International Journal of Approximate Reasoning 53 (2012) 1137–1154

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Likelihood-based Imprecise Regression

Marco E.G.V. Cattaneo *, Andrea Wiencierz

Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany

ARTICLE INFO

Article history:
Available online 28 June 2012

Keywords:
Imprecise data
Likelihood inference
Imprecise probability
Complex uncertainty
Robust regression
Quantile estimation

ABSTRACT

We introduce a new approach to regression with imprecisely observed data, combining likelihood inference with ideas from imprecise probability theory, and thereby taking different kinds of uncertainty into account. The approach is very general: it provides a uniform theoretical framework for regression analysis with imprecise data, where all kinds of relationships between the variables of interest may be considered and all types of imprecisely observed data are allowed.

Furthermore, we propose a regression method based on this approach, where no parametric distributional assumption is needed and likelihood-based interval estimates of quantiles of the residuals distribution are used to identify a set of plausible descriptions of the relationship of interest. Thus, the proposed regression method is very robust and yields a set-valued result, whose extent is determined by the amounts of both kinds of uncertainty involved in the regression problem with imprecise data: statistical uncertainty and indetermination.

In addition, we apply our robust regression method to an interesting question in the social sciences by analyzing data from a social survey. As result we obtain a large set of plausible relationships, reflecting the high uncertainty inherent in the analyzed data set.

1. Introduction

Data are often available only with limited precision. That is, they contain only the information that the values of interest lie in certain subsets of the observation space. For example, technical measuring instruments usually provide a precise value and an assessment of the measurement uncertainty, which translates the measurement into an interval of possible values. However, up to now there is no standard methodology for analyzing the relationships between imprecisely observed variables. Only few general approaches have been proposed so far, which fall mainly in two categories. One of them consists of approaches suggesting to apply standard regression methods to all possible precise data compatible with the observations, and to consider the range of outcomes as the imprecise result [17,2,26,18]. The approaches in the second category consist in representing the imprecise observations by few precise values (for example, representing intervals by center and width), and in applying standard regression methods to those values [15,14,5,24,16,6].

In the present paper, we propose a new, completely different approach, in which the regression problem with imprecise data is not reduced to few or many regression problems with precise data. Instead, we introduce a general methodology for likelihood inference with imprecisely observed data, on the basis of which we develop a new regression methodology directly applicable to the imprecise data. We call our approach Likelihood-based Imprecise Regression (LIR).

LIR combines likelihood inference with ideas from imprecise probability theory, allowing to take into account different kinds of uncertainty involved in the regression problem with imprecise data. On the one hand there is statistical uncertainty,

* Corresponding author.
E-mail addresses: cattaneo@stat.uni-muenchen.de (M. Cattaneo), andrea.wiencierz@stat.uni-muenchen.de (A. Wiencierz).

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2012.06.010

as one usually has only a finite sample of observations, and on the other hand there is indetermination, due to the fact that the data are only imprecisely observed. The resulting complex uncertainty can be described by an imprecise probability model, consisting of all probability measures that were sufficiently good in predicting the observed imprecise data (that is, all probability measures whose likelihood exceeds a given threshold). This imprecise probability model then provides set-valued estimates for any characteristic of the distribution of the (unobserved) precise data in which one is actually interested. Such characteristics can be for example a quantile of the distribution or the probability of a particular event. The extent of the estimated sets depends on both types of uncertainty involved. This methodology is part of the introduced theoretical framework for likelihood inference with imprecisely observed data and it is thoroughly presented in Section 2.

The general framework is then applied to the case of regression as a problem of statistical inference. Here, the imprecise observations may be arbitrary subsets of the Cartesian product of the observation spaces of the dependent and of the (possibly multiple) independent variables. The characteristics of interest are characteristics of the distribution of the (unobserved) precise residuals for some possible regression function describing the relationship between the (unobserved) precise variables of interest. Thanks to the general inference methodology that we introduce in Section 2, we obtain a set-valued estimate for the chosen characteristic of the residuals distribution (e.g., a certain quantile) for each considered regression function. The regression problem then reduces to a decision problem where the possible actions are the considered regression functions and the (imprecise) loss is given by the set-valued estimates. We suggest to consider as the regression's result all regression functions that are not strictly dominated by another one. In this way, we obtain an imprecise result, consisting of the set of all regression functions that cannot be excluded on the basis of the likelihood inference. The mathematical details of the regression methodology are set out in Section 3. In the present paper, we focus on the setting without parametric distributional assumptions and where quantiles of the residuals distribution are used to evaluate the possible descriptions of the relationship of interest. We derive the LIR method for this case, which turns out to be a very robust regression method due to the absence of sensitive distributional assumptions and to the use of quantiles. Furthermore, it is important to mention that the theoretical framework of LIR allows for any kind of relationship between the variables of interest, and for all types of imprecisely observed data, including missing data and actually precise data as special cases. Hence, our approach is neither restricted to linear regression nor to interval-censored data.

The alternative general approaches falling in the two categories mentioned above are more restrictive in their assumptions, and their results often reflect only the uncertainty related to the imprecision of the data. In contrast to this, the major aim of our LIR methodology is to describe the whole uncertainty about which of the considered regression functions best describes the relationship of interest in the light of the (possibly) imprecisely observed data. This is achieved by considering all plausible regression functions as the set-valued result of a LIR analysis, which can be seen as a confidence region for the true relationship and thus reflects the complex uncertainty of the regression problem. A more thorough and illustrative comparison of the new LIR method with the alternative regression methods is described in [10].

In this paper, which is an extended and refined version of [9], we set out the mathematical details of the theoretical framework of our new approach. Moreover, we suggest a first implementation of the proposed LIR method illustrated with an application example. Of course, the computational issues related to the new methodology have to be examined in further detail, but this goes beyond the scope of the present paper. Some of these aspects in the special case of simple linear regression with interval data are studied in [11], where we also suggest an improved implementation of the robust LIR method.

The present paper is organized as follows. First, we introduce the general methodology for likelihood inference with imprecise data in Section 2. Then, in Section 3, we develop in detail the theoretical framework of the LIR analysis for the case without distributional assumptions. In addition to the theoretical results, in Section 4 we apply the method to analyze an interesting question in the social sciences. We investigate the relationship between age and income on the basis of survey data. The source of data used in this paper is "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) — German General Social Survey" of 2008. The data is provided by GESIS — Leibniz Institute for the Social Sciences.

2. Imprecise data

Before considering the specific problem of regression with imprecisely observed variables, in the present section we derive some general results about likelihood inference with imprecise data. Let $V_1, \ldots, V_n$ be n random objects taking values in a set $\mathcal{V}$, and let $V_1^*, \ldots, V_n^*$ be n random sets taking values in a set $\mathcal{V}^* \subseteq 2^{\mathcal{V}}$, such that the events $V_i \in V_i^*$ are measurable. We are actually interested in the data V_i, but we can only observe the imprecise data V_i^*. The connection between precise and imprecise data is established by the following assumptions about the probability measures considered as models of the situation.

For each $\varepsilon \in [0, 1]$, let $\mathcal{P}_\varepsilon$ be the set of all probability measures P such that the n random objects $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ are independent and identically distributed and satisfy

$$P(V_i \in V_i^*) \geq 1 - \varepsilon \tag{1}$$

(where, as usual, probability measures and random objects are defined on an underlying measurable space). We assume that the precise and imprecise data can be modeled by a probability measure P included in a particular set $\mathcal{P} \subseteq \mathcal{P}_\varepsilon$, for some $\varepsilon \in [0, 1]$. Each $P \in \mathcal{P}$ can be identified with a particular joint distribution for V_i and V_i^* (that is, the precise and imprecise data, respectively) satisfying condition (1). In particular, $\mathcal{P} = \mathcal{P}_\varepsilon$ corresponds to the fully nonparametric assumption that

any joint distribution for V_i and V_i^* satisfying condition (1) is a possible model of the situation (this is the assumption we consider in Sections 3 and 4). The usual choice for the value of ε is 0 (see for example [12,33]), which corresponds to an assumption of correctness of the imprecise data: $V_i^* = A$ implies $V_i \in A$ (a.s.). However, this assumption is often too strong: some imprecise data can be incorrect, in the sense that $V_i^* = A$, but $V_i \notin A$. This is for example the case, when the imprecise data represent the classification of the precise data into categories, and some observations are misclassified. By choosing a positive value for ε, we allow each imprecise observation to be incorrect with probability at most ε.

The set $\mathcal{V}^*$ describes which imprecise data $V_i^* = A$ are considered as possible. As extreme cases of imprecise data we have the actually precise data (when A is a singleton) and the missing data (when $A = \mathcal{V}$). In general, the fully nonparametric assumption $\mathcal{P} = \mathcal{P}_\varepsilon$ does not exclude informative coarsening (see for example [38]): parametric models or uninformative coarsening can be imposed by a stronger assumption $\mathcal{P} \subset \mathcal{P}_\varepsilon$. However, it is important to note that the set $\mathcal{P}_\varepsilon$ depends strongly on the choice of $\mathcal{V}^*$. For example, when $\varepsilon = 0$, the choice of a set $\mathcal{V}^*$ such that its elements build a partition of $\mathcal{V}$ implies the assumption that the coarsening is deterministic and uninformative, because each possible precise data value is contained in exactly one possible imprecise observation $A \in \mathcal{V}^*$.

Example 1. Let $\mathcal{V} = \{0, 1\}$ and $\mathcal{V}^* = 2^{\mathcal{V}}$, and assume $\mathcal{P} = \mathcal{P}_\varepsilon$ for some $\varepsilon \in [0, 1]$. In this case, each (unobserved) variable V_i assumes either the value 0 or 1, but we observe only the imprecise data $V_i^* = A$, with $A \subseteq \{0, 1\}$. When $A = \{0\}$ or $A = \{1\}$, the observation is actually precise (but possibly incorrect): if it is correct, then we have $V_i = 0$ or $V_i = 1$, respectively. When $A = \mathcal{V}$, the data V_i is in fact missing: we did not learn anything about it by observing $V_i^* = \{0, 1\}$. Finally, when $A = \varnothing$, the imprecise observation does not tell us anything about V_i, because it is certainly incorrect; therefore, condition (1) implies $P(V_i^* = \varnothing) \leq \varepsilon$.

To exemplify the subtle difference between the cases $A = \mathcal{V}$ and $A = \varnothing$, consider that V_i describes the correct, unobserved answer of the individual "i" to a particular survey question, which can be answered with either "yes" or "no" (encoded by $V_i = 1$ and $V_i = 0$, respectively). When the actual, possibly incorrect answer of the individual "i" to the survey question is "yes" or "no", it can be described by the imprecise observation $V_i^* = \{1\}$ or $V_i^* = \{0\}$, respectively. When the individual "i" does not answer the question, the missing answer can be described by the imprecise observation $V_i^* = \mathcal{V}$, while when the actual answer is for instance "blue" (that is, neither "yes" nor "no"), it can be described by the imprecise observation $V_i^* = \varnothing$. In both latter cases, we did not learn anything about the correct answer V_i: the only difference is that the imprecise observation $V_i^* = \varnothing$ describes an actual answer that is certainly incorrect, and according to assumption (1), the probability of an incorrect observation is bounded above by ε.

2.1. Complex uncertainty

In general, we are uncertain about which of the probability measures in $\mathcal{P}$ is the best model of the reality under consideration. Our uncertainty is composed of two parts. On the one hand, we are uncertain about the distribution of the imprecise data V_i^*: this uncertainty decreases when we observe more and more (imprecise) data; we call it statistical uncertainty. On the other hand, even if we (asymptotically) knew the distribution of the imprecise data V_i^*, we would still be uncertain about the distribution of the (unobserved) precise data V_i: this uncertainty is unavoidable; we call it indetermination. To formulate this mathematically, let P_V and P_{V^*} be the marginal distributions of V_i and V_i^*, respectively, corresponding to the probability measure $P \in \mathcal{P}$. There is statistical uncertainty about P_{V^*} in the set $\mathcal{P}_{V^*} := \{P'_{V^*} : P' \in \mathcal{P}\}$, but even if P_{V^*} were known, there would still be the (unavoidable) indetermination of P_V in the set

$$[P_{V^*}] := \{P'_V : P' \in \mathcal{P},\ P'_{V^*} = P_{V^*}\}.$$

The sets $[P_{V^*}]$ with $P_{V^*} \in \mathcal{P}_{V^*}$ are the identification regions for P_V in the terminology of [25]. Each of them consists of all the distributions for the precise data V_i compatible with a particular distribution for the imprecise data V_i^*. Hence, each set $[P_{V^*}]$ can be interpreted as an imprecise probability distribution on $\mathcal{V}$. By observing the realizations of the imprecise data V_i^*, we learn something about which of the imprecise probability distributions $[P_{V^*}]$ is the best model for the (unobserved) precise data V_i.

Example 2. In the situation of Example 1, the only condition on the marginal distribution of the imprecise data V_i^* is $P(V_i^* = \varnothing) \leq \varepsilon$. Hence, $\mathcal{P}_{V^*}$ is the set of all probability distributions on $2^{\{0,1\}}$ such that the probability of $\varnothing$ is at most ε. The only condition on the joint distribution of V_i and V_i^* is given by assumption (1), which is equivalent to $P(V_i \notin V_i^*) \leq \varepsilon$, and thus can be written as

$$P\left(V_i = 0,\ V_i^* \in \{\varnothing, \{1\}\}\right) + P\left(V_i = 1,\ V_i^* \in \{\varnothing, \{0\}\}\right) \leq \varepsilon.$$

Therefore, for each $P_{V^*} \in \mathcal{P}_{V^*}$, the identification region $[P_{V^*}]$ is the set of all probability distributions on $\{0, 1\}$ such that the probability of 1 lies in the interval $[\underline{P}_{V^*}\{1\},\ \overline{P}_{V^*}\{1\}]$, with

$$\overline{P}_{V^*}\{1\} = P_{V^*}\{\{1\}, \mathcal{V}\} + \min\left(P_{V^*}\{\varnothing, \{0\}\},\ \varepsilon\right) = \min\left(P_{V^*}\{\{1\}, \mathcal{V}\} + \varepsilon,\ 1\right),$$
$$\underline{P}_{V^*}\{1\} = 1 - \left(P_{V^*}\{\{0\}, \mathcal{V}\} + \min\left(P_{V^*}\{\varnothing, \{1\}\},\ \varepsilon\right)\right) = \max\left(P_{V^*}\{\varnothing, \{1\}\} - \varepsilon,\ 0\right).$$

In particular, when $\varepsilon = 0$, the imprecise probability distribution $[P_{V^*}]$ corresponds to the belief function on $\mathcal{V}$ with basic probability assignment P_{V^*} (see for example [31]), in the sense that $[P_{V^*}]$ is the set of all probability distributions on $\{0, 1\}$ dominating that belief function.

2.2. Likelihood

The likelihood function is a central concept in statistical inference (see for example [29]). For parametric probability models, it is usually expressed as a function of the parameters: here we consider the more general formulation (as a function of the probability measures), which is applicable also to nonparametric models. The observed (imprecise) data $V_1^* = A_1, \ldots, V_n^* = A_n$ induce the (normalized) likelihood function $lik : \mathcal{P} \to [0, 1]$ defined by

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

for all $P \in \mathcal{P}$. The likelihood function describes the relative ability of the probability measures P in predicting the observed (imprecise) data. Therefore, the value $lik(P)$ depends only on the marginal distribution P_{V^*} of the imprecise data V_i^*. The likelihood function can be interpreted as the second level of a hierarchical model for imprecise probabilities, with $\mathcal{P}$ as first level (see for example [7,8]). In particular, for any $\beta \in (0, 1)$, the likelihood function can be used to reduce $\mathcal{P}$ to the set

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

of all the probability measures that were sufficiently good in predicting the observed (imprecise) data.

Let g be a multivalued mapping from $\mathcal{P}$ to a set $\mathcal{G}$, describing a particular characteristic (in which we are interested) of the models considered (mathematically, $g : \mathcal{P} \to 2^{\mathcal{G}} \setminus \{\varnothing\}$, but g is interpreted as an "imprecise" mapping from $\mathcal{P}$ to $\mathcal{G}$). For example, g can be the multivalued mapping from $\mathcal{P}$ to $\mathbb{R}$ assigning to each probability measure P the p-quantile of the distribution of $h(V_i)$ under P, for some $p \in (0, 1)$ and some measurable function $h : \mathcal{V} \to \mathbb{R}$; that is, $g(P) = \{q \in \mathbb{R} : P(h(V_i) < q) \leq p \leq P(h(V_i) \leq q)\}$ for all $P \in \mathcal{P}$. This is the kind of mapping g we consider in Sections 3 and 4: it is multivalued, because in general quantiles are not uniquely defined (a p-quantile of the distribution of $h(V_i)$ is any value $q \in \mathbb{R}$ such that $P(h(V_i) < q) \leq p \leq P(h(V_i) \leq q)$). For each $\beta \in (0, 1)$, the set

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

is called likelihood-based confidence region with cutoff point β for the values of the multivalued mapping g. This confidence region consists of all values that the characteristic described by g takes on the set $\mathcal{P}_{>\beta}$ of all the probability measures that were sufficiently good in predicting the observed (imprecise) data. The unique function $lik_g : \mathcal{G} \to [0, 1]$ describing these confidence regions, in the sense that

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

for all $\beta \in (0, 1)$, is called (normalized) profile likelihood function induced by the multivalued mapping g.

Lemma 1. *For all $\gamma \in \mathcal{G}$,*

$$lik_g(\gamma) = \sup_{P \in \mathcal{P} : \gamma \in g(P)} lik(P)$$

(where the supremum is 0 when no P satisfies the condition).

Proof. For all $\beta \in (0, 1)$,

$$lik_g(\gamma) > \beta \Leftrightarrow \gamma \in \mathcal{G}_{>\beta} \Leftrightarrow \exists P \in \mathcal{P}_{>\beta} : \gamma \in g(P) \Leftrightarrow \exists P \in \mathcal{P} : \gamma \in g(P) \wedge lik(P) > \beta \Leftrightarrow \sup_{P \in \mathcal{P} : \gamma \in g(P)} lik(P) > \beta,$$

from which the result follows, since both sides of the equation take values in $[0, 1]$. □

Example 3. In the situation of Examples 1 and 2, let $\varepsilon = 0$, and assume that the imprecise data $\{0\}$, $\{1\}$, and $\mathcal{V}$ have been observed n_0, n_1, and n_{01} times, respectively, where n_0, n_1, and n_{01} are positive integers. In this case the likelihood function $lik : \mathcal{P} \to [0, 1]$ satisfies, for all $P \in \mathcal{P}$,

$$lik(P) = \frac{P_{V^*}(\{\{0\}\})^{n_0} \; P_{V^*}(\{\{1\}\})^{n_1} \; P_{V^*}(\{\mathcal{V}\})^{n_{01}}}{\sup_{P' \in \mathcal{P}} P'_{V^*}(\{\{0\}\})^{n_0} \; P'_{V^*}(\{\{1\}\})^{n_1} \; P'_{V^*}(\{\mathcal{V}\})^{n_{01}}}.$$

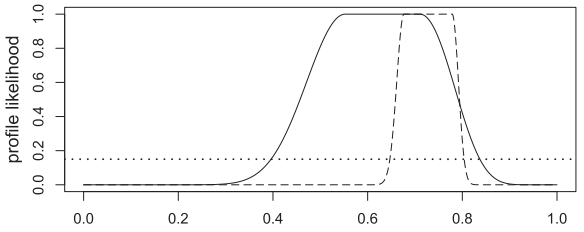

Fig. 1. Profile likelihood functions from Examples 3 and 4.

Consider now the mapping g from $\mathcal{P}$ to $[0, 1]$ assigning to each probability measure P the probability $P_V\{1\}$ that a precise data value V_i is 1 (before observing the corresponding imprecise data value V_i^*; as a multivalued mapping, g is defined by $g(P) = \{P_V\{1\}\}$ for all $P \in \mathcal{P}$). The induced profile likelihood function lik_g on $[0, 1]$ is plotted in Fig. 1 for the cases with $(n_0, n_1, n_{01}) = (11, 21, 6)$ and $(n_0, n_1, n_{01}) = (213, 651, 98)$: solid and dashed lines, respectively (a detailed calculation of lik_g is given in Example 4).

In these two cases, the likelihood-based confidence regions with cutoff point $\beta = 0.15$ for the probability $P_V\{1\}$ are approximately the intervals $[0.39, 0.84]$ and $[0.65, 0.80]$, respectively (the cutoff point $\beta = 0.15$ is represented by the dotted line in Fig. 1). They are (conservative) confidence intervals of approximate level 95% (see for example [23]).

2.3. Likelihood for imprecise data models

In the situation we consider, we are actually interested in the (unobserved) precise data V_i. In this case, the characteristic of interest (described by g) depends only on the marginal distribution P_V of the precise data V_i; that is, we can write $g(P) =: g'(P_V)$ for all $P \in \mathcal{P}$. For example, the p-quantile of the distribution of $h(V_i)$ depends only on the distribution of V_i. By contrast, as noted at the beginning of Subsection 2.2, the value $lik(P)$ depends only on the marginal distribution P_{V^*} of the imprecise data V_i^*. By writing $lik(P) = lik^*(P_{V^*})$ for all $P \in \mathcal{P}$, we define a function $lik^* : \mathcal{P}_{V^*} \to [0, 1]$, which can be interpreted as the likelihood function on $\mathcal{P}_{V^*}$.

In order to obtain the profile likelihood function lik_g, it can be useful to consider the multivalued mapping g^* from $\mathcal{P}_{V^*}$ to $\mathcal{G}$ defined by

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

for all $P_{V^*} \in \mathcal{P}_{V^*}$. The multivalued mapping g^* assigns to each P_{V^*} all the values that the characteristic described by g' takes on the set $[P_{V^*}]$ of all distributions for the precise data V_i compatible with the distribution P_{V^*} for the imprecise data V_i^*. That is, g^* can be interpreted as an imprecise version of g', assigning to each imprecise probability distribution $[P_{V^*}]$ the corresponding imprecise value of g'.

We can now define the function $lik^*_{g^*} : \mathcal{G} \to [0, 1]$ in analogy with the expression for the profile likelihood function lik_g given in Lemma 1:

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

for all $\gamma \in \mathcal{G}$ (where the supremum is 0 when no P_{V^*} satisfies the condition). The function $lik^*_{g^*}$ can be interpreted as the profile likelihood function induced by the multivalued mapping g^*, when lik^* is considered as the likelihood function on $\mathcal{P}_{V^*}$. This profile likelihood function is particularly interesting in connection with the discussion of Subsection 2.1, because lik^* describes the statistical uncertainty about the distribution P_{V^*} of the imprecise data V_i^*, which decreases when we observe more and more (imprecise) data, while g^* describes the (unavoidable) indetermination of the values of g (in the terminology of [25], the values of g^* are the identification regions for the values of g). Thanks to the following result, the profile likelihood function $lik^*_{g^*}$ is not only interesting from a conceptual point of view, but also useful in order to calculate the likelihood-based confidence regions for the values of g.

Lemma 2.

$$lik_g = lik^*_{g^*}$$

Proof. From Lemma 1 and the above definitions it follows that for all $\gamma \in \mathcal{G}$,

$$lik_g(\gamma) = \sup_{P \in \mathcal{P} : \gamma \in g'(P_V)} lik^*(P_{V^*}) = \sup_{P_{V^*} \in \mathcal{P}_{V^*} : \exists P' \in \mathcal{P} : P'_{V^*} = P_{V^*} \wedge \gamma \in g'(P'_V)} lik^*(P_{V^*})$$
$$= \sup_{P_{V^*} \in \mathcal{P}_{V^*} : \exists P_V \in [P_{V^*}] : \gamma \in g'(P_V)} lik^*(P_{V^*}) = lik^*_{g^*}(\gamma). \quad \square$$

Example 4. The imprecise version g^* of the mapping g of Example 3 is the multivalued mapping from $\mathcal{P}_{V^*}$ to $[0, 1]$ assigning to each P_{V^*} the set $\{P_V\{1\} : P_V \in [P_{V^*}]\}$. In Example 2 we have seen that, since now $\varepsilon = 0$, this set is the interval

$$[\underline{P}_{V^*}\{1\},\ \overline{P}_{V^*}\{1\}] = [P_{V^*}\{\{1\}\},\ 1 - P_{V^*}\{\{0\}\}].$$

That is, $g^*(P_{V^*})$ is the interval probability that a precise data value V_i is 1 (before observing the corresponding imprecise data value V_i^*) according to the imprecise probability distribution $[P_{V^*}]$ (i.e., the belief function on $\mathcal{V}$ with basic probability assignment P_{V^*}).

As seen in Example 2, the only condition on the marginal distributions $P_{V^*} \in \mathcal{P}_{V^*}$ is $P_{V^*}\{\varnothing\} = 0$ (since now $\varepsilon = 0$). That is, $\mathcal{P}_{V^*}$ corresponds to the set of all probability distributions on the set $\{\{0\}, \{1\}, \mathcal{V}\}$, and can thus be parametrized by the 2-dimensional simplex

$$\mathcal{S}_2 = \left\{p = (p_0, p_1, p_{01}) \in [0, 1]^3 : p_0 + p_1 + p_{01} = 1\right\}.$$

Therefore, $lik^* : \mathcal{P}_{V^*} \to [0, 1]$ corresponds to a (normalized) multinomial likelihood function, and we obtain

$$lik^*_{g^*}(\gamma) = \max_{p \in \mathcal{S}_2 : \gamma \in [p_1,\ 1-p_0]} \frac{p_0^{n_0}\, p_1^{n_1}\, p_{01}^{n_{01}}}{\hat{p}_0^{n_0}\, \hat{p}_1^{n_1}\, \hat{p}_{01}^{n_{01}}}$$

for all $\gamma \in [0, 1]$, where

$$\hat{p} = \left(\frac{n_0}{n_0 + n_1 + n_{01}},\ \frac{n_1}{n_0 + n_1 + n_{01}},\ \frac{n_{01}}{n_0 + n_1 + n_{01}}\right)$$

is the maximum likelihood estimate of the parameter $p \in \mathcal{S}_2$. Hence, in particular, $lik^*_{g^*}(\gamma) = 1$ for all $\gamma \in [\hat{p}_1,\ 1 - \hat{p}_0]$. Moreover, it can be easily proved that if $\gamma < \hat{p}_1$, then

$$p = \left(\hat{p}_0\, \frac{1 - \gamma}{1 - \hat{p}_1},\ \gamma,\ \hat{p}_{01}\, \frac{1 - \gamma}{1 - \hat{p}_1}\right)$$

maximizes $p_0^{n_0}\, p_1^{n_1}\, p_{01}^{n_{01}}$ among all $p \in \mathcal{S}_2$ such that $p_1 \leq \gamma$. Symmetrically, if $1 - \gamma < \hat{p}_0$, then

$$p = \left(1 - \gamma,\ \hat{p}_1\, \frac{\gamma}{1 - \hat{p}_0},\ \hat{p}_{01}\, \frac{\gamma}{1 - \hat{p}_0}\right)$$

maximizes $p_0^{n_0}\, p_1^{n_1}\, p_{01}^{n_{01}}$ among all $p \in \mathcal{S}_2$ such that $p_0 \leq 1 - \gamma$. Altogether, thanks to Lemma 2, we obtain the following expression for the profile likelihood function induced by the multivalued mapping g (see also [39]):

$$lik_g(\gamma) = lik^*_{g^*}(\gamma) = \begin{cases} \left(\dfrac{\gamma}{\hat{p}_1}\right)^{n_1} \left(\dfrac{1 - \gamma}{1 - \hat{p}_1}\right)^{n_0 + n_{01}} & \text{if } \ 0 \leq \gamma < \hat{p}_1, \\ 1 & \text{if } \ \hat{p}_1 \leq \gamma \leq 1 - \hat{p}_0, \\ \left(\dfrac{1 - \gamma}{\hat{p}_0}\right)^{n_0} \left(\dfrac{\gamma}{1 - \hat{p}_0}\right)^{n_1 + n_{01}} & \text{if } \ 1 - \hat{p}_0 < \gamma \leq 1. \end{cases}$$

The profile likelihood function $lik_g = lik^*_{g^*}$ on $[0, 1]$ is plotted in Fig. 1 for the two cases considered in Example 3. In the case with 38 data (solid line) there is (statistical) uncertainty also about the distribution P_{V^*} of the imprecise data V_i^*, while in the case with 962 data (dashed line) almost only the (unavoidable) indetermination described by g^* remains, in the sense that $lik^*_{g^*}$ is almost equal to the indicator function of an identification region for $P_V\{1\}$ (more precisely, the indicator function of the probability interval $g^*(\hat{P}_{V^*}) = [\hat{p}_1,\ 1 - \hat{p}_0]$ corresponding to the maximum likelihood estimate of $P_{V^*} \in \mathcal{P}_{V^*}$).

3. Regression

We now apply the results of Section 2 to the problem of regression with imprecisely observed variables. Hence, we assume that the (unobservable) precise data are pairs $V_i = (X_i, Y_i)$, where $X_1, \ldots, X_n$ are n random objects taking values in a set $\mathcal{X}$, and $Y_1, \ldots, Y_n$ are n random variables, with $\mathcal{V} = \mathcal{X} \times \mathbb{R}$. For some $\mathcal{V}^* \subseteq 2^{\mathcal{X}\times\mathbb{R}}$ and some $\varepsilon \in [0, 1]$, we consider the fully nonparametric assumption $\mathcal{P} = \mathcal{P}_\varepsilon$. This means that we do not assume anything about the joint distribution of X_i and Y_i, while the only condition on the joint distribution of the (unobserved) precise data V_i and their imprecise observations V_i^* is given by assumption (1). In the remainder of the paper, we focus on this setting.

We want to describe the relation between X_i and Y_i by means of a function $f \in \mathcal{F}$, where $\mathcal{F}$ is a particular set of (measurable) functions $f : \mathcal{X} \to \mathbb{R}$. In order to assess the quality of the description by means of f, we define the (absolute) residuals

$$R_{f,i} := |Y_i - f(X_i)| .$$

The n random variables $R_{f,1}, \ldots, R_{f,n} \in [0, +\infty)$ are independent and identically distributed: the more their distribution is concentrated near 0, the better is the description by means of f.

In order to compare the quality of the descriptions by means of different functions $f \in \mathcal{F}$, we need to compare the concentration near 0 of the distributions of the corresponding residuals $R_{f,i}$. Usual choices of measures for this concentration are the second and first moments $E(R_{f,i}^2)$ and $E(R_{f,i})$, respectively. However, the moments of the distribution of the residuals cannot be really estimated in the fully nonparametric setting we consider, because moments are too sensitive to small variations in the distribution (see also Subsection 4.2). In fact, if $\varepsilon > 0$ or the set

$$\mathcal{R}_f := \{|y - f(x)| : (x, y) \in A,\ A \in \mathcal{V}^*\}$$

(i.e., the set of all possible values of $R_{f,i}$ when $V_i \in V_i^*$) is unbounded, then the likelihood-based confidence region for any particular moment of the distribution of the residuals is unbounded (even when only the distributions with finite moments are considered), independently of the cutoff point and of the observed (imprecise) data.

By contrast, the quantiles of the distribution of the residuals can in general be estimated even in the fully nonparametric setting we consider. Therefore, we propose to use the p-quantile of the distribution of the residuals $R_{f,i}$ as a measure of the concentration near 0 of this distribution, for some $p \in (0, 1)$. The technical details of the estimation of such quantiles are given in Subsections 3.1 and 3.2.

The minimizations of the second and first moments of the distribution of the residuals can be interpreted as the theoretical counterparts of the methods of least squares and least absolute deviations, respectively. In the same sense, the minimization of the p-quantile of the distribution of the residuals can be interpreted as the theoretical counterpart of the method of least quantile of squares (or absolute deviations), introduced in [30] as a generalization of the method of least median of squares (corresponding to the choice $p = 0.5$). The method of least quantile of squares leads to robust regression estimators, with breakdown point $\min\{p, 1 - p\}$ (that is, the highest possible breakdown point 50% is reached when $p = 0.5$). By contrast, the methods of least squares and least absolute deviations lead to regression estimators with breakdown point 0, since they cannot even handle a single outlier (including leverage points; see for example [19,27,22]).

In the location problem (that is, when $\mathcal{F}$ is the set of all constant functions $f : \mathcal{X} \to \mathbb{R}$), the values of the constant functions f minimizing the second and first moments of the distribution of the residuals $R_{f,i}$ are the mean and median of the distribution of Y_i, respectively (when these exist and are unique). The value of the constant function f minimizing the p-quantile of the distribution of the residuals $R_{f,i}$ is the p-center of the distribution of Y_i (that is, the center of the shortest interval containing Y_i with probability at least p), when this exists and is unique. The p-center can be interpreted as a generalization of the mode of a distribution, since under some regularity conditions the mode corresponds to the limit of the p-center when p tends to 0. The p-center of a symmetric, strictly unimodal distribution corresponds to its median and mean (when this exists), independently of p. Therefore, the minimizations of the p-quantile, first moment, and second moment of the distribution of the residuals lead to the same (correct) regression function, under the usual assumptions for the error distribution: see for example [34].

Example 5. We consider a problem of simple linear regression: that is, $\mathcal{X} = \mathbb{R}$ (and thus $\mathcal{V} = \mathbb{R}^2$), and $\mathcal{F} = \{f_{a,b} : a, b \in \mathbb{R}\}$ is the set of all linear functions $f_{a,b}$ defined by $f_{a,b}(x) = a + b\,x$ for all $x \in \mathbb{R}$. The left plot of Fig. 2 shows $n = 99$ precise data points $V_1, \ldots, V_n$. However, we assume that the pairs of precise values $V_i = (X_i, Y_i) \in \mathbb{R}^2$ are not known. Instead, for given partitions of the real line in intervals, we only know in which intervals X_i^* and Y_i^* lie X_i and Y_i, respectively. That is, we assume that only the imprecise data $V_1^*, \ldots, V_n^*$ are observed, where $V_i^* = X_i^* \times Y_i^*$, and the elements of $\mathcal{V}^*$ (i.e., the possible imprecise observations V_i^*) build a partition of $\mathcal{V} = \mathbb{R}^2$ (hence, in particular, $\mathcal{R}_f = [0, +\infty)$ for all $f \in \mathcal{F}$). The relevant part of this partition is represented in the left plot of Fig. 2 (gray lines), while the right plot is the corresponding two-dimensional histogram of the data set (where a darker shade of gray indicates a higher frequency of the imprecise observation).

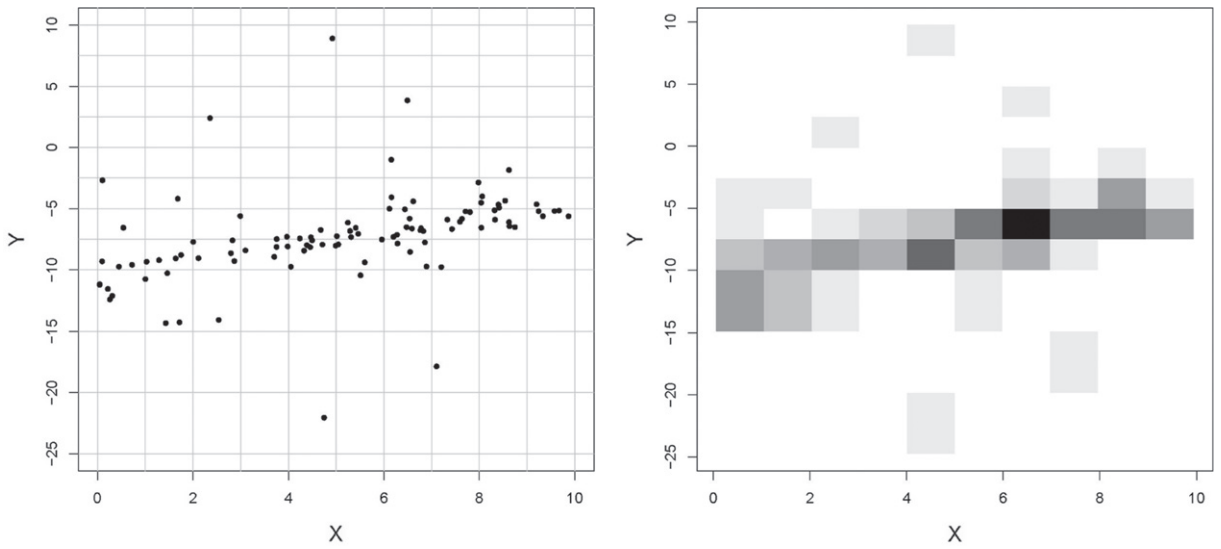

Fig. 2. Data set from Examples 5–10: (unobserved) precise data $V_i = (X_i, Y_i) \in \mathbb{R}^2$ and partition of $\mathbb{R}^2$ (left), and corresponding two-dimensional histogram representing the observed (imprecise) data $V_i^* = X_i^* \times Y_i^* \subset \mathbb{R}^2$.

3.1. Determination of profile likelihood functions for quantiles of residuals

We want to determine the likelihood-based confidence regions for the quantiles of the distribution of the residuals: for this purpose, we calculate the profile likelihood function for such quantiles. Let $p \in (0, 1)$, and for each function $f \in \mathcal{F}$, let $\mathcal{Q}_f$ be the interval defined by $\mathcal{Q}_f = \mathcal{L}_f \cap \mathcal{U}_f$, with

$$\mathcal{L}_f = \bigcup_{r \in \mathcal{R}_f} [r, +\infty)$$

when $p > \varepsilon$ and $\mathcal{L}_f = [0, +\infty)$ otherwise, while

$$\mathcal{U}_f = \bigcup_{r \in \mathcal{R}_f} [0, r]$$

when $p < 1 - \varepsilon$ and $\mathcal{U}_f = [0, +\infty)$ otherwise. The definition of $\mathcal{Q}_f$ can be interpreted as follows: if $\varepsilon < p < 1 - \varepsilon$, then $\mathcal{Q}_f$ is the smallest interval containing $\mathcal{R}_f$, while if $p \leq \varepsilon$, then this interval is extended to the left until 0 (included), and if $p \geq 1 - \varepsilon$, then it is extended to the right until $+\infty$ (not included). Therefore, $\mathcal{Q}_f$ is the set of all possible values for the p-quantile of the distribution of the residuals $R_{f,i}$, since $P(R_{f,i} \notin \mathcal{R}_f) \leq \varepsilon$ follows from assumption (1).

For each $f \in \mathcal{F}$, let Q_f be the multivalued mapping from $\mathcal{P}$ to $\mathcal{Q}_f$ assigning to each probability measure P the p-quantile of the distribution of the residuals $R_{f,i}$ under P. As noted in Subsection 2.2, the mapping Q_f is multivalued, because in general quantiles are not uniquely defined. We want to determine the profile likelihood function $lik_{Q_f} : \mathcal{Q}_f \to [0, 1]$ induced by the multivalued mapping Q_f. It is important to note that we would obtain the same results by considering only the distributions for which the p-quantile is unique (that is, the vagueness in the definition of quantiles has no influence on the resulting likelihood-based confidence regions).

Assume that the (imprecise) data $V_1^* = A_1, \ldots, V_n^* = A_n$ are observed, where $A_1, \ldots, A_n \in \mathcal{V}^* \setminus \{\varnothing\}$. In order to obtain the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of the residuals $R_{f,i}$, we define for each function $f \in \mathcal{F}$ and each distance $q \in [0, +\infty)$ the bands

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

and the functions $\overline{k}_f, \underline{k}_f$ on $[0, +\infty)$ such that

$$\overline{k}_f(q) = \#\{i \in \{1, \ldots, n\} : A_i \cap \overline{B}_{f,q} \neq \varnothing\},$$
$$\underline{k}_f(q) = \#\{i \in \{1, \ldots, n\} : A_i \subseteq \underline{B}_{f,q}\}$$

for all $q \in [0, +\infty)$ (where $\#A$ denotes the cardinality of a set A). That is, $\overline{k}_f(q)$ is the number of imprecise data intersecting $\overline{B}_{f,q}$, while $\underline{k}_f(q)$ is the number of imprecise data completely contained in $\underline{B}_{f,q}$. Therefore, in particular, $\overline{k}_f$ and

$\underline{k}_f$ are monotonically increasing functions of q, and $\underline{k}_f(q) \leq \overline{k}_f(q)$ for all $q \in [0, +\infty)$. Finally, we define the function $\lambda : [0, 1] \times (0, 1) \to (0, 1]$ as follows, for all $s \in [0, 1]$ and all $t \in (0, 1)$:

$$\lambda(s,t) = \begin{cases} 1-t & \text{if } \ s = 0, \\ \left(\frac{t}{s}\right)^s \left(\frac{1-t}{1-s}\right)^{1-s} & \text{if } \ 0 < s < 1, \\ t & \text{if } \ s = 1. \end{cases}$$

Example 6. Consider the problem of simple linear regression introduced in Example 5. For each $f \in \mathcal{F}$, we have $\mathcal{R}_f = [0, +\infty)$, and therefore $\mathcal{Q}_f = [0, +\infty)$ as well, independently of the values of $p \in (0, 1)$ and $\varepsilon = [0, 1]$. The region between the two dashed lines in the left plot of Fig. 4 corresponds to the closed band $\overline{B}_{f,q}$ (when the points on the dashed lines are included) or to the open band $\underline{B}_{f,q}$ (when the points on the dashed lines are excluded), where $f \in \mathcal{F}$ is the linear function represented by the solid line, and $q \approx 4.47$ is the half of the vertical width of the bands. In this case, $\underline{k}_f(q) = 82$ imprecise data are completely contained in $\underline{B}_{f,q}$, and $\overline{k}_f(q) = 92$ imprecise data intersect $\overline{B}_{f,q}$.

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 $q \in \mathcal{Q}_f$:*

$$lik_{Q_f}(q) = \begin{cases} \lambda\left(\frac{\overline{k}_f(q)}{n},\ p-\varepsilon\right)^n & \text{if } \ \overline{k}_f(q) < (p-\varepsilon)\, n, \\ 1 & \text{if } \ \left[\underline{k}_f(q),\ \overline{k}_f(q)\right] \cap \left[(p-\varepsilon)\, n,\ (p+\varepsilon)\, n\right] \neq \varnothing, \\ \lambda\left(\frac{\underline{k}_f(q)}{n},\ p+\varepsilon\right)^n & \text{if } \ \underline{k}_f(q) > (p+\varepsilon)\, n. \end{cases} \tag{2}$$

Proof. In order to prove expression (2), we use Lemma 2, which tells us that $lik_{Q_f}(q) = lik^*_{Q_f^*}(q)$ for all $q \in \mathcal{Q}_f$, where lik^* and Q_f^* are defined on the set $\mathcal{P}_{V^*}$ of all possible distributions P_{V^*} for the imprecise data V_i^*. The function lik^* assigns to each P_{V^*} the corresponding likelihood value: in particular, it has a unique maximum in the empirical distribution (of the imprecise data) $\hat{P}_{V^*}$. The multivalued mapping Q_f^* assigns to each P_{V^*} all p-quantiles of the residuals $R_{f,i}$ for all distributions of the precise data V_i compatible with P_{V^*}.

We first consider the empirical distribution (of the imprecise data) $\hat{P}_{V^*}$: we know that $lik^*(\hat{P}_{V^*}) = 1$, and we want to determine $Q_f^*(\hat{P}_{V^*})$. Each joint distribution of V_i and V_i^* with marginal distribution $\hat{P}_{V^*}$ can be described by the conditional distributions of V_i given $V_i^* = A_j$ (for each imprecise observation A_j), since $\hat{P}_{V^*}\{A_1, \ldots, A_n\} = 1$. In particular, for each $q \in \mathcal{Q}_f$ we can construct joint distributions of V_i and V_i^* as follows: for each one of the $\overline{k}_f(q) - \underline{k}_f(q)$ imprecise observations A_j such that $(\overline{B}_{f,q} \setminus \underline{B}_{f,q}) \cap A_j \neq \varnothing$, we can choose the conditional distribution of V_i given $V_i^* = A_j$ to be concentrated on $(\overline{B}_{f,q} \setminus \underline{B}_{f,q}) \cap A_j$, while for all other imprecise observations A_j, we can choose the conditional distributions of V_i given $V_i^* = A_j$ in such a way that as much probability as possible is given to $\overline{B}_{f,q} \setminus \underline{B}_{f,q}$, according to assumption (1). The resulting probability distributions satisfy

$$\frac{\underline{k}_f(q)}{n} \leq P(R_{f,i} < q) = P_V(\underline{B}_{f,q}) \geq \max\left(\frac{\underline{k}_f(q)}{n} - \varepsilon,\ 0\right) = \min_{P'_V \in [\hat{P}_{V^*}]} P'_V(\underline{B}_{f,q}),$$

$$\frac{\overline{k}_f(q)}{n} \leq P(R_{f,i} \leq q) = P_V(\overline{B}_{f,q}) \leq \min\left(\frac{\overline{k}_f(q)}{n} + \varepsilon,\ 1\right) = \max_{P'_V \in [\hat{P}_{V^*}]} P'_V(\overline{B}_{f,q}),$$

and therefore,

$$q \in Q_f^*(\hat{P}_{V^*}) \Leftrightarrow \frac{\underline{k}_f(q)}{n} - \varepsilon \leq p \leq \frac{\overline{k}_f(q)}{n} + \varepsilon \Leftrightarrow \left[\underline{k}_f(q),\ \overline{k}_f(q)\right] \cap \left[(p-\varepsilon)\, n,\ (p+\varepsilon)\, n\right] \neq \varnothing.$$

This proves the second case of expression (2).

We now prove the first case of expression (2), and thus assume that $q \in \mathcal{Q}_f$ satisfies $\overline{k}_f(q) < (p-\varepsilon)\, n$. If q is a p-quantile according to $P \in \mathcal{P}$, then $P(V_i \in \overline{B}_{f,q}) = P(R_{f,i} \leq q) \geq p$, and assumption (1) implies $P(V_i \in V_i^* \cap \overline{B}_{f,q}) \geq p - \varepsilon$. This is more than what the empirical distribution $\hat{P}_{V^*}$ assigns to the $\overline{k}_f(q)$ imprecise data intersecting $\overline{B}_{f,q}$, and it can be easily proved that all marginal distributions $P_{V^*} \in \mathcal{P}_{V^*}$ maximizing lik^* among the ones satisfying $q \in Q_f^*(P_{V^*})$ can be expressed as

$$P_{V^*} = (p-\varepsilon)\, P'_{V^*} + (1-p+\varepsilon)\, P''_{V^*}, \tag{3}$$

where $P''_{V^*} \in \mathcal{P}_{V^*}$ is the empirical distribution obtained when only the $n - \overline{k}_f(q)$ imprecise data not intersecting $\overline{B}_{f,q}$ are considered, and if $\overline{k}_f(q) > 0$, then $P'_{V^*} \in \mathcal{P}_{V^*}$ is the empirical distribution obtained when only the $\overline{k}_f(q)$ imprecise data intersecting $\overline{B}_{f,q}$ are considered. In this case, P_{V^*} is unique, while if $\overline{k}_f(q) = 0$, then $P'_{V^*} \in \mathcal{P}_{V^*}$ can be any distribution assigning the whole probability to elements of $\mathcal{V}^*$ intersecting $\overline{B}_{f,q}$. Such elements of $\mathcal{V}^*$ exist, because $p > \varepsilon$ (since $\overline{k}_f(q) < (p-\varepsilon)\, n$), and therefore the definition of $\mathcal{Q}_f$ implies that there is an $r \in \mathcal{R}_f$ such that $r \leq q$. If $\overline{k}_f(q) > 0$, then for the unique marginal distribution P_{V^*} of the form (3) we have

$$lik^*(P_{V^*}) = \frac{\prod_{i=1}^n P_{V^*}\{A_i\}}{\prod_{i=1}^n \hat{P}_{V^*}\{A_i\}} = \frac{\left(\frac{p-\varepsilon}{\overline{k}_f(q)}\right)^{\overline{k}_f(q)} \left(\frac{1-p+\varepsilon}{n-\overline{k}_f(q)}\right)^{n-\overline{k}_f(q)}}{\left(\frac{1}{n}\right)^n} = \left(\frac{p-\varepsilon}{\frac{\overline{k}_f(q)}{n}}\right)^{\overline{k}_f(q)} \left(\frac{1-(p-\varepsilon)}{1-\frac{\overline{k}_f(q)}{n}}\right)^{n-\overline{k}_f(q)}$$
$$= \lambda\left(\frac{\overline{k}_f(q)}{n},\ p-\varepsilon\right)^n,$$

while if $\overline{k}_f(q) = 0$, then for all marginal distributions P_{V^*} of the form (3) we have

$$lik^*(P_{V^*}) = \frac{\prod_{i=1}^n P_{V^*}\{A_i\}}{\prod_{i=1}^n \hat{P}_{V^*}\{A_i\}} = \frac{\left(\frac{1-p+\varepsilon}{n}\right)^n}{\left(\frac{1}{n}\right)^n} = (1-(p+\varepsilon))^n = \lambda\,(0,\ p-\varepsilon)^n.$$

These two expressions for $lik^*(P_{V^*})$ are valid also when some of the imprecise observations $A_1, \ldots, A_n$ are equal, because in this case additional factors appear in the numerator as well as in the denominator of the fractions expressing the likelihood ratio of P_{V^*} and $\hat{P}_{V^*}$ (see for instance also [28, Section 2.3]). This proves the first case of expression (2), the third one can be proved analogously. □

The expression for lik_{Q_f} given in Theorem 1 is very general, but rather involved. To obtain a simpler result about lik_{Q_f}, we define, for each function $f \in \mathcal{F}$ and each imprecise data A_i, the infimum $\underline{r}_{f,i}$ and the supremum $\overline{r}_{f,i}$ of the set of all possible values of the residual $R_{f,i}$ when $V_i \in A_i$ (i.e., when the imprecise observation $V_i^* = A_i$ is correct):

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

As usual in statistics, $\underline{r}_{f,(i)}$ and $\overline{r}_{f,(i)}$ denote then the ith smallest infimum and supremum, respectively, so that with $\underline{r}_{f,(0)} := \overline{r}_{f,(0)} := \inf \mathcal{Q}_f$ and $\underline{r}_{f,(n+1)} := \overline{r}_{f,(n+1)} := \sup \mathcal{Q}_f$ 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 $\underline{i} := \max\left(\lceil (p-\varepsilon)\, n \rceil,\ 0\right)$ and $\overline{i} := \min\left(\lfloor (p+\varepsilon)\, n \rfloor,\ n\right) + 1$, so that $\underline{i} \in \{0, \ldots, n\}$ and $\overline{i} \in \{1, \ldots, n+1\}$, with $\underline{i} \leq \overline{i}$.

Lemma 3. *The points of discontinuity of the restriction of $\overline{k}_f$ to $\mathcal{Q}_f$, including the endpoints of $\mathcal{Q}_f$, are (in ascending order, with possible repetitions) $\underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(n+1)}$, and for all other values of $q \in \mathcal{Q}_f$ we have $\overline{k}_f(q) = i$ if $\underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n\}$.*

The points of discontinuity of the restriction of $\underline{k}_f$ to $\mathcal{Q}_f$, including the endpoints of $\mathcal{Q}_f$, are (in ascending order, with possible repetitions) $\overline{r}_{f,(0)}, \ldots, \overline{r}_{f,(n+1)}$, and for all other values of $q \in \mathcal{Q}_f$ we have $\underline{k}_f(q) = i$ if $\overline{r}_{f,(i)} < q < \overline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n\}$.

Proof. The points of discontinuity of the restrictions of $\overline{k}_f$, $\underline{k}_f$ to $\mathcal{Q}_f$, possibly including the endpoints of $\mathcal{Q}_f$, are (for all imprecise data A_i)

$$\inf\{q \in \mathcal{Q}_f : A_i \cap \overline{B}_{f,q} \neq \varnothing\} = \inf\{q \in \mathcal{Q}_f : \exists (x,y) \in A_i : |y - f(x)| \leq q\} = \inf\{|y - f(x)| : (x,y) \in A_i\} = \underline{r}_{f,i},$$
$$\sup\{q \in \mathcal{Q}_f : A_i \not\subseteq \underline{B}_{f,q}\} = \sup\{q \in \mathcal{Q}_f : \exists (x,y) \in A_i : |y - f(x)| \geq q\} = \sup\{|y - f(x)| : (x,y) \in A_i\} = \overline{r}_{f,i},$$

respectively, because $(x,y) \in A_i$ implies $|y - f(x)| \in \mathcal{R}_f \subseteq \mathcal{Q}_f$. Hence, if $\underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n\}$, then there are exactly i imprecise data intersecting $\overline{B}_{f,q}$ (i.e., $\overline{k}_f(q) = i$). Analogously, if $\underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)}$ with $i \in \{0, \ldots, n\}$, then there are exactly i imprecise data completely contained in $\underline{B}_{f,q}$ (i.e., $\underline{k}_f(q) = i$). □

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)

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

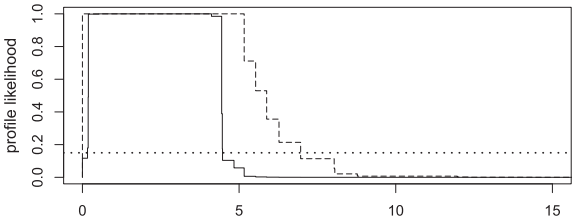

Fig. 3. Profile likelihood functions from Examples 7 and 8.

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

$$lik_{Q_f}(q) = \begin{cases} \lambda\left(\frac{i}{n},\ p-\varepsilon\right)^n & \text{if } \ \underline{r}_{f,(i)} < q < \underline{r}_{f,(i+1)} \ \text{ with } \ i \in \{0,\dots,\underline{i}-1\} \ \text{ (when } \underline{i} \geq 1\text{)}, \\ 1 & \text{if } \ \underline{r}_{f,(\underline{i})} < q < \overline{r}_{f,(\bar{i})}, \\ \lambda\left(\frac{i}{n},\ p+\varepsilon\right)^n & \text{if } \ \overline{r}_{f,(i)} < q < \overline{r}_{f,(i+1)} \ \text{ and } \ i \in \{\bar{i},\dots,n\} \ \text{ (when } \bar{i} \leq n\text{)}. \end{cases} \tag{4}$$

Proof. The function lik_{Q_f} can take at most $n+2$ different values, because in the first case of expression (2) the possible values of $\overline{k}_f(q)$ are the integers k such that $0 \leq k < (p-\varepsilon)\,n$, while in the third case the possible values of $\underline{k}_f(q)$ are the integers k such that $(p+\varepsilon)\,n < k \leq n$ (hence, in these two cases taken together, lik_{Q_f} can take at most $n+1$ different values).

If $\underline{i} \geq 1$, then $0 \leq \underline{i}-1 < (p-\varepsilon)\,n$, and if $\bar{i} \leq n$, then $n \geq \bar{i} > (p+\varepsilon)\,n$. Hence, expression (4) is well-defined, and in order to prove the second part of the corollary, it suffices to show that it holds for all $q \in \mathcal{Q}_f$. This is easily done, since expression (4) is a direct consequence of Theorem 1 and Lemma 3. In the first case of expression (4), Lemma 3 implies $\overline{k}_f(q) = i < (p-\varepsilon)\,n$, in the third case it implies $\underline{k}_f(q) = i > (p-\varepsilon)\,n$, while in the second case it implies $\overline{k}_f(q) \geq \underline{i} \geq (p-\varepsilon)\,n$ and $\underline{k}_f(q) \leq \bar{i}-1 \leq (p+\varepsilon)\,n$. □

Example 7. In the problem of simple linear regression introduced in Example 5, let $f \in \mathcal{F}$ be the linear function plotted in Fig. 4 (left, solid line). In this situation, the sets of all possible values of the residuals $R_{f,i}$ (when the imprecise observations $V_i^* = X_i^* \times Y_i^*$ are correct) are intervals, and their endpoints $\underline{r}_{f,i}, \overline{r}_{f,i}$ can be easily calculated. They can then be used in expression (4), which determines the values of the profile likelihood function lik_{Q_f} for the p-quantile of the distribution of $R_{f,i}$ (apart in its points of discontinuity, including the endpoint 0 of $\mathcal{Q}_f = [0,+\infty)$). The function lik_{Q_f} with $p = 0.75$ is plotted in Fig. 3 for the cases with $\varepsilon = 0$ (solid line) and $\varepsilon = 0.1$ (dashed line).

3.2. Determination of confidence intervals for quantiles of residuals

Thanks to Theorem 1, we can now calculate, for each cutoff point $\beta \in (0,1)$, the likelihood-based confidence regions for the quantiles of the distribution of the residuals $R_{f,i}$. We obtain in particular the following result.

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

$$(\max\{p, 1-p\} + \varepsilon)^n \leq \beta \tag{5}$$

holds, then

$$\underline{k} := \max\left\{k \in \{0,\dots,\underline{i}-1\} : \lambda\left(\frac{k}{n},\ p-\varepsilon\right) \leq \sqrt[n]{\beta}\right\},$$
$$\overline{k} := \min\left\{k \in \{\bar{i},\dots,n\} : \lambda\left(\frac{k}{n},\ p+\varepsilon\right) \leq \sqrt[n]{\beta}\right\}$$

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*

$$\mathcal{C}_f := \left\{q \in [0,+\infty) : \left[\underline{k}_f(q),\ \overline{k}_f(q)\right] \cap (\underline{k},\ \overline{k}) \neq \varnothing\right\},$$

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

Proof. Assumption (5) implies in particular $(p-\varepsilon)\,n > 0$ and $\lambda(0,\ p-\varepsilon) = 1-p+\varepsilon \leq \sqrt[n]{\beta}$, and therefore $\underline{k}$ is well-defined, since $\underline{i}-1 \geq 0$ and $k=0$ satisfies the condition of the maximum. Analogously, $\overline{k}$ is well-defined, because $\overline{i} \leq n$ and $k=n$ satisfies the condition of the minimum, since $(p+\varepsilon)\,n < n$ and $\lambda(1,\ p+\varepsilon) = p+\varepsilon \leq \sqrt[n]{\beta}$ follow from assumption (5). The definitions of $\underline{k}$ and $\overline{k}$ imply in particular the inequalities $0 \leq \underline{k} < (p-\varepsilon)\,n$ and $(p+\varepsilon)\,n < \overline{k} \leq n$.

We now prove that $\mathcal{C}_f$ is the likelihood-based confidence region with cutoff point β for the p-quantile of the distribution of the residuals $R_{f,i}$ (that is, for the values of the multivalued mapping Q_f). From the definition of profile likelihood function given in Subsection 2.2 it follows that this confidence region is the set of all $q \in \mathcal{Q}_f$ such that $lik_{Q_f}(q) > \beta$. We can thus use Theorem 1 to determine the confidence region. It can be easily proved that for each $t \in (0,1)$ considered as a constant, λ is a continuous function of $s \in [0,1]$, monotonically increasing on $[0,\ t]$ and monotonically decreasing on $[t,\ 1]$. Therefore, in the first case of expression (2) we have $lik_{Q_f}(q) > \beta$ if and only if $\overline{k}_f(q) > \underline{k}$, while in the third case we have $lik_{Q_f}(q) > \beta$ if and only if $\underline{k}_f(q) < \overline{k}$. Altogether, we obtain that $lik_{Q_f}(q) > \beta$ if and only if $[\underline{k}_f(q),\ \overline{k}_f(q)] \cap (\underline{k},\ \overline{k}) \neq \varnothing$, since $[(p-\varepsilon)\,n,\ (p+\varepsilon)\,n] \subset (\underline{k},\ \overline{k})$. It remains to show that $q \in \mathcal{C}_f$ implies $q \in \mathcal{Q}_f$. If $q \in \mathcal{C}_f$, then $\overline{k}_f(q) > 0$, and so there is an $r \in \mathcal{R}_f$ such that $r \leq q$. Analogously, if $q \in \mathcal{C}_f$, then $\underline{k}_f(q) < n$, and so there is an $r \in \mathcal{R}_f$ such that $r \geq q$. Hence, $q \in \mathcal{C}_f$ implies $q \in \mathcal{Q}_f$, and therefore $\mathcal{C}_f$ is the desired confidence region.

The set $\mathcal{C}_f$ is an interval, since the functions $\overline{k}_f$, $\underline{k}_f$ are monotonically increasing, and $\underline{k}_f(q) \leq \overline{k}_f(q)$ for all $q \in [0,+\infty)$. Moreover, the definition of likelihood function implies that there is a probability measure $P \in \mathcal{P}$ such that $lik(P) > \beta$, and therefore $\mathcal{C}_f$ is not empty, because $Q_f(P) \subseteq \mathcal{C}_f$ follows from the definition of likelihood-based confidence region. Finally, Lemma 3 implies

$$\inf \mathcal{C}_f = \inf\{q \in [0,+\infty) : \overline{k}_f(q) > \underline{k}\} = \underline{r}_{f,(\underline{k}+1)},$$
$$\sup \mathcal{C}_f = \sup\{q \in [0,+\infty) : \underline{k}_f(q) < \overline{k}\} = \overline{r}_{f,(\overline{k})},$$

since $\overline{k}_f(q) = n$ for all $q \in [0,+\infty) \setminus \mathcal{U}_f$, and $\underline{k}_f(q) = 0$ for all $q \in [0,+\infty) \setminus \mathcal{L}_f$. □

The interval $\mathcal{C}_f$ defined in Corollary 2 consists of all $q \in [0,+\infty)$ such that the band $\overline{B}_{f,q}$ intersects at least $\underline{k}+1$ imprecise data, and the band $\underline{B}_{f,q}$ contains at most $\overline{k}-1$ imprecise data. When $\varepsilon = 0$, the interval $\mathcal{C}_f$ is asymptotically a (conservative) confidence interval of level $F_{\chi^2}(-2\,\log\beta)$ for the p-quantile of the distribution of the residuals $R_{f,i}$, where F_{χ^2} is the cumulative distribution function of the chi-square distribution with 1 degree of freedom (see for example [28]). The finite-sample level of the (conservative) confidence interval $\mathcal{C}_f$ can be obtained directly from its definition, by means of simple combinatorial arguments (also when $\varepsilon > 0$), but this goes beyond the scope of the present paper.

It is important to note that the confidence intervals $\mathcal{C}_f$ do not depend on the choice of the set $\mathcal{V}^*$ of possible imprecise data (as far as the observed ones, $A_1, \ldots, A_n$, are contained in it). This can be surprising, since the set $\mathcal{P} = \mathcal{P}_\varepsilon$ of probability measures considered depends strongly on $\mathcal{V}^*$, as noted at the beginning of Section 2. However, the independence of the confidence intervals $\mathcal{C}_f$ from the choice of the set $\mathcal{V}^*$ is not so surprising when one considers that the intervals $\mathcal{C}_f$ are likelihood-based confidence regions, and that likelihood inference is always conditional on the data (that is, independent of considerations about which other data could have been observed). This can be considered as a sort of robustness against misspecification of the set $\mathcal{V}^*$ of possible imprecise data. The practical advantage is that it is not necessary to think about which other imprecise data could have been observed, besides the ones that were actually observed (that is, $A_1, \ldots, A_n$).

Example 8. In the situation of Example 7, the confidence interval $\mathcal{C}_f$ with $\beta = 0.15$ is approximately $[0.16, 4.47]$ when $\varepsilon = 0$ (implying $\underline{k} = 66$ and $\overline{k} = 83$), and $[0, 6.96]$ when $\varepsilon = 0.1$ (implying $\underline{k} = 55$ and $\overline{k} = 91$); the cutoff point $\beta = 0.15$ is represented by the dotted line in Fig. 3.

3.3. Regression as a decision problem

The problem of minimizing the p-quantile of the distribution of the residuals $R_{f,i}$ can be described as a statistical decision problem: the set of probability measures considered is $\mathcal{P} = \mathcal{P}_\varepsilon$, the set of possible decisions is $\mathcal{F}$, and the loss function $L : \mathcal{P} \times \mathcal{F} \to [0,\infty)$ is defined by

$$L(P,f) = Q_f(P)$$

for all $P \in \mathcal{P}$ and all $f \in \mathcal{F}$. That is, the p-quantile of the distribution of the residuals $R_{f,i}$ is interpreted as the loss we incur when we choose the function f. In fact, the loss function L is multivalued, since in general the p-quantile is not unique: $L(P,f)$ could be reduced to a single value by taking for example the upper p-quantile of the distribution of the residuals $R_{f,i}$.

The information provided by the observed (imprecise) data is described by the likelihood function lik on $\mathcal{P}$. A very simple way of using this information consists in reducing $\mathcal{P}$ to the set $\mathcal{P}_{>\beta}$ for some cutoff point $\beta \in (0,1)$. The resulting set $\mathcal{P}_{>\beta}$ can be interpreted as an imprecise probability measure, on which we can base our choice of f. For each $f \in \mathcal{F}$, the set of all possible values of the loss $L(P,f)$ when P varies in $\mathcal{P}_{>\beta}$ can be interpreted as the imprecise p-quantile of the residuals $R_{f,i}$ under the imprecise probability measure $\mathcal{P}_{>\beta}$. It corresponds to the interval $\mathcal{C}_f$, when condition (5) is satisfied.

Assume that condition (5) is satisfied. In order to choose a function f, we can minimize the supremum of $\mathcal{C}_f$. This approach is similar to the Γ-minimax decision criterion with respect to the imprecise probability measure $\mathcal{P}_{>\beta}$, and is called LRM (Likelihood-based Region Minimax) criterion in [7]. When there is a unique $f \in \mathcal{F}$ minimizing $\sup \mathcal{C}_f$ (i.e., minimizing $\overline{r}_{f,(\underline{k})}$), it can be denoted by f_{LRM}, and $\sup \mathcal{C}_f$ can be denoted by $\overline{q}_{LRM}$. In this case, f_{LRM} is characterized geometrically by the fact that $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q}$ containing at least $\underline{k}$ imprecise data, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$. Therefore, in order to find the function f_{LRM}, it suffices to adapt to the case of imprecise data the algorithms for the method of least quantile of squares (see for example [30,37,3]), but this goes beyond the scope of the present paper.

An interesting description of the uncertainty about the optimal choice of $f \in \mathcal{F}$ is obtained by considering interval dominance for the imprecise p-quantiles of the residuals $R_{f,i}$ under the imprecise probability measure $\mathcal{P}_{>\beta}$. When f_{LRM} exists, the undominated functions $f \in \mathcal{F}$ are those such that $\mathcal{C}_f$ intersects $\mathcal{C}_{f_{LRM}}$. In particular, when $\overline{q}_{LRM} \in \mathcal{C}_{f_{LRM}}$ (that is, $\mathcal{C}_{f_{LRM}}$ is right-closed), the undominated functions $f \in \mathcal{F}$ are characterized geometrically by the fact that $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k} + 1$ imprecise data. In general, the set of undominated functions f can be interpreted as the imprecise result of the regression: it describes the complex uncertainty about the optimal choice of $f \in \mathcal{F}$. When we observe more and more (imprecise) data, the statistical uncertainty diminishes, but the set of undominated functions does not necessarily tend to reduce to a singleton, because of the (unavoidable) indetermination discussed in Subsection 2.1 (see also [10] for a more detailed analysis).

Example 9. Consider the problem of simple linear regression introduced in Example 5, with $\varepsilon = 0$ (that is, the classification of the precise data into the elements of $\mathcal{V}^*$ is assumed to be correct), $p = 0.75$, and $\beta = 0.15$. The thinnest band of the form $\overline{B}_{f,q}$ (for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$) containing at least $\underline{k} = 83$ imprecise data is represented by the dashed lines in the left plot of Fig. 4. It is the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$, where f_{LRM} is also plotted in Fig. 4 (left, solid line), while $\overline{q}_{LRM} = \sup \mathcal{C}_{f_{LRM}} = \overline{r}_{f_{LRM},(83)} \approx 4.47$, as we have seen in Example 8. The right plot of Fig. 4 shows the undominated functions $f \in \mathcal{F}$ (gray lines), which are characterized by the fact that the band $\overline{B}_{f,\,4.47}$ intersects at least $\underline{k} + 1 = 67$ imprecise data.

3.4. Prediction

Consider the case in which (instead of n) we have $n + 1$ pairs (V_i, V_i^*) of precise and imprecise data $V_i = (X_i, Y_i)$ and V_i^*, respectively. We want to predict the realization of the precise data value V_{n+1} on the basis of the realization of the n imprecise data $V_1^*, \ldots, V_n^*$. Choose $k \in \{1, \ldots, n\}$, and assume that for each possible realization of the $n+1$ imprecise data $V_1^*, \ldots, V_{n+1}^*$, there is a distance $q' \in [0, +\infty)$ such that for some $f' \in \mathcal{F}$ (not necessarily unique), $\overline{B}_{f',q'}$ is a thinnest band of the form $\overline{B}_{f,q}$ containing at least k of the $n + 1$ imprecise data, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$. Because of symmetry, the probability that V_{n+1}^* is included in a band $\overline{B}_{f,q'}$ containing at least k of the $n + 1$ imprecise data (for some $f \in \mathcal{F}$) is at least $k/(n+1)$. Hence, when $\overline{B}_{f'',q''}$ is a thinnest band of the form $\overline{B}_{f,q}$ containing at least k of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$), the probability that V_{n+1}^* is included in the union $\mathcal{B}$ of all bands $\overline{B}_{f,q''}$ containing at least

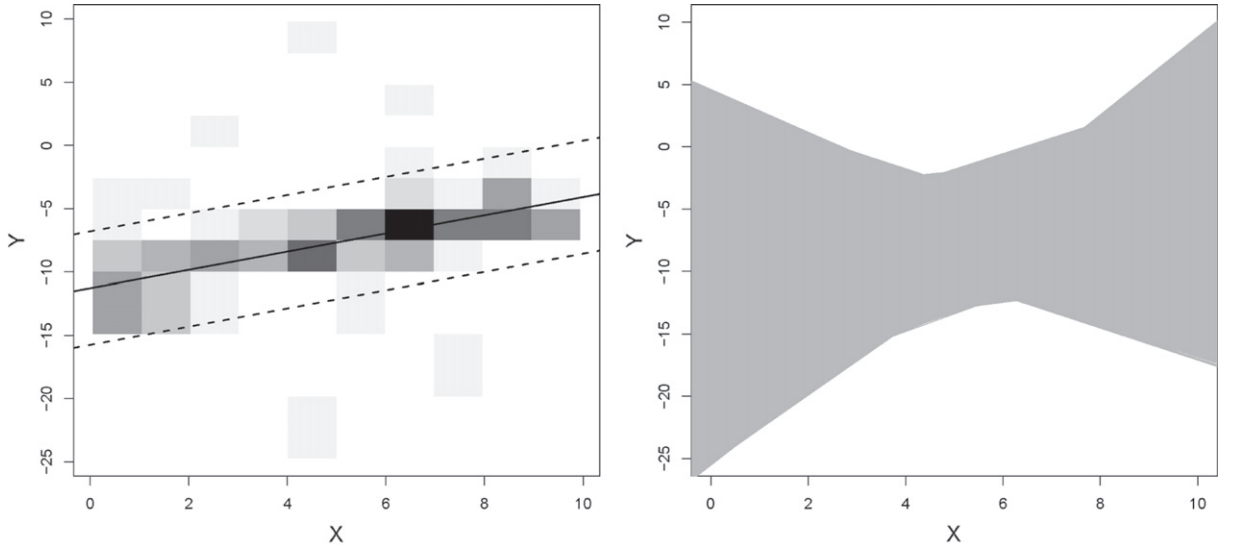

Fig. 4. Function f_{LRM} (left, solid line), band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (left, dashed lines), and set of undominated functions (right, gray lines) from Example 9.

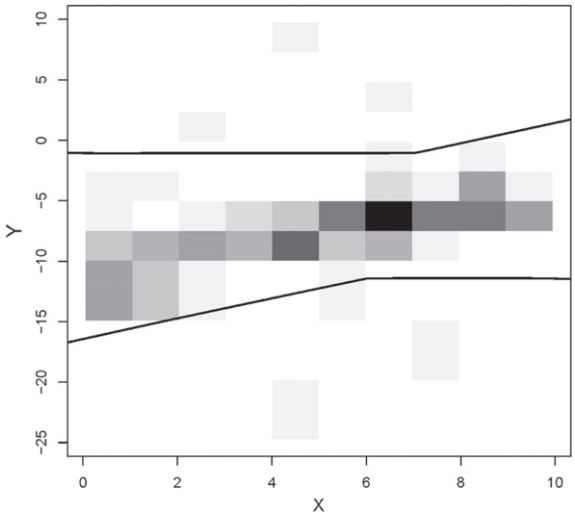

Fig. 5. Prediction region from Example 10.

$k-1$ of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$) is at least $k/(n+1)$. That is, $\mathcal{B}$ is a (conservative) prediction region of level $k/(n+1) - \varepsilon$ for the precise data value V_{n+1}.

In particular, when condition (5) is satisfied and f_{LRM} exists, the union $\mathcal{B}$ of all bands $\overline{B}_{f,\bar{q}_{LRM}}$ containing at least $\overline{k} - 1$ of the n imprecise data $V_1^*, \ldots, V_n^*$ (for all $f \in \mathcal{F}$) is a (conservative) prediction region of level $\overline{k}/(n+1) - \varepsilon$ for the precise data value V_{n+1}. Prediction regions of this form can sometimes be reduced to smaller regions thanks to the assumption that V_{n+1}^* takes values in $\mathcal{V}^*$. When besides the realization of the n imprecise data $V_1^*, \ldots, V_n^*$, also the (precise or imprecise) realization of X_{n+1} has been observed, the realization of Y_{n+1} can be predicted for example by using the idea of conformal prediction (see [36]), but this goes beyond the scope of the present paper.

Example 10. In the situation of Example 9, the union $\mathcal{B}$ of all bands $\overline{B}_{f,\,4.47}$ containing at least $\overline{k} - 1 = 82$ imprecise data (for all $f \in \mathcal{F}$) corresponds to the region between the two curves in Fig. 5. It is a (conservative) prediction region of level $\overline{k}/(n+1) - \varepsilon = 0.83$ for a future precise data point.

4. Example of application

In this section, we apply the proposed regression method to socioeconomic data from the ALLBUS (German General Social Survey). Data collection in surveys is subject to many different influences that may cause various biases in the data set (see for example [4]). Therefore, it is often reasonable to assume that the actual value lies rather in some interval around the observed value. Furthermore, data on sensitive quantities is sometimes only available in categories that form a partition of the space of possible values. A simple, ad hoc approach to analyze this kind of data is to reduce the intervals to their central values and to apply usual regression methods to the reduced, precise data. However, such an approach in general produces biased results (see [32,2,13]). In contrast to this, we suggest to analyze directly the interval-valued data by means of the regression method proposed in Section 3.

Here, we investigate how personal income varies with age, which is a fundamental relationship in the social sciences and a typical example in textbooks on social research methods (see for example [1]). Income is a key demographic variable in socioeconomic research questions, but asking for income in an interview is a sensitive question that some respondents refuse to answer. Therefore, survey data on personal income often include missing values. One way to make the question less sensitive and thus to obtain better response rates is to present predefined income categories (forming a partition of the range of possible income values) to the respondent according to which the personal income shall be classified. In the ALLBUS study, income data is collected by means of a two-step design with the open question for income as first step and the presentation of a category scheme as second step. As a result, the data set contains at the same time precise values for some individuals and interval-valued observations for others. Yet, even if the respondents answer the open question, they usually give only a rough estimate of their exact income, like a rounded or a heaped income value (see [20]), where heaping refers to irregular rounding behavior (see for example [21]). Therefore, it is more reliable to regard also the precise income values as intervals, e.g., in considering as actual observations the income classes in which the precise values lie.

Data on the age of the respondents are more easily obtained, but these data are usually of limited precision. Often, the age is measured on a discrete scale, i.e., $age \in \mathbb{N}$. In that case, the information contained in a data value is that the actual age of the respondent lies in the interval $[age, age + 1)$. Furthermore, also age data are sometimes provided as a

categorical variable taking values in a set of disjoint age classes forming a partition of the observation space of the continuous variable *age*.

4.1. ALLBUS data and regression model

We analyze the ALLBUS data set of 2008 containing 3 469 interviews constituting a disproportional sample of the adult population living in Germany, where Eastern Germany is deliberately over-represented for particular reasons (see [35]). We disregard this disproportion here, as our major purpose is the illustration of the proposed regression method, knowing that our results will not be representative of the German adult population. The variables we consider in our analysis are *personal income* (on average per month in euros) and *age*. We use the categorized income variable *v*389 with 22 possible income classes and the discrete age variable *v*154. (Detailed information about the data set can be found in [35].) Although the data set contains 1 063 precise income values and our regression method could also be applied to a data set with some precise and some imprecise observations (see [10]), we prefer to use the categorized income variable for the reasons mentioned above. Moreover, the age data are interpreted as intervals of length 1. Thus, for each individual $i \in \{1, \ldots, 3\,469\}$ we consider observations $V_i^* = X_i^* \times Y_i^*$, where $X_i^* = [age_i, age_i + 1)$ is the interval covering the age of respondent i and $Y_i^* = [\underline{y}_i, \overline{y}_i)$ is the interval of the corresponding income category. In the given data set, there are 682 missing income values and 12 missing age values. Missing values are replaced by the entire observation space of each variable, i.e., $X_i^* = \mathcal{X} := [18, 100)$ or $Y_i^* = \mathcal{Y} := [0, +\infty)$, respectively. A two-dimensional histogram of the data set is given in Fig. 7 (left), where a darker shade of gray indicates a higher frequency of the imprecise observation.

The relationship between *age* and *income* is usually modeled by a quadratic function in *age* (see for example [1]). Thus, the set of regression functions we consider here is $\mathcal{F} = \{f_{a,b_1,b_2} : a, b_1, b_2 \in \mathbb{R}\}$, where each function f_{a,b_1,b_2} is defined by $f_{a,b_1,b_2}(x) = a + b_1\,x + b_2\,x^2$ for all $x \in \mathcal{X}$. We choose to minimize the median of the distribution of the absolute residuals (i.e., $p = 0.5$), which is the choice of p implying the most robust results (see the beginning of Section 3). As regards the cutoff point of the likelihood, we use a very high value: $\beta = 0.9999$. This choice of β corresponds to the special case of LIR where we consider maximum likelihood (ML) estimates to evaluate the regression functions $f_{a,b_1,b_2} \in \mathcal{F}$ (i.e., $\underline{k} = \underline{i} - 1$ and $\overline{k} = \overline{i}$). Note that in the present analysis the ML estimates $\mathcal{C}_{f_{a,b_1,b_2}}$ of the median of the absolute residuals are intervals, since the analyzed data set consists of proper sets (implying $\underline{r}_{f,i} < \overline{r}_{f,i}$ for each imprecise observation $A_i = X_i^* \times Y_i^*$). Choosing the ML intervals means to ignore the statistical uncertainty of the regression problem. A lower cutoff point β would imply a higher confidence level of the intervals $\mathcal{C}_{f_{a,b_1,b_2}}$ and lead to a more imprecise result. In the present analysis, the resulting set of undominated functions would change only a little, because there is not much statistical uncertainty given the relatively large number of observations. Finally, as we consider only the income classes, we assume that the imprecise observations are correct and set $\varepsilon = 0$. The effect of different choices of β and ε on the result of a LIR analysis has been studied thoroughly in [10].

For the present regression problem, we have implemented the LIR analysis as a grid search over the parameter space $\mathbb{R}^3$: First, the likelihood-based confidence regions $\mathcal{C}_{f_{a,b_1,b_2}}$ are computed for all regression functions corresponding to the parameter values (a, b_1, b_2) on a predefined grid. Then, we identify the parameter combination among these that minimizes the upper end point of $\mathcal{C}_{f_{a,b_1,b_2}}$. The function corresponding to this parameter combination is the function f_{LRM} which is optimal according to the LRM criterion (see Subsection 3.3). Finally, the upper end point $\overline{q}_{LRM}$ of $\mathcal{C}_{f_{LRM}}$ is used to determine the set of undominated regression functions.

4.2. Results

We considered a grid of combinations of parameter values where $a \in [-10\,000,\ 12\,000]$, $b_1 \in [-200,\ 250]$, and $b_2 \in [-10,\ 10]$. Corresponding to the set of undominated functions, we find the set of undominated parameter combinations displayed in Fig. 6. This set is clearly not convex. Moreover, in the case considered here, the parameters are not independent from each other, in the sense that many different combinations of parameter values (a, b_1, b_2) may lead to very similar shapes of f_{a,b_1,b_2} over $\mathcal{X}$. Thus, there are actually infinitely many undominated parameter combinations, but the associated curves are similar to those we find within the considered grid.

The parameter combination implying the smallest upper endpoint of the ML interval for the 0.5-quantile of the residuals is (600, 5, 0) with $\mathcal{C}_{f_{600,5,0}} \approx [270,\ 680]$. Thus, the function f_{LRM} is a slightly increasing line. One interpretation of this function is given by the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ limited by the functions $f_{LRM} - \overline{q}_{LRM}$ and $f_{LRM} + \overline{q}_{LRM}$: Among all bands (of any width) constructed around all considered functions, $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest one that contains at least $\overline{k} = 1\,735$ imprecise observations.

The function f_{LRM} and the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ are presented in Fig. 7 (right, black lines), besides the undominated functions (right, gray curves). As we considered ML estimates, no statistical uncertainty is reflected in the regression's result, thus, the extent of the set of undominated functions is only due to the imprecision of the data. It can be seen that among the undominated functions there is a large variety of shapes of the *age-income* profile, including straight lines, convex parabolic curves as well as concave ones. From a social scientist's point of view this result may be unsatisfying because it does not support only one form of the relationship between *age* and *income*. However, it is reasonable to consider all shapes consistent with the imprecise data as possible *age-income* profiles. If the observed intervals were overlapping or if they constituted

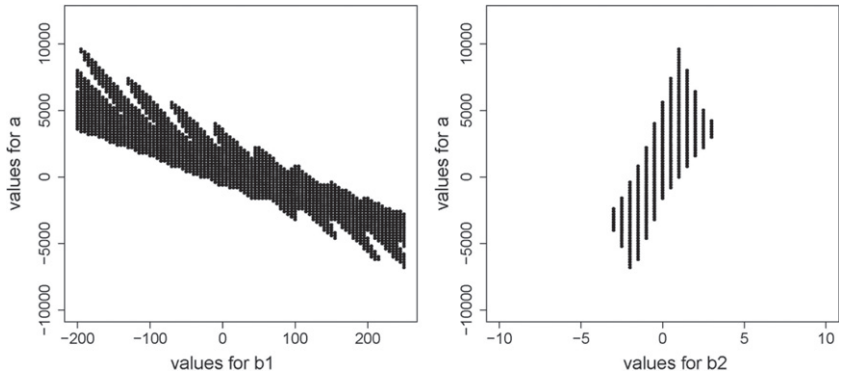

Fig. 6. Two-dimensional projections of the set of undominated parameter values.

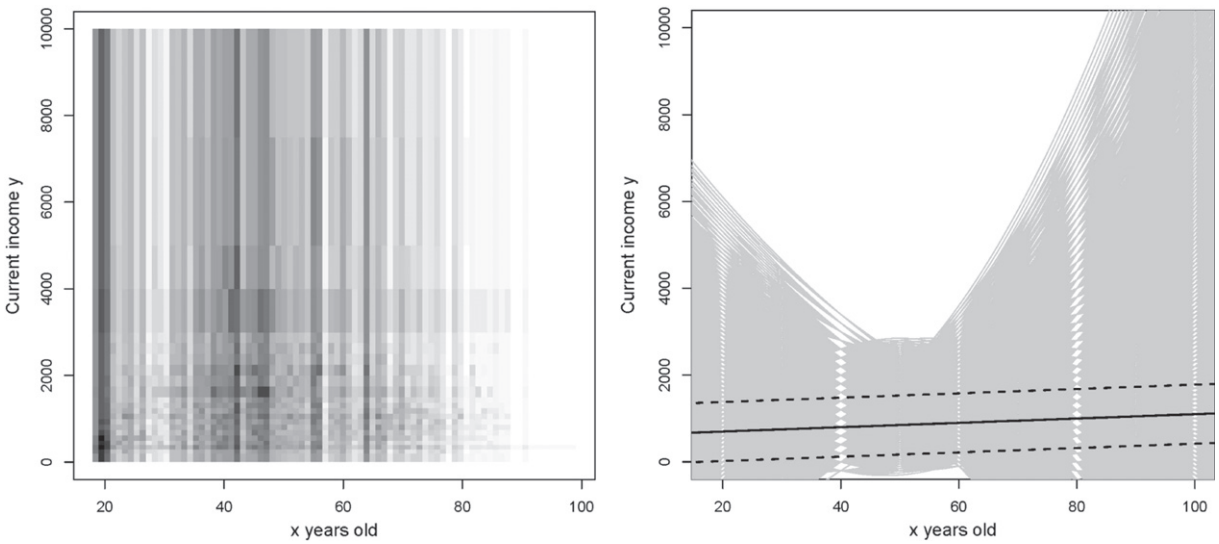

Fig. 7. Two-dimensional histogram of the analyzed data set (left) and set of undominated functions (right, gray curves), minimax function f_{LRM} (right, black solid line) and band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (right, black dashed lines).

a finer partition of the space of possible observations, the set of undominated functions would be smaller. The effect of different degrees of imprecision of the data on the regression's result was studied in [10], where different versions of the ALLBUS data set were analyzed and their results compared. In the present analysis, the set of undominated functions can be interpreted as the set of all plausible descriptions of the *age-income* profile that reflects at the same time the indetermination induced by the imprecise data.

The common, simple method to analyze this kind of interval data is to conduct a quadratic least squares (LS) regression based on the interval centers ignoring the indetermination induced by the imprecision of the data. In this case, an upper limit for the highest income class $[7\,500,\ +\infty)$ has to be set in order to compute the interval centers. Of course, the choice of this upper limit has an impact on the estimates of the LS regression. The effect of two different choices of the upper income limit is illustrated in Fig. 8 (black dashed curves). The LS curves displayed there are based on interval centers with upper income limits 10 000 and 50 000, respectively. In contrast to the LS regression based on the interval midpoints, the regression method proposed in this paper can also be applied to unbounded data. Since in the LIR method the evaluation of the regression functions is based on quantiles of the distribution of the absolute residuals, the result is not sensitive to the extremes. If there were less than $\overline{k}$ bounded data, e.g., if there were more than 50% missing observations in the present data set, the result would be the entire set $\mathcal{F}$ of considered regression functions.

An improvement of the approach of an LS regression based on the interval centers could be achieved by considering a robust variant of this approach, in which least median of squares (LMS) estimation is used. In this case, an upper income limit has to be fixed, but the estimated regression function is insensitive to the choice of the extreme values, since the regression is based on the median of the absolute residuals. The LMS curve estimated on the basis of the interval centers with upper income limit 50 000 (black dotted line) and the function f_{LRM} obtained from the LIR analysis (black solid line) are also shown

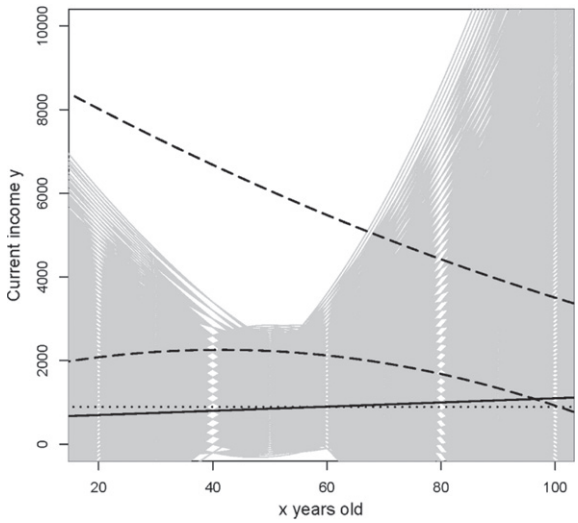

Fig. 8. Set of undominated functions (gray curves) and f_{LRM} (black solid line) of the LIR analysis versus LS curves based on interval centers with upper limits 10 000 (lower black dashed curve) and 50 000 (upper black dashed curve), respectively, and LMS curve (black dotted line) based on interval centers with upper limit 50 000.

in Fig. 8. These lines are similar to each other, which is not surprising as the proposed regression method can be seen as a generalization of the LMS regression to the case of imprecise data.

The proposed LIR method permits to identify plausible descriptions of the relationship between the socioeconomic characteristics *age* and *income*. Given the imprecise data, many different shapes of the age-income profile are plausible. Further computations indicated that our findings hold for transformed income data on the logarithmic scale, too. The results are not very informative, but reflect the indetermination induced by the imprecision of the data. One idea to obtain more informative results from categorized data could be to use many different category schemes during the income data collection and thereby obtain a data set with overlapping categories.

5. Conclusion

In this paper, we introduced a new regression method for imprecise data, in which the error distribution is not constrained to a particular parametric family. The regression method is very robust and can be adapted to a wide range of practical settings, since it can be applied to all kinds of imprecisely observed data, covering among others interval data, precise data, and missing data. In our method, the imprecise data are interpreted as the result of a coarsening process which can be informative, and even wrong with a certain probability.

The proposed method is derived from a novel general approach to regression with imprecise data, which we call Likelihood-based Imprecise Regression. It consists in identifying by means of likelihood inference all sufficiently plausible regression curves, which are considered as the imprecise result of the regression analysis. The extent of the imprecise result reflects both kinds of uncertainty involved in a regression problem with imprecise data: statistical uncertainty and indetermination.

In this paper, we developed the theoretical framework of LIR. First, we introduced a general methodology for likelihood inference with imprecise data and then we applied it to the statistical problem of regression. Focusing on the setting without parametric distributional assumptions and where quantiles of the residuals distribution are used to evaluate the possible descriptions of the relationship of interest, we derived the above mentioned robust regression method for imprecise data and set out the mathematical details of this LIR method. Moreover, we suggested a first implementation of the proposed LIR method illustrated with an application example. Of course, the computational issues related to the new methodology have to be examined in further detail. Some of these aspects in the special case of simple linear regression with interval data are studied in [11], where we also suggest an improved implementation of the robust LIR method.

In future work, we intend to further amend the implementation of the robust LIR method, and to study the computational issues in more detail. We also want to further examine its statistical properties as well as to develop criteria to evaluate the performance of a LIR analysis. Moreover, we plan to investigate the consequences of stronger assumptions about the error distribution and the coarsening process, and the possibility of replacing in the decision problem the quantiles of the residuals by other loss functions.

Acknowledgements

The authors wish to thank Thomas Augustin and the anonymous referees for their helpful comments.

References

[1] P.D. Allison, Multiple Regression, Pine Forge Press, 1998.
[2] A.E. Beaton, D.B. Rubin, J.L. Barone, The acceptability of regression solutions: Another look at computational accuracy, J. Am. Stat. Assoc. 71 (1976) 158–168.
[3] T. Bernholt, Computing the least median of squares estimator in time $O(n^d)$, in: O. Gervasi, M.L. Gavrilova, V. Kumar, A. Laganà, H.P. Lee, Y. Mun, D. Taniar, C.J.K. Tan (Eds.), Computational Science and Its Applications — ICCSA 2005, Springer, 2005, pp. 697–706.
[4] P.P. Biemer, L.E. Lyberg, Introduction to Survey Quality, Wiley, 2003.
[5] L. Billard, E. Diday, Regression analysis for interval-valued data, in: H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen, M. Schader (Eds.), Data Analysis, Classification, and Related Methods, Springer, 2000, pp. 369–374.
[6] A. Blanco-Fernández, N. Corral, G. González-Rodríguez, Estimation of a flexible simple linear model for interval data based on set arithmetic, Comput. Stat. Data Anal. 55 (2011) 2568–2578.
[7] M. Cattaneo, Statistical Decisions Based Directly on the Likelihood Function, Ph.D. thesis, ETH Zurich, 2007.
[8] M. Cattaneo, Fuzzy probabilities based on the likelihood function, in: D. Dubois, M.A. Lubiano, H. Prade, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Soft Methods for Handling Variability and Imprecision, Springer, 2008, pp. 43–50.
[9] M. Cattaneo, A. Wiencierz, Regression with imprecise data: A robust approach, in: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (Eds.), ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications, SIPTA, 2011, pp. 119–128.
[10] M. Cattaneo, A. Wiencierz, Robust regression with imprecise data, Technical Report 114, Department of Statistics, LMU Munich, 2011.
[11] M. Cattaneo, A. Wiencierz, On the implementation of LIR: the case of simple linear regression with interval data, Technical Report 127, Department of Statistics, LMU Munich, 2012.
[12] G. de Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Artif. Intell. 159 (2004) 75–125.
[13] A.P. Dempster, D.B. Rubin, Rounding error in regression: The appropriateness of Sheppard's corrections, J. R. Stat. Soc. Ser. B 45 (1983) 51–59.
[14] P. Diamond, Least squares fitting of compact set-valued data, J. Math. Anal. Appl. 147 (1990) 351–362.
[15] M.A.O. Domingues, R.M.C.R. de Souza, F.J.A. Cysneiros, A robust method for linear regression of symbolic interval data, Pattern Recognit. Lett. 31 (2010) 1991–1996.
[16] M.B. Ferraro, R. Coppi, G. González-Rodríguez, A. Colubi, A linear regression model for imprecise response, Int. J. Approx. Reason. 51 (2010) 759–770.
[17] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, L. Ginzburg, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, Technical Report, SAND2007-0939, Sandia National Laboratories, 2007.
[18] F. Gioia, C.N. Lauro, Basic statistical methods for interval data, Ital. J. Appl. Stat. 17 (2005) 75–104.
[19] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, 1986.
[20] J.U. Hanisch, Rounded responses to income questions, Allg. Stat. Arch. 89 (2005) 39–48.
[21] D.F. Heitjan, D.B. Rubin, Ignorability and coarse data, Ann. Stat. 19 (1991) 2244–2253.
[22] P.J. Huber, E.M. Ronchetti, Robust Statistics, 2nd ed., Wiley, 2009.
[23] D.J. Hudson, Interval estimation from the likelihood function, J. R. Stat. Soc. Ser. B 33 (1971) 256–262.
[24] E.A. Lima Neto, F.A.T. de Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Stat. Data Anal. 52 (2008) 1500–1515.
[25] C.F. Manski, Partial Identification of Probability Distributions, Springer, 2003.
[26] M. Marino, F. Palumbo, Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression, Ital. J. Appl. Stat. 14 (2002) 277–291.
[27] R.A. Maronna, D.R. Martin, V.J. Yohai, Robust Statistics: Theory and Methods, Wiley, 2006.
[28] A.B. Owen, Empirical Likelihood, Chapman & Hall/CRC, 2001.
[29] Y. Pawitan, All Likelihood, Oxford University Press, 2001
[30] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, 1987.
[31] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[32] W.F. Sheppard, On the calculation of the most probable values of frequency-constants for data arranged according to equidistant divisions of a scale, Lond. M.S. Proc. 29 (1898) 353–380.
[33] V. Strassen, Meßfehler und information, Z. Wahrscheinlichkeitstheorie 2 (1964) 273–305.
[34] D. Tasche, Unbiasedness in least quantile regression, in: R. Dutter, P. Filzmoser, U. Gather, P.J. Rousseeuw (Eds.), Developments in Robust Statistics, Physica-Verlag, 2003, pp. 377–386.
[35] M. Terwey, S. Baltzer, ALLBUS Datenhandbuch 2008, GESIS, 2009.
[36] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, 2005.
[37] G.A. Watson, On computing the least quantile of squares estimate, SIAM J. Sci. Comput. 19 (1998) 1125–1138.
[38] M. Zaffalon, E. Miranda, Conservative inference rule for uncertain reasoning under incompleteness, J. Artif. Intell. Res. (JAIR) 34 (2009) 757–821.
[39] Z. Zhang, Profile likelihood and incomplete data, Int. Stat. Rev. 78 (2010) 102–116.