}
&\cup \left\{ \frac{(\overline{y}_i + 2\,\overline{q}_{LRM}) - \underline{y}_j}{\overline{x}_i - \underline{x}_j} \;:\; (i,j) \in \{1,\dots,n\}^2 \text{ and } \right. \\
&\qquad \left. \overline{x}_i > \underline{x}_j \text{ and } (\overline{y}_i + 2\,\overline{q}_{LRM}) < \underline{y}_j \right\} \\
&\cup \left\{ \frac{\underline{y}_i - (\overline{y}_j + 2\,\overline{q}_{LRM})}{\overline{x}_i - \underline{x}_j} \;:\; (i,j) \in \{1,\dots,n\}^2 \text{ and } \right. \\
&\qquad \left. \overline{x}_i > \underline{x}_j \text{ and } \underline{y}_i > (\overline{y}_j + 2\,\overline{q}_{LRM}) \right\} \Bigg) \\
&\cap\ \mathbb{R}.
\end{aligned}
$$

Then, the union set $\tilde{\mathcal{B}} \cup \check{\mathcal{B}} \cup \{0\}$ contains all possible locations $b \in \mathbb{R}$ of the vertices of $\mathcal{U}'$. The set $\mathcal{B}_{\mathcal{U}'}$ of all relevant slopes can thus be written as

$$
\mathcal{B}_{\mathcal{U}'} = \tilde{\mathcal{B}} \cup \check{\mathcal{B}} \cup \Big\{ 0, -\eta, \eta, \min(\tilde{\mathcal{B}} \cup \check{\mathcal{B}} \cup \{0\}) - \omega, \max(\tilde{\mathcal{B}} \cup \check{\mathcal{B}} \cup \{0\}) + \omega \Big\},
$$

where η is a small value between zero and $\min\{|b| \;:\; b \in \tilde{\mathcal{B}} \cup \check{\mathcal{B}} \text{ and } b \neq 0\}$ and ω is an arbitrary positive number. In examining the points of the $n - \underline{k}$ functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ at all $b \in \mathcal{B}_{\mathcal{U}'}$ the exact shape of $\mathcal{U}'$ can be determined.

The precise description of the set $\mathcal{U}'$ can furthermore be exploited to identify a suitable range $[\underline{b}, \overline{b}] \subseteq \mathbb{R}$ of slope values when the set $\mathcal{U}$ shall be approximated. Considering the values $b \in \mathcal{B}_{\mathcal{U}'}$ ordered by their size and starting from $\min \mathcal{B}_{\mathcal{U}'}$, one can find $\underline{b}$ as the first $b \in \mathcal{B}_{\mathcal{U}'}$ for which the corresponding set $\mathcal{A}_b \neq \varnothing$. Analogously, starting from $\max \mathcal{B}_{\mathcal{U}'}$ and descending in $\mathcal{B}_{\mathcal{U}'}$, the upper endpoint $\overline{b}$ is the first $b \in \mathcal{B}_{\mathcal{U}'}$ such that the corresponding set $\mathcal{A}_b \neq \varnothing$. If we have $\underline{b} = \min \mathcal{B}_{\mathcal{U}'}$ or $\overline{b} = \max \mathcal{B}_{\mathcal{U}'}$ or both, then, the set of undominated functions is unbounded with respect to b.

Computational complexity

In this section, the first exact algorithm to determine the result of the robust LIR method in the case of simple linear regression with interval data was presented. The algorithm is composed of two parts, the first of which is devoted to finding $\overline{q}_{LRM}$, which is needed in the second part for identifying all undominated functions, either in the form of the set $\mathcal{U}$ of all

undominated linear functions or in representing it by the corresponding set $\mathcal{U}'$ of parameter values. Finally, the worst-case time complexity of the algorithm has to be investigated.

The first part of the algorithm serves to identify $\overline{q}_{LRM}$, which is obtained by determining the shortest of the intervals $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ with $j \in \{1, \ldots, n-\overline{k}+1\}$ for each $b \in \mathcal{B}$. This can be done in time $O(n \log n)$. The set $\mathcal{B}$ contains at most $2\binom{n}{2}+1$ elements, because each pair of data fulfills at most two of the four conditions given in the definition of $\mathcal{B}$. Therefore, the first part of the algorithm has the complexity $O(n^3 \log n)$.

In the second part of the algorithm, if we want to describe $\mathcal{U}'$ precisely, we have to determine $\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $\underline{z}_{b,(j)} + \overline{q}_{LRM}$ for all $j \in \{1, \ldots, n-\underline{k}\}$ and all $b \in \mathcal{B}_{\mathcal{U}'}$. For each slope b this can be done in time $O(n \log n)$, because we only need to compute and sort the two lists $\underline{z}_{b,1}, \ldots, \underline{z}_{b,n}$ and $\overline{z}_{b,1}, \ldots, \overline{z}_{b,n}$. As each pair of data fulfills at most two of the four possibilities given in their definitions, the sets $\tilde{\mathcal{B}}$ and $\check{\mathcal{B}}$ have each at most $2\binom{n}{2}$ elements. Hence, $\mathcal{B}_{\mathcal{U}'}$ consists of at most $4\binom{n}{2}+5$ slope values, and the second part of the algorithm has the complexity $O(n^3 \log n)$.

As both independent parts of the algorithm have the worst-case time complexity $O(n^3 \log n)$, also the computational complexity of the entire algorithm is $O(n^3 \log n)$, that is, it is of the same order as the complexity of the initial algorithm for LQS regression (see, e.g., Steele and Steiger, 1986). The latter algorithm was further optimized to reach a better computational efficiency and it was generalized to multiple linear regression. Likewise, the exact algorithm for robust simple linear LIR with interval data can be adapted to the case of multiple linear regression, and moreover, to more general types of imprecise data.

4.3.2 R package `linLIR`

We implemented the presented algorithm in the statistical software environment R (R Core Team, 2013). It is part of the package `linLIR` (Wiencierz, 2013), designed for the implementation of LIR methods for the case of linear regression with interval data. The available version of the `linLIR` package includes a function to create a particular data object

for interval-valued observations (`idf.create`), the function `s.linlir` to perform the robust LIR analysis for two variables out of the data object, as well as associated methods for the generic functions `print`, `summary`, and `plot`.

Both parts of the exact algorithm are incorporated in the `s.linlir` function. In the current version of the `linLIR` package, the first step of the `s.linlir` function consists in finding $\overline{q}_{LRM}$ and the corresponding parameter combinations of the LRM functions. Then, the range of slope values for which there are undominated functions is identified in the way described at the end of the previous subsection. Finally, $\mathcal{U}'$ is approximated by determining the corresponding sets of intercept values over a fine grid across this range of slope values and retrieving a fixed number of parameter combinations $(a, b) \in \mathcal{A}_b \times \{b\}$. In case that $\mathcal{U}'$ is unbounded, the set of undominated functions is approximated only over a coarse grid of slope values ranging at most from -10^9 to 10^9, if unbounded on both sides. The `s.linlir` function then returns an object of the class "`s.linlir`" consisting of a list whose elements include the ranges of slope and intercept values in $\mathcal{U}'$, a data frame containing the intercept-slope combinations that represent the approximation of the set $\mathcal{U}'$, the bound $\overline{q}_{LRM}$, the analyzed data set, the used LIR settings, $\underline{k}$ and $\overline{k}$, etc. The `linLIR` package provides a `print` method and a `summary` method for these `s.linlir`-objects.

Furthermore, there is a `plot` method associated with the `s.linlir` function providing tools to visualize the LIR results. There are three options: 1) to plot only the LRM regression functions (`typ="lrm"`), 2) to plot a draft of the set of undominated functions $\mathcal{U}$ (`typ="func"`), or 3) to plot the entire set $\mathcal{U}'$ (`typ="para"`). When the first option is chosen together with the option `pl.dat=TRUE`, the LRM functions are drawn into the data plot. To visualize the set $\mathcal{U}$ of undominated functions, a random selection of a chosen number of these functions is drawn. The selection of functions is obtained by randomly choosing parameter combinations (a, b) from the discrete approximation of $\mathcal{U}'$. The default option `para.typ="polygon"` for the plot of the set of parameter values associated with the undominated functions is based on the precise description of $\mathcal{U}'$ as the union of $n-\underline{k}$ poly-

gons, which was derived in the previous subsection. As described there, the plot is obtained in drawing the polygons determined by the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ for each $j \in \{1, \ldots, n - \underline{k}\}$ using the R function `polygon()`. To be able to choose a particular section of the set $\mathcal{U}'$ to be displayed, the functions are evaluated at many equally spaced points within a range of slope values that has to be handed over to the `plot` function.

The `linLIR` package provides a ready-to-use first implementation of the robust LIR method for simple linear regression with interval data. However, the current version of the `s.linlir` function is not optimized for computational speed yet.

4.4 Statistical properties of the robust LIR method

In this section, some properties of the robust LIR method are discussed. In particular, we deal with the confidence level of the imprecise result of this regression method and examine its robustness in more detail.

4.4.1 Confidence level of the set of undominated functions

As the imprecise result of the robust LIR method consists of all functions $f \in \mathcal{F}$ that are plausible in the light of the imprecise data, it can be interpreted as a confidence set for the function that best describes the relationship of interest. In Subsection 4.1.3, we furthermore showed that the result $\mathcal{U}$ of a robust LIR analysis always covers the set $\mathcal{T}$ of corresponding standard LQS regression functions resulting from precise data sets that are compatible with the imprecise data. This is a desirable property, because it is intuitive to require that the imprecise result should not contradict the results that would be obtained, if the data were precisely observed at locations within the observed sets. However, the set $\mathcal{T}$ is not based on a statistical model for inference with imprecise data and its extent neither reflects statistical uncertainty nor does it account for the possibility of wrong coarsening. Yet, it seems reasonable to generally require that regression methods for imprecise data generalizing standard regression methods

should yield a result that is compatible with the imprecise data in this sense. Moreover, we showed in Section 4.2 that the sets of undominated functions for different levels of the confidence regions $\mathcal{C}_f$ are nested. The confidence regions $\mathcal{C}_f$ for a certain p-quantile, with $p \in (0,1)$, of the distribution of the residuals associated with the functions $f \in \mathcal{F}$ constitute the set-valued decision criteria of the decision problem corresponding to the robust LIR method. Following the general LIR methodology, these confidence regions are obtained by cutting the graphs of the corresponding normalized profile likelihood functions at a chosen cutoff point $\beta \in (0,1)$. As discussed in Chapter 3, the thus obtained confidence regions have asymptotically at least a coverage probability of $F_{\chi^2_1}(-2\log(\beta))$, where $F_{\chi^2_1}$ is the cumulative distribution function of the χ^2-distribution with one degree of freedom. The lower the cutoff point β, the higher the confidence level. According to Corollary 1, for each function $f \in \mathcal{F}$, the profile likelihood function lik_{Q_f} is a unimodal step function that is monotonically increasing until the quantile value(s) with maximal likelihood are reached and monotonically decreasing afterwards. Therefore, the confidence regions $\mathcal{C}_f$ for different levels are nested intervals, which implies that also the corresponding sets of undominated functions are nested. This means that the confidence level of the set-valued result of the robust LIR method is related to the coverage probability of the set-valued decision criteria of the regression problem. However, it cannot be easily derived how these confidence levels are related to each other.

To gain some insights regarding the coverage probability of the overall result $\mathcal{U}$, we perform several simulations. For simplicity, we consider the case of simple linear regression here, and without loss of generality, the function describing the relationship between $X \in \mathcal{X} \subseteq \mathbb{R}$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is chosen to be the constant function at zero. As for all functions $f \in \mathcal{F}$, the confidence regions $\mathcal{C}_f$ only get wider as the variables are imprecisely observed, we focus on the special case of actually precisely observed variables to estimate the largest lower bound to the coverage probability of $\mathcal{U}$ under two different assumptions about the distribution of the analyzed random variables, namely a standard normal distribution in the first con-

sidered scenario and a standard Cauchy distribution in the second. To generate the data, we furthermore adopt the assumption of a strictly unimodal and symmetrical error distribution like in the standard regression model. If the underlying probability measure P_V satisfies this assumption, the optimal regression function of the robust LIR method coincides with the optimal regression function of standard LS regression, which is the function describing the conditional expectation of Y given $X = x$, for all $x \in \mathcal{X}$. This is because the regression function minimizing a p-quantile of the residuals' distribution corresponds to the center of the shortest interval that covers at least probability p of the conditional distribution of Y given X. In general, the mode of a univariate distribution can be defined as the limit for $p \to 0$ of the center of the shortest interval covering at least probability p. Therefore, the optimal function of the robust LIR method can be interpreted as describing a generalized mode of the conditional distribution of Y given $X = x$, for all $x \in \mathcal{X}$. If the conditional distribution is strictly unimodal and symmetrical, for each $p \in (0, 1)$, the generalized mode is identical with the mode of this probability distribution. As moreover expected value and mode of the conditional distribution coincide in this case, both regression methods aim at the same optimal function.

Thanks to this, we can base the simulations on data generated by the standard regression model $Y = 0 + 0\,X + W$, where W is a random error that is independent of X and identically distributed as X. Since also the random variable Y has the same distribution, it suffices to simulate pairs of i.i.d. random variables for X and Y and to test if the $(\underline{k}+1)$-th smallest of the simulated realizations of the response variable is not larger than the $\overline{k}$-th smallest of the residuals corresponding to the LQS line estimated from the simulated data. Furthermore, we choose $p = 1/2$ and we consider two different assumptions about the possibility of wrong observations, namely $\varepsilon \in \{0, 0.1\}$, in addition to the two distributional scenarios mentioned above. For each of the resulting four scenarios, we estimate the lower coverage probability of $\mathcal{U}$ for three different values of $\beta \in \{0.15, 0.5, 0.75\}$, each on the basis of 10 000 simulation runs. All computations are realized in the statistical software environment R (R Core Team, 2013).

β	Asymptotic level of $\mathcal{C}_f$	n	Estimated coverage probability of $\mathcal{U}$
0.75	0.5519	20	0.3895
		100	0.6107
		1 000	0.9106
0.50	0.7610	20	0.8284
		100	0.9359
		1 000	0.9979
0.15	0.9486	20	0.9986
		100	0.9995
		1 000	1.0000

Table 4.1: Results of the simulations based on normally distributed variables with expectation zero and variance one, with $\varepsilon = 0$.

First, we investigate the case where there is no doubt about the correctness of the observations, and therefore, it is assumed that $\varepsilon = 0$. The corresponding simulation-based estimates of the lower coverage probability of $\mathcal{U}$ for each distributional scenario are displayed in Tables 4.1 and 4.2, respectively. As expected, we observe that the sharp lower bound to

β	Asymptotic level of $\mathcal{C}_f$	n	Estimated coverage probability of $\mathcal{U}$
0.75	0.5519	20	0.7205
		100	0.9222
		1 000	0.9981
0.50	0.7610	20	0.9697
		100	0.9967
		1 000	1.0000
0.15	0.9486	20	0.9998
		100	1.0000
		1 000	1.0000

Table 4.2: Results of the simulations based on variables following each a Cauchy distribution with location parameter zero and scale parameter one, with $\varepsilon = 0$.

the coverage probability of $\mathcal{U}$ increases as β decreases and as more data are observed. Moreover, most of the estimated coverage probabilities are close to one, regardless of β and often already for $n = 100$. Comparing the two different distributional scenarios, we find that the lower coverage probability is considerably higher when the variables are generated by a standard Cauchy distribution. As in this case the data are more dispersed, more diverse functions can be undominated than in the case of the normally distributed variables. Therefore, the sets of undominated functions resulting from the data generated by the standard Cauchy distribution are less informative, which implies that they have a higher probability of including the true regression function.

Now, we consider the situation in which it is assumed that the observations may not cover the true values with probability at most 0.1. The corresponding percentages of simulation runs in which the set of undominated regression functions covered the true regression function are shown in Tables 4.3 and 4.4. Again, we observe that the estimated lower cover-

β	Asymptotic level of $\mathcal{C}_f$	n	Estimated coverage probability of $\mathcal{U}$
0.75	0.5519	20	0.9767
		100	1.0000
		1 000	1.0000
0.50	0.7610	20	0.9986
		100	1.0000
		1 000	1.0000
0.15	0.9486	20	1.0000
		100	1.0000
		1 000	1.0000

Table 4.3: Results of the simulations based on normally distributed variables with expectation zero and variance one, with $\varepsilon = 0.1$.

age probability of $\mathcal{U}$ increases as n increases and as β decreases. Compared to the situation considered before, the coverage is even higher and there is no considerable difference between the distributional scenarios anymore.

β	Asymptotic level of $\mathcal{C}_f$	n	Estimated coverage probability of $\mathcal{U}$
0.75	0.5519	20	0.9979
		100	1.0000
		1 000	1.0000
0.50	0.7610	20	0.9998
		100	1.0000
		1 000	1.0000
0.15	0.9486	20	1.0000
		100	1.0000
		1 000	1.0000

Table 4.4: Results of the simulations based on variables following each a Cauchy distribution with location parameter zero and scale parameter one, with $\varepsilon = 0.1$.

The simulations' results indicate that the confidence level of the set $\mathcal{U}$ of undominated functions is generally rather high, even if the observations are precise and correct. Therefore, we conclude that the robust LIR method provides very cautious inferences about the relationship of interest.

4.4.2 Breakdown point

In the present subsection, we discuss the robustness of the LIR method presented in Section 4.1. According to Huber (1981, page 1), robustness of a statistical method means, "insensitivity to small deviations from the assumptions." The assumptions mainly referred to in this definition of robustness are different choices of the set of probability measures that are considered as possible models of the analyzed situation. The main motivation for the development of robust statistical methods was that, in most practical settings, the assumption of normally distributed random quantities underlying many standard methods is too idealistic and deviations from this assumption may lead to very unreliable results (see, e.g., Stigler, 2010). As observations that are much different from the majority of the data, so-called outliers, can be the result of a deviation from the normality assumption, for example, in the form of a long-tailed distribution

or of a mixture between a normal distribution and another distribution, statistical methods were proposed whose result is not much influenced by such outlying observations. Hence, the goal of a robust regression method is to describe the relationship between the variables of interest as it is revealed by the majority of the data. To evaluate how insensitive a regression method is to outliers, several robustness measures were proposed. We here focus on the so-called breakdown point for the assessment of the robustness of the robust LIR method. For more details on robust regression methods and on general concepts of Robust Statistics, see, for example, Maronna et al. (2006); Rousseeuw and Leroy (1987); Huber (1981).

In the beginning of Section 4.1, we pointed out that the robust LIR method generalizes LQS regression in several ways. LQS regression is a very robust regression method in terms of a high breakdown point. The breakdown point is a robustness measure for statistical methods indicating which fraction of outliers in the data a statistical method can support without yielding a meaningless result (see, e.g., Rousseeuw and Leroy, 1987, Section 1.2). For example, in the particular case of linear regression with precise data, following Rousseeuw and Leroy (1987, Section 1.2), the breakdown point can be formally expressed as follows. Let f_θ be the linear function modeling the relationship between X and Y, where f_θ is defined for all $x_i \in \mathcal{X}$ by $f_\theta(x_i) = \theta_0 + \theta_1 x_{i,1} + \ldots + \theta_d x_{i,d}$, with $d \in \mathbb{N}$ and $(\theta_0, \theta_1, \ldots, \theta_d)^\mathrm{T} = \theta \in \mathbb{R}^{d+1}$. The vector $\hat{\theta} \in \mathbb{R}^{d+1}$ is the estimate of θ resulting from the investigated regression method on the basis of a precise data set containing $n \in \mathbb{N}$ observations that comply with the distributional assumptions. Furthermore, we denote by $\tilde{\Theta}^{(m)} \subseteq \mathbb{R}^{d+1}$ the set of all possible results of the regression method when $m \in \{0, \ldots, n\}$ observations in the data set are replaced by arbitrary values. Then, the finite-sample breakdown point of the linear regression estimator $\hat{\theta}$ for any possible data set of size n can be defined as the number

$$\max\left\{\frac{m}{n} \,:\, m \in \{0,\ldots,n\} \text{ and } \sup_{\tilde{\theta}\in\tilde{\Theta}^{(m)}} \|\tilde{\theta} - \hat{\theta}\| < +\infty\right\}. \tag{4.5}$$

To have a robustness evaluation that is independent of the sample size

n, we focus on the asymptotic definition of the breakdown point, which corresponds to the limit of (4.5) for $n \to +\infty$. For any statistical method, the highest possible breakdown point is $1/2$, because by definition there cannot be more than 50% outliers in a data set. As shown by Rousseeuw and Leroy (1987, Section 3.4) for the case of linear regression, the LQS regression method yields a breakdown point of $\min\{p, 1-p\}$, given the chosen proportion $p \in (0,1)$. This is also true for more general regression problems, as discussed below. Hence, the highest possible breakdown point $1/2$ is achieved by the LQS method for $p = 1/2$.

Now, to evaluate the robustness of the LIR method presented in Section 4.1 in terms of its breakdown point, we first have to discuss the notion of outlier in the context of regression analysis with imprecise data and the meaning of breakdown for a regression method that generally yields an imprecise result. Recall that the robust LIR method is based on the assumption $(V, V^*) \sim P \in \mathcal{P}_\varepsilon$, where $\mathcal{P}_\varepsilon$ contains all probability measures P' with $P'(V \in V^*) \geq 1 - \varepsilon$, for some $\varepsilon \in [0, 1/2)$. One could argue that the nonparametric probability model implies that every kind of distribution of the data is allowed, and therefore, robustness is not necessary and even inconsistent with the underlying assumption. Yet, basing the analysis on a nonparametric assumption essentially means that the possible probability models are not restricted to, for example, a parametric family of probability distributions for the precise data together with a (probabilistic) coarsening scheme relating the imprecise data to them. Since we are concerned with regression analysis, of course, we presume that there is a relationship between the explanatory variables X and the response variable Y, which means that we expect the joint distribution of X and Y to be concentrated around some function $f \in \mathcal{F}$. Therefore, it is sensible to require that the result of the regression method should not be too much affected by observations that are far away from the majority of the data. In particular, when analyzing (partially) unbounded imprecise data, the precise values of interest can be arbitrarily far away from the bounded observations. Thus, in this situation, robustness is a necessary property to have a chance to obtain informative results. Finally, we restrict the

investigation to the type of imprecise data that is the most relevant for statistical practice, namely interval data, where each imprecise observation is the Cartesian product of $d + 1$ (possibly unbounded) intervals.

By direct generalization of the meaning of outlier given above to the situation in which the data are only imprecisely observed, we consider each imprecise observation that allows the corresponding precise value to be far away from the majority of the (unobserved) precise data as an outlier. If there is no doubt about the correctness of the imprecise data, an imprecise observation that is far away from the majority of the imprecise data implies that the contained precise value is also far away from the majority of the precise data, and therefore, it is regarded as an outlier. Moreover, an imprecise observation that is (partially) unbounded or corresponds to the entire observation space can allow the precise value to be much different from the majority of the precise data. Hence, (partially) unbounded or completely uninformative imprecise data are also considered as outliers according to this definition. If we allow the possibility of wrong observations with probability at most $\varepsilon > 0$ in the probability model underlying the robust LIR method, asymptotically up to $\varepsilon\,100\%$ of the imprecise data do not cover the corresponding precise values. As we make no further assumptions about the coarsening mechanism, the corresponding precise values can be far away from the majority of the precise data. Therefore, the robust LIR method inherently accounts for the possibility of $\varepsilon\,100\%$ of the data to be outliers, without revealing which ones. Hence, in this situation, the breakdown point measures only the supplementary fraction of tolerated outliers in addition to ε.

To define the notion of breakdown for a regression method that yields a set-valued result, we consider again the hypothetical setting in which the fraction m/n of a given data set of size n is replaced by arbitrary imprecise data or is already uninformative in a certain sense. We qualify as breakdown of an imprecise regression method the situation in which the union set $\mathcal{U}^{(m)} \subseteq \mathcal{F}$ of all corresponding results is a superset of the set $\tilde{\mathcal{U}} \subseteq \mathcal{F}$ of all functions $f \in \mathcal{F}$ that intersect the observation space, that is, $\tilde{\mathcal{U}} = \{f \in \mathcal{F} : G_f \cap \mathcal{V} \neq \varnothing\}$, where $G_f = \{(x, y) \in \mathcal{X} \times \mathbb{R} : y = f(x)\}$

denotes the graph of a function $f \in \mathcal{F}$. Hence, breakdown means that the set of undominated functions obtained by analyzing a data set of size n with m outliers can contain any function that intersects the observation space, and thus, the result can be completely meaningless. Recall that the set of undominated functions is composed of all functions $f \in \mathcal{F}$ whose associated closed bands $\overline{B}_{f,\overline{q}_{LRM}}$ of width $2\,\overline{q}_{LRM}$ intersect at least $\underline{k}+1$ imprecise data. If the response variable is such that $\mathcal{Y} = \mathbb{R}$, we have that $\tilde{\mathcal{U}} = \mathcal{F}$, otherwise, $\tilde{\mathcal{U}}$ is a proper subset of $\mathbb{R}$.

For the investigation of the finite-sample breakdown point of the robust LIR method, we consider separately the two parts of the determination of the LIR result that were distinguished in Section 4.3.1, which are the determination of $\overline{q}_{LRM}$ and the identification of the set $\mathcal{U}$ of all undominated regression functions on the basis of $\overline{q}_{LRM}$. Furthermore, three cases depending on the observation space $\mathcal{Y}$ of the response variable are distinguished.

At first, the case $\mathcal{Y} = \mathbb{R}$ is investigated. We consider again hypothetical data sets $V_1^* = A_1, \ldots, V_n^* = A_n$ where a certain number $m \in \{0, \ldots, n\}$ of imprecise data is replaced by arbitrary observations or corresponds to uninformative interval data and all possible resulting sets of undominated regression functions. Then, the maximal m such that the union set $\mathcal{U}^{(m)}$ of all possible results is a proper subset of $\mathcal{F}$ is identified. The set $\mathcal{U}^{(m)}$ equals $\mathcal{F}$ due to the first part of the determination of the LIR result, if on the basis of the hypothetical data sets, it can happen that $\overline{q}_{LRM} = +\infty$, because in this case, the vertical bands $\overline{B}_{f,\overline{q}_{LRM}}$ around all functions $f \in \mathcal{F}$ intersect all data. As $\overline{q}_{LRM}$ is given by the $\overline{k}$-th smallest upper residual associated with some function $f \in \mathcal{F}$, this occurs if there are less than $\overline{k}$ imprecise data that are bounded in the Y-dimension or cannot take arbitrary values, i.e., that are no outliers in the sense defined above. Therefore, there can be at most $n-\overline{k}$ outliers. In addition, when $\overline{k} \leq n-\overline{k}$ and there are $n-\overline{k}$ imprecise observations that can take arbitrary values, all $\overline{k}$ data that determine $\overline{q}_{LRM}$ can be such outliers. As the outliers can be located anywhere in $\mathcal{V}$ and $\underline{k}+1 \leq \overline{k}$, every function $f \in \mathcal{F}$ can be undominated for some of the hypothetical data sets, which implies that

$\mathcal{U}^{(\overline{k})} = \mathcal{F}$. Therefore, there have to be less than $\overline{k}$ outliers. Hence, the number of outliers that can be supported in the first part when $\mathcal{Y} = \mathbb{R}$ is $\min\{\overline{k}-1, n-\overline{k}\}$. As to the second part, we investigate what the largest m is for which the union set of all possible results is not $\mathcal{F}$, provided that there is no breakdown in the first part. Since the graph of every function $f \in \mathcal{F}$ can intersect each arbitrary observation and each completely uninformative observation, i.e., $A_i = \mathcal{V}$, there must be less than $\underline{k} + 1$ such outliers to prevent breakdown. However, if $n - \underline{k} \leq \underline{k}$ and there are $\underline{k}$ outliers, some of the outliers are used to determine the set of undominated functions, which makes the result somewhat arbitrary. Yet, there is no breakdown in the above defined sense, because the graphs of the functions in the resulting sets are each close to at least one of the fixed and bounded observations. Hence, in the case where $\mathcal{Y} = \mathbb{R}$, the maximum number of supported outliers in the second part is $\underline{k}$. Taking the minimum over both parts, the number of outliers that are tolerated is $\min\{\underline{k}, n - \overline{k}\}$.

Now, consider the situation in which $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. Again, the investigation is based on a hypothetical data set of size n where m imprecise data are replaced by arbitrary interval data or correspond to the entire observation space, i.e., $A_i = \mathcal{V}$, and the aim is to identify the largest m such that the union set $\mathcal{U}^{(m)}$ of all possible results is a proper subset of the set $\tilde{\mathcal{U}}$ of all functions that intersect the observation space $\mathcal{V}$. In the first part of the determination of the LIR result, if there are at least $\overline{k}$ outliers, $\overline{q}_{LRM}$ can be completely determined by those arbitrary observations. As furthermore $\underline{k} + 1 \leq \overline{k}$, every function $f \in \tilde{\mathcal{U}}$ can be undominated for some of the hypothetical data sets. Therefore, there cannot be more than $\overline{k} - 1$ outliers. If the number of outliers is larger than $n - \overline{k}$, the upper endpoint of the confidence regions is always determined by an outlier, and therefore, it can take arbitrary values within a certain range depending on $\mathcal{F}$ and $\mathcal{Y}$. For example, if $\mathcal{F}$ contains all constant functions, the highest possible value for $\overline{q}_{LRM}$ is $1/2\,(\max \mathcal{Y} - \min \mathcal{Y})$, which corresponds to the cases in which each outlier is either such that $[\underline{y}_i, \overline{y}_i] = \mathcal{Y}$ or completely uninformative. However, there is not necessarily breakdown in the above defined sense. This can only occur if the number of outliers is at the same

time at least $\underline{k}+1$, which is excluded by the restrictions for the second part. Hence, to prevent breakdown in the first part when $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, there can be at most $\overline{k}-1$ outliers. Regarding the second part, we can repeat the arguments of the previous case for all $f \in \tilde{\mathcal{U}}$. Hence, also in the case where $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$, the maximum number of supported outliers in the second part is $\underline{k}$. As $\underline{k} \leq \overline{k}-1$, we here obtain $\underline{k}/n$ as the maximal fraction of supported outliers.

Finally, for the case where $\mathcal{Y}$ is bounded in one direction and unbounded in the other direction, the first part is analogous to the first case, while the second part is the same as in the second case. This leads again to the maximal number of $\min\{\underline{k}, n-\overline{k}\}$ outliers.

Putting the results together over all three cases, the determination of the LIR result is insensitive to at most $\min\{\underline{k}, n-\overline{k}\}$ outliers. Hence, we obtain as the overall finite-sample breakdown point of the robust LIR method $1/n \, \min\{\underline{k}, n-\overline{k}\}$. As to the asymptotic breakdown point, consider again the definitions of $\underline{k}$ and $\overline{k}$ given in Corollary 2. It is easy to derive from these definitions that, for $n \to +\infty$, we have $\underline{k}/n \to (p-\varepsilon)$ and $\overline{k}/n \to (p+\varepsilon)$. Thus, for a given configuration of $p \in (0,1)$ and $\varepsilon \in [0, 1/2)$, the breakdown point of the robust LIR method is given by

$$\lim_{n \to +\infty} \frac{\min\{\underline{k}, n-\overline{k}\}}{n} = \min\{p, 1-p\} - \varepsilon.$$

For the choice of $p = 1/2$ and if wrong observations are not accounted for in the probability model, i.e., $\varepsilon = 0$, the robust LIR method yields the highest possible breakdown point of $1/2$. If we consider $\varepsilon > 0$, the robust LIR method inherently accounts for the possibility that $\varepsilon\,100\%$ of the data are outliers. Therefore, in this case, the breakdown point measures only the fraction of outliers that can be supported in addition to ε. That is, for $p = 1/2$ and $\varepsilon > 0$, the breakdown point is $1/2 - \varepsilon$, but altogether the robust LIR method tolerates 50% outliers, which again is the highest possible fraction. Hence, the LIR method presented in Section 4.1 is very robust in the sense that it takes a large fraction of outliers to cause the regression method to yield a meaningless result.

Chapter 5

Support Vector Regression with interval data

This chapter is devoted to a thorough examination of the regression methodology for interval data proposed in Utkin and Coolen (2011). This regression methodology consists in a modification of standard Support Vector Regression (SVR), which is a specific class of regularized kernel-based methods for data analysis. These methods originated in the field of Machine Learning and were a popular research topic in diverse areas during the past 20 years. In particular, in the context of Computer Science and Engineering, various kernel-based methods were developed (see, e.g., Schölkopf and Smola, 2002; Suykens et al., 2002; Müller et al., 2001), which are mainly based on the framework introduced by Vapnik (1998, 1995). In recent years, there is also a growing interest in kernel-based methods in the field of Statistics (see, e.g., Hable, 2012; Christmann and Hable, 2012; Christmann et al., 2009; Hofmann et al., 2008). Steinwart and Christmann (2008) provide a comprehensive overview of regularized kernel-based methods for the statistical problems of classification and regression in a unified formulation and deduce important results regarding their statistical properties in mathematical detail.

Standard SVR is based on a fully nonparametric probability model and permits analyzing many different kinds of relationships, including, for ex-

ample, linear regression and the estimation of a general smooth regression function. The corresponding estimators are under suitable regularity conditions consistent and they can be robust if an appropriate loss function is chosen. A generalization of SVR to imprecise data preserving these characteristics would be a desirable achievement and could provide a competitive alternative to the robust LIR method presented in Section 4.1, since we aim at such a universal and ideally robust regression method for imprecise data. Moreover, the solution of a standard SVR analysis can be efficiently computed also in the case of a general relationship, which appears to be very difficult to realize for the robust LIR method.

Whether the regression methodology proposed by Utkin and Coolen (2011) achieves these goals is investigated in Section 5.2, after a review of the core elements of the standard SVR framework in the following section.

5.1 Standard SVR

In this section, the standard SVR methodology for precise data is presented. We present the core elements of the theoretical framework of SVR by employing the same notions that were used to describe the LIR methodology in Chapter 3 and the robust LIR method in Section 4.1. Hence, the regression problem is formalized as a decision problem with loss function L assigning to each pair $(f, P_V) \in \mathcal{F} \times \mathcal{P}_V$ an evaluation of the error resulting from describing the relationship of interest by f if the quantities of interest $(X, Y) = V \in \mathcal{V}$, with $\mathcal{V} = \mathcal{X} \times \mathcal{Y}$, are distributed according to P_V. For simplicity, we here assume that $\mathcal{X} \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, is compact and that $\mathcal{Y} \subseteq \mathbb{R}$ is closed, although many of the following definitions and statements apply also to more general cases. As $\mathcal{P}_V$, the set of all probability measures on $\mathcal{V}$ is considered. Thus, like the robust LIR method, SVR is based on a nonparametric probability model. However, in contrast to the robust LIR method, the loss $L(f, P_V)$ assigned to a possible regression function f and a distribution P_V is not given by a quantile of the distribution of the associated residual R_f (under P_V), but by the so-called risk functional, that is, by the corresponding expectation of a function of the residual.

5.1.1 Theoretical framework of SVR methods

Presupposing measurability, the risk functional can be defined for each $P_V \in \mathcal{P}_V$ as

$$E_{P_V} : \mathcal{F} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}, \text{ with } f \mapsto E_{P_V}(f) = \mathbb{E}(\psi(R_f)), \tag{5.1}$$

where ψ is a convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ satisfying $\psi(0) = 0$. For example, if ψ is defined by $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, the loss associated with a pair (f, P_V) is given by $L(f, P_V) = E_{P_V}(f) = \mathbb{E}(R_f^2)$. Thus, we obtain the loss function corresponding to LS regression. Another famous example is the function defined by $\psi(r) = \max\{0, r - \nu\}$, for all $r \in \mathbb{R}_{\geq 0}$ and some $\nu \geq 0$, which was introduced in Vapnik (1995, Section 6.1) and represents the so-called ν-insensitive loss. The convexity of the mapping ψ implies convexity of the risk functional E_{P_V}, that is, the risk functional satisfies for each $\rho \in [0, 1]$

$$E_{P_V}(\rho\, f + (1 - \rho)\, f') \leq \rho\, E_{P_V}(f) + (1 - \rho)\, E_{P_V}(f'),$$

for all $f, f' \in \mathcal{F}$ (see also Steinwart and Christmann, 2008, Lemma 2.13). As discussed later, this property amongst others guarantees that there exists a unique optimal regression function, as long as the true probability distribution P_V is not such that $E_{P_V}(f) = +\infty$ for all $f \in \mathcal{F}$.

In the SVR framework, the set $\mathcal{F}$ of considered regression functions is supposed to be a particular kind of Hilbert space, a so-called Reproducing Kernel Hilbert Space (RKHS). A Hilbert space over $\mathbb{R}$ is a normed vector space over $\mathbb{R}$ with a scalar product that is furthermore complete with respect to the norm induced by the scalar product. For example, the space $\mathbb{R}^2$ with the standard scalar product $\langle \cdot, \cdot \rangle : \mathbb{R}^2 \to \mathbb{R}$, with $\langle w, w' \rangle = w^{\mathrm{T}} w'$ for all $w, w' \in \mathbb{R}^2$, is a Hilbert space, because $\mathbb{R}^2$ is complete with respect to the Euclidean norm given by $\|w\| = \sqrt{\langle w, w \rangle}$ for each $w \in \mathbb{R}^2$. In the context of regression analysis, we consider Hilbert spaces of functions from $\mathcal{X}$ to $\mathbb{R}$. We denote by $\mathcal{H}$ such a function Hilbert space over $\mathcal{X}$ and by $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}$ the associated scalar product. To explain the special

case of an RKHS, we furthermore need to clarify the notion of a kernel function and the particular reproducing property characterizing an RKHS. A kernel function κ is a positive semi-definite function on $\mathcal{X} \times \mathcal{X}$, that is, $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j) \geq 0$, for all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $n \in \mathbb{N}$. If κ is the reproducing kernel function of an RKHS $\mathcal{H}$, for each $x \in \mathcal{X}$ we have that $\kappa(\cdot, x) \in \mathcal{H}$ and

$$f(x) = \langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}},$$

for all $f \in \mathcal{H}$. From this property called reproducing property follows that $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x') \rangle_{\mathcal{H}}$, for all $x, x' \in \mathcal{X}$. Applying general results from Functional Analysis, it can be shown that, for a given RKHS, there exists a unique kernel function satisfying the reproducing property and vice versa. Hence, there is a one-to-one correspondence between reproducing kernel function and RKHS. For example, the RKHS of the linear kernel defined by $\kappa(x, x') = \langle x, x' \rangle + 1$, for all $x, x' \in \mathcal{X}$, is the Hilbert space of all (affine) linear functions from $\mathcal{X}$ to $\mathbb{R}$. Another common kernel function is the so-called Gaussian kernel, which is defined for all $x, x' \in \mathcal{X}$ by $\kappa(x, x') = \exp\left\{-1/\sigma^2 \, \|x - x'\|^2\right\}$, with $\sigma > 0$. The associated RKHS is a very large function space that is dense in the space of all continuous (real-valued) functions on $\mathcal{X}$. For more details on kernels and RKHSs, see, for example, Steinwart and Christmann (2008, Chapter 4).

Besides an RKHS considered as $\mathcal{F}$, the decision problem of the regression analysis in the SVR framework involves as $\mathcal{P}_V$ the set of all probability measures on $\mathcal{V}$. Given the true probability distribution of the quantities of interest, P_V, the best description of the relationship between X and Y is the function minimizing the expected error E_{P_V} (under P_V). Yet, to avoid too wiggly functions as descriptions of the relationship of interest when the regression analysis is based on a finite sample of observations, also a modified decision problem is considered, in which an additive penalty term for the complexity of the functions $f \in \mathcal{F}$ is included in the loss function. In the modified decision problem, instead of E_{P_V} the regularized

risk functional $E_{P_V,\lambda}$ is minimized, which is defined for all $f \in \mathcal{F}$ by

$$E_{P_V,\lambda}(f) = E_{P_V}(f) + \lambda \, \|f\|_{\mathcal{F}}^2 ,$$

where $\lambda > 0$ is a fixed parameter regulating the penalization and $\|\cdot\|_{\mathcal{F}}$ is the norm induced by the scalar product in $\mathcal{F}$. The regularization can be interpreted as minimizing E_{P_V} under the restriction $\|f\|_{\mathcal{F}}^2 \leq c$, for some $c \in \mathbb{R}$, but instead of choosing the bound c explicitly, we fix the value of the corresponding Lagrange multiplier λ in the constrained optimization problem. As the functional $f \mapsto \lambda \, \|f\|_{\mathcal{F}}^2$ is strictly convex by general properties of norms in Hilbert spaces and E_{P_V} is convex because of ψ, we have that $E_{P_V,\lambda}$ is also a strictly convex functional on $\mathcal{F}$. Exploiting the strict convexity of $E_{P_V,\lambda}$, it can be shown that such an optimal function always exists and is unique, provided that some regularity conditions are fulfilled (see, e.g., Steinwart and Christmann, 2008, Lemma 5.1 and Theorem 5.2).

Of course, usually, the true distribution P_V is unknown, but an i.i.d. sample of observations $V_1 = v_1, \ldots, V_n = v_n$ is available. In contrast to the robust LIR method where the inference is based on all probability measures that are plausible in the light of the data, in the SVR methodology, P_V is estimated by the empirical distribution $\hat{P}_V$ of the observations, before the best regression function $f_{\hat{P}_V,\lambda}$ for this probability distribution is identified by minimizing $E_{\hat{P}_V,\lambda}$, for some $\lambda > 0$. Hence, SVR uses only the information associated with the maximum of the likelihood function lik_V induced by the observations on the set of considered probability distributions. Like in the general case before, there always exists a unique minimizer of the regularized risk for $\hat{P}_V$. Moreover, the so-called Representer Theorem states that the unique function $f_{\hat{P}_V,\lambda}$ can be represented as the linear combination of the corresponding functions $\kappa(\cdot, x_1), \ldots, \kappa(\cdot, x_n)$, that is, there exist weights $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f_{\hat{P}_V,\lambda}(x) = \sum_{j=1}^{n} \alpha_j \, \kappa(x, x_j), \tag{5.2}$$

for all $x \in \mathcal{X}$ (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5).

Equation (5.2) is sometimes called support vector expansion of $f_{\hat{P}_V,\lambda}$ and the optimal function $f_{\hat{P}_V,\lambda}$ is often referred to as a Support Vector Machine (SVM). This term has historical reasons, because Vapnik (1998, 1995) proposed to use functions for ψ that have the property that some of the resulting $\alpha_1, \ldots, \alpha_n$ are zero. The vectors x_j for which $\alpha_j \neq 0$ are called support vectors, whence the notion SVM. One example for such a representing function ψ is the function associated with the ν-insensitive loss mentioned before. Nevertheless, in general, SVMs are not sparse in this sense (see, e.g., Steinwart and Christmann, 2008, Section 11.1).

If $\mathcal{F}$ is a large RKHS of arbitrary smooth regression functions, for example, if $\mathcal{F}$ is the RKHS corresponding to the Gaussian kernel, it can be shown that under suitable regularity conditions $f_{\hat{P}_V,\lambda}$ is risk consistent. That is, provided that the conditions are fulfilled, for $n \to +\infty$, we have that $E_{P_V}(f_{\hat{P}_V,\lambda^{(n)}}) \to \inf_{f\in\mathcal{F}} E_{P_V}(f)$ in probability, where $(\lambda^{(n)})_{n\in\mathbb{N}}$ is a sequence of penalty parameters with $\lambda^{(n)} \to 0$ (not too fast) and $P_V \in \mathcal{P}_V$ is the unknown distribution underlying the observations $V_1 = v_1, \ldots, V_n = v_n$ (see, e.g., Steinwart and Christmann, 2008, Theorem 9.1). If less general regression functions are considered, often, other consistency results can be derived. For example, if a linear kernel function is considered together with $\psi(r) = r^2$, for all $r \in \mathbb{R}_{\geq 0}$, SVR is equivalent to penalized linear LS regression, which is also called Ridge regression (see, e.g., Hoerl and Kennard, 1970). This special case of SVR is discussed in detail in the next subsection. To this configuration of SVR the theorem about risk consistency does not apply, but consistency statements can be derived in some different ways, as shown in the following. Moreover, if the loss function is chosen to be such that ψ is Lipschitz continuous, i.e., $\exists \varsigma > 0$ such that $|\psi(r) - \psi(r')| \leq \varsigma\, |r - r'|$, for all $r, r' \in \mathbb{R}_{\geq 0}$, the SVR estimator is robust in a certain sense (see, e.g., Steinwart and Christmann, 2008, Section 10.4). For example, the representing function of the ν-insensitive loss is Lipschitz continuous, but the one of LS regression is not, thus, the corresponding SVR method is not robust, while SVR with the ν-insensitive loss is.

Finally, the determination of the optimal regression function for a given data set $V_1 = v_1, \ldots, V_n = v_n$ and a fixed $\lambda > 0$ is straightforward. Thanks

to the Representer Theorem expressed in (5.2), we know that $f_{\hat{P}_V,\lambda}$ is an element of the set $\mathcal{F}_n \subset \mathcal{F}$, with $\mathcal{F}_n = \left\{\sum_{j=1}^n \alpha_j\, \kappa(\cdot, x_j) \,:\, \alpha_1, \ldots, \alpha_n \in \mathbb{R}\right\}$. For all functions $f_\alpha \in \mathcal{F}_n$, with $\alpha = (\alpha_1, \ldots, \alpha_n)^\mathrm{T} \in \mathbb{R}^n$, the squared norm is given by $\|f_\alpha\|_{\mathcal{F}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i\, \alpha_j\, \kappa(x_i, x_j)$. Hence, the regularized risk associated with $\hat{P}_V$ can be written for each $f_\alpha \in \mathcal{F}_n$ as

$$E_{\hat{P}_V,\lambda}(f_\alpha) = \\ \frac{1}{n}\sum_{i=1}^n \psi\big(\big|y_i - \textstyle\sum_{j=1}^n \alpha_j\, \kappa(x_i, x_j)\big|\big) + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i\, \alpha_j\, \kappa(x_i, x_j). \quad (5.3)$$

As $E_{\hat{P}_V,\lambda}$ is convex, the SVM $f_{\hat{P}_V,\lambda}$ can be obtained by solving a convex optimization problem over $\alpha \in \mathbb{R}^n$, for which there are numerous efficient algorithms (see, e.g., Reinhardt et al., 2013). For the selection of an appropriate regularization parameter $\lambda > 0$ and of other hyper-parameters like the parameter σ of the Gaussian kernel, different strategies can be applied, for instance, cross-validation (see, e.g., Steinwart and Christmann, 2008, Section 11.3).

5.1.2 Ridge regression as a special case of SVR

Ridge regression was introduced in Hoerl and Kennard (1970) as a regularization of standard linear LS regression for the situation in which some of the explanatory variables in X are strongly correlated. In standard linear regression, we suppose that the relationship between the variables of interest, X and Y, can be described by a linear function f_θ defined for all $x_i \in \mathcal{X}$ by $f_\theta(x_i) = \theta_0 + \theta_1\, x_{i,1} + \ldots + \theta_d\, x_{i,d}$, with unknown coefficients $\theta \in \mathbb{R}^{d+1}$. Furthermore, it is usually assumed that for each possible realization x of X the deviation of Y from $f_\theta(x)$ can be described by means of an uncorrelated random quantity with expectation zero and finite variance $\tau^2 < +\infty$. That is, f_θ models the conditional expectation of Y given $X = x$, from which Y deviates with the same variance τ^2 for all $x \in \mathcal{X}$. This constitutes a strong assumption about the probability distribution P_V of the analyzed variables, and consequently, the set $\mathcal{P}_V$ of probability measures that are considered as possible models of the data situation is

much smaller in standard linear regression than in SVR and in the robust LIR method.

The loss function of the decision problem corresponding to LS regression is given by the expectation $\mathbb{E}(R_f^2)$ of the squared residual. If P_V is known, the best description of the linear relationship of interest is the function f_θ. As already mentioned in Section 3.1, this is due to the fact that the function f minimizing the expectation of the squared residual (under P_V) is the one that fulfills $f(x) = \mathbb{E}(Y|x)$ for all $x \in \mathcal{X}$ and according to the model assumptions $f_\theta(x) = \mathbb{E}(Y|x)$ for all $x \in \mathcal{X}$. Recall that $\mathbb{E}(R_f^2)$ corresponds to the risk functional E_{P_V} defined in (5.1), where ψ is defined by $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$. To estimate the LS regression function on the basis of a data set $V_1 = v_1, \ldots, V_n = v_n$, the risk $E_{\hat{P}_V}$ with respect to the empirical distribution $\hat{P}_V$ is minimized. Let D denote the design matrix of the linear regression model, that is, D is an $n \times (d+1)$ matrix that comprises as rows the observed vectors of the explanatory variables, $x_1, \ldots, x_n$, each supplemented by the value one in the first column to model the intercept θ_0, and let $y = (y_1, \ldots, y_n)$. The best regression function is the linear function $f_{\hat{\theta}_{LS}}$ associated with the vector $\hat{\theta}_{LS}$ that minimizes the expression $1/n\,(y - D\,\theta)^{\mathrm{T}}(y - D\,\theta)$, which corresponds to $E_{\hat{P}_V}(f_\theta)$. Equivalently, the residual sum of squares, given by $n\,E_{\hat{P}_V}(f_\theta)$, can be considered as the criterion to be minimized. The corresponding minimization problem can be solved analytically and the unique minimizer is given by

$$\hat{\theta}_{LS} = (D^{\mathrm{T}} D)^{-1}\, D^{\mathrm{T}} y.$$

Thus, $\hat{\theta}_{LS}$ is the LS estimator of the vector of regression coefficients and $f_{\hat{\theta}_{LS}}$ is the corresponding LS regression function.

It can easily be seen that, under the assumptions of the standard linear regression model, $\hat{\theta}_{LS}$ is unbiased, that is, $\mathbb{E}(\hat{\theta}_{LS}) = \theta$, and has the variance matrix $\tau^2\,(D^{\mathrm{T}} D)^{-1}$. According to the Gauss-Markov Theorem, the variances of the components of $\hat{\theta}_{LS}$ are the smallest among all possible unbiased estimators that are linear functions of y. Hence, $\hat{\theta}_{LS}$ is the best linear unbiased estimator for θ (see, e.g., Casella and Berger, 2002,

Section 11.3). Moreover, the LS estimator of the regression coefficients is consistent. Roughly speaking, a consistent estimator $\hat{\theta}$ for a parameter (vector) θ is a function of the sample variables $V_1, \ldots, V_n$ that converges to θ as the sample size $n \in \mathbb{N}$ increases. This property can be expressed by different mathematical definitions, considering different types of convergence for random variables. A possible definition of consistency is the definition based on convergence in probability, which is usually called weak consistency. The sequence of estimators $(\hat{\theta}^{(n)})_{n\in\mathbb{N}}$, where $\hat{\theta}^{(n)}$ is the estimator $\hat{\theta}$ for a sample of size n, is said to converge to θ in probability if for each $\delta > 0$, $\lim_{n\to+\infty} P_V(\|\hat{\theta}^{(n)} - \theta\| > \delta) = 0$. The estimator $\hat{\theta}$ is weakly consistent, if it converges to θ in probability. Another definition of consistency involves convergence in terms of the Mean Squared Error (MSE) of the estimator. The MSE of an estimator $\hat{\theta}$ for an unknown parameter $\theta \in \mathbb{R}^{d+1}$ is defined by

$$MSE(\hat{\theta}) = \sum_{l=1}^{d+1} \mathbb{E}((\hat{\theta}_l - \theta_l)^2) = \sum_{l=1}^{d+1} (\mathbb{E}(\hat{\theta}_l - \theta_l))^2 + \mathbb{E}((\hat{\theta}_l - \mathbb{E}(\hat{\theta}_l))^2),$$

where $(\mathbb{E}(\hat{\theta}_l - \theta_l))^2$ is the squared bias of the l-th component of $\hat{\theta}$ and $\mathbb{E}((\hat{\theta}_l - \mathbb{E}(\hat{\theta}_l))^2)$ is its variance. With respect to the MSE, the sequence of estimators $(\hat{\theta}^{(n)})_{n\in\mathbb{N}}$ converges to θ if its MSE converges to zero, that is, if $\lim_{n\to+\infty} MSE(\hat{\theta}^{(n)}) = 0$. Hence, the estimator $\hat{\theta}$ is in this sense consistent, if its MSE converges to zero. This notion of consistency states a stronger property than weak consistency, because convergence in terms of the MSE implies convergence in probability (see, e.g., Schervish, 1995, Section 7.1). Hence, to investigate the consistency of the LS estimator, we consider the asymptotic behavior of its MSE. As $\hat{\theta}_{LS}$ is unbiased, we have that $MSE(\hat{\theta}_{LS}) = \sum_{l=1}^{d+1} \mathbb{E}((\hat{\theta}_{LS,l} - \mathbb{E}(\hat{\theta}_{LS,l}))^2) = \tau^2\, tr((D^\top D)^{-1})$, where $tr(M)$ denotes the trace of a square matrix M. Under the rather uncritical assumption that the sequence of matrices $((D^{(n)\top} D^{(n)})^{-1})_{n\in\mathbb{N}}$ converges to the $(d+1)\times(d+1)$ zero matrix as n increases, where $D^{(n)}$ is the design matrix of the linear regression model for a sample of size n, we obtain that $MSE(\hat{\theta}_{LS}^{(n)}) \to 0$ as $n \to +\infty$ (see, e.g., Fahrmeir et al., 2013,

Section 3.2). Hence, the LS estimator is consistent for θ with respect to the MSE, which implies its weak consistency.

For a given data set, the expression for $\hat{\theta}_{LS}$ is only well-defined if the matrix $D^{\mathrm{T}}D$ is invertible, which requires that the columns of D have to be linearly independent. In the presence of multicollinearity, that is, if two or more explanatory variables in X are strongly correlated, $D^{\mathrm{T}}D$ is usually still invertible but the inverse matrix $(D^{\mathrm{T}}D)^{-1}$ can have large diagonal elements, due to the fact that the determinant of $D^{\mathrm{T}}D$ is relatively small when some of its columns are strongly correlated. Hence, in this situation, the LS estimators $\hat{\theta}_{LS,l}$ of the regression coefficients θ_l, with $l \in \{0, \ldots, d\}$, can have very large variances. This means that the obtained estimates of the LS regression can differ a lot from the true coefficients of interest. Therefore, Hoerl and Kennard (1970) proposed to add a small number $\lambda > 0$ on the diagonal of $D^{\mathrm{T}}D$ alleviating the multicollinearity, in order to reduce the variance of the regression estimator in this situation. The resulting Ridge estimator, for a fixed $\lambda > 0$, is given by

$$\hat{\theta}_{R,\lambda} = (D^{\mathrm{T}}D + \lambda\, I_{d+1})^{-1}\, D^{\mathrm{T}}y,$$

where I_{d+1} is the $(d+1)$-dimensional identity matrix. Since for each fixed $\lambda > 0$, we have $\mathbb{E}(\hat{\theta}_{R,\lambda}) = (D^{\mathrm{T}}D + \lambda\, I_{d+1})^{-1}\, D^{\mathrm{T}}D\,\theta$, the Ridge estimator is biased. However, $\hat{\theta}_{R,\lambda}$ is a consistent estimator for the vector of regression coefficients in the standard linear regression model. To see this, assume again that $((D^{(n)\,\mathrm{T}}D^{(n)})^{-1})_{n\in\mathbb{N}}$ converges to the zero matrix, which implies that the diagonal elements of the matrix $D^{(n)\,\mathrm{T}}D^{(n)}$, given by n^2 and by $\sum_{i=1}^{n} x_{i,l}^2$ for all $l \in \{1, \ldots, d\}$, increase as n gets larger. Therefore, the effect of the regularization parameter λ added to each of these diagonal elements vanishes in the limit and $\hat{\theta}_{R,\lambda}$ behaves asymptotically like the LS estimator. Thus, for the sequence $(\hat{\theta}_{R,\lambda}^{(n)})_{n\in\mathbb{N}}$ associated with a fixed $\lambda > 0$, we have that $MSE(\hat{\theta}_{R,\lambda}^{(n)}) \to 0$ as $n \to +\infty$. In a practical analysis, however, to fix the regularization parameter at an appropriate value is a crucial problem. Hoerl and Kennard (1970, Theorem 4.3) showed that there always exists a $\lambda > 0$ such that the MSE of the Ridge estimator is

smaller than the MSE of the LS estimator. Hence, despite the bias introduced by λ, if λ is appropriately chosen, the reduction of the estimator's variance is so large that $\hat{\theta}_{R,\lambda}$ is more efficient than $\hat{\theta}_{LS}$. In fact, there is a trade-off between a small bias if λ is small and a small variance of the estimator for large values of λ. Yet, which $\lambda > 0$ corresponds to an efficient Ridge estimator, is usually unknown in a practical setting, nevertheless, λ has to be fixed a priori. A common technique to select the regularization parameter is to apply a cross-validation scheme, but there are many other approaches (see, e.g., Hastie et al., 2009, Chapter 7; Draper and van Nostrand, 1979). As the variances of the components of the Ridge estimator are in general smaller than those of the LS estimator, the coefficients are effectively shrunk in their size. Since its introduction by Hoerl and Kennard (1970), Ridge regression was generalized in many ways by using different shrinkage methods for the regression coefficients. Moreover, the general idea of regularized estimation can be applied to various statistical problems and numerous statistical methods employing this idea emerged during the past few years. For more details on Ridge regression methods and modern regularization techniques, see, e.g., Fahrmeir et al. (2013); Hastie et al. (2009); Draper and van Nostrand (1979).

The Ridge estimator $\hat{\theta}_{R,\lambda}$ can be derived as the minimizer of the modified LS criterion $(y - D\,\theta)^{\mathrm{T}}(y - D\,\theta) + \lambda\,\theta^{\mathrm{T}}\theta$, which is composed of the residual sum of squares and a penalty for the length of the vector of regression coefficients, for a fixed $\lambda > 0$. Equivalently, the corresponding penalized risk $E_{\hat{P}_V}(f_\theta) + {}^{\lambda}/_{n}\,\theta^{\mathrm{T}}\theta$ can be considered as the criterion to be minimized. The latter expression looks very similar to the regularized risk of (5.3) that is minimized in SVR. In fact, the Ridge estimator can alternatively be derived as the solution of an SVR based on the LS loss and on the linear kernel. More precisely, the SVR formulation of the linear regression problem corresponds to the dual problem of the minimization problem associated with the penalized LS criterion leading to the Ridge estimator, as shown in the following. However, the interpretation of the results is not the same in both contexts due to the different underlying probability models.

Consider the minimization problem of the penalized LS criterion in the following form

$$\min_{\theta \in \mathbb{R}^{d+1}} \ \frac{1}{2} (y - D\,\theta)^{\mathrm{T}} (y - D\,\theta) + \frac{\lambda}{2} \theta^{\mathrm{T}} \theta,$$

where the criterion is divided by 2 to make the computations more convenient. Directly solving this problem leads to the Ridge estimator $\hat{\theta}_{R,\lambda}$. Alternatively, the minimization problem can be reformulated as

$$\min_{(\xi,\theta) \in \mathbb{R}^n \times \mathbb{R}^{d+1}} \ \frac{1}{2\,\lambda} \xi^{\mathrm{T}} \xi + \frac{1}{2} \theta^{\mathrm{T}} \theta, \quad \text{subject to } \xi = y - D\,\theta, \tag{5.4}$$

where $\xi \in \mathbb{R}^n$. That is, the unconstrained minimization problem is transformed to a minimization problem with n equality constraints. The constrained minimization problem can be solved by applying the Lagrange multiplier rule (see, e.g., Reinhardt et al., 2013, Sections 2.2 and 4.1). The corresponding Lagrangian function $\Lambda : \mathbb{R}^n \times \mathbb{R}^{d+1} \times \mathbb{R}^n \to \mathbb{R}$ with Lagrange multipliers $\alpha \in \mathbb{R}^n$ for the constraints is given by

$$\Lambda(\xi, \theta, \alpha) = \frac{1}{2\,\lambda} \xi^{\mathrm{T}} \xi + \frac{1}{2} \theta^{\mathrm{T}} \theta + \alpha^{\mathrm{T}} (y - D\,\theta - \xi),$$

for all $(\xi, \theta, \alpha) \in \mathbb{R}^n \times \mathbb{R}^{d+1} \times \mathbb{R}^n$. As the minimum under the restrictions is attained at the saddle point of Λ, a possible solution $(\check{\xi}, \check{\theta}, \check{\alpha})$ must satisfy $\nabla\Lambda(\check{\xi}, \check{\theta}, \check{\alpha}) = 0 \in \mathbb{R}^n \times \mathbb{R}^{d+1} \times \mathbb{R}^n$. The partial derivatives of Λ with respect to the primal variables, θ and ξ, imply the following conditions for the critical point of the Lagrangian

$$\xi = \lambda\,\alpha \quad \text{and} \quad \theta = D^{\mathrm{T}} \alpha.$$

We can use these conditions to express $\check{\xi}$ and $\check{\theta}$ as functions of the dual variables α and to replace the primal variables in Λ by these expressions. Then, we only need to maximize $\Lambda_{\check{\xi},\check{\theta}}$ defined by $\Lambda_{\check{\xi},\check{\theta}}(\alpha) = \Lambda(\check{\xi}, \check{\theta}, \alpha)$ for all $\alpha \in \mathbb{R}^n$. This corresponds to the dual problem of the minimization

problem in (5.4). The dual solution $\check{\alpha}$ is given by

$$\check{\alpha} = (D\,D^{\mathrm{T}} + \lambda\,I_n)^{-1}\,y.$$

Finally, we get $\hat{\theta}_{R,\lambda}$ by substituting $\check{\alpha}$ for α in the primal condition of θ

$$\hat{\theta}_{R,\lambda} = D^{\mathrm{T}}\check{\alpha} = D^{\mathrm{T}}(D\,D^{\mathrm{T}} + \lambda\,I_n)^{-1}\,y,$$

which is a slightly different (but equivalent) expression for the Ridge estimator than the one introduced by Hoerl and Kennard (1970). The estimated function $f_{\hat{\theta}_{R,\lambda}}$ can then be written for all $x_i \in \mathcal{X}$ as

$$f_{\hat{\theta}_{R,\lambda}}(x_i) = (1, x_{i,1}, \ldots, x_{i,d})\,D^{\mathrm{T}}\check{\alpha} = \sum_{j=1}^{n} \check{\alpha}_j(1 + \textstyle\sum_{l=1}^{d} x_{i,l}\,x_{j,l}).$$

This expression, in fact, corresponds to a support vector expansion of the linear function $f_{\hat{\theta}_{R,\lambda}}$. To see this, consider (5.2) with x replaced by x_i and κ by the linear kernel, given by $\kappa(x_i, x_j) = 1 + \langle x_i, x_j\rangle = 1 + \sum_{l=1}^{d} x_{i,l}\,x_{j,l}$, for all $x_i, x_j \in \mathcal{X}$. To verify the equivalence of $f_{\hat{\theta}_{R,\lambda}}$ with the solution of an SVR with linear kernel and LS loss, we consider the corresponding optimization problem in the following.

In an SVR with linear kernel, the corresponding RKHS $\mathcal{F}$ of considered regression functions contains all (affine) linear functions on $\mathcal{X}$. The loss function of the regression problem is the risk functional E_{P_V} with ψ defined by $\psi(r) = r^2$, for all $r \in \mathbb{R}_{\geq 0}$. Given some observations $V_1 = v_1, \ldots, V_n = v_n$, the optimal regression function $f_{\hat{P}_V,\tilde{\lambda}}$ is the function that minimizes the regularized risk associated with the empirical distribution of the data, defined in (5.3), for some $\tilde{\lambda} > 0$. According to the Representer Theorem, the set $\mathcal{F}_n \subset \mathcal{F}$ of candidates for the SVM contains functions of the form $f_\alpha(\cdot) = \sum_{j=1}^{n} \alpha_j\,(1 + \langle \cdot, x_j\rangle)$, with $\alpha \in \mathbb{R}^n$. Hence, the corresponding minimization problem can be written as

$$\min_{\alpha\in\mathbb{R}^n}\ \frac{1}{2\,n}\,(y - D\,D^{\mathrm{T}}\alpha)^{\mathrm{T}}(y - D\,D^{\mathrm{T}}\alpha) + \frac{\tilde{\lambda}}{2}\,\alpha^{\mathrm{T}}D\,D^{\mathrm{T}}\alpha,$$

which is again divided by 2 for convenience. This minimization problem can also be transformed to a constrained minimization problem, whose Lagrangian $\tilde{\Lambda}$ with multipliers $\mu \in \mathbb{R}^n$ is defined by

$$\tilde{\Lambda}(\xi, \alpha, \mu) = \frac{1}{2\,n\,\tilde{\lambda}}\,\xi^{\mathrm{T}}\xi + \frac{1}{2}\,\alpha^{\mathrm{T}} D\,D^{\mathrm{T}}\alpha + \mu^{\mathrm{T}}(y - D\,D^{\mathrm{T}}\alpha - \xi),$$

for all $(\xi, \alpha, \mu) \in \mathbb{R}^{3\,n}$. The partial derivatives of $\tilde{\Lambda}$ with respect to the primal variables, ξ and α, imply the conditions

$$\xi = \tilde{\lambda}\,n\,\mu \quad \text{and} \quad \alpha = \mu.$$

Inserting these in $\tilde{\Lambda}$, we obtain the corresponding dual problem, given as

$$\max_{\alpha \in \mathbb{R}^n} \; \frac{\tilde{\lambda}\,n}{2}\,\alpha^{\mathrm{T}}\alpha + \frac{1}{2}\,\alpha^{\mathrm{T}} D\,D^{\mathrm{T}}\alpha + \alpha^{\mathrm{T}}(y - D\,D^{\mathrm{T}}\alpha - \tilde{\lambda}\,n\,\alpha),$$

whose objective function equals the function $\Lambda_{\tilde{\xi},\tilde{\theta}}$, if we set $\tilde{\lambda} = \lambda/n$. Hence, for $\tilde{\lambda} = \lambda/n$, the optimization problems of SVR and Ridge regression are equivalent and consequently have the same solution $f_{\hat{P}_V,\tilde{\lambda}} = f_{\hat{\theta}_{R,\lambda}}$. Thus, Ridge regression can be considered as a special case of SVR.

The interpretations of the estimates, however, are different. In the case of Ridge regression, the model assumptions imply that the linear regression function models the conditional expectation of Y given $X = x$ for all $x \in \mathcal{X}$, from which Y deviates for all $x \in \mathcal{X}$ with the same variance τ^2. In this setting, the best regression function minimizes this conditional variance τ^2. In the SVR framework, the consideration of the LS loss implies that the optimal regression function describes the conditional expectation $\mathbb{E}(Y|x)$ for all $x \in \mathcal{X}$, yet, without further assumptions about the random behavior of Y. Here, the optimal line minimizes the mean conditional variance of Y given $X = x$ over $\mathcal{X}$, that is, the marginal variance of R_f. Therefore, $f_{\hat{P}_V,\tilde{\lambda}}$ has a more general interpretation than $f_{\hat{\theta}_{R,\lambda}}$. Of course, this only applies, if the model is correctly specified, that is, if the conditional expectation of Y given $X = x$ is a linear function according to the true probability measure P_V. If the conditional expectation is not

given by a linear function, we can still define as function of interest the line minimizing the second moment of the distribution of the residuals, which is asymptotically given by $f_{\hat{\theta}_{LS}}$, and thus, can also be consistently estimated by $f_{\hat{P}_V,\tilde{\lambda}}$. However, the meaning of this function is less clear. In addition, in most practical settings, the assumption of a linear conditional expectation appears too idealistic. Therefore, LS regression would be more interesting if no particular form of the possible regression functions has to be imposed. In the SVR framework, this can easily be done by considering kernel functions with very large RKHSs like, for instance, the Gaussian kernel. Nevertheless, the representing function ψ of the LS loss is not Lipschitz continuous, which prevents the corresponding SVR methods from being robust. A better configuration of SVR in this regard would be, for example, the one employing the Gaussian kernel and the representing function of the ν-insensitive loss.

Regarding the asymptotic behavior of the SVM in the setting of linear regression with LS loss, we consider again the notion of risk consistency discussed in the context of more general SVR methods in the previous subsection. Risk consistency of an SVM $f_{\hat{P}_V,\lambda}$ means that $E_{P_V}(f_{\hat{P}_V,\lambda^{(n)}}) \to \inf_{f\in\mathcal{F}} E_{P_V}(f)$ in probability as $n \to +\infty$ and $\lambda^{(n)} \to 0$. In the case of the LS loss, we know that the infimal risk of the limit is attained by the function that minimizes $\mathbb{E}(R_f^2)$. By definition, the linear function $f_{\hat{\theta}_{LS}}$ that is associated with the LS estimator for the regression coefficients minimizes the risk functional $E_{\hat{P}_V}$ associated with the empirical distribution of a finite sample of observations $V_1 = v_1, \ldots, V_n = v_n$. As more and more data are observed, $\hat{P}_V$ converges to the unknown probability distribution P_V, which implies that the sequence $(E_{P_V}(f_{\hat{\theta}_{LS}^{(n)}}))_{n\in\mathbb{N}}$ converges in probability to $\inf_{f\in\mathcal{F}} E_{P_V}(f)$. Furthermore, when we consider a sequence of penalty parameters, $(\lambda^{(n)})_{n\in\mathbb{N}}$ with $\lim_{n\to+\infty} \lambda^{(n)} = 0$, in the limit, the corresponding sequence of SVMs $(f_{\hat{P}_V,\lambda^{(n)}})_{n\in\mathbb{N}}$ behaves like the sequence of functions associated with the LS estimator. Thus, it can be shown that $E_{P_V}(f_{\hat{P}_V,\lambda^{(n)}}) \to \inf_{f\in\mathcal{F}} E_{P_V}(f)$ in probability as $n \to +\infty$ and $\lambda^{(n)} \to 0$, that is, the SVR estimator based on the linear kernel together with the LS loss is risk consistent.

5.2 Adaptation of SVR to interval data

Utkin and Coolen (2011) proposed a generalization of the SVR methodology to the situation in which the response variable Y is observed as bounded intervals. That is, Utkin and Coolen (2011) consider the precise variables $(X, Y) = V \in \mathcal{V}$, where $\mathcal{V} = \mathcal{X} \times \mathcal{Y}$ is a compact subset of $\mathbb{R}^{d+1}$, with $d \in \mathbb{N}$. Instead of V, only the random set $V^* \subseteq \mathcal{V}$ can be observed, whose possible realizations are of the form $V^* = \{X\} \times [\underline{Y}, \overline{Y}]$, with $X \in \mathcal{X} \subset \mathbb{R}^d$ and $\underline{Y}, \overline{Y} \in \mathcal{Y} \subset \mathbb{R}$ such that $\underline{Y} \leq \overline{Y}$. Moreover, it is assumed that $(V, V^*) \sim P \in \mathcal{P}$, where $\mathcal{P}$ entails all probability measures P' satisfying $P'(V \in V^*) = 1$, which corresponds to the probability model underlying the robust LIR method with $\varepsilon = 0$ in (3.1). Hence, like it is done in most approaches to analyzing imprecise data, Utkin and Coolen (2011) assume that the imprecise data always contain the precise values of interest.

Since the precise variables are not observable, it is impossible to evaluate the considered regression functions $f \in \mathcal{F}$, where $\mathcal{F}$ is an RKHS, by the associated empirical risk $E_{\hat{P}_V}(f)$. However, the marginal distribution of the imprecise data can be estimated on the basis of data. When the probability distribution P_{V^*} of the imprecise data is known, the only available information about the unknown probability distribution P_V of the precise data is that $P_V \in [P_{V^*}]$, where $[P_{V^*}] \subset \mathcal{P}_V$ is the set of all marginal distributions P'_V of the precise data corresponding to models $P' \in \mathcal{P}$ with $P'_{V^*} = P_{V^*}$. As explained in Section 3.2, since we assume here that $P(V \in V^*) = 1$, the set $[P_{V^*}]$ consists of all probability measures on $\mathcal{V}$ that satisfy the inequalities in (3.4). Hence, the unknown probabilities $P_V(V \in A)$ of all measurable events $A \subseteq \mathcal{V}$ are bounded by

$$\begin{aligned} &P_V(V \in A) \geq P_{V^*}(V^* \in \{A' \in \mathcal{V}^* \,:\, A' \subseteq A\}) \quad \text{and} \\ &P_V(V \in A) \leq P_{V^*}(V^* \in \{A' \in \mathcal{V}^* \,:\, A' \cap A \neq \varnothing\}). \end{aligned} \tag{5.5}$$

By consequence, for all $f \in \mathcal{F}$, the unknown risk $E_{P_V}(f)$ lies in the interval

$[\underline{E}_{P_{V^*}}(f), \overline{E}_{P_{V^*}}(f)]$, where

$$\underline{E}_{P_{V^*}}(f) = \min_{P'_V \in [P_{V^*}]} E_{P'_V}(f) \quad \text{and} \quad \overline{E}_{P_{V^*}}(f) = \max_{P'_V \in [P_{V^*}]} E_{P'_V}(f).$$

Hence, in contrast to standard SVR, the regression functions $f \in \mathcal{F}$ cannot be directly evaluated by a precise value here, even if the probability distribution of the observable data is known. Therefore, in the decision problem corresponding to SVR with imprecise response, the set $[\underline{E}_{P_{V^*}}(f), \overline{E}_{P_{V^*}}(f)]$ of all possible evaluations is considered for each $f \in \mathcal{F}$. Of course, it is in general impossible to directly determine an SVM with respect to this imprecise optimization criterion. The central idea of the regression methodology proposed by Utkin and Coolen (2011) is to use the minimin or the minimax rule to solve the decision problem, that is, to minimize either the lower risk $\underline{E}_{P_{V^*}}$ or the upper risk $\overline{E}_{P_{V^*}}$ in order to identify a single optimal regression function.

To derive expressions of the lower and upper risk, Utkin and Coolen (2011) describe, for each regression function $f \in \mathcal{F}$, the set of compatible probability distributions of the residual R_f given P_{V^*} by a so-called p-box and apply results from Utkin and Destercke (2009). Introduced by Ferson et al. (2003, Section 2), the notion p-box designates a convex set of probability measures for a univariate random quantity that is bounded by a lower and an upper cumulative distribution function. In the situation considered here, given P_{V^*}, also the marginal distribution of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, where $\underline{R}_f = \min_{(x,y)\in V^*} |y - f(x)|$ and $\overline{R}_f = \max_{(x,y)\in V^*} |y - f(x)|$, is known for each $f \in \mathcal{F}$. According to (5.5), the marginal distribution of the imprecise residual implies lower and upper bounds to the probabilities of all measurable events associated with the marginal distribution of the precise residual R_f. If we consider these lower and upper bounds for all events of the form $[-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, we obtain a lower and an upper cumulative distribution function that constitute a p-box. As the p-box covers all probability distributions of R_f that comply with the bounds at least for the intervals $[-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, some of the probability measures included in the p-box may not satisfy

(5.5) for all measurable events, and thus, may be incompatible with the marginal distribution of the imprecise residual. However, the p-box obtained in the described way from the random set $[\underline{R}_f, \overline{R}_f]$, with $f \in \mathcal{F}$, is the tightest outer approximation by a p-box of the set of probability distributions of R_f implied by this random set (see, e.g., Destercke et al., 2008). In fact, in the present situation, for each $f \in \mathcal{F}$, the upper bound of the associated p-box corresponds to the cumulative distribution function of the lower endpoint of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, while the lower bound of the p-box corresponds to the cumulative distribution function of the upper endpoint. This can be seen by considering the corresponding bounds to the probabilities of the events $[-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, used to derive the p-box for all $f \in \mathcal{F}$, that is,

$$\begin{aligned} P_V(R_f \leq r) &\geq P_{V^*}([\underline{R}_f, \overline{R}_f] \in \{[\underline{r}, \overline{r}] \subset \mathbb{R}_{\geq 0} \,:\, [\underline{r}, \overline{r}] \subseteq [-\infty, r]\}) \\ &= P_{V^*}(\overline{R}_f \leq r) \quad \text{and} \\ P_V(R_f \leq r) &\leq P_{V^*}([\underline{R}_f, \overline{R}_f] \in \{[\underline{r}, \overline{r}] \subset \mathbb{R}_{\geq 0} \,:\, [\underline{r}, \overline{r}] \cap [-\infty, r] \neq \varnothing\}) \\ &= P_{V^*}(\underline{R}_f \leq r). \end{aligned}$$

It can easily be checked that the probability distributions corresponding to the bounds of the p-box comply with (5.5) for arbitrary measurable events, and thus, are elements of $[P_{V^*}]$. Since, according to (5.1), the risk functional E_{P_V} is the expectation of a convex function in R_f with minimum at zero, it is straightforward to conclude that $\underline{E}_{P_{V^*}}$ and $\overline{E}_{P_{V^*}}$ coincide with the expected errors associated with the marginal distributions of the lower and of the upper residual, that is, of $\underline{R}_f$ and of $\overline{R}_f$, respectively (see also Utkin and Destercke, 2009, Proposition 3).

As the true probability distribution P_{V^*} is typically unknown, it is usually estimated on the basis of an i.i.d. sample of imprecise observations $V_1^* = A_1, \ldots, V_n^* = A_n$. By analogy with standard SVR, P_{V^*} is estimated by the empirical distribution $\hat{P}_{V^*}$ of the imprecise data, i.e., by the ML estimate, and furthermore, the complexity of the estimated functions is restricted by an additive penalty term. Hence, the optimization criteria

considered in the modified decision problems corresponding to the minimin and the minimax rule are the regularized lower and upper risk, respectively. For a fixed $\lambda > 0$, the regularized lower and upper risks associated with the empirical distribution $\hat{P}_{V^*}$ are for each $f \in \mathcal{F}$ given by

$$\underline{E}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n} \sum_{i=1}^{n} \min_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \, \|f\|_{\mathcal{F}}^2 \quad \text{and}$$

$$\overline{E}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n} \sum_{i=1}^{n} \max_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \, \|f\|_{\mathcal{F}}^2 \,,$$

where as before, ψ is the convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ representing the chosen loss. Following the same steps as in standard SVR, Utkin and Coolen (2011) deduce from these expressions solvable formulations of the optimization problems corresponding to both strategies in the special case of linear regression for different choices of the loss function. We do not restrict the approach to this special case here and continue to consider more general RKHSs of regression functions. However, before deriving formulations of the optimization problems based on the support vector expansion of the possible solution, it has to be verified that the Representer Theorem applies to or that its statements can be transferred to the setting considered here. Only in this case, the simple expression (5.2) can be used for the optimal regression function in (5.3), which provides a favorable starting point for solving the corresponding optimization problems.

As mentioned in Subsection 5.1.1, the Representer Theorem implies that if an SVR analysis of a data set $V_1 = v_1, \ldots, V_n = v_n$ with empirical distribution $\hat{P}_V$ is based on a convex representing function ψ, then, for all $\lambda > 0$, there exists a unique function minimizing $E_{\hat{P}_V,\lambda}$, which can be represented as (5.2) (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5). In the proof of this theorem as it is presented in Steinwart and Christmann (2008, Theorem 5.5), the first steps are to show strict convexity and continuity of $E_{\hat{P}_V,\lambda}$, which provide existence and uniqueness of the minimizing function $f_{\hat{P}_V,\lambda} \in \mathcal{F}$, by the corresponding arguments of the proofs of Theorem 5.2 and Lemma 5.1 of Steinwart and Christmann

(2008), respectively. Then, the representation of $f_{\hat{P}_V,\lambda}$ as the kernel expansion of (5.2) is derived by exploiting properties of the function spaces $\mathcal{F}_n$ and $\mathcal{F}$ in addition to the existence and the uniqueness of the function $f_{\hat{P}_V,\lambda}$. The generalized SVR methods discussed in this section differ from the standard SVR methods only in the expressions of their risks. In case that the summands of $\underline{E}_{\hat{P}_{V^*}}$ and $\overline{E}_{\hat{P}_{V^*}}$ are convex functions, those are also continuous, and thus, the continuity of $\underline{E}_{\hat{P}_{V^*},\lambda}$ and $\overline{E}_{\hat{P}_{V^*},\lambda}$ can be derived by slightly adapting the first argument of the existence part of the proof of the Representer Theorem. Therefore, the critical aspect of transferring the arguments of the proof of the Representer Theorem to the present situation appears to be the convexity of $\underline{E}_{\hat{P}_{V^*},\lambda}$ and $\overline{E}_{\hat{P}_{V^*},\lambda}$ and their components.

In the standard SVR methodology as set out in Subsection 5.1.1, the strict convexity of $E_{\hat{P}_V,\lambda}$ follows from the convexity of ψ, because a weighted sum of convex functions is convex, and from the fact that the mapping $f \mapsto \lambda \, \|f\|_{\mathcal{F}}^2$ is strictly convex by general properties of $\mathcal{F}$. In the situation considered here, $\overline{E}_{\hat{P}_{V^*},\lambda}$ is the sum of maxima over sets of convex functions and the strictly convex penalty term. As, in general, a function defined as the maximum of convex functions is convex, we can derive that $\overline{E}_{\hat{P}_{V^*},\lambda}$ is strictly convex. To show that $\underline{E}_{\hat{P}_{V^*},\lambda}$ is also strictly convex, however, requires some more effort.

Since the sum of convex functions is convex and the penalty term is strictly convex, the regularized lower risk associated with $\hat{P}_{V^*}$ is convex if, for each possible $A \in \mathcal{V}^*$, the mapping $f \mapsto \min_{(x,y)\in A} \psi(|y - f(x)|)$ is convex. Hence, it has to be verified that for every $\{x\} \times [\underline{y}, \overline{y}] = A \in \mathcal{V}^*$ the inequality

$$\begin{aligned} &\min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \psi(|y - (\rho\, f + (1-\rho)\, f')(x)|) \leq \\ &\rho \min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \psi(|y - f(x)|) + (1-\rho) \min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \psi(|y - f'(x)|) \end{aligned} \tag{5.6}$$

holds for all $f, f' \in \mathcal{F}$ and all $\rho \in [0,1]$. As ψ is a nonnegative convex

mapping with $\psi(0) = 0$, the minimum of $\psi(|y - f(x)|)$ over some interval $[\underline{y}, \overline{y}] \subseteq \mathcal{Y}$, for a given $x \in \mathcal{X}$ and some $f \in \mathcal{F}$, is attained either at $y = \underline{y}$, at $y = \overline{y}$, or at $y = f(x)$ if $f(x) \in [\underline{y}, \overline{y}]$. Hence, for each $f \in \mathcal{F}$ and $\{x\} \times [\underline{y}, \overline{y}] \in \mathcal{V}^*$ we have that

$$\min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \psi(|y - f(x)|) = \begin{cases} \psi(|\underline{y} - f(x)|) & \text{if} \quad f(x) < \underline{y}, \\ 0 & \text{if} \quad f(x) \in [\underline{y}, \overline{y}], \\ \psi(|\overline{y} - f(x)|) & \text{if} \quad f(x) > \overline{y}. \end{cases} \tag{5.7}$$

Now, starting from the expression on the left-hand side of (5.6), the convexity of the mapping ψ implies that

$$\min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \psi(|y - (\rho\, f + (1-\rho)\, f')(x)|) \leq \min_{(x,y)\in\{x\}\times[\underline{y},\overline{y}]} \rho\, \psi(|y - f(x)|) + (1-\rho)\, \psi(|y - f'(x)|).$$

Obviously, the right-hand side of this inequality is in general larger or equal the right-hand side of (5.6). Therefore, we cannot derive Inequality (5.6) directly from the convexity of ψ. One possibility for Inequality (5.6) to hold is that both sides are equal. In fact, it can be shown that the equality generally holds, by considering all 3^2 different cases resulting from the distinctions in (5.7) for both functions, f and f', for the expression on the right-hand side and verifying that the left-hand side yields the same value. For example, consider the situation in which $f(x) \in [\underline{y}, \overline{y}]$ and $f'(x) \in [\underline{y}, \overline{y}]$, and thus, the right-hand side is zero. In this case, also the convex combination $\rho\, f(x) + (1-\rho)\, f'(x)$ is for all $\rho \in [0, 1]$ contained in $[\underline{y}, \overline{y}]$, and thus, the left-hand side is always zero, too. Hence, for this case, the equality in (5.6) is verified. Finally, the remaining $3^2 - 1$ cases have to be checked, which can easily be done.

Thus, the regularized lower and upper risks associated with the empirical distribution of the imprecise data are indeed strictly convex. Therefore, the statements of the Representer Theorem can be transferred to the SVR methods proposed by Utkin and Coolen (2011) and $f(x_i)$ can be replaced

by $\sum_{j=1}^{n} \alpha_j \, \kappa(x_i, x_j)$, where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and κ is the reproducing kernel function of $\mathcal{F}$, in the expressions of $\underline{E}_{\hat{P}_{V^*},\lambda}$ and $\overline{E}_{\hat{P}_{V^*},\lambda}$. On the basis of the simplified expressions, solvable formulations of the corresponding optimization problems can be derived for different configurations of loss and kernel function, which provide the basis for an implementation of the associated generalized SVR methods. For example, Utkin and Coolen (2011) deduce the optimization problems associated with the minimin and the minimax rule for different loss functions combined with the linear kernel function.

5.3 Discussion

In the adaptation of the standard SVR methodology to interval-valued observations of the response variable suggested in Utkin and Coolen (2011), each regression function is evaluated by the interval of the risk values associated with all probability measures that are compatible with the probability distribution of the imprecise data. To solve the decision problem, the minimin or the minimax decision rule is applied, yielding a single estimated regression function. The proposed regression methodology can be seen as a generalization of standard SVR, because the suggested methods reduce to standard SVR methods if the data are in fact precisely observed. However, it is not clear if the obtained functions constitute meaningful results for the regression problem with imprecise data.

As discussed in Section 4.4, we consider as a basic requirement for a precise SVR estimator based on imprecise data that it should yield a result that could be obtained by the corresponding standard SVR method with a precise data set that is compatible with the observed imprecise data. Since the estimated function resulting from the minimin method has the smallest regularized lower risk associated with $\hat{P}_{V^*}$ over all $f \in \mathcal{F}$, the configuration of precise data $(x_1, y_1), \ldots, (x_n, y_n)$, with $(x_i, y_i) \in \{x_i\} \times [\underline{y}_i, \overline{y}_i]$ for all $i \in \{1, \ldots, n\}$, corresponding to the minimal risk for this function yields a higher regularized risk for any other function $f \in \mathcal{F}$. Therefore, the regression function obtained from the minimin method corresponds to the

standard SVM for this particular precise data set. Hence, the generalized SVR method employing the minimin decision rule yields a result that is meaningful in the sense mentioned above. If the same holds for the SVM obtained by the minimax method, however, remains an open question. Nevertheless, the generalized SVR method based on the minimax decision rule yields a result that is plausible in another sense, as we discuss in the following section.

Furthermore, compared with the robust LIR method, the considered data situation is more restricted, because it is assumed that $\mathcal{X} \times \mathcal{Y} = \mathcal{V}$ is a compact subset of $\mathbb{R}^{d+1}$. The assumption that $\mathcal{X}$ is compact is transferred from standard SVR. The compactness assumption for $\mathcal{Y}$ is necessary because a single unbounded observation would imply that $\overline{E}_{\hat{P}_{V^*},\lambda}(f) = +\infty$ for all $f \in \mathcal{F}$, and thus, would cause the minimax method to break down. This is due to the fact that the SVR methodology is based on the expected error, which is not a robust centrality measure for the distribution of $\psi(R_f)$. However, in most practical settings, the range of possible values of the response variable is not naturally bounded and there is typically not enough information to justify a particular choice of the lower and upper bounds, which have a strong effect on the obtained result. Moreover, a generalization of the approach to imprecisely observed explanatory variables X appears to be very challenging, because in this situation the regression functions cannot be represented as the linear combination of kernel functions as implied by the Representer Theorem, since the imprecise observations cannot each be identified with a single function $\kappa(\cdot, x_j)$, for all $j \in \{1, \ldots, n\}$. These aspects clearly limit the applicability of SVR methods based on the approach proposed by Utkin and Coolen (2011).

Finally, in the context of the statistical analysis of imprecise data, methods yielding precise results are in general problematic, because a reasonable statistical method should reflect the imprecision of the data in its result. In addition, a responsible statistical analysis should always take the involved statistical uncertainty into account. The LIR methodology allows expressing both types of uncertainty by the extent of the set-valued result of the regression analysis. In fact, it can easily be shown that, for

each $f \in \mathcal{F}$, the interval $[\underline{E}_{\hat{P}_{V^*}}(f), \overline{E}_{\hat{P}_{V^*}}(f)]$ is the ML estimate of $E_{P_V}(f)$ in the situation considered by Utkin and Coolen (2011). Therefore, for this particular kind of imprecise data, we can alternatively derive a generalization of SVR within the LIR framework.

5.4 A LIR method for SVR with interval data

In this section, we propose an alternative adaptation of SVR based on the LIR methodology described in Chapter 3 for the particular setting considered in Utkin and Coolen (2011). Hence, we suppose that $\mathcal{V}$ is a compact subset of $\mathbb{R}^{d+1}$, with $d \in \mathbb{N}$, and that the response variable is observed as a bounded interval, i.e., V^* is of the form $\{X\} \times [\underline{Y}, \overline{Y}]$, with $X \in \mathcal{X} \subset \mathbb{R}^d$ and $\underline{Y}, \overline{Y} \in \mathcal{Y} \subset \mathbb{R}$ such that $\underline{Y} \leq \overline{Y}$. Furthermore, we consider the fully nonparametric probability model $\mathcal{P} = \mathcal{P}_\varepsilon$ with $\varepsilon = 0$ in Assumption (3.1).

Following the LIR methodology, we regard the regression problem with imprecise data as a decision problem on $\mathcal{F} \times \mathcal{P}$. Since the aim of a LIR analysis is to identify those functions in the RKHS $\mathcal{F}$ that well describe the relationship between the precise variables, the loss function of the decision problem is a characteristic that depends only on the probability distribution of the precise data. On the other hand, as the variables are only imprecisely observed, the likelihood function used to derive a set-valued decision criterion for the regression problem depends only on the marginal distribution of the imprecise data. Like in standard SVR, we consider as loss function the risk functional $E_{P_V}(f)$, assigning to each pair $(f, P) \in \mathcal{F} \times \mathcal{P}$ the expected error implied by f under the corresponding marginal distribution P_V of the precise variables. Adopting the terminology of Section 3.2, we define a function-specific loss function E_f for each $f \in \mathcal{F}$ by $E_f(P) = E_{P_V}(f)$, for all $P \in \mathcal{P}$. We can express E_f also as a function on $\mathcal{P}_V$ by writing $E'_f(P_V)$ instead of $E_f(P)$, for all $P \in \mathcal{P}$. Furthermore, we define an imprecise version E^*_f of E_f assigning to each marginal distribution P_{V^*} of the imprecise data the set $E^*_f(P_{V^*}) = \bigcup_{P_V \in [P_{V^*}]} E'_f(P_V)$ of all compatible risk values. Obviously,

for every function f, we have that $E^*_f(P_{V^*}) = [\underline{E}_{P_{V^*}}(f), \overline{E}_{P_{V^*}}(f)]$, for all $P_{V^*} \in \mathcal{P}_{V^*}$. The (normalized) likelihood function lik induced by an i.i.d. sample of imprecise data $V_1^* = A_1, \ldots, V_n^* = A_n$ assigns to each probability measure $P \in \mathcal{P}$ the probability with which P had predicted the observations relative to the highest possible one, see also (3.2). As only the imprecise data are available, lik is entirely determined by their marginal distribution P_{V^*}, hence, we write $lik(P) = lik^*(P_{V^*})$, for all $P \in \mathcal{P}$.

Within the LIR framework, the information provided by the likelihood function is used to determine likelihood-based confidence regions for the loss E_f associated with a function $f \in \mathcal{F}$, which constitute the imprecise decision criterion of the regression problem. The confidence regions $\mathcal{E}_{f,>\beta}$ can be expressed by means of the (normalized) profile likelihood function lik_{E_f} and the threshold $\beta \in (0,1)$ as

$$\mathcal{E}_{f,>\beta} = \{e \in \mathbb{R}_{\geq 0} \,:\, lik_{E_f}(e) > \beta\},$$

where lik_{E_f} is for all $e \in \mathbb{R}_{\geq 0}$ given by

$$lik_{E_f}(e) = \sup_{P_{V^*} \in \mathcal{P}_{V^*} \,:\, e \in E^*_f(P_{V^*})} lik^*(P_{V^*}),$$

see also (3.3), (3.5), and (3.6). As $\mathcal{V}$ is compact here, the support of the distribution of $\psi(R_f)$ is also a closed and bounded subset of $\mathbb{R}_{\geq 0}$, for each $f \in \mathcal{F}$. In this case, informative confidence regions for the expected value of a random quantity can be obtained under the fully nonparametric probability assumption, which are moreover intervals (see, e.g., Owen, 2001, Section 2.5).

For each function $f \in \mathcal{F}$, the confidence region $\mathcal{E}_{f,>\beta}$ contains all risk values associated with f that are plausible to certain degree, which is determined by the choice of $\beta \in (0,1)$. If the cutoff point β is chosen close enough to one such that only the empirical distribution $\hat{P}_{\hat{V}^*}$ exceeds the likelihood threshold, we have that $\mathcal{E}_{f,>\beta} = E^*_f(P_{\hat{V}^*})$. Thus, the interval $[\underline{E}_{\hat{P}_{V^*}}(f), \overline{E}_{\hat{P}_{V^*}}(f)]$ corresponds to the ML estimate of the risk $E_{P_V}(f)$

in the considered regression problem. According to the standard SVR methodology, we furthermore replace the risk E_{P_V} by the regularized risk $E_{P_V,\lambda}$, for some $\lambda > 0$, and consider the corresponding modified decision problem for the estimation. For some $\lambda > 0$ and $\beta \in (0,1)$, we obtain as confidence intervals for the regularized risk

$$\mathcal{E}_{f,\lambda,>\beta} = \{e + \lambda \left\|f\right\|_{\mathcal{F}}^2 : e \in \mathbb{R}_{\geq 0} \text{ and } lik_{E_f}(e) > \beta\},$$

for all $f \in \mathcal{F}$. The confidence regions $\mathcal{E}_{f,\lambda,>\beta}$ then constitute the decision criterion of the LIR method for SVR with interval data. If the cutoff point β is chosen close enough to one such that $\mathcal{E}_{f,\lambda,>\beta}$ is the ML estimate of the regularized risk $E_{P_V,\lambda}(f)$, for all $f \in \mathcal{F}$, the minimax method proposed by Utkin and Coolen (2011) corresponds to applying the LRM rule discussed in Section 3.2, which aims at a single optimal regression function. In contrast to this approach, in the LIR methodology, we apply the dominance principle to the imprecise decision criterion in order to identify the set of all regression functions that are plausible in the light of the observations. Hence, all functions $f \in \mathcal{F}$ for which

$$\inf \mathcal{E}_{f,\lambda,>\beta} \leq \inf_{f' \in \mathcal{F}} \sup \mathcal{E}_{f',\lambda,>\beta}$$

is satisfied are considered as the set-valued result of the regression analysis. This set consists of all regression functions that are plausible given the imprecise data. In Subsection 4.1.3, we discussed that the result of the robust LIR method always contains all compatible precise LQS regression functions. Likewise, the set of undominated regression functions obtained here is always a superset of the set of all accordingly configurated SVMs obtained from precise data sets that are compatible with the imprecise data. To see this, consider a precise data set $(x_1, y_1), \ldots, (x_n, y_n)$ with $(x_i, y_i) \in \{x_i\} \times [\underline{y}_i, \overline{y}_i]$, for all $i \in \{1, \ldots, n\}$, and the corresponding SVM $f = f_{\hat{P}_V,\lambda} = \arg\inf_{f' \in \mathcal{F}} E_{\hat{P}_V,\lambda}(f')$. For all regression functions $f' \in \mathcal{F}$, the regularized risk $E_{\hat{P}_V,\lambda}(f')$ associated with the empirical distribution of this compatible data set is larger or equal $E_{\hat{P}_V,\lambda}(f)$. Therefore, for each $f' \in \mathcal{F}$, the upper endpoint $\sup \mathcal{E}_{f',\lambda,>\beta}$ of the confidence in-

terval is also larger or equal $E_{\hat{P}_V,\lambda}(f)$, which implies that $E_{\hat{P}_V,\lambda}(f) \leq \inf_{f' \in \mathcal{F}} \sup \mathcal{E}_{f',\lambda,>\beta}$. As furthermore $E_{\hat{P}_V,\lambda}(f) \geq \inf \mathcal{E}_{f,\lambda,>\beta}$, we obtain that $\inf \mathcal{E}_{f,\lambda,>\beta} \leq \inf_{f' \in \mathcal{F}} \sup \mathcal{E}_{f',\lambda,>\beta}$, and thus, f is an undominated regression function. Finally, as the LRM function is always included in the set of undominated functions and since the sets of undominated functions are nested for different levels of β because the confidence intervals $\mathcal{E}_{f,\lambda,>\beta}$ are nested for each $f \in \mathcal{F}$, we can conclude that the result of the minimax method by Utkin and Coolen (2011) is meaningful in the sense that the obtained regression function is always an undominated regression function.

How to obtain the set of all undominated regression functions in a practical analysis, however, is a difficult question, because it appears to be difficult to derive an analytical expression of lik_{E_f}. Maybe it is possible to adapt some computation method proposed by Owen (2001, Section 2.9) to compute the confidence regions in the situation considered here. In the special case where $\mathcal{E}_{f,\lambda,>\beta}$ is the ML estimate of the risk for all $f \in \mathcal{F}$, an implementation of the LIR method for SVR can be based on the minimax method by Utkin and Coolen (2011). The minimax method allows determining the smallest regularized upper risk, which is necessary to identify the functions whose lower risk does not exceed this value. Thus, given the smallest upper bound, a random search over $\mathcal{F}_n$ can be performed to approximately determine the set of all undominated regression functions.

Chapter 6

Applications

In this chapter, the regression methods discussed in the previous chapters are applied to study two interesting questions in the contexts of social sciences and winemaking, respectively. At first, we investigate the relationship between the income and the perceived overall well-being of a person by means of the robust LIR method presented in Section 4.1. After this, the determination of the sensory quality of a particular variety of Portuguese red wine by its alcohol level is analyzed in employing the SVR methods discussed in Chapter 5, whose results are finally compared with those obtained by the robust LIR method.

6.1 Analysis of subjective well-being with the robust LIR method

In recent years, there has been a lively interest in analyzing subjective well-being in various disciplines of the social and behavioral sciences. In this context, one important question is how an increase in income translates to subjective well-being (see, e.g., Deaton, 2012; Clark et al., 2008; Diener and Biswas-Diener, 2002). Empirical studies in this field often use global measures of subjective well-being, which are obtained from a single survey question about the overall satisfaction with life. These global measures are indicators of the state of an individual's well-being, and therefore, it is

sensible to use them to analyze subjective well-being (Deaton, 2008), although, of course, they do not capture the entire complexity of the concept of well-being (Huppert et al., 2009). As single-item measures are usually measured on a discrete scale, they can be considered as coarse observations of the latent, continuous variable of interest *degree of subjective well-being*. The coarseness of the discrete values can be represented by intervals, thus, the LIR approach is suitable to analyze this kind of data. Moreover, when investigating the relationship between income and subjective well-being, sometimes also the income data are only available as classes, which represent, in fact, intervals that form a partition of the associated observation space $\mathbb{R}_{\geq 0}$. Finally, as the relationship between income and subjective well-being is usually assumed to be log-linear (see, e.g., Deaton, 2012; Diener and Biswas-Diener, 2002), we can conduct a linear LIR analysis with the logarithm of income as independent variable X and subjective well-being as dependent variable Y to analyze the relationship of interest, accounting for the imprecision of the data.

To this end, we use the robust LIR method presented in detail in Section 4.1, which is implemented for the analysis of interval data in the `linLIR` package (Wiencierz, 2013) for the statistical software environment R (R Core Team, 2013). Thus, all computations and graphs in this section are made with the `linLIR` package. We analyze data from the fifth round of the European Social Survey (ESS, Norwegian Social Science Data Services, 2010). The ESS is a biennial multi-country survey established to monitor changing attitudes and behavior of people in Europe. The data collected for the ESS are available free of charge on the website `www.europeansocialsurvey.org`.

Previous empirical studies indicated that the relationship between income and subjective well-being on the individual level is not the same in rich countries as in poor countries, and furthermore, that there may be different relationships for men than for women (see, e.g., Clark et al., 2005; Diener and Biswas-Diener, 2002). For these reasons, we choose Finland and Bulgaria as representatives for the groups of rich and poor European countries, respectively, and we analyze only the corresponding subsets of

the ESS data set. Furthermore, for each country, we perform separate LIR analyses for the subpopulations of women and men. From the variables included in the ESS data set, we retrieve the following ones: *household income* (net per month, in categories corresponding to the decile classes of the income distribution in each country) and *overall satisfaction with life* (on a discrete scale from 0 – *extremely dissatisfied* to 10 – *extremely satisfied*). In a data preprocessing step, the income classes are replaced by the corresponding intervals, then the interval endpoints are divided by the household size, and finally, the logarithmic transformation is made. The data on subjective well-being are changed from discrete values $0, 1, \ldots, 9, 10$ to intervals $[0, 0.5], [0.5, 1.5], \ldots, [8.5, 9.5], [9.5, 10]$. Hence, the independent and dependent precise quantities whose relationship is investigated by the linear LIR analysis are the logarithmic monthly net household income per capita in euros and the subjective well-being on a latent, continuous scale from 0 to 10, respectively.

The resulting four data frames contain each four columns: two for each of the analyzed variables, one column for the lower interval endpoint and one for the upper endpoint, which is the required data format for the `linLIR` package. Applying the function `idf.create` to these data frames, we create so-called interval data frame (`idf`) objects, which consist of a list of data frames, each containing the corresponding two columns of interval endpoints of one variable. For these `idf`-objects, the `linLIR` package provides a `summary` method as well as a `plot` method with two options. Figures 6.1 and 6.2 show the data plots of the four data sets we analyze. As the data sets consist of roughly 1 000 observations each, we used the two-dimensional histogram plot by choosing the option `typ="hist"` in the `plot` function. As expected, we notice that the marginal distribution of subjective well-being is concentrated at a higher level in Finland compared to Bulgaria, but there appear to be no big differences between men and women within the countries. Moreover, we can see that there are many observations that are unbounded with respect to X. This is partly caused by the high number of observations in the lowest and highest income classes. In addition to this, there is a significant percentage of completely missing

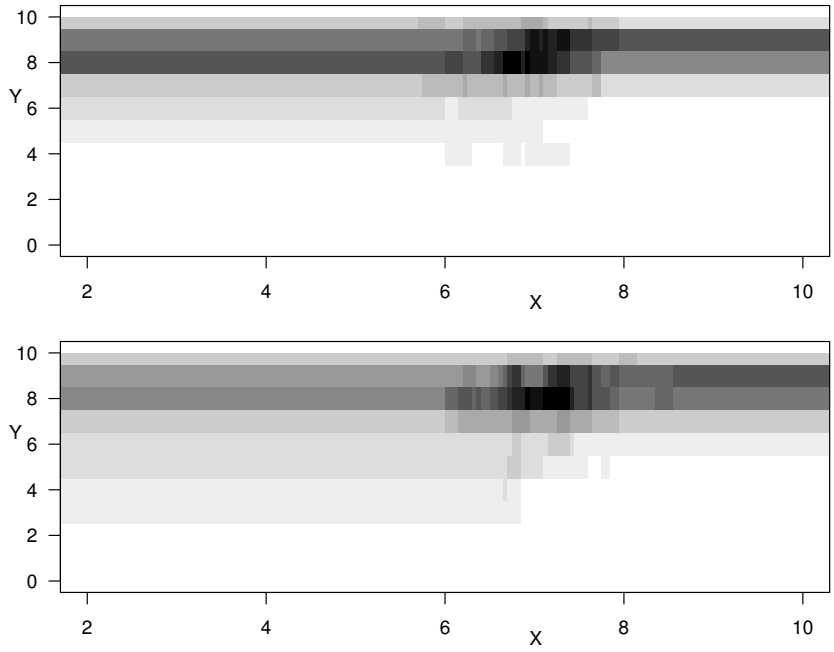

Figure 6.1: Histogram plots of the Finnish data sets: women on the top ($n = 967$), men on the bottom ($n = 911$). The darker a rectangle the more observations overlap this rectangle.

income values (Finland 5–10%, Bulgaria 15–20%), which are represented in the data set as intervals $[\underline{x}_i, \overline{x}_i] = [-\infty, +\infty]$. Given the high degree of data imprecision, we can expect to obtain rather uninformative results from the LIR analyses, reflecting the high uncertainty induced by the interval data. It can be argued that using $-\infty$ as lower endpoint of the range of the logarithmic income (instead of using, e.g., zero) entails too much unnecessary data uncertainty. However, the results of the LIR analyses are affected only a little by this, because the used LIR method is very robust.

Before conducting the linear LIR analyses, we have to set up the probability model by selecting the only model parameter ε characterizing the considered probability measures in terms of Assumption (3.1). Furthermore, we need to choose the quantile to be considered as loss function in the robust LIR method and the cutoff point β. For simplicity, we here

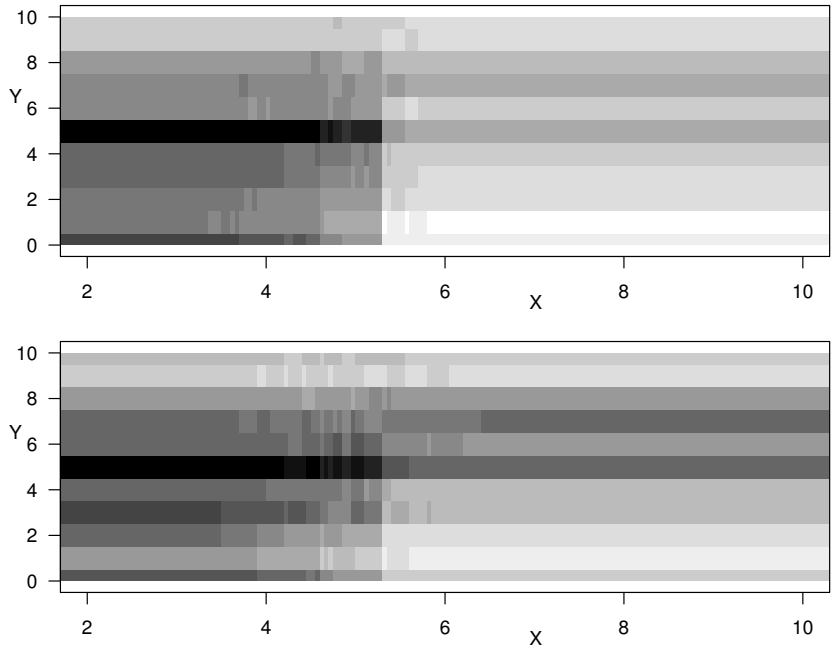

Figure 6.2: Histogram plots of the Bulgarian data sets: women on the top ($n = 1\,370$), men on the bottom ($n = 1\,064$). The darker a rectangle the more observations overlap this rectangle.

assume that the imprecise data are correct in the sense that the observed rectangles contain the correct precise values with probability one, i.e., we consider $\varepsilon = 0$. If we had concerns about the data quality or if we wanted to account for possibly wrong coarsening, a positive ε could be considered in the nonparametric probability model characterized by (3.1). As shown in Section 4.2, this would lead to more imprecise results of the LIR analyses, reflecting the fact that there is additional uncertainty. As the residual's quantile to be minimized we consider the median, that is, $p = 1/2$, which is the most robust choice of p. Finally, we choose $\beta = 0.8$ as cutoff point for the likelihood-based confidence regions $\mathcal{C}_f$ with $f \in \mathcal{F}$. This choice of β satisfies Condition (4.1) and corresponds to an asymptotic confidence level of at least approximately 50% for each $\mathcal{C}_f$.

The model parameter ε, the LIR settings p and β, as well as the `idf`-object to be analyzed are handed over to the `s.linlir` function of the R package `linLIR`, which determines the set $\mathcal{U}'$ by the exact algorithm. As we already mentioned at the end of Subsection 4.3.2, the current version of the function `s.linlir` is not optimized for computational speed. The computations for the present analysis took about two to ten minutes on a standard desktop computer. Most of the time is needed for the first part of the algorithm, where $\overline{q}_{LRM}$ is determined. To display the results of the conducted linear LIR analyses, we use the type `typ="para"` of the associated `plot` method with the default option `para.typ="polygon"` and obtain Figures 6.3 and 6.4, where the black points indicate the LRM regression functions.

The sets $\mathcal{U}'$ resulting from the LIR analyses of the data sets of women and men in Finland are displayed in Figure 6.3. Both sets of parameter values are bounded and have a similar shape, admitting both lines with positive and negative slopes ranging approximately from -9.5 to 12.

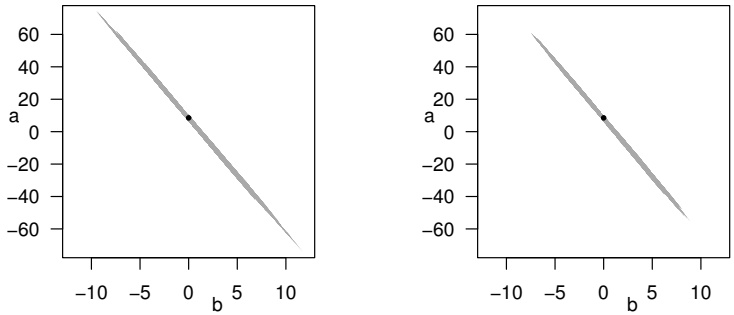

Figure 6.3: Sets $\mathcal{U}'$ for Finland: women on the left ($n = 967$), men on the right ($n = 911$).

For the sample of Bulgarian women, the shape of the obtained set $\mathcal{U}'$ is much different, as shown in the left part of Figure 6.4. In this particular data set, there are 687 observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that $\underline{x}_i = -\infty$ and $[\underline{y}_i, \overline{y}_i] \neq \mathbb{R}$. A line with an arbitrarily high slope always goes through these observations at the lower end of the income range as long as the intercept is not too low, and conversely, a line with a negative slope always intersects

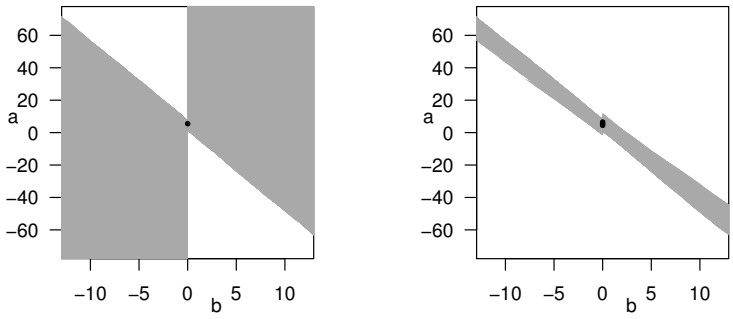

Figure 6.4: Sets $\mathcal{U}'$ for Bulgaria: women on the left ($n = 1\,370$), men on the right ($n = 1\,064$).

these observations if the intercept is not too high. As here $\underline{k} + 1 = 673$, all lines intersecting these 687 observations are undominated. Therefore, the obtained set of undominated functions is unbounded, reflecting the high degree of imprecision inherent in this data set. Furthermore, we here observe the particular data situation in which the set $\mathcal{U}'$ is not closed, that is, the borders at $b = 0$ are not included (see Section 4.3.1). In the LIR results for the sample of men in Bulgaria, $\mathcal{U}'$ is not unbounded, but large, which is to some extent due to the almost 20% of missing income values. In the right part of Figure 6.4, we displayed only the middle section of $\mathcal{U}'$. Interestingly, in this LIR analysis, we find three LRM regression lines. These lines can be characterized geometrically by the fact that the closed bands of width $2\,\overline{q}_{LRM} = 4$ around them completely include at least $\overline{k} = 543$ observations. In the present data set, there are only 500 observations bounded with respect to X, therefore, only the band around a horizontal line can contain at least 543 observations. Hence, each of the three LRM functions has slope 0.

The results of the LIR analyses do not give a clear answer to the question of how an increase in income translates to subjective well-being. However, the obtained results are more or less in line with current research in this field, as there is no clear evidence about the direct relationship between these two variables. Some empirical studies in rich countries found only very weak positive effects of income on subjective well-being, while

others even suggested a negative effect at the upper end of the income distribution (Diener and Biswas-Diener, 2002). These two possibilities are also admitted by the LIR results for the Finnish data sets, containing increasing and decreasing functions. In poorer countries, several studies found a strong positive effect, reflecting the fact that in these countries an increase in income is more often used to fulfill basic material needs that are clearly improving the individual living standard (Diener and Biswas-Diener, 2002). The LIR result for the sample of Bulgarian men admits more extreme slope and intercept values, while the data of the sample of Bulgarian women are too imprecise to obtain informative results.

6.2 Analysis of wine quality with generalized SVR methods

In this section, we analyze a data set collected to study the quality of Vinho Verde wines from Portugal. The data were obtained from wine samples that were tested by the official certification entity of the system of protected designation of origin of the Vinho Verde wines from May 2004 to February 2007. For each of the included 1 599 red and 4 898 white wines, 11 physicochemical characteristics and an evaluation of the sensory quality are available. The data set was initially analyzed by Cortez et al. (2009) and is freely available from the UC Irvine Machine Learning Repository (Bache and Lichman, 2013). Here, we concentrate on the subsample of red wines. An important determinant of the taste of red wine is its alcohol level. Therefore, in this section, we investigate the relationship between the alcohol content and the taste of red Vinho Verde wine by means of the generalized SVR methods introduced in Chapter 5 and we compare the results with those obtained by the robust LIR method.

The sensory quality of the wine is measured by the median evaluation of the wine over at least three test persons assessing the taste of the wine on a discrete scale ranging from 0 – *very bad* to 10 – *excellent*. Similar to the data on subjective well-being in the previous section, the discrete quality measurements can be considered as coarse observations of

an underlying continuous variable taking values in $[0, 10]$. Therefore, we base the regression analysis on the imprecise quality data where the discrete values $0, 1, \ldots, 9, 10$ are replaced by the intervals $[0, 0.5], [0.5, 1.5], \ldots, [8.5, 9.5], [9.5, 10]$. As explanatory variable of the regression analysis, we here consider the alcohol content of the wine. In the given data set, this quantity is given by measurements of the volume percent of alcohol in the wine that we assume to be sufficiently accurate. Hence, we analyze the relationship between the precisely observed alcohol content and the imprecisely observed sensory quality of the red Vinho Verde wine, which corresponds to the data situation required for the generalized SVR methods. The analyzed data set is displayed in the left graph of Figure 6.5, where X is the alcohol level in percent by volume and Y corresponds to the sensory quality. Again, all graphs and computations are realized in the statistical software environment R (R Core Team, 2013), resorting amongst others to functions provided by the packages `kernlab` (Karatzoglou et al., 2004) and `quadprog` (Turlach and Weingessel, 2013).

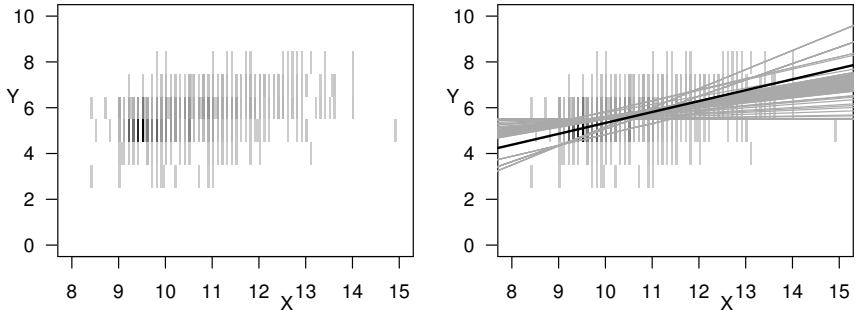

Figure 6.5: Histogram plot of the red wine data set (left, $n = 1\,599$) and minimax function (black line) together with draft of the set of compatible precise SVMs (right, 1 000 randomly chosen functions). The darker a line segment the more observations overlap this line segment.

Since the data suggest a positive linear relationship, the linear kernel function is chosen for the SVR methods. Furthermore, the identity map is considered as function ψ, which corresponds to considering the expected value of the (absolute) residual as risk. Finally, the regularization param-

eter λ is set to 0.0001. The black line in the right graph of Figure 6.5 indicates the regression line obtained by the generalized SVR method of Utkin and Coolen (2011) when adopting the minimax rule. Moreover, we consider the set of all standard SVMs (with the same configuration of κ, ψ, and λ) based on precise data sets that are compatible with the imprecise data. This set is sketched by the gray lines in the same graph and gives a first impression of the uncertainty about the relationship of interest associated with the imprecision of the data. The minimax function and the compatible SVMs include no decreasing lines. Hence, these results confirm the surmise of a positive linear relationship between alcohol content and taste of red Vinho Verde wines.

Now, we consider the LIR method for SVR described in Section 5.4 with the choice of β for which the confidence intervals cover only the maximum likelihood estimate of the regularized risk. The associated set of undominated regression functions is approximated in the way outlined at the end of Section 5.4. The left graph of Figure 6.6 shows the obtained results, including increasing and decreasing lines with slopes ranging approximately from -1.2 to 2. By definition, the LRM function of this LIR method with the here chosen cutoff point corresponds to the minimax function displayed above. However, the result obtained by the LIR method for SVR comprises also decreasing functions. As explained in Section 5.4, the set of undominated regression functions is in general a superset of the set of compatible SVMs. Hence, although we here consider only the most plausible regression functions, a decreasing relationship cannot be excluded on the basis of the likelihood inference underlying the LIR method for SVR.

For comparison, we analyzed the wine quality data set also by applying the robust LIR method presented in Section 4.1 with $p = 1/2$. This can easily be done by means of the `linLIR` package (Wiencierz, 2013), because we, in fact, consider a simple linear regression problem here. To make the results comparable, we furthermore assume $\varepsilon = 0$ and choose $\beta = 0.9999$, which implies that the confidence intervals constituting the decision criterion of the regression problem encompass only the maximum likelihood estimate of the median of the residuals' distribution. The results obtained

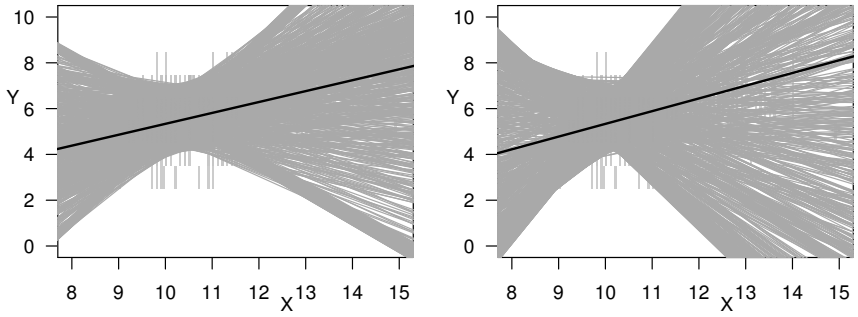

Figure 6.6: Drafts of the sets of undominated regression functions resulting from applying the LIR method for SVR (left, 1 000 randomly chosen functions) and the robust LIR method (right, 1 000 randomly chosen functions). The black lines indicate the corresponding LRM functions.

from the robust LIR analysis are displayed in the right graph of Figure 6.6. The extent of the associated set of undominated regression functions is visibly larger than the set corresponding to the LIR method for SVR, that is, steeper functions in both directions are allowed. This might be due to the penalization involved in SVR and to the fact that the robust LIR method always leads to very cautious inferences.

Chapter 7

Conclusion and outlook

In this thesis, the statistical problem of analyzing the relationship between a response variable and one or more explanatory variables when these quantities are only imprecisely observed was studied. The goal was to find a regression method for imprecise data that is general in the sense that it does not impose restrictive assumptions about the form of the imprecise observations, about the underlying probability distribution, and about the shape of the relationship between the variables of interest.

After a review of different approaches proposed in the literature constituting Chapter 2, a new likelihood-based approach to regression with imprecisely observed variables named LIR was introduced in Chapter 3. The LIR methodology consists in determining likelihood-based confidence regions for the loss of the regression problem on the basis of imprecise data and in regarding the set of all regression functions that are not strictly dominated as the imprecise result of the regression analysis. Hence, a LIR analysis usually yields an imprecise result, which can be interpreted as a confidence set for the unknown regression function. In Chapter 4, a robust regression method was derived from the general LIR methodology, where quantiles of the residuals' distribution are considered as loss. At first, the formal framework of the robust LIR method was presented in mathematical detail and further explained by means of illustrative examples. Moreover, an exact algorithm to implement this regression method for the

special case of simple linear regression with interval data was developed and implemented in an R package. Finally, selected statistical properties of the robust LIR method were thoroughly investigated. Chapter 5 dealt with an alternative regression methodology proposed by Utkin and Coolen (2011) for situations where only the response variable is imprecisely observed. This approach is based on SVR and was discussed in detail, before an alternative adaptation of SVR was developed by following the LIR approach, which further generalizes the methods suggested by Utkin and Coolen (2011). Finally, the discussed regression methods were applied to investigate two practical questions in the contexts of social sciences and winemaking, respectively, in Chapter 6. In both cases, the LIR analyses provided very cautious inferences.

The robust LIR method introduced in Chapter 4 meets all the targets outlined in the beginning of this thesis. The formal framework of this regression method encompasses all kinds of imprecise data and of possible regression functions and it imposes no considerable constraints on the set of probability measures considered as possible models of the analyzed situation. The only restriction is the assumption that the imprecise data contain the unobserved precise values with probability at least $1 - \varepsilon$, for some $\varepsilon \in [0, 1/2)$. Hence, in contrast to most alternative approaches to analyzing imprecise data, the LIR methodology permits accounting for coarsening errors and even allows informative coarsening in the nonparametric setting underlying the robust LIR method. Moreover, as found in Section 4.4, this LIR method is robust in terms of a high breakdown point and it yields highly reliable results in the sense that the coverage probability of the resulting set of regression functions seems to be generally rather high. Despite all these desirable features, the implementation of the robust LIR method poses a big challenge. The exact algorithm developed in Section 4.3.1 for the special case of simple linear regression with interval data, in principle, can be generalized to multiple linear regression. In more general regression problems, however, it is yet to be investigated whether there is a better implementation of the robust LIR method than an inner approximation of the set of undominated functions

by a random search over the set of considered regression functions, which can be computationally very demanding.

The regression methods discussed in Chapter 5 generalize standard SVR methods, whose results can be efficiently computed, even when a very large space of regression functions is considered as the set of possible descriptions of the relationship of interest. However, the applicability of the generalized SVR methods is much more limited compared to the robust LIR method, because only data situations with imprecisely observed responses but precisely observed explanatory variables can be considered. To adapt the generalized SVR methods to imprecisely observed explanatory variables appears to be very challenging, because in this case, the imprecise observations cannot each be identified with a single kernel function. A possible solution to this problem could be to consider fixed kernel functions similar to the basis functions at fixed knots considered in spline-based regression methods. However, such a modification would require a thorough investigation of its impact on the properties of the SVR estimators. Another possibility to obtain a feasible regression method when the shape of the analyzed relationship is not restricted could be to develop a LIR method that directly generalizes a standard regression method based on splines.

According to the general LIR methodology, the imprecise result of a LIR analysis consists of all regression functions that are plausible in the light of the imprecise data and its extent reflects the whole uncertainty about the relationship of interest. In practice, also prediction is an important goal of a regression analysis. Usually, (a region for) the value of the dependent variable given a future observation of the explanatory variables is predicted on the basis of a single estimated regression function. Yet, in the situation considered here, the additional observation of the explanatory variables is in general set-valued, while the set of undominated functions contains all plausible descriptions of the relationship between the precise quantities of interest. How to adapt the standard idea of prediction to this situation, is one of the fundamental questions that have to be answered, before it is possible to develop prediction techniques for LIR methods. In

Cattaneo and Wiencierz (2012), a joint prediction region for a complete future observation (including response and explanatory variables) was derived in the context of the robust LIR method. The topic of predicting (a region for) the response variable given an imprecise observation of the vector of explanatory variables will be addressed in the future.

Furthermore, the second application in Chapter 6 suggested that not only the robust LIR method but also the LIR method for SVR yields very reliable results. This is partly due to the fact that in both cases the set of possible probability distributions is not much restricted. Yet, for statistical practice, the obtained inferences may be too cautious and there may be more information about the behavior of the analyzed random quantities that should be taken into account. Therefore, the impact of stronger distributional assumptions on the robust LIR method will be investigated in future research. In addition to this, the possibility of deriving other LIR methods based on more restrictive probability models will be addressed.

Finally, the LIR methodology permits considering the possibility of wrong coarsening, even if the variables are in fact precisely observed. As the only necessary specification of this possible error is an upper bound to the probability of an observation not containing the correct value, the LIR methodology could provide a framework for very general measurement error methods. This potential of the LIR methodology is yet to be investigated.

Notation

Miscellaneous

$\lvert\mathcal{S}\rvert$	cardinality of a set $\mathcal{S}$
$2^{\mathcal{S}}$	power set of a set $\mathcal{S}$
$\mathcal{S}_1 \subset \mathcal{S}_2$, $\mathcal{S}_2 \supset \mathcal{S}_1$	$\mathcal{S}_1$ is a proper subset of the set $\mathcal{S}_2$, while $\mathcal{S}_2$ is a proper superset of the set $\mathcal{S}_1$
$\mathcal{S}_1 \subseteq \mathcal{S}_2$, $\mathcal{S}_2 \supseteq \mathcal{S}_1$	$\mathcal{S}_1$ is a general subset of the set $\mathcal{S}_2$, while $\mathcal{S}_2$ is a general superset of the set $\mathcal{S}_1$
$\mathcal{S}_1 \backslash \mathcal{S}_2$	set difference, i.e., $\mathcal{S}_1$ excluding $\mathcal{S}_1 \cap \mathcal{S}_2$
$\mathbb{I}_{\mathcal{S}}$	indicator function of a set $\mathcal{S}$
$[\underline{c}, \overline{c}]$	closed and possibly unbounded interval, with $\underline{c} \leq \overline{c}$ and $\underline{c}, \overline{c} \in \mathbb{R} \cup \{-\infty, +\infty\}$
$(\underline{c}, \overline{c})$	open bounded interval, with $\underline{c} < \overline{c}$ and $\underline{c}, \overline{c} \in \mathbb{R}$
$[\underline{c}, \overline{c})$, $(\underline{c}, \overline{c}]$	bounded intervals whose lower and upper endpoint, respectively, belongs to the interval while the other does not, with $\underline{c} < \overline{c}$ and $\underline{c}, \overline{c} \in \mathbb{R}$
log	natural logarithm
∇	gradient of a function
I_u	u-dimensional identity matrix, with $u \in \mathbb{N}$
tr	trace of a square matrix
w^{T}, M^{T}	transpose of a vector w and of a matrix M
$\mathbb{E}$	expectation operator

d	number of explanatory variables in X, with $d \in \mathbb{N}$
n	number of observations in a sample, with $n \in \mathbb{N}$
p	proportion associated with the $p\,100\%$ quantile of a probability distribution, with $p \in (0,1)$
β	cutoff point of the (normalized profile) likelihood function, with $\beta \in (0,1)$
ε	upper bound to the probability $P(V \notin V^*)$ for the considered probability models, with $\varepsilon \in [0, 1/2)$

Spaces and sets

$\varnothing$	empty set
$\mathbb{N}$	set of positive integers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_{>0}$, $\mathbb{R}_{\geq 0}$	sets of positive and nonnegative real numbers, respectively
$\mathbb{R}_{\neq 0}$	set of real numbers excluding zero
$\mathbb{R}_{<0}$	set of negative real numbers
$\mathcal{X}$	observation space of the vector X of explanatory variables, with $\mathcal{X} \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$
$\mathcal{Y}$	observation space of the response variable Y, with $\mathcal{Y} \subseteq \mathbb{R}$
$\mathcal{V}$	observation space of the joint random vector $V = (X, Y)$
$\mathcal{V}^*$	observation space of the random set V^* describing the imprecise observation of V, with $\mathcal{V}^* \subseteq 2^{\mathcal{V}}$
$\mathcal{F}$	(considered) set of regression functions $f : \mathcal{X} \to \mathbb{R}$
$\mathcal{G}$	domain of some characteristic g of the considered probability measures, with $\mathcal{G} \subseteq \mathbb{R}$
$\mathcal{G}_{>\beta}$	likelihood-based confidence region for the characteristic g, with $\mathcal{G}_{>\beta} = \{\gamma \in \mathcal{G} : lik_g(\gamma) > \beta\}$, for some $\beta \in (0,1)$
$\mathcal{C}_f$, $\mathcal{C}_{f,>\beta}$	likelihood-based confidence region for L_f, for some $\beta \in (0,1)$ and $f \in \mathcal{F}$

Random variables and realizations

X	vector of d explanatory variables, with $X \in \mathcal{X} \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$
Y	response variable, with $Y \in \mathcal{Y} \subseteq \mathbb{R}$
V	vector of regression variables $V = (X, Y)$
V^*	random set describing the imprecise observation of V
R_f	(absolute) residual for some regression function $f \in \mathcal{F}$, with $R_f = \lvert Y - f(X) \rvert$
v_i	(unobserved) realization of V, with $i \in \{1, \ldots, n\}$
A_i	realization of V^*, with $i \in \{1, \ldots, n\}$
$r_{f,i}$	(unobserved) realization of R_f, with $i \in \{1, \ldots, n\}$ and for some $f \in \mathcal{F}$
$\underline{r}_{f,i},\ \overline{r}_{f,i}$	infimal and supremal residuals related to an imprecise observation $V_i^* = A_i$, with $i \in \{1, \ldots, n\}$ and for some $f \in \mathcal{F}$

Probabilities

P	probability distribution of the joint random object (V, V^*)
P_V	marginal probability distribution of V
P_{V^*}	marginal probability distribution of V^*
$\hat{P}_{V^*}$	empirical distribution of an observed sample $V_1^* = A_1$, $\ldots, V_n^* = A_n$; when only precise observations are considered, we denote by $\hat{P}_V$ the empirical distribution of $V_1 = v_1, \ldots, V_n = v_n$
$\mathcal{P}$	(considered) set of probability measures P on $\mathcal{V} \times \mathcal{V}^*$
$\mathcal{P}_\varepsilon$	set of all probability measures on $\mathcal{V} \times \mathcal{V}^*$ that satisfy $P(V \in V^*) \geq 1 - \varepsilon$, for some $\varepsilon \in [0, 1/2)$
$\mathcal{P}_V$	(considered) set of marginal probability measures P_V on $\mathcal{V}$ (corresponding to probability measures $P \in \mathcal{P}$)
$\mathcal{P}_{V^*}$	(considered) set of marginal probability measures P_{V^*} on $\mathcal{V}^*$ (corresponding to probability measures $P \in \mathcal{P}$)
$[P_{V^*}]$	set of all probability distributions P'_V of the precise data corresponding to probability measures $P' \in \mathcal{P}$ with marginal distribution $P'_{V^*} = P_{V^*}$ for the imprecise data

$\mathcal{P}_{>\beta}$	set of plausible probability measures after the observation of data, with $\mathcal{P}_{>\beta} = \{P \in \mathcal{P} : lik(P) > \beta\}$, for some $\beta \in (0,1)$; when only precise observations are considered, we have $\mathcal{P}_{V,>\beta} = \{P_V \in \mathcal{P}_V : lik_V(P_V) > \beta\}$

Functions

f	regression function $f : \mathcal{X} \to \mathbb{R}$
L	loss function considered in the regression problem, with $L : \mathcal{F} \times \mathcal{P} \to 2^{\mathbb{R}_{\geq 0}}$ (possibly multi-valued mapping); defined on $\mathcal{F} \times \mathcal{P}_V$ when only precise observations are considered
L_f	function-specific loss function for some $f \in \mathcal{F}$, with $L_f(P) = L(f,P)$ for all $P \in \mathcal{P}$; defined on $\mathcal{P}_V$ when only precise observations are considered
lik	(normalized) likelihood function, with $lik : \mathcal{P} \to [0,1]$; when only precise observations are considered, we have $lik_V : \mathcal{P}_V \to [0,1]$
g	characteristic of the probability distributions in $\mathcal{P}$, with $g : \mathcal{P} \to 2^{\mathcal{G}}$ (possibly multi-valued mapping); defined on $\mathcal{P}_V$ when only precise observations are considered
lik_g	(normalized) profile likelihood function for the characteristic g, with $lik_g : \mathcal{G} \to [0,1]$
g'	characteristic g only depending on the marginal distribution of the precise variables expressed as a function on $\mathcal{P}_V$, with $g'(P_V) = g(P)$ for all $P_V \in \mathcal{P}_V$
lik^*	(normalized) likelihood function on $\mathcal{P}_{V^*}$, defined by $lik^*(P_{V^*}) = lik(P)$ for all $P_{V^*} \in \mathcal{P}_{V^*}$
g^*	characteristic g only depending on the marginal distribution of the precise variables expressed as a function on $\mathcal{P}_{V^*}$, with $g^*(P_{V^*}) = \bigcup_{P_V \in [P_{V^*}]} g'(P_V)$ for all $P_{V^*} \in \mathcal{P}_{V^*}$
$lik^*_{g^*}$	(normalized) profile likelihood function associated with g^*

Technicalities of the robust LIR method

Q_f	p-quantile of the residual's distribution considered as loss, for some $p \in (0,1)$ and $f \in \mathcal{F}$
$\mathcal{Q}_f$	domain of the p-quantile of the residual, for some $p \in (0,1)$ and $f \in \mathcal{F}$, with $\mathcal{Q}_f \subseteq \mathbb{R}_{\geq 0}$
lik_{Q_f}	profile likelihood function for the p-quantile of the residual's distribution, with $lik_{Q_f} : \mathcal{Q}_f \to [0,1]$, for some $p \in (0,1)$ and $f \in \mathcal{F}$
$\overline{B}_{f,q}$	closed band of vertical bandwidth $2q$ around a function f, for some $q \in \mathbb{R}_{\geq 0}$ and $f \in \mathcal{F}$
$\underline{B}_{f,q}$	open band of vertical bandwidth $2q$ around a function f, for some $q \in \mathbb{R}_{\geq 0}$ and $f \in \mathcal{F}$
$\overline{k}_f$	number of imprecise data intersecting $\overline{B}_{f,q}$, for some $q \in \mathbb{R}_{\geq 0}$ and $f \in \mathcal{F}$
$\underline{k}_f$	number of imprecise data completely included in $\underline{B}_{f,q}$, for some $q \in \mathbb{R}_{\geq 0}$ and $f \in \mathcal{F}$
h	function introduced to express lik_{Q_f} in a simpler way, with $h : [0,1] \times (0,1) \to (0,1]$
$\underline{i}$, $\overline{i}$	integers introduced to express the points of discontinuity of lik_{Q_f}, for some $f \in \mathcal{F}$
$\underline{k}$, $\overline{k}$	integers introduced to express $\mathcal{C}_f$, for some $f \in \mathcal{F}$
$\overline{q}_{LRM}$	smallest upper endpoint of the confidence regions $\mathcal{C}_f$ over all $f \in \mathcal{F}$
f_{LRM}	regression function providing $\overline{q}_{LRM}$, if it is unique
$\mathcal{U}$	set of all undominated regression functions, with $\mathcal{U} \subseteq \mathcal{F}$
$\mathcal{T}$	set of all LQS functions for the $\overline{k}/n$-quantile based on precise data sets that are compatible with the imprecise data
$\mathcal{U}'$	set of parameters associated with the undominated functions in the case of simple linear regression, with $\mathcal{U}' \subseteq \mathbb{R}^2$

Technicalities of the exact algorithm for LIR

$\mathcal{D}$	set of indices of the bounded data, with $\mathcal{D} \subseteq \{1, \ldots, n\}$
$\mathcal{B}$	set of candidate slopes for the LRM functions, with $\mathcal{B} \subseteq \mathbb{R}$
$\underline{z}_{b,i},\ \overline{z}_{b,i}$	interval endpoints of the slope-adjusted imprecise data $[\underline{z}_{b,i},\ \overline{z}_{b,i}] \subseteq \mathbb{R}$, for some $i \in \{1, \ldots, n\}$
$\overline{z}_{b,[j]}$	the $\overline{k}$-th smallest value among those $\overline{z}_{b,i}$ for which the corresponding $\underline{z}_{b,i} \geq \underline{z}_{b,(j)}$, for some $j \in \{1, \ldots, n-\overline{k}+1\}$ and $b \in \mathbb{R}$
$\mathcal{I}$	set of indices introduced to define $\mathcal{A}_b$, with $\mathcal{I} \subseteq \{1, \ldots, n\}$
$\mathcal{A}_b$	set of intercept values of undominated functions with slope b, for some $b \in \mathbb{R}$
$\tilde{\mathcal{B}}$	set of all $b \in \mathbb{R}$ at which two functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \underline{z}_{b,j} - \overline{q}_{LRM}$ or two functions $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,j} + \overline{q}_{LRM}$ intersect, for some $(i,j) \in \{1, \ldots, n\}^2$ with $i \neq j$
$\check{\mathcal{B}}$	set of all $b \in \mathbb{R}$ at which the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,j} + \overline{q}_{LRM}$ intersect, for some $(i,j) \in \{1, \ldots, n\}^2$
$\mathcal{B}_{\mathcal{U}'}$	set of all relevant slopes for the precise description of $\mathcal{U}'$
η	small number between zero and $\min\{\lvert b\rvert : b \in \tilde{\mathcal{B}} \cup \check{\mathcal{B}}$ and $b \neq 0\}$ used in the definition of $\mathcal{B}_{\mathcal{U}'}$, with $\eta > 0$
ω	arbitrary positive number used in the definition of $\mathcal{B}_{\mathcal{U}'}$, with $\omega > 0$

Technicalities of the investigation of the statistical properties

θ	coefficients vector in standard linear regression, with $\theta \in \mathbb{R}^{d+1}$
G_f	graph of a function $f \in \mathcal{F}$, with $G_f = \{(x,y) \in \mathcal{X} \times \mathbb{R} : y = f(x)\}$
$\tilde{\mathcal{U}}$	set of all functions $f \in \mathcal{F}$ whose graphs intersect $\mathcal{V}$, with $\tilde{\mathcal{U}} = \{f \in \mathcal{F} : G_f \cap \mathcal{V} \neq \varnothing\}$

Technicalities of SVR

ψ	function introduced to express different loss functions of the regression problem, with $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$
E_{P_V}	risk functional considered as loss function, assigning to each $f \in \mathcal{F}$ the expectation $\mathbb{E}(\psi(R_f))$ under P_V, with $E_{P_V} : \mathcal{F} \to \mathbb{R}_{\geq 0}$, for some $P_V \in \mathcal{P}_V$ and ψ
κ	kernel function, with $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	scalar product in an RKHS $\mathcal{H}$, with $\langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x)$, for all $f \in \mathcal{H}$, where κ is the reproducing kernel of $\mathcal{H}$
$\|\cdot\|_{\mathcal{H}}$	norm in an RKHS $\mathcal{H}$ induced by its scalar product, with $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$
λ	regularization parameter, with $\lambda > 0$
$E_{P_V,\lambda}$	regularized risk functional considered as loss function in the estimation problem, with $E_{P_V,\lambda} : \mathcal{F} \to \mathbb{R}_{\geq 0}$, $f \mapsto E_{P_V}(f) + \lambda \|f\|_{\mathcal{F}}^2$, for some $P_V \in \mathcal{P}_V$ and $\lambda > 0$
$f_{\hat{P}_V,\lambda}$	SVM minimizing $E_{\hat{P}_V,\lambda}$ given a sample of precise observations $V_1 = v_1, \ldots, V_n = v_n$, for some $\lambda > 0$
α	vector of weights in the linear combination of kernel functions constituting $f_{\hat{P}_V,\lambda}$, with $\alpha \in \mathbb{R}^n$
$\mathcal{F}_n$	subset of the considered RKHS of functions, containing all linear combinations of kernel functions associated with the observed $X_1 = x_1, \ldots, X_n = x_n$, with $\mathcal{F}_n \subset \mathcal{F}$
D	design matrix of the standard linear regression model for an observed sample $V_1 = v_1, \ldots, V_n = v_n$, with i-th row given by $(1, x_{i,1}, \ldots, x_{i,d})$, for all $i \in \{1, \ldots, n\}$
τ^2	error variance in the standard linear regression model, with $\tau^2 \in \mathbb{R}_{\geq 0}$
$\hat{\theta}_{LS}$	LS estimator for the coefficients vector in standard linear regression, with $\hat{\theta}_{LS} \in \mathbb{R}^{d+1}$
$\hat{\theta}_{R,\lambda}$	Ridge estimator for the coefficients vector in standard linear regression, with $\hat{\theta}_{R,\lambda} \in \mathbb{R}^{d+1}$

Technicalities of generalized SVR

$\underline{E}_{P_{V^*}}$, $\overline{E}_{P_{V^*}}$	lower and upper risk functional, assigning to each $f \in \mathcal{F}$ the minimal and maximal risk, respectively, over all $P'_V \in [P_{V^*}]$, for some $P_{V^*} \in \mathcal{P}_{V^*}$
$\underline{R}_f$, $\overline{R}_f$	random quantities describing the lower and upper endpoint, respectively, of the interval-valued residual associated with V^*, for some $f \in \mathcal{F}$
$\underline{E}_{\hat{P}_{V^*},\lambda}$, $\overline{E}_{\hat{P}_{V^*},\lambda}$	lower and upper regularized risk functional associated with the empirical distribution $\hat{P}_{V^*}$ of the imprecise data, for some $\lambda > 0$

Technicalities of the LIR method for SVR

E_f	risk functional considered as function-specific loss function, for each $f \in \mathcal{F}$ defined by $E_f(P) = E_{P_V}(f)$, for all $P \in \mathcal{P}$
E'_f	E_f expressed as a function on $\mathcal{P}_V$, with $E'_f(P_V) = E_f(P)$, for all $P_V \in \mathcal{P}_V$ and some $f \in \mathcal{F}$
E^*_f	imprecise version of E_f on $\mathcal{P}_{V^*}$, for each $f \in \mathcal{F}$ defined by $E^*_f(P_{V^*}) = \bigcup_{P_V \in [P_{V^*}]} E'_f(P_V)$, for all $P_{V^*} \in \mathcal{P}_{V^*}$
lik_{E_f}	(normalized) profile likelihood function for E_f, for some $f \in \mathcal{F}$
$\mathcal{E}_{f,>\beta}$	likelihood-based confidence region for E_f, for some $f \in \mathcal{F}$ and $\beta \in (0,1)$
$\mathcal{E}_{f,\lambda,>\beta}$	likelihood-based confidence region for the regularized risk, for some $f \in \mathcal{F}$, $\lambda > 0$, and $\beta \in (0,1)$

Bibliography

Alexandrov, A. (2005). *Convex Polyhedra*. Springer.

Bache, K. and M. Lichman (2013). UCI Machine Learning Repository.

Beaton, A., D. Rubin, and J. Barone (1976). The acceptability of regression solutions: Another look at computational accuracy. *Journal of the American Statistical Association 71*, 158–168.

Beresteanu, A., I. Molchanov, and F. Molinari (2012). Partial identification using random set theory. *Journal of Econometrics 166*, 17–32.

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

Blanco-Fernández, A., N. Corral, and G. González-Rodríguez (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis 55*, 2568–2578.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.). Chapman & Hall/CRC.

Casella, G. and R. Berger (2002). *Statistical Inference* (2nd ed.). Duxbury.

Cattaneo, M. (2007). *Statistical Decisions Based Directly on the Likelihood Function*. Ph. D. thesis, ETH Zurich.

Cattaneo, M. (2013). Likelihood decision functions. *Electronic Journal of Statistics 7*, 2924–2946.

Cattaneo, M. and A. Wiencierz (2011). Robust regression with imprecise data. Technical Report 114, Department of Statistics, LMU Munich.

Cattaneo, M. and A. Wiencierz (2012). Likelihood-based Imprecise Regression. *International Journal of Approximate Reasoning 53*, 1137–1154.

Cattaneo, M. and A. Wiencierz (2013). On the implementation of LIR: the case of simple linear regression with interval data. Accepted for publication in *Computational Statistics*.

Cheney, E. (1982). *Introduction to Approximation Theory* (2nd ed.). Chelsea Publishing.

Christmann, A. and R. Hable (2012). Consistency of support vector machines using additive kernels for additive models. *Computational Statistics & Data Analysis 56*, 854–873.

Christmann, A., A. Van Messem, and I. Steinwart (2009). On consistency and robustness properties of Support Vector Machines for heavy-tailed distributions. *Statistics and Its Interface 2*, 311–327.

Clark, A., F. Etilé, F. Postel-Vinay, C. Senik, and K. Van der Straeten (2005). Heterogeneity in Reported Well-Being: Evidence from Twelve European Countries. *The Economic Journal 115*, C118–C132.

Clark, A., P. Frijters, and M. Shields (2008). Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles. *Journal of Economic Literature 46*, 95–144.

Coppi, R., P. D'Urso, P. Giordani, and A. Santoro (2006). Least squares estimation of a linear regression model with LR fuzzy response. *Computational Statistics & Data Analysis 51*, 267–286.

Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems 47*, 547–553.

Deaton, A. (2008). Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives 22*, 53–72.

Deaton, A. (2012). The financial crisis and the well-being of Americans. *Oxford Economic Papers 64*, 1–26.

Dempster, A. (1968). Upper and Lower Probabilities Generated by a Random Closed Interval. *The Annals of Mathematical Statistics 39*, 957–966.

Dempster, A. and D. Rubin (1983). Rounding Error in Regression: The Appropriateness of Sheppard's Corrections. *Journal of the Royal Statistical Society (Series B) 45*, pp. 51–59.

Destercke, S., D. Dubois, and E. Chojnacki (2008). Unifying practical uncertainty representations: I. Generalized p-boxes. *International Journal of Approximate Reasoning 49*, 649–663.

Diamond, P. (1990). Least squares fitting of compact set-valued data. *Journal of Mathematical Analysis and Applications 147*, 351–362.

Diener, E. and R. Biswas-Diener (2002). Will money increase subjective well-being? *Social Indicators Research 57*, 119–169.

Domingues, M., R. de Souza, and F. Cysneiros (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters 31*, 1991–1996.

Draper, N. and C. van Nostrand (1979). Ridge Regression and James-Stein Estimation: Review and Comments. *Technometrics 21*, 451–466.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer.

Ferraro, M., R. Coppi, G. González-Rodríguez, and A. Colubi (2010). A linear regression model for imprecise response. *International Journal of Approximate Reasoning 51*, 759–770.

Ferson, S., V. Kreinovich, L. Ginzburg, D. Myers, and K. Sentz (2003). Constructing Probability Boxes and Dempster-Shafer Structures. Technical Report SAND2002-4015, Sandia National Laboratories.

Ferson, S., V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg (2007). Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Technical Report SAND2007-0939, Sandia National Laboratories.

Gioia, F. and C. Lauro (2005). Basic statistical methods for interval data. *Italian Journal of Applied Statistics 17*, 75–104.

Gómez, G., A. Espinal, and S. Lagakos (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine 22*, 409–425.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis 106*, 92–117.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Heitjan, D. and D. Rubin (1990). Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping. *Journal of the American Statistical Association 85*, 304–314.

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

Hoerl, A. and R. Kennard (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics 12*, 55–67.

Hofmann, T., B. Schölkopf, and A. Smola (2008). Kernel Methods in Machine Learning. *The Annals of Statistics 36*, 1171–1220.

Horowitz, J. and C. Manski (1995). Identification and Robustness with Contaminated and Corrupted Data. *Econometrica 63*, 281–302.

Huber, P. (1981). *Robust Statistics*. Wiley.

Huppert, F., N. Marks, A. Clark, J. Siegrist, A. Stutzer, J. Vittersø, and M. Wahrendorf (2009). Measuring Well-being Across Europe: Description of the ESS Well-being Module and Preliminary Findings. *Social Indicators Research 91*, 301–315.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software 11*, 1–20.

Körner, R. and W. Näther (1998). Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. *Information Sciences 109*, 95–118.

Lima Neto, E. and F. de Carvalho (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis 52*, 1500–1515.

Lindsey, J. (1998). A study of interval censoring in parametric regression models. *Lifetime Data Analysis 4*, 329–354.

Little, R. and D. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

Manski, C. (2003). *Partial Identification of Probability Distributions*. Springer.

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

Marino, M. and F. Palumbo (2002). Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. *Italian Journal of Applied Statistics 14*, 277–291.

Maronna, R., D. Martin, and V. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley.

Müller, K., S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks 12*, 181–201.

Nguyen, H. and B. Wu (2006). Random and fuzzy sets in coarse data analysis. *Computational Statistics & Data Analysis 51*, 70–85.

Norwegian Social Science Data Services (2010). *ESS Round 5: European Social Survey Round 5 Data*. Data file edition 3.0.

Owen, A. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika 75*, 237–249.

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

Pötter, U. (2008). Statistical Models of Incomplete Data and their Use in the Social Sciences. Habilitation thesis.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Used R version 2.15.2.

Reinhardt, R., A. Hoffmann, and T. Gerlach (2013). *Nichtlineare Optimierung: Theorie, Numerik und Experimente*. Springer.

Rousseeuw, P. and A. Leroy (1987). *Robust Regression and Outlier Detection*. Wiley.

Rubin, D. (1976). Inference and Missing Data. *Biometrika 63*, 581–592.

Salibian-Barrera, M. and V. Yohai (2008). High breakdown point robust regression with censored data. *The Annals of Statistics 36*, 118–146.

Schervish, M. (1995). *Theory of Statistics*. Springer.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Schollmeyer, G. and T. Augustin (2013). On Sharp Identification Regions for Regression Under Interval Data. In F. Cozman, T. Denoeux, S. Destercke, and T. Seidfenfeld (Eds.), *ISIPTA '13, Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*, pp. 285–294. SIPTA.

Sheppard, W. (1898). On the Calculation of the most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Division of a Scale. *Proceedings of the London Mathematical Society 29*, 353–380.

Smets, P. (2005). Belief functions on real numbers. *International Journal of Approximate Reasoning 40*, 181–223.

Steele, J. and W. Steiger (1986). Algorithms and complexity for least median of squares regression. *Discrete Applied Mathematics 14*, 93–100.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. Springer.

Stigler, S. (2010). The Changing History of Robustness. *The American Statistician 64*, 277–281.

Stromberg, A. (1993). Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal on Scientific Computing 14*, 1289–1299.

Suykens, J., T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific.

Svejdar, V., H. Küchenhoff, L. Fahrmeir, and J. Wassermann (2011). External forcing of earthquake swarms at Alpine regions: example from a seismic meteorological network at Mt. Hochstaufen SE-Bavaria. *Nonlinear Processes in Geophysics 18*, 849–860.

Turlach, B. and A. Weingessel (2013). *quadprog: Functions to solve Quadratic Programming Problems*. R package version 1.5-5.

Utkin, L. and F. Coolen (2011). Interval-valued Regression and Classification Models in the Framework of Machine Learning. In F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger (Eds.), *ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pp. 371–380. SIPTA.

Utkin, L. and S. Destercke (2009). Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning 50*, 778–798.

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

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Watson, G. (1998). On computing the least quantile of squares estimate. *SIAM Journal on Scientific Computing 19*, 1125–1138.

Wiencierz, A. (2013). *linLIR: linear Likelihood-based Imprecise Regression*. R package version 1.1-1.

Wiencierz, A. and M. Cattaneo (2012). An Exact Algorithm for Likelihood-Based Imprecise Regression in the Case of Simple Linear Regression with Interval Data. In R. Kruse, M. Berthold, C. Moewes, M. Gil, P. Grzegorzewski, and O. Hryniewicz (Eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, Volume 190 of *Advances in Intelligent Systems and Computing*, pp. 293–301. Springer.

Wilks, S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics 9*, 60–62.

Wunder, C., A. Wiencierz, J. Schwarze, and H. Küchenhoff (2013). Well-Being over the Life Span: Semiparametric Evidence from British and

German Longitudinal Data. *Review of Economics and Statistics 95*, 154–167.

Zhang, Z. (2009). Likelihood-based confidence sets for partially identified parameters. *Journal of Statistical Planning and Inference 139*, 696–710.

Zhang, Z. (2010). Profile Likelihood and Incomplete Data. *International Statistical Review 78*, 102–116.

Eidesstattliche Versicherung

gemäß § 8 Abs. 2 Punkt 5 der Promotionsordnung

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 30. Oktober 2013

(Andrea Wiencierz)