A Minimum Distance Estimator in an Imprecise Probability Model – Computational Aspects and Applications

Robert Hable
University of Bayreuth, Germany
Robert.Hable@uni-bayreuth.de

Abstract

The present article considers estimating a parameter θ in an imprecise probability model $(\overline{P}_\theta)_{\theta\in\Theta}$ which consists of coherent upper previsions $\overline{P}_\theta$. After the definition of a minimum distance estimator in this setup and a summarization of its main properties, the focus lies on applications. It is shown that approximate minimum distances on the discretized sample space can be calculated by linear programming. After a discussion of some computational aspects, the estimator is applied in a simulation study consisting of two different models. Finally, the estimator is applied on a real data set in a linear regression model.

Keywords. Imprecise probabilities, coherent lower previsions, minimum distance estimator, empirical measure, R Project for Statistical Computing.

1 Introduction

1.1 Motivation

In classical statistics, it is common to assume complete knowledge about a statistical model which consists of a (smooth parametric) family $(P_\theta)_{\theta\in\Theta}$ of (precise) probability measures. The task is to make assertions about the true parameter $\theta_0 \in \Theta$. Most often, it is assumed that such assertions can be based on data $x_1, \ldots, x_n$ from random variables which are independent identically distributed according to the true distribution P_{θ_0}. That is, the data analyst already knows that the real distribution P_0 can only be a member of a very special family of probability measures $(P_\theta)_{\theta\in\Theta}$ and the only thing which is not one hundred percent sure is the correct parameter $\theta_0 \in \Theta$. Since this assumption is much to strong for many real applications, generalizations of this probabilistic setup are needed. Suitable generalizations of the concept of probability have been developed, among others, by [12] (coherent lower/upper prevision) and [15] (F-probability). Here, the probability of an event is no longer a number $p \in [0,1]$ but an interval $[\underline{p}, \overline{p}] \subset [0,1]$. In order to generalize the setup of classical statistics to a (more realistic) imprecise probability setup, it is natural to replace the precise model $(P_\theta)_{\theta\in\Theta}$ by an imprecise model $(\overline{P}_\theta)_{\theta\in\Theta}$ which consists of such coherent upper previsions $\overline{P}_\theta$.

The classical frequentist theory of statistics is, in large part, concerned with hypothesis testing (in the sense of Neyman-Pearson) and estimating a parameter. While Neyman-Pearson testing under imprecise probabilities has been extensively studied (cf. e.g. [1] and [2]), estimating a parameter has hardly been considered explicitly within the theory of coherent lower previsions so far. There are a few articles which are concerned with it in Bayesian models (primarily associated with Walley's Imprecise Dirichlet Model), e.g. [13], [9], [7] and [14]. In addition, there are a few articles which address very special applications, e.g. [8] (climate projections) and [3] (prediction of the next influenza pandemic). However, general investigations about frequentist estimation of a parameter using coherent lower/upper previsions are still missing. A first attempt is made in [6] where a minimum distance estimator is developed, and its asymptotic properties are investigated.

The present article focuses on applications of this estimator; for the theoretical validation of the estimator, it is referred to [6]. After a recollection of the definition and the basic properties of the minimum distance estimator in Section 2, Section 3 investigates the concrete calculation of the estimator. At first, the sample space has to be suitable discretized, then the distances between the empirical measure and the coherent upper previsions can be approximately calculated by linear programming. An explicit linear program is developed in Subsection 3.2. The minimum distance estimator is already implemented in the (open source) statistical programming language R; it is publicly available as R-package "imprProbEst" [5]. Subsection 3.3 explains some details about this implementation in R. Next, Section 4 presents a simulation

study where the estimator is applied in two different models and compared to classical estimators. This simulation study exemplifies that the proposed estimator can also be calculated for large sample sizes. This meets objections that, due to high computational costs, imprecise probabilities could not be used for practical purposes. Finally, the minimum distance estimator is applied on a real data set in Section 5. Section 6 contains some concluding remarks.

1.2 Setup and Notation

Let $\mathcal{X}$ be a set with σ-algebra $\mathcal{B}$. Then, $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ denotes the set of all bounded, $\mathcal{B}$-measurable real functions $f : \mathcal{X} \to \mathbb{R}$. The supremum norm on $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ is denoted by $\|f\| = \sup_{x\in\mathcal{X}} |f(x)|$. The set of all bounded, finitely additive, signed measures is denoted by $\mathrm{ba}(\mathcal{X},\mathcal{B})$ and can be identified with the dual space of $\mathcal{L}_\infty(\mathcal{X},\mathcal{B})$; cf. [4, Theorem IV.5.1]. Finally, $\mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ denotes the set of all finitely additive probability measures. Integrals with respect to $\mu \in \mathrm{ba}(\mathcal{X},\mathcal{B})$ are denoted by $\mu[f]$.
In accordance with [12, § 2.5.1], a coherent upper prevision on $(\mathcal{X},\mathcal{B})$ is a map

$$\overline{P} \;:\; \mathcal{L}_\infty(\mathcal{X},\mathcal{B}) \;\to\; \mathbb{R}, \qquad f \;\mapsto\; \overline{P}[f]$$

such that there is a (non-empty) set $\mathcal{V} \subset \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ and $\overline{P}[f] = \sup_{P\in\mathcal{V}} P[f]$ for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$; cf. also [12, § 3.3.3] and [6, § 2.3]. The non-empty set $\mathcal{M} := \left\{P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \;\middle|\; P[f] \le \overline{P}[f] \;\; \forall f\right\}$ is called *credal set of* $\overline{P}$ then.
A coherent upper prevision $\overline{P}$ is called *finitely generated* if there is a finite set $\{f_1,\dots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ such that $\overline{P}$ is the natural extension of a coherent upper prevision on $\{f_1,\dots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$. That is, $P \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$ is in the credal set of $\overline{P}$ if and only if $P[f_j] \le \overline{P}[f_j]$ for every $j \in \{1,\dots,s\}$. Such coherent upper previsions naturally arise in applications whenever a practitioner is only able to specify upper (or lower) bounds on the probability or expectation of a finite number of events or functions respectively. A finitely generated, coherent upper prevision $\overline{P}$ is called *regular* if, in addition, $\overline{P}[f_j] > \underline{P}[f_j] \;\; \forall j \in \{1,\dots,s\}$ where $\underline{P}$ denotes the coherent lower prevision corresponding to $\overline{P}$; i.e. $\underline{P}[f] = -\overline{P}[-f] = \inf_{P\in\mathcal{M}} P[f]$ for every $f \in \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$.

2 A minimum distance estimator for imprecise models

2.1 Assumptions

In order to state the definition of the minimum distance estimator, the following fixings and assumptions are made. These are valid throughout the rest of the article:

$(\mathcal{X},\mathcal{B})$ is a measurable space and Θ is a finite[1] index set. The data $x_1,\dots,x_n$ stem from random variables which are independent identically distributed according to a probability measure P_0. For every $\theta \in \Theta$, let $\overline{P}_\theta$ be a coherent upper previsions on $(\mathcal{X},\mathcal{B})$ with credal set $\mathcal{M}_\theta$; $(\overline{P}_\theta)_{\theta\in\Theta}$ is called *imprecise model*. It is only assumed that the true probability measure P_0 is contained in $\mathcal{M}_{\theta_0}$ where $\theta_0 \in \Theta$ is the unknown true parameter. The task is to estimate θ_0.[2]

The following fundamental assumptions on the coherent upper previsions are made:

There is a finite subset $\mathcal{K} = \{f_1,\dots,f_s\} \subset \mathcal{L}_\infty(\mathcal{X},\mathcal{B})$ such that

$$\mathcal{M}_\theta \;=\; \left\{P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \;\middle|\; P_\theta[f_j] \le \overline{P}_\theta[f_j] \;\; \forall f_j \in \mathcal{K}\right\}$$

for every $\theta \in \Theta$. Furthermore, it is assumed that

$$\overline{P}_\theta[f_j] - \underline{P}_\theta[f_j] \;>\; 0 \qquad \forall f_j \in \mathcal{K} \tag{1}$$

where $\underline{P}_\theta$ is the corresponding lower coherent prevision. In particular, each $\overline{P}_\theta$ is a regular, finitely generated coherent upper previsions.[3]
In the following, it is always assumed that each $f_j \in \mathcal{K}$ is standardized; i.e. $\inf f_j = 0$ and $\sup f_j = 1$. Of course, this is no restriction since every bounded, non-constant function f' can be standardized by

$$f \;:=\; \frac{f' - \inf f'}{\sup f' - \inf f'}$$

and, for every $P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$, we have

$$P_\theta[f] \;\le\; \overline{P}_\theta[f] \qquad \Leftrightarrow \qquad P_\theta[f'] \;\le\; \overline{P}_\theta[f']$$

2.2 Definition and asymptotic properties of the minimum distance estimator

The idea of the minimum distance estimator developed in [6, § 6] is very simple: The data $x_1, \dots, x_n$ are used to build the empirical measure $\mathbb{P}^{(n)}$. Then, the minimum distance estimator is that $\hat{\theta} \in \Theta$ such

[1] Finiteness of the index set is not crucial for the definition and basic properties of the estimator (see [6, § 6]) but the algorithm which calculates the estimator is based on this assumption (see § 3).

[2] This approach corresponds to the use of the type-2 product of coherent upper previsions [12, § 9.3.5]. The type-2 product of coherent upper previsions is consistent with a strict sensitivity analyst's point of view on imprecise probabilities.

[3] Though credal sets may also contain elements P which are not σ-additive, the above assumptions include that P_0 is σ-additive. In case of regular, finitely generated coherent upper previsions, this assumption is justified by [6, Prop. 6.4] which states that these previsions can be represented by sets of (σ-additive) probability measures.

that $\mathbb{P}^{(n)}$ lies next to $\mathcal{M}_{\hat{\theta}}$. That is, we calculate the distance between $\mathbb{P}^{(n)}$ and $\mathcal{M}_\theta$ for every $\theta \in \Theta$ and pick that $\hat{\theta}$ where the distance is minimal.

The empirical measure $\mathbb{P}^{(n)}$ is defined to be the map

$$\mathbb{P}^{(n)} : \mathcal{X}^n \to \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}), \quad x \mapsto \mathbb{P}_x^{(n)} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$$

where $x = (x_1,\ldots,x_n)$ and δ_{x_i} denotes the Dirac measure in $x_i \in \mathcal{X}$. Appropriately to the sensitivity analyst's point of view, the distance between a measure P' and a coherent upper prevision $\overline{P}$ is defined to be

$$\left\|P' - \overline{P}\right\| := \inf_{P\in\mathcal{M}} \|P' - P\| \tag{2}$$

where $\mathcal{M}$ denotes the credal set of $\overline{P}$ and $\|P' - P\|$ the operator norm

$$\|P' - P\| = \sup_{f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{B})} \frac{\left| P'[f] - P[f] \right|}{\|f\|}$$

The *minimum distance estimator* $\hat{\theta}_n$ is defined to be

$$\hat{\theta}_n : x \mapsto \arg\min_{\theta\in\Theta} \left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\right\|$$

Note that the minimizing θ is not always unique; in this case, the minimum distance estimator may pick any minimizing θ.

Now, let us turn over to the asymptotic properties of the minimum distance estimator according to [6, § 6.4]. Firstly, note that the use of the operator norm together with the empirical measure is not unproblematic in classical statistics: Though several distances d provide the desirable property that

$$d\big(\mathbb{P}_x^{(n)}, P_0\big) \xrightarrow[n\to\infty]{} 0 \tag{3}$$

almost surely, this is not necessarily true for the operator norm (e.g. in case of the standard normal distribution). However, this annoying difficulty totally disappears in our imprecise probability setup (Subsections 1.2 and 2.1). If we replace P_0 by a regular, finitely generated coherent upper prevision $\overline{P}$, we get that

$$\left\|\mathbb{P}_x^{(n)} - \overline{P}\right\| \xrightarrow[n\to\infty]{} 0 \tag{4}$$

almost surely if P_0 lies in the credal set $\mathcal{M}$ of $\overline{P}$; cf. [6, Theorem 6.6].

A true parameter θ_0 is any $\theta_0 \in \Theta$ such that

$$P_0 \in \mathcal{M}_{\theta_0}$$

Since it is not assumed that the credal sets are disjoint, there may be several true parameters.

According to [6, Theorem 6.10], the probability of the event

$$\left\{x \in \mathcal{X}^n \;\middle|\; P_0 \notin \mathcal{M}_{\hat{\theta}_n(x)}\right\} \tag{5}$$

tends to zero for increasing sample size n if the index set Θ is finite.

The mathematically rigorous statements about these asymptotic properties are more involved and have to be formulated in terms of random variables and image measures. This is because the expressions in (4) and (5) will not be measurable in general. For the treatment of unmeasurable maps in asymptotic statistics, confer e.g. [11, §18].

3 Calculation of the minimum distance estimator

3.1 Discretization of the sample space

As seen in the previous section, it is not necessary to discretize the sample space in order to define the minimum distance estimator based on the total variation norm in a sensible way. Since this is not possible for precise probabilities, going over to imprecise probabilities, in a sense, turns out to be a simplification.
Of course, if we want to calculate the estimator by use of computers, the sample space has to be discretized – at least implicitly. However, it is one of the most striking properties of the above presented minimum distance estimator, that this is only a practical need which is irrelevant for theoretical investigations. That is, we can also deal with infinite sample spaces $(\mathcal{X},\mathcal{B})$. In case of precise probabilities, discretization would even be part of the definition of the minimum distance estimator.

Recall our assumptions given in Subsection 2.1. In order to calculate the minimum distance estimator, we have to calculate

$$\left\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\right\| = \inf_{P_\theta\in\mathcal{M}_\theta} \sup_{f\in\mathcal{L}_\infty(\mathcal{X},\mathcal{B})} \frac{\left| \mathbb{P}_x^{(n)}[f] - P_\theta[f] \right|}{\|f\|}$$

for $\theta \in \Theta$. Though $\mathcal{M}_\theta$ is a large subset of $\mathrm{ba}_1^+(\mathcal{X},\mathcal{B})$, these values can nevertheless be calculated with arbitrary accuracy as explained in the following:

At first, fix any accuracy $\varepsilon > 0$. Then, the sample space $(\mathcal{X},\mathcal{B})$ may be discretized as follows:

For $\theta \in \Theta$, let $\mathcal{K}_\theta$ be the smallest subset of $\mathcal{K}$ such that

$$\mathcal{M}_\theta = \left\{P_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{B}) \;\middle|\; P_\theta[f_j] \le \overline{P}_\theta[f_j] \;\forall\, f_j \in \mathcal{K}_\theta\right\}$$

and put $\mathcal{I}_\theta = \big\{j \in \{1,\dots,s\}\big|\, f_j \in \mathcal{K}_\theta\big\}$. That is, $\mathcal{K}_\theta = \{f_j \in \mathcal{K} | j \in \mathcal{I}_\theta\}$. Furthermore, put

$$\varepsilon_\theta^{(j)} \;:=\; \frac{\overline{P}_\theta[f_j] - \underline{P}_\theta[f_j]}{2s} \cdot \varepsilon \qquad \forall\, j \in \mathcal{I}_\theta$$

and choose simple functions $h_\theta^{(j)}$ such that

$$f_j \;\le\; h_\theta^{(j)} \;\le\; f_j + \varepsilon_\theta^{(j)} \qquad \forall\, j \in \mathcal{I}_\theta \tag{6}$$

Then, let $\mathcal{C}_\theta$ be the smallest σ-algebra on $\mathcal{X}$ such that the simple functions $h_\theta^{(j)}$, $j \in \mathcal{I}_\theta$, are $\mathcal{C}_\theta/\mathbb{B}$-measurable. Note that $\mathcal{C}_\theta$ is a finite subset of $\mathcal{B}$. So, there is a finite partition $\big\{C_\theta^{(1)},\dots,C_\theta^{(r)}\big\}$ of $\mathcal{X}$ such that every event $C \in \mathcal{C}_\theta$ is a (finite) union of elements of the partition $\big\{C_\theta^{(1)},\dots,C_\theta^{(r)}\big\}$.

Now, let $\overline{Q}_\theta$ be the coherent upper prevision on $\mathcal{L}_\infty(\mathcal{X},\mathcal{C}_\theta)$ which corresponds to the credal set

$$\mathcal{N}_\theta \;=\; \left\{ Q_\theta \in \mathrm{ba}_1^+(\mathcal{X},\mathcal{C}_\theta) \;\middle|\; \begin{array}{c} Q_\theta[h_\theta^{(j)}] \;\le\; \overline{P}_\theta[f_j] + \varepsilon_\theta^{(j)} \\ \forall\, j \in \mathcal{I}_\theta \end{array} \right\}$$

According to [6, Theorem 6.11], we have the following inequalities for every $x \in \mathcal{X}^n$:

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;\le\; \big\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\big\| \;\le\; \big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| + \varepsilon \tag{7}$$

3.2 Approximate calculation of the distance by linear programming

According to (7), it is possible to calculate

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| = \inf_{Q_\theta \in \mathcal{N}_\theta} \sup_{f \in \mathcal{L}(\mathcal{X},\mathcal{C}_\theta)} \frac{\big|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\big|}{\|f\|} \tag{8}$$

in order to approximately calculate $\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\|$, where $\overline{Q}_\theta$ is a coherent upper prevision on the finite space $(\mathcal{X},\mathcal{C}_\theta)$. So, we have to minimize the convex function

$$\mathcal{N}_\theta \;\to\; \mathbb{R}\,, \qquad Q_\theta \;\mapsto\; \sup_{f \in \mathcal{L}(\mathcal{X},\mathcal{C}_\theta)} \frac{\big|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]\big|}{\|f\|}$$

Though this is a convex optimization problem, the optimal solution can be found by solving one single linear program.

In order to formulate this linear program, choose any $c_j \in C_\theta^{(j)}$ for every element $C_\theta^{(j)}$ of the partition $\{C_\theta^{(1)},\dots,C_\theta^{(r)}\}$ of $\mathcal{X}$ which generates $\mathcal{C}_\theta$. Furthermore, put

$$N_j \;=\; \big\{i \in \{1,\dots,n\} \;\big|\; x_i \in C_\theta^{(j)}\big\}$$

and let n_j be the number of elements in N_j for every $j \in \{1,\dots,r\}$. In addition, put

$$\mathcal{J}_0 \;=\; \big\{j \in \{1,\dots,r\} \;\big|\; n_j = 0\big\}$$

and

$$\mathcal{J}_1 \;=\; \big\{j \in \{1,\dots,r\} \;\big|\; n_j > 0\big\}$$

Now, consider the following linear program:

$$\sum_{j \in \mathcal{J}_1} q_j - \gamma_j \quad\longrightarrow\quad \max! \tag{9}$$

where

$$\sum_{j=1}^{r} q_j \;=\; 1\,, \tag{10}$$

$$\sum_{j=1}^{r} q_j h_\theta^{(k)}(c_j) \;\le\; \overline{P}_\theta[f_k] + \varepsilon_\theta^{(k)} \quad \forall\, k \in \mathcal{I}_\theta \tag{11}$$

and

$$q_j - \gamma_j \;\le\; \frac{n_j}{n} \qquad \forall\, j \in \mathcal{J}_1 \tag{12}$$

for the variables

$$(q_1,\dots,q_r) \;\in\; \mathbb{R}^r\,, \quad q_j \ge 0 \quad \forall\, j \in \{1,\dots,r\} \tag{13}$$

and

$$(\gamma_j)_{j \in \mathcal{J}_1} \;\subset\; \mathbb{R}\,, \qquad \gamma_j \ge 0 \quad \forall\, j \in \mathcal{J}_1 \tag{14}$$

Let β_θ be the optimal value of the above linear program. Then, Proposition 3.1 below shows that

$$\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\| \;=\; 2 \cdot \big(1 - \beta_\theta\big) \tag{15}$$

Hence, it is, in fact, enough to solve one single linear program in order to obtain the distance $\big\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\big\|$. Of course, this was useless in applications if this linear program would tend to be unsolvable because of exceedingly high computational costs. So let us take a closer look on the size of the above linear program:

Since the number of elements in $\mathcal{J}_1$ is not larger than $\min\{r,n\}$, we have the following upper bounds:

Number of variables: $\quad r + \min\{r,n\}$

Number of inequalities: $\quad 2 + \sharp(\mathcal{K}_\theta) + \min\{r,n\}$

Similar to the discretization method presented in [6, § 5.4] in data-based decision theory, r can – in general – exceed beyond all reasonable bounds but will stay within a reasonable order of magnitude in most applications. In particular, the latter statement is true if the functions $f_j \in \mathcal{K}_\theta$ are convex, concave or indicator functions of (finite unions of) intervals; confer [6, Prop. 5.16]. Though the number n of observations may be very large, it will hardly reach astronomical orders of magnitude in real applications. The size of

the number of elements in $\mathcal{K}_\theta$ (i.e. the number of elements in $\mathcal{I}_\theta$) will usually be negligible.

Note that a very large r will usually result from small values $\varepsilon_\theta^{(j)}$. However, in most real applications, $\overline{P}_\theta$ cannot be specified so accurately that too small values $\varepsilon_\theta^{(j)}$ are meaningful. Furthermore, such small values $\varepsilon_\theta^{(j)}$ indicates that the imprecise model $(\overline{P}_\theta)_{\theta\in\Theta}$ is in danger of being instable – confer [6, § 5.2]. In this case it might be justified to replace $\varepsilon_\theta^{(j)}$ by a larger value. In doing so, we end up with a linear program of a smaller size but, then, it is not guaranteed that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ still is an approximation of $\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\|$. However, replacing $\varepsilon_\theta^{(j)}$ by a larger value corresponds to a more conservative proceeding. If this has a large effect on $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$, this means that small changes of $\overline{P}_\theta[f_j]$, $j \in \mathcal{I}_\theta$, have large effects on $\overline{P}_\theta[f]$ for some $f \notin \mathcal{K}_\theta$. In this unstable case, it seems to be a good idea to be more conservative because this may save from arbitrary results.[4]

The following proposition says that $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ can indeed be calculated by solving the linear program given by (9) – (14):

Proposition 3.1 *Let β_θ be the optimal value of the linear program given by (9) – (14). Then, $\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|$ is given by (15).*

Proof:
STEP 1: Firstly, it is shown that, for every $Q \in \mathcal{N}_\theta$,

$$\|\mathbb{P}_x^{(n)} - Q\| = 2\sum_{j\in\mathcal{J}_1}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \qquad (16)$$

To this end, fix any $Q \in \mathcal{N}_\theta$ and note that – due to finiteness of $\mathcal{C}_\theta$ – the total variation distance is equal to

$$\|\mathbb{P}_x^{(n)} - Q\| \;=\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\big| \qquad (17)$$

Since $\{C_\theta^{(1)},\ldots,C_\theta^{(r)}\}$ is a partition of $\mathcal{X}$, we have

$$\begin{aligned} 0 \;&=\; \mathbb{P}_x^{(n)}(\mathcal{X}) - Q(\mathcal{X}) \;=\; \sum_{j=1}^{r}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) \\ &=\; \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ - \\ &\qquad - \sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^- \end{aligned}$$

Hence,

$$\begin{aligned} \|\mathbb{P}_x^{(n)} - Q\| \;&\overset{(17)}{=}\; \sum_{j=1}^{r}\big|\,\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\big| \\ &=\; 2\cdot\sum_{j=1}^{r}\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \end{aligned}$$

Note that $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) = 0$ if $j \notin \mathcal{J}_1$ and, therefore,

$$\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \;=\; 0 \qquad \forall\, j \notin \mathcal{J}_1$$

This proves (16).

STEP 2: Secondly, it is shown that, for every $Q \in \mathcal{N}_\theta$ and every $j \in \mathcal{J}_1$,

$$\begin{aligned} &\Big(\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)})\Big)^+ \;= \\ &\quad=\; \inf_{\gamma_j\in\Gamma_j(Q)} \mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) + \gamma_j \qquad (18) \end{aligned}$$

where

$$\Gamma_j(Q) := \left\{\gamma_j \in \mathbb{R}\;\middle|\; \begin{array}{c}\gamma_j \geq 0\,,\\ Q(C_\theta^{(j)}) - \gamma_j \;\leq\; \mathbb{P}_x^{(n)}(C_\theta^{(j)})\end{array}\right\}$$

In case of $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) \geq Q(C_\theta^{(j)})$, it is easy to see that the infimum is attained in $\tilde{\gamma}_j = 0 \in \Gamma_j(Q)$ and, therefore, (18) is fulfilled.
In case of $\mathbb{P}_x^{(n)}(C_\theta^{(j)}) < Q(C_\theta^{(j)})$, it is easy to see that the infimum is attained in $\tilde{\gamma}_j = Q(C_\theta^{(j)}) - \mathbb{P}_x^{(n)}(C_\theta^{(j)}) \in \Gamma_j(Q)$ and, therefore, (18) is again fulfilled.

STEP 3: Finally, put

$$\mathbb{M} \;=\; \left\{(Q,\gamma) \in \mathcal{N}_\theta \times \mathbb{R}^{\sharp(\mathcal{J}_1)}\;\middle|\; \begin{array}{c}\gamma = (\gamma_j)_{j\in\mathcal{J}_1}\,,\\ \gamma_j \in \Gamma_j(Q)\;\forall\, j \in \mathcal{J}_1\end{array}\right\}$$

Then, it follows from *STEP 1* and *STEP 2* that

$$\begin{aligned} &\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| \;= \qquad (19) \\ &\quad=\; 2\cdot\inf_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) - Q(C_\theta^{(j)}) + \gamma_j \end{aligned}$$

The definition of $\mathcal{J}_1$ implies $\sum_{j\in\mathcal{J}_1}\mathbb{P}_x^{(n)}(C_\theta^{(j)}) = 1$. Hence,

$$\begin{aligned} &\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| \;= \\ &\quad\overset{(19)}{=}\; 2\cdot\Big(1 - \sup_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\big(Q(C_\theta^{(j)}) - \gamma_j\big)\Big) \end{aligned}$$

For every $j \in \{1,\ldots,r\}$, identify $Q(C_\theta^{(j)})$ with the variable q_j in the linear program. Then, it follows from the definitions of $\mathcal{N}_\theta$ and $\mathbb{M}$ that

$$\sup_{(Q,\gamma)\in\mathbb{M}}\sum_{j\in\mathcal{J}_1}\big(Q(C_\theta^{(j)}) - \gamma_j\big) \;=\; \beta_\theta$$

and, therefore, $\inf_{Q\in\mathcal{N}_\theta}\|\mathbb{P}_x^{(n)} - Q\| = 2\cdot(1-\beta_\theta)$. □

[4] Confer [6, § 5.2] for more details on the stability of coherent upper previsions and the potential instability of the natural extension.

3.3 Implementation in the statistical programming language R

The minimum distance estimator is implemented in the (open source) statistical programming language R [10] and is publicly available as R-package "imprProbEst" [5]. In order to calculate the estimator, the program has to do the following steps:

1. for "some" $\theta \in \Theta$, (approximately) calculate the distance $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$, i.e.
   - discretize the sample space
   - solve the linear program given by (9)- (14)
2. choose that $\hat{\theta}$ which minimizes $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$

The inputs are the observations $x = (x_1, \ldots, x_n)$ and the imprecise model given by the (standardized) functions $f_j \in \mathcal{K}_\theta$ and the previsions $\overline{P}_\theta[f_j]$, $f_j \in \mathcal{K}_\theta$, for every $\theta \in \Theta$.

Note that we do not assume any condition of regularity for the map $\theta \mapsto \overline{P}_\theta$. Therefore, one might suppose that we have to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ for every $\theta \in \Theta$ in order to find the minimizing $\hat{\theta}$. Though this is possible since Θ is assumed to be finite here, such a proceeding is very cumbersome because the calculation of $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ is computationally costly. Fortunately, it usually suffices to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ only for very few elements of Θ: Put

$$t(\theta) \;=\; 2 \cdot \big(\max_{j \in \mathcal{I}_\theta} \mathbb{P}_x^{(n)}[h_\theta^{(j)}] - \overline{P}_\theta[f_j] - \varepsilon_\theta^{(j)} \big)$$

and $\Theta = \{\theta_1, \ldots, \theta_m\}$. Then, for every $\theta_l \in \Theta$,

$$\begin{aligned} &\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\| \;\geq \\ &\overset{(*)}{\geq} \max_{j \in \mathcal{I}_\theta} \mathbb{P}^{(n)}\big[f_j - (1 - f_j)\big] - \overline{Q}_{\theta_l}\big[f_j - (1 - f_j)\big] \\ &= \; 2 \cdot \big(\max_{j \in \mathcal{I}_\theta} \mathbb{P}^{(n)}[f_j] - \overline{Q}_{\theta_l}[f_j] \big) \overset{(**)}{\geq} t(\theta_l) \end{aligned}$$

where $(*)$ is valid since the standardization of f_j implies $\|f_j - (1 - f_j)\| = 1$, and $(**)$ follows from the definition of $t(\theta_l)$ and (6). Hence, the algorithm only has to calculate the subsequent value $\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\|$ if

$$t(\theta_l) \;\leq\; \min_{k \in \{1, \ldots, l-1\}} \|\mathbb{P}^{(n)} - \overline{Q}_{\theta_k}\| \tag{20}$$

is fulfilled. If (20) is not fulfilled, we do not have to calculate $\|\mathbb{P}^{(n)} - \overline{Q}_{\theta_l}\|$ because, in this case, it follows from the above calculation that θ_l is already known to be not a minimizer. The simulation studies described in Section 4 showed that, in this way, usually only a very small number of distances $\|\mathbb{P}^{(n)} - \overline{Q}_\theta\|$ has to be calculated.

4 A simulation study

4.1 Model 1: A first example

Model 1 is intended to demonstrate two aspects of the proposed estimator: Firstly, the estimator can really be calculated even for large numbers of observations. In the simulation study, the estimator is applied for sample sizes $n = 30$, $n = 100$, $n = 1000$, $n = 10000$. For each number of observations, the estimator is evaluated 500 times. Secondly, the estimator can provide good results even though it is developed for the rather large imprecise models given by finitely generated coherent upper previsions. In order to demonstrate this, the imprecise Model 1 contains a nice precise parametric model so that the estimator can be compared with a maximum likelihood estimator. While the maximum likelihood estimator is applied by using complete knowledge of the precise parametric model, our minimum distance estimator is only based on the knowledge of a large imprecise model. Since the simulated data exactly stem from the ideal parametric model, this is a rather unequal situation which favors the maximum likelihood estimator and, therefore, the maximum likelihood estimator should clearly beat our estimator. Nevertheless, the performance of our estimator appears to be almost as good as the one of the maximum likelihood estimator in the simulation study. In this way, it can be seen that going over to a large imprecise model does not necessarily mean to loose a lot of efficiency even if the ideal parametric model was precisely true.

Here is a detailed description of Model 1: The sample space is $(\mathcal{X}, \mathcal{B})$ where $\mathcal{X}$ is equal to $[0, 1]$ and $\mathcal{B}$ is the Borel-σ-algebra. The precise parametric model $(P_\theta)_{\theta \in \Theta}$ is given by $dP_\theta = p_\theta \, d\lambda$, $\theta \in \Theta := [-2, 2]$ where the Lebesgue-densities p_θ are

$$p_\theta(x) = 1 + \theta\big(x - 0.5\big) I_{[0,0.5]}(x) + \theta\big(0.75 - x\big) I_{(0.5,1]}(x)$$

for every $x \in [0, 1]$. Despite of this confusing formula, the densities p_θ are very simple and natural as can be seen from Figure 1. In order to define the imprecise model, the parameter set Θ is discretized as follows:

$$\Theta_0 := \big\{ \theta \in \Theta \,\big|\, \theta = -2 + 0.1k - 0.05\,,\; k \in \{1, \ldots, 40\} \big\}$$

That is, $\theta_0 \in \Theta_0$ corresponds to the interval $(\theta_0 - 0.05\,,\, \theta_0 + 0.05]$ with center θ_0. The imprecise model $(\overline{P}_\theta)_{\theta \in \Theta_0}$ is given by credal sets

$$\mathcal{M}_\theta \;= \big\{ Q_\theta \;\big|\; Q_\theta[f_j] \;\leq\; \overline{P}_\theta[f_j] \;\; \forall\, f_j \in \mathcal{K} \big\} \qquad \forall\, \theta \in \Theta_0$$

Here, $\mathcal{K}$ is the finite set $\mathcal{K} = \big\{ f_1, \ldots, f_{10} \big\}$ which consists of the (rather arbitrarily chosen) functions $f_j : [0, 1] \to \mathbb{R}$, $x \mapsto f_j(x)$ given by

$$f_1(x) \;=\; x\,, \qquad f_2(x) \;=\; 1 - x\,, \qquad f_3(x) \;=\; x^2\,,$$

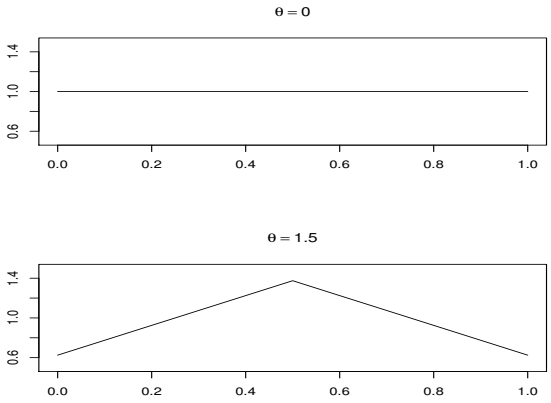

Figure 1: Graphs of p_θ for $\theta = 0$ (the uniform distribution) and $\theta = 1.5$ in Model 1

$$f_4(x) = x^3, \quad f_5(x) = I_{\left[\frac{1}{4},\frac{3}{4}\right]}(x), \quad f_6(x) = I_{\left[0,\frac{1}{4}\right]}(x),$$
$$f_7(x) \;=\; I_{\left[\frac{3}{4},1\right]}(x)\,, \qquad f_8(x) \;=\; \sqrt{x}\,,$$
$$f_9(x) \;=\; x + \tfrac{1}{2} I_{\left[\frac{1}{4},\frac{1}{2}\right]}(x)\,, \qquad f_{10}(x) \;=\; 4(x - x^2)$$

and the upper previsions on these functions are defined by

$$\overline{P}_{\theta_0}[f_j] \;=\; \sup_{\theta \in [\theta_0 - 0.05\,,\, \theta_0 + 0.05]} \int_0^1 f_j(x) p_\theta(x)\, \lambda(dx)$$

for every $j \in \{1, \ldots, 10\}$ and $\theta_0 \in \Theta_0$.

In the simulation study, the data $x_1, \ldots, x_n$ stem from the uniform distribution $P_0 = \mathrm{Unif}\big([0,1]\big)$. That is, $\theta = 0$ is the true parameter which has to be estimated.

For the estimation, the proposed minimum distance estimator and the maximum likelihood estimator

$$\hat{\theta}_{n,\mathrm{MaxLikelihood}}(x_1, \ldots, x_n) \;=\; \arg \max_{\theta \in [-2,2]} \prod_{i=1}^{n} p_\theta(x_i)$$

are applied. Note that – due to the discretization of Θ – our minimum distance estimator does not specify a precise value θ as an estimation but an interval $[\theta_0 - 0.05, \theta_0 + 0.05]$. In order to compare the results between both estimators, these intervals $[\theta_0 - 0.05, \theta_0 + 0.05]$ are recorded by their center θ_0.

Table 1 shows the empirical mean squared error (MSE)

$$\frac{1}{500} \sum_{j=1}^{500} \big(\hat{\theta}_n^{(j)} \;-\; 0\big)^2$$

of the estimations $\hat{\theta}_n^{(j)}$ calculated over all runs $j = 1, \ldots, 500$ for the proposed minimum distance estimator (MinDistance) and the classical maximum likelihood estimator (MaxLikelihood); these values are

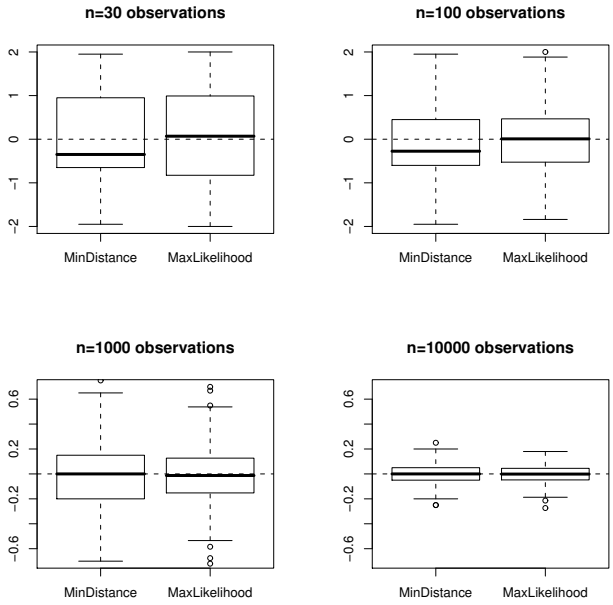

Figure 2: Boxplots of the estimations obtained in 500 runs for each number of observations in Model 1

n	MinDistance	MaxLikelihood
30	1.29943	1.35598
100	0.59675	0.49674
1000	0.06753	0.04692
10000	0.00711	0.00482

Table 1: Empirical mean squared error calculated over the estimations obtained in 500 runs for each number of observations in Model 1

similar for both estimators. Figure 2 shows the boxplots of the estimations. These results demonstrate that, in Model 1, the maximum likelihood estimator is not much better than the minimum distance estimator even though the unequal situation of Model 1 highly privilege the maximum likelihood estimator as explained above.

4.2 Model 2: Approximate Poisson distributions

In Model 2, the sample space is $(\mathbb{N}_0, 2^{\mathbb{N}_0})$ and it is assumed that the data "approximately" stem from a Poisson distribution $\mathrm{Poi}(\theta)$ where the parameter set is $\Theta = (0, 50]$. The parameter set is again discretized:

$$\Theta_0 := \big\{\theta \in \Theta \,\big|\, \theta = 0.1 + 0.05k\,,\; k \in \{0, \ldots, 998\}\big\}$$

The imprecise model $(\overline{P}_\theta)_{\theta \in \Theta_0}$ is given by credal sets

$$\mathcal{M}_\theta \;=\; \big\{Q_\theta \;\big|\; Q_\theta[f_\theta^{(j)}] \;\leq\; \overline{P}_\theta[f_\theta^{(j)}] \;\; \forall\, f_\theta^{(j)} \in \mathcal{K}_\theta\big\}$$

and $\mathcal{K}_\theta$ is the finite set $\mathcal{K}_\theta = \{f_\theta^{(1)}, \ldots, f_\theta^{(56)}\}$ which consists of the following functions:

$$f_\theta^{(j)} \;=\; I_{\{4(j-1)\,,\ldots,\,4j-1\}} \qquad \forall\, j \in \{1,\ldots,25\}$$

$$f_\theta^{(25+j)} \;=\; 1 - f_\theta^{(j)} \qquad \forall\, j \in \{1,\ldots,25\}$$

$$f_\theta^{(51)}(x) = \frac{x}{100} I_{\{0,\ldots,100\}}(x)\,, \quad f_\theta^{(52)} = 1 - f_\theta^{(51)}$$

$$f_\theta^{(53)}(x) = \Big(\frac{x}{100}\Big)^2 I_{\{0,\ldots,100\}}(x)\,, \quad f_\theta^{(54)} = 1 - f_\theta^{(53)}$$

$$f_\theta^{(55)} = I_{(\theta-1,\theta]}\,, \quad f_\theta^{(56)} = 1 - f_\theta^{(55)}$$

The upper previsions on these functions are defined by

$$\overline{P}_{\theta_0}[f_{\theta_0}^{(j)}] = (1-r) \sup_{\theta\in[\theta_0-0.025\,,\,\theta_0+0.025]} \mathrm{Pois}(\theta)[f_{\theta_0}^{(j)}] + r$$

for every $j \in \{1,\ldots,56\}$ and $\theta_0 \in \Theta_0$. In the simulation study, we put $r = 0.01$. [5]

For the estimation, our minimum distance estimator and the maximum likelihood estimator

$$(x_1,\ldots,x_n) \;\mapsto\; \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \mathrm{Pois}(\theta)\big(\{x_i\}\big)\,,$$

are applied. The simulation study consists of 500 runs with different sample sizes $n = 20$, $n = 100$ and $n = 250$. The real distribution which generates the data is equal to

$$P_0 \;=\; (1-c)\mathrm{Pois}(12.5) + c\,\mathrm{Unif}\big(\{0,\ldots,100\}\big)$$

for $c = 0$, $c = 0.01$ and $c = 0.1$ where $c = 0$ is the "ideal situation" and $c \in \{0.01; 0.1\}$ stands for (very) small deviations of the "ideal situation". Figure 3 shows the boxplots for $c = 0$ and $c = 0.01$ (only sample sizes $n = 20$ and $n = 250$); Figure 4 shows the boxplots for $c = 0.1$. Table 2 gives the empirical mean squared errors. In the ideal situation, the maximum likelihood estimator is only slightly better than the (imprecise probability) minimum distance estimator. However, very small deviations from the ideal situation are enough so that the minimum distance estimator beats the maximum likelihood estimator. In particular, this is true even for $c = 0.01$ and $n = 20$ though, in this case, most samples $x_1,\ldots,x_{20}$ will not contain any "wrong" observation – i.e. will be "ideal".

[5] Though this looks very similar to contamination neighborhoods (which are quite common in robust statistics), these upper previsions lead to much bigger credal sets than contamination neighborhoods. This is because, here, the definition of the upper previsions only involves a finite number of functions $f_{\theta_0}^{(j)}$, while the definition of contamination neighborhoods involves all functions $f \in \mathcal{L}_\infty(\mathbb{N}_0, 2^{\mathbb{N}_0})$.

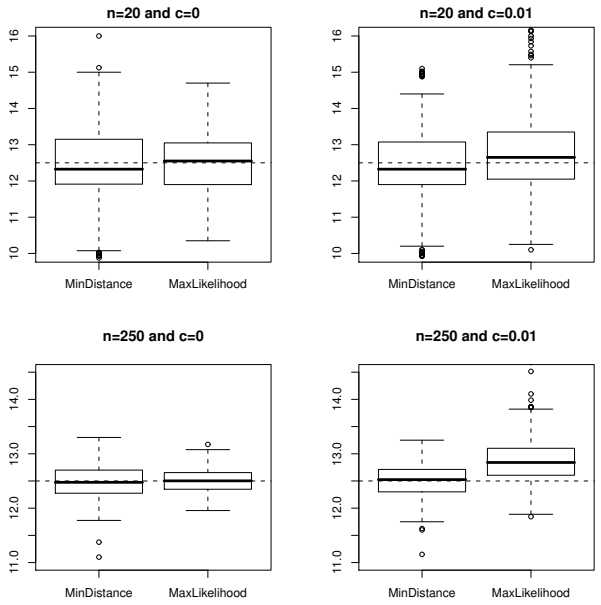

Figure 3: boxplots of the estimations obtained in 500 runs for sample size $n = 20$ and $n = 250$ in Model 2

$n = 20$	c=0	c=0.01	c=0.10
MinDistance	1.22	1.15	1.20
MaxLikelihood	0.65	1.88	24.99
$n = 100$	c=0	c=0.01	c=0.10
MinDistance	0.24	0.29	0.22
MaxLikelihood	0.12	0.52	16.24
$n = 250$	c=0	c=0.01	c=0.10
MinDistance	0.10	0.10	0.12
MaxLikelihood	0.05	0.29	15.27

Table 2: Empirical mean squared error calculated over the estimations obtained in 500 runs in Model 2

5 Application on a real data set

Finally, the estimator is applied on a real data set for linear regression. The data set consists of 200 data

$$x_i = (y_i, z_i) \;\in\; [0,\infty) \times [160,\infty)\,, \quad i \in \{1,\ldots,200\}$$

from the *National Health and Nutrition Examination Survey* (NHANES) from the years 2005–2006 which records the health and nutritional status of adults and children in the United States of America. [6] Every observation x_i corresponds to a person where y_i specifies the person's weight (in kilograms) and z_i specifies the person's height (in centimeters). [7] The following rela-

[6] The data are publicly available in the Internet on the website of the *Centers for Disease Control and Prevention*: `http://www.cdc.gov/nchs/nhanes.htm`

[7] The original data set contains many additional variables which have been omitted here. The 200 persons whose data

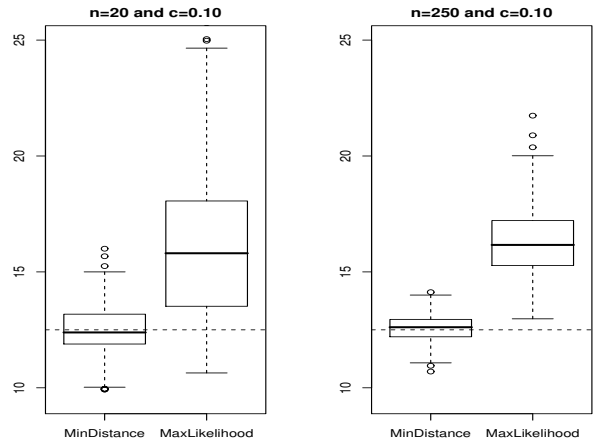

Figure 4: Boxplots of the estimations obtained in 500 runs for sample size $n = 20$ and $n = 250$ in Model 2

tion is assumed

$$y_i \;=\; \theta_1 + \theta_2(z_i - 160) + \varepsilon_i\,, \qquad i \in \{1, \ldots, 200\}$$

for persons with a height of at least 160 cm. Accordingly, only persons have been considered who fulfill this condition. The set of possible parameters is bounded and may be given by $\Theta = [25, 100] \times [0.5, 1.5]$ In order to apply the minimum distance estimator, Θ is again discretized:

$$\Theta_0 \;=\; \left\{ (\theta_1, \theta_2) \;\middle|\; \begin{array}{l} \theta_1 \in \{25, 26, \ldots, 100\}\,, \\ \theta_2 \in \{0.5, 0.55, 0.6, \ldots, 1.45, 1.5\} \end{array} \right\}$$

As an imprecise distribution of the i.i.d errors ε_i, we take the coherent upper prevision $\overline{E}_\sigma$, which is based on the normal distribution $\mathcal{N}(0, \sigma^2)$ in the following way: Take $h_0 = I_{(-\infty, -20]}$,

$$h_1 = I_{(-20, -15]},\;\; h_2 = I_{(-15, -10]},\;\; \ldots\;,\;\; h_{12} = I_{(35\,,\,40]}$$

$$h_{12+j} \;=\; 1 - h_j \qquad \forall\, j \in \{1, \ldots, 12\}$$

and $h_{25} \;=\; I_{(40, \infty)}$. Put $S_0 = \{1, 2, \ldots, 30\}$. The error distribution $\overline{E}_{\sigma_0}$ is assumed to be the coherent upper prevision whose credal set consists of all probability charges E on $\mathbb{R}$ such that for every $j \in \{0, \ldots, 25\}$

$$E[h_j] \;\leq\; (1 - r) \sup_\sigma \mathcal{N}(0, \sigma^2)[h_j] + r \sup h_j I_{(0, \infty)}$$

where the supremum is over $\sigma \in [\sigma_0 - 0.5, \sigma_0 + 0.5]$, $r = 0.05$ and $\sigma_0 \in S$. (Roughly speaking, this means that E is "approximately" a normal distribution but overweight is more likely than underweight. Then, the imprecise model is given by

$$\overline{P}_{\theta_0, \sigma_0} \;=\; \overline{S}_{\sigma_0}[f_{\theta_0}^{(j)}] \qquad \forall\, j \in \{0, \ldots, 25\}$$

are analyzed here have been randomly picked out of the data from the National Health and Nutrition Examination Survey.

	MinDistance	LeastSquares
θ_1	59	67.8
θ_2	0.95	1.03
σ_0	17	—

Table 3: Results of the estimators for the real data set NHANES; the nuisance parameter σ_0 is only estimated by the minimum distance estimator

where $f_{\theta_0}^{(j)} : (y, z) \mapsto h_j\big(y - \theta_1 - \theta_2(z_i - 160)\big)$. The parameter of interest is $\theta_0 = (\theta_1, \theta_2)$; σ_0 is a nuisance parameter.

Our minimum distance estimator is compared to the classical least-squares estimator. The results are given in Table 3, and Figure 5 illustrates the corresponding regression lines. By definition, the least-squares estimator fits the data best with respect to the squared residuals. However, this also leads to the fact that this estimator is sensitive to outliers. This effect is also visible in Figure 5: The least-squares estimator seems to be more influenced by a relatively small number of considerably overweight persons than the minimal distance estimator.

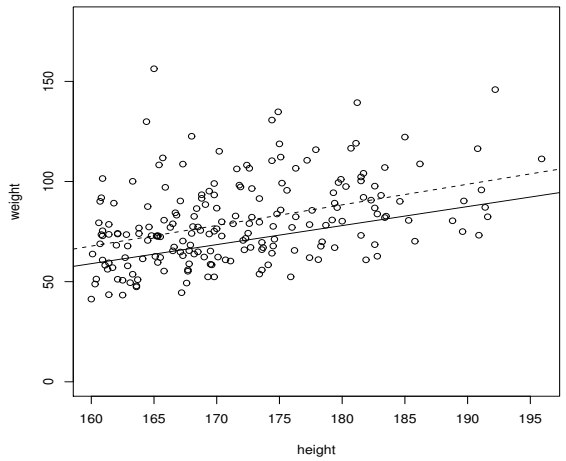

Figure 5: Regression lines for the real data set NHANES obtained by the minimum distance estimator (solid line) and by the least-squares estimator (dashed line)

6 Concluding remarks

The present article considers estimating a parameter in an imprecise probability model – a topic which has hardly been considered explicitly within the theory of coherent upper previsions so far. In this setup, a minimum distance estimator is presented and an algorithm for calculating the estimator is given which is based on

linear programming. The applicability of the estimator is verified by a simulation study and on a real data set. In particular, the simulation study shows that the proposed estimator can even be used for large sample sizes and may, in fact, lead to good results in realistic situations. This meets objections that imprecise probabilities could not be used for practical purposes. The estimator has been programmed in R and has already been made publicly available as (open source) R package "imprProbEst"; cf. [5]. However, future research should also develop alternative estimators so that the proposed minimum distance estimator can be compared to other estimators under imprecise probabilities.

Acknowledgments

I would like to thank Thomas Augustin for valuable suggestions and Matthias Kohl for his help with programming in R. In addition, I thank reviewers for helpful comments.

References

[1] T. Augustin. *Optimale Tests bei Intervallwahrscheinlichkeit*. Vandenhoeck & Ruprecht, Göttingen, 1998.

[2] T. Augustin. Neyman-Pearson testing under interval probability by globally least favorable pairs – reviewing Huber-Strassen theory and extending it to general interval probability. *Journal of Statistical Planning and Inference*, 105:149–173, 2002.

[3] M. Bickis and U. Bickis. Predicting the next pandemic: An exercise in imprecise hazards. In G. de Cooman, J. Vejnarová, and M. Zaffalon, editors, *Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications*, pages 41–46. SIPTA, Prague, 2007.

[4] N. Dunford and J.T. Schwartz. *Linear operators. I. General theory*. Wiley-Interscience Publishers, New York, 1958.

[5] R. Hable. *imprProbEst: Minimum distance estimation in an imprecise probability model*, 2008. Contributed R-Package on CRAN, Version 1.0, 2008-10-23; maintainer Hable, R.

[6] R. Hable. *Data-based decisions under complex uncertainty*. PhD thesis, Ludwig-Maximilians-Universität (LMU) Munich, 2009. http://edoc.ub.uni-muenchen.de/9874/.

[7] M. Hutter. Practical robust estimators for the imprecise Dirichlet model. *International Journal of Approximate Reasoning*, 50:231–242, 2009.

[8] E. Kriegler and H. Held. Climate projections for the 21st century using random sets. In J.M. Bernard, T. Seidenfeld, and M. Zaffalon, editors, *ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, Lugano*, pages 345–360. Carleton Scientific, Waterloo, 2003.

[9] E. Quaeghebeur and G. de Cooman. Imprecise probability models for inference in exponential families. In F.G. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, Pittsburg*, pages 287–296. SIPTA, Manno, 2005.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[11] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998.

[12] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman & Hall, London, 1991.

[13] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):3–57, 1996. With discussion and a reply by the author.

[14] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice: Special Issue on Imprecision*, 3:255–271, 2009.

[15] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24:149–170, 2000.