Journal of Statistical Planning and Inference 140 (2010) 461–479

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Minimum distance estimation in imprecise probability models

Robert Hable

University of Bayreuth, Department of Mathematics, 95440 Bayreuth, Germany

ARTICLE INFO

Article history:
Received 23 January 2009
Received in revised form
24 July 2009
Accepted 24 July 2009
Available online 3 August 2009

MSC:
62F10
62G05
62G20
62G30
62G35

Keywords:
Robust statistics
Imprecise probabilities
Coherent lower previsions
Minimum distance estimator
Empirical measure

ABSTRACT

The present article considers estimating a parameter θ in an imprecise probability model $(\overline{P}_\theta)_{\theta\in\Theta}$. This model consists of coherent upper previsions $\overline{P}_\theta$ which are given by finite numbers of constraints on expectations. A minimum distance estimator is defined in this case and its asymptotic properties are investigated. It is shown that the minimum distance can be approximately calculated by discretizing the sample space. Finally, the estimator is applied in a simulation study and on a real data set.

1. Introduction

1.1. Motivation

In statistical evaluations, assuming that the data precisely stem from one of a very small couple of smooth parametric models (e.g. normal distributions) is extremely popular. However, such assumptions are very strong so that they can hardly be justified well and, in this case, statistical conclusions are in danger of being arbitrary. These facts are well known for several decades and have lead to the raise of robust statistics. For a motivation of robust statistics, see e.g. Tukey (1960), Huber (1965), Huber (1981, Section 1.1), Hampel et al. (1986, Section 1.2), Marazzi (1993, Introduction), Huber (1997, Section 1), and Kohl (2005, Introduction). In robust statistics, it is often assumed that the data are approximately distributed according to an ideal, smooth parametric model (see e.g. the definition of robust statistics suggested in Hampel et al., 1986, p. 7). Though assuming a known ideal parametric model can often be justified e.g. by use of the central limit, this is not always possible. In order to deal also with more general kinds of ambiguity, the theory of imprecise probabilities has recently been developed, among others, by Walley (1991) (coherent lower previsions) and Weichselberger (2001) (interval probabilities).

The present article deals with the concept of coherent lower previsions—or, equivalently, coherent upper previsions—due to Walley. Such imprecise probabilities may be interpreted in various ways. The so-called sensitivity analyst's point of view

E-mail address: Robert.Hable@uni-bayreuth.de.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.07.025

considers upper previsions $\overline{P}(A)$ and $\overline{P}[f]$ as upper bounds on the true probability $P(A)$ or on the true expectation $\mathbb{E}_P f$, respectively. That is, in contrast to classical statistics, practitioners are not expected to precisely specify correct probabilities; they only have to give lower and upper bounds on probabilities and expectations. Even more, it is assumed in the present article that a practitioner is only able to give lower/upper bounds for a *finite* number of events A or real functions f. Though this assumption is restrictive, it is very often fulfilled—especially in real applications. In particular, this is true for expert systems since, there, it is a natural proceeding to ask some experts about their prevision (or expectation) on some specific events, experiments, gambles, assets, etc. and this can only be done for a finite number of such objects.

While hypothesis testing under imprecise probabilities has been extensively studied—e.g. in Augustin (1998) on base of the Huber–Strassen theory (see also Augustin, 2002, for a review of the work following Huber and Strassen, 1973), 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. Walley (1996), Quaeghebeur and de Cooman (2005), Hutter (2009) and Walter and Augustin (2009). In addition, there are a few articles which address very special applications, e.g. Kriegler and Held (2003) (climate projections) and Bickis and Bickis (2007) (prediction of the next influenza pandemic). However, investigations of general frequentist estimation of a parameter using coherent lower/upper previsions are still missing.

In short, the present article considers the following setup: we are faced with a random sample $x_1, \ldots, x_n$ but we do not assume a known precise statistical model $(P_\theta)_{\theta\in\Theta}$. Instead, we only assume knowledge of an imprecise statistical model $(\overline{P}_\theta)_{\theta\in\Theta}$ which consists of coherent upper previsions. The present article develops a minimum distance estimator which is based on the following simple idea: the data are used to build the empirical measure; then, the minimum distance estimator is that $\hat{\theta} \in \Theta$ such that $\mathbb{P}^{(n)}$ lies next to the credal set $\mathscr{M}_{\hat{\theta}}$ of $\overline{P}_{\hat{\theta}}$. The credal set $\mathscr{M}_\theta$ of a coherent upper prevision $\overline{P}_\theta$ is the set of all precise probabilities which are in accordance with the bounds given by $\overline{P}_\theta$. Minimum distance estimators have already attracted attention in robust statistics (see e.g. Parr and Schucany, 1980; Millar, 1981; Donoho and Liu, 1988; Rieder, 1994, Section 6; Öztürk and Hettmansperger, 1998) but our setup considerably differs from those ones usually used in robust frequentist estimation. In particular, our credal sets are not in accordance with common neighborhood systems.

A rigorous definition of credal sets is given in Section 1.2 which recalls the basic setup and some notation concerning coherent upper previsions. Section 2 develops the minimum distance estimator: Section 2.1 provides a mathematical description of the imprecise statistical model and lists the assumptions; Section 2.2 contains the exact definition of the minimum distance estimator. Asymptotic properties of the estimator are investigated in Section 3; in particular, it is proven that the estimator is consistent. A short glance at the definition of the estimator shows that concrete calculations are a matter of its own—just as usual within the theory of imprecise probabilities. In Section 4, it is proven that distances between the empirical measures and coherent upper previsions can be approximately calculated by going over to a suitable finite sample space which enables calculations by means of linear programming; finally, the minimum distance estimator is applied in a small simulation study and on a real data set in Section 5. Section 6 contains some concluding remarks.

1.2. Setup and notation

Let $\mathscr{X}$ be a set with σ-algebra $\mathscr{B}$. Then, $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ denotes the set of all bounded, $\mathscr{B}$-measurable real functions $f : \mathscr{X} \to \mathbb{R}$. The supremum norm on $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ is denoted by $\|f\| = \sup_{x\in\mathscr{X}}|f(x)|$. The set of all bounded, finitely additive, signed measures is denoted by $\mathrm{ba}(\mathscr{X},\mathscr{B})$ and can be identified with the dual space of $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$; cf. Dunford and Schwartz (1958, Theorem IV.5.1). Furthermore, $\mathrm{ca}(\mathscr{X},\mathscr{B})$ denotes the set of all bounded (σ-additive) signed measures. Finally, $\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ and $\mathrm{ca}_1^+(\mathscr{X},\mathscr{B})$ denote the set of all finitely additive probability measures and the set of all probability measures, respectively. For every $f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$, integrals are defined with respect to any $\mu \in \mathrm{ba}(\mathscr{X},\mathscr{B})$—confer e.g. Dunford and Schwartz (1958, Section III) or Bhaskara Rao and Bhaskara Rao (1983, Section 4). This integral is denoted by

$$\mu[f] := \int f d\mu.$$

According to Dunford and Schwartz (1958, Section V.3), the $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$-topology on $\mathrm{ba}(\mathscr{X},\mathscr{B})$ is the weakest topology on $\mathrm{ba}(\mathscr{X},\mathscr{B})$ such that, for every $f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$, the map $\mu \mapsto \mu[f]$ is continuous.

In accordance with Walley (1991, Section 2.5.1), a coherent upper prevision on $(\mathscr{X},\mathscr{B})$ is a map

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

such that there is a (non-empty) set $\mathscr{V} \subset \mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ and $\overline{P}[f] = \sup_{P\in\mathscr{V}} P[f]$ for every $f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$; cf. also Walley (1991, Section 3.3.3) and Hable (2009a, Section 2.3). The non-empty set

$$\mathscr{M} := \left\{ P \in \mathrm{ba}_1^+(\mathscr{X},\mathscr{B}) \,\middle|\, P[f] \le \overline{P}[f]\ \forall f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B}) \right\}$$

is called *credal set of* $\overline{P}$ then. Of course, we have $\sup_{P\in\mathscr{M}} P[f] = \overline{P}[f]$. Every coherent upper prevision corresponds to a coherent lower prevision defined by $\underline{P}[f] = \inf_{P\in\mathscr{M}} P[f] = -\overline{P}[-f]$. A subset $\mathscr{M} \subset \mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ is the credal set of a coherent upper prevision if and only if it is convex and compact with respect to the $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$-topology; confer Walley (1991, Theorem 3.6.1) and Hable (2009a, Corollary 2.16).

A coherent upper prevision $\overline{P}$ is called *finitely generated* if there is a finite set $\{f_1,\dots,f_s\} \subset \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ such that the credal set of $\overline{P}$ is equal to

$$\mathscr{M} = \left\{ P \in \mathrm{ba}_1^+(\mathscr{X},\mathscr{B}) \,\middle|\, P[f_j] \le \overline{P}[f_j] \ \ \forall j \in \{1,\dots,s\} \right\}. \tag{1}$$

According to Hable (2009a, Proposition 2.13), $\overline{P}$ is the natural extension of a coherent upper prevision on $\{f_1,\dots,f_s\} \subset \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$. 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 to (1), $\overline{P}[f_j] > \underline{P}[f_j]\,\forall j \in \{1,\dots,s\}$ where $\underline{P}$ denotes the coherent lower prevision corresponding to $\overline{P}$. So, the credal set of a regular, finitely generated coherent upper prevision is not a singleton.

2. A minimum distance estimator for imprecise models

2.1. Assumptions

In order to define the estimator in a mathematical rigorous way, the following fixings and assumptions are made. These are valid throughout the rest of the article:

$(\Omega,\mathscr{A})$ and $(\mathscr{X},\mathscr{B})$ are measurable spaces, Θ is an index set. Let U_0 be a probability measure on $(\Omega,\mathscr{A})$. The observations $x_1,\dots,x_n$ are modeled via random variables

$$X_i : (\Omega,\mathscr{A},U_0) \longrightarrow (\mathscr{X},\mathscr{B}), \quad i \in \{1,\dots,n\}.$$

It is assumed that the random variables $X_1,\dots,X_n$ are independent identically distributed according to a probability measure $P_0 = X_i(U_0)$. Next, $(\overline{P}_\theta)_{\theta\in\Theta}$ is an imprecise model which consists of coherent upper previsions on $(\mathscr{X},\mathscr{B})$. For every $\theta \in \theta$, let $\mathscr{M}_\theta$ be the credal set of $\overline{P}_\theta$. It is only assumed that the true probability measure P_0 is contained in $\mathscr{M}_{\theta_0}$ where $\theta_0 \in \Theta$ is the unknown true parameter. The task is to estimate θ_0. This approach exactly corresponds to the use of the well-known type-2 product of coherent upper previsions (Walley, 1991, Section 9.3.5) but is formulated in terms of random variables and image measures in order to deal with more elaborated stochastic concepts such as empirical processes. The type-2 product of coherent upper previsions is consistent with a strict sensitivity analyst's point of view on imprecise probabilities. Accordingly, an analog to the type-2 product of coherent upper previsions is also common in robust statistics where credal sets are replaced by neighborhoods about parametric models. The use of the type-2 product makes it possible to use arguments from classical asymptotic statistics. This is fortunate because a theory of asymptotic statistics under imprecise probabilities is still at an early stage. A starting point of such an asymptotic theory is given by laws of large numbers for imprecise probabilities which have recently been developed e.g. in De Cooman and Miranda (2008).

The following fundamental assumptions on the coherent upper previsions are made: there is a *finite* subset $\mathscr{K} = \{f_1,\dots,f_s\} \subset \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ such that

$$\mathscr{M}_\theta = \left\{ P_\theta \in \mathrm{ba}_1^+(\mathscr{X},\mathscr{B}) \,\middle|\, P_\theta[f] \le \overline{P}_\theta[f] \ \ \forall f \in \mathscr{K} \right\} \tag{2}$$

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

$$\overline{P}_\theta[f] - \underline{P}_\theta[f] > 0 \quad \forall f \in \mathscr{K}, \tag{3}$$

where $\underline{P}_\theta$ is the corresponding lower coherent prevision. In particular, each $\overline{P}_\theta$ is a regular, finitely generated coherent upper previsions.

Since U_0 is assumed to be a (σ-additive) probability measure, the image $P_0 = X_i(U_0)$ is a (σ-additive) probability measure as well. However, σ-additivity is commonly dropped in the theory of imprecise probabilities. Accordingly, $\mathscr{M}_{\theta_0}$ may also contain probability charges which are not σ-additive and, therefore, the above assumption needs some justification—the more so as, in general, it can completely counteract the imprecision of the probability assessments since $\mathscr{M}_{\theta_0} \cap \mathrm{ba}_1^+(\mathscr{X},\mathscr{B}) = \{P_0\}$ is possible. However, Theorem 2.1 shows that this cannot happen in case of regular, finitely generated coherent upper previsions—even more, these previsions can be represented by (σ-additive) probability measures.

Theorem 2.1. *Let $(\mathscr{X},\mathscr{B})$ be a measurable space and let $\overline{P}$ be a regular, finitely generated coherent upper prevision on $(\mathscr{X},\mathscr{B})$ with credal set $\mathscr{M}$. Then,*

$$\overline{P}[f] = \sup_{P\in\mathscr{P}} P[f] \quad \forall f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B}) \tag{4}$$

for $\mathscr{P} := \mathscr{M} \cap \mathrm{ca}_1^+(\mathscr{X},\mathscr{B})$.

Proof. Fix any $P_0 \in \mathscr{M}$, any $f_0 \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})\backslash\{0\}$ and any $\varepsilon > 0$. Then, in order to prove (4), it is enough to show that there is a $P_c \in \mathscr{P}$ such that

$$P_c[f_0] > P_0[f_0] - \varepsilon. \tag{5}$$

According to the assumptions, there is a set $\mathscr{K}=\{f_1,\ldots,f_s\}$ such that

$$\mathscr{M}=\left\{P\in \mathrm{ba}_1^+(\mathscr{X},\mathscr{B})\,\middle|\, P[f]\le \overline{P}[f]\ \ \forall f\in\mathscr{K}\right\}$$

and

$$\forall j\in\{1,\ldots,s\}\quad \exists P_j\in\mathscr{M}\quad \text{such that } P_j[f_j]<\overline{P}[f_j]. \tag{6}$$

Put

$$P_\alpha=(1-\alpha)P_0+\frac{\alpha}{s}\sum_{j=1}^{s}P_j$$

for $\alpha=\min\left\{\varepsilon(4\|f_0\|)^{-1},1\right\}$. Of course, convexity of $\mathscr{M}$ implies $P_\alpha\in\mathscr{M}$ but, even more, a simple calculation shows that $P_\alpha[f_i]<\overline{P}[f_j]$ for every $j\in\{1,\ldots,s\}$ and, therefore,

$$\varepsilon_0:=\min\left\{\frac{\varepsilon}{2},\overline{P}[f_1]-P_\alpha[f_1],\ldots,\overline{P}[f_s]-P_\alpha[f_s]\right\}>0.$$

Furthermore, the definition of P_α implies

$$\left|P_\alpha[f_0]-P_0[f_0]\right|\ \le 2\alpha\|f_0\|\le\frac{\varepsilon}{2}. \tag{7}$$

For every $j\in\{0,1,\ldots,s\}$, put

$$\Lambda_j:\mathrm{ba}(\mathscr{X},\mathscr{B})\to\mathbb{R},\quad \mu\mapsto\mu[f_j].$$

Since these maps are $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ - continuous,

$$B_0:=\bigcap_{j=1}^{s}\Lambda_j^{-1}\left((-\infty,P_\alpha[f_j]+\varepsilon_0)\right)\cap\Lambda_0^{-1}\left((P_\alpha[f_0]-\varepsilon_0,\infty)\right)$$

is an $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$-open neighborhood of P_α. This implies the existence of some $P_c\in B_0\cap\mathrm{ca}_1^+(\mathscr{X},\mathscr{B})$ because $\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ is the $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$-closure of $\mathrm{ca}_1^+(\mathscr{X},\mathscr{B})$ (see e.g. Hable, 2009a, Theorem 2.11). Then, it follows from the definition of B_0 that $P_c\in\mathscr{M}\cap\mathrm{ca}_1^+(\mathscr{X},\mathscr{B})=\mathscr{P}$ and that

$$P_c[f_0]>P_\alpha[f_0]-\varepsilon_0\overset{(7)}{\ge}P_0[f_0]-\frac{\varepsilon}{2}-\frac{\varepsilon}{2}=P_0[f_0]-\varepsilon$$

i.e. (5) is fulfilled, and this proves (4). $\square$

In particular, Theorem 2.1 implies that every regular, finitely generated coherent upper prevision is an upper expectation in the sense of Buja (1984).

2.2. Definition of the minimum distance estimator

We are faced with a random sample $x_1,\ldots,x_n$ from a precise distribution P_0 which is unknown. It is only known that P_0 is contained in a credal set $\mathscr{M}_{\theta_0}$ which belongs to a family of credal sets $(\mathscr{M}_\theta)_{\theta\in\Theta}$. The true parameter θ_0 is also unknown and should be estimated. As stated above, the idea of the presented minimum distance estimator is very simple: the data $x_1,\ldots,x_n$ are used to build the empirical measure $\mathbb{P}^{(n)}$. Then, the minimum distance estimator is that $\hat\theta\in\Theta$ such that $\mathbb{P}^{(n)}$ lies next to $\mathscr{M}_{\hat\theta}$. That is, we calculate the distance between $\mathbb{P}^{(n)}$ and $\mathscr{M}_\theta$ for every $\theta\in\Theta$ and pick that $\hat\theta$ where the distance is minimal. Now, the minimum distance estimator can be mathematically rigorously defined by use of the fixings and assumptions given in Section 2.1.

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

$$\mathbb{P}^{(n)}:\Omega\to\mathrm{ba}_1^+(\mathscr{X},\mathscr{B}),\quad \omega\mapsto\mathbb{P}_\omega^{(n)}=\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i(\omega)},$$

where δ_x denotes the Dirac measure in $x\in\mathscr{X}$. Note that

$$\mathbb{P}^{(n)}[f]:\Omega\to\mathbb{R},\quad \omega\mapsto\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i(\omega)}[f]=\frac{1}{n}\sum_{i=1}^{n}f(X_i(\omega))$$

is a (bounded) random variable for every $f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ and

$$\Omega \times \mathscr{B} \to \mathbb{R}, \quad (\omega, B) \mapsto \mathbb{P}_\omega^{(n)}[I_B]$$

is a Markov kernel. The following notation will also be used: for every $x = (x_1, \dots, x_n) \in \mathscr{X}^n$, the probability measure on $(\mathscr{X},\mathscr{B})$ defined by

$$\mathbb{P}_x^{(n)}[f] := \frac{1}{n}\sum_{i=1}^{n} f(x_i) \quad \forall f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})$$

is denoted by $\mathbb{P}_x^{(n)}$.

Next, we have to choose a suitable notion of "distance" between a measure P_0 and a coherent upper prevision $\overline{P}$ on $(\mathscr{X},\mathscr{B})$. Appropriately to the sensitivity analyst's point of view, the distance is defined to be $\inf_{P\in\mathscr{M}} d(P_0, P)$ where d is a suitable metric on $\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ and $\mathscr{M}$ is the credal set of $\overline{P}$.

Since bounded charges $\mu \in \mathrm{ba}(\mathscr{X},\mathscr{B})$ are mainly regarded as bounded linear operators on $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$ within the theory of imprecise probabilities, it seems to be most natural to choose the operator norm for d; that is,

$$d(P_0, P) = \|P_0 - P\| = \sup_{f\in\mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{|P_0[f] - P[f]|}{\|f\|}$$

and we put

$$\|P_0 - \overline{P}\| := \inf_{P\in\mathscr{M}} \|P_0 - P\|. \tag{8}$$

A map $\hat{\theta}_n : \mathscr{X}^n \to \Theta$ is called *minimum distance estimator* if

$$\|\mathbb{P}_x^{(n)} - \overline{P}_{\hat{\theta}_n(x)}\| = \inf_{\theta\in\Theta} \|\mathbb{P}_x^{(n)} - \overline{P}_\theta\| \quad \forall x \in \mathscr{X}.$$

By a slight abuse of notation, the map $\hat{\theta}_n \circ (X_1, \dots, X_n) : \Omega \to \Theta$ will also be denoted by $\hat{\theta}_n$ in the following.

According to Dunford and Schwartz (1958, Theorem IV.5.1), the operator norm in $\mathrm{ba}(\mathscr{X},\mathscr{B})$ is equal to the total variation. Therefore, the minimum distance estimator is based on the total variation norm. Though the total variation norm has some annoying properties in conjunction with the empirical measure in classical statistics, these problems completely disappear in the above developed setup based on imprecise probabilities, as will be seen in the following section.

Note that, in general, estimators such as maximum likelihood estimators or minimum distance estimators need neither be unique nor exist. Just as in case of precise probabilities, these properties depend on regularity assumptions about the parametrization $\theta \mapsto \overline{P}_\theta$. Minimum distance estimators always exist if Θ is finite. If a minimum distance estimator does not exist, it is possible to consider ε_n—versions of minimum distance estimators where

$$\|\mathbb{P}_x^{(n)} - \overline{P}_{\hat{\theta}_n(x)}\| \leq \inf_{\theta\in\Theta} \|\mathbb{P}_x^{(n)} - \overline{P}_\theta\| + \varepsilon_n \quad \forall x \in \mathscr{X}$$

and $\varepsilon_n \searrow 0$ for $n \to \infty$.

3. Asymptotic properties of the estimator

The present section investigates asymptotic properties of the minimum distance estimator; the fixings and assumptions of Section 2.1 are still valid. As already mentioned before, the use of the total variation norm together with the empirical measure is not unproblematic in classical statistics: several distances d provide the desirable property that

$$d(\mathbb{P}^{(n)}, P) \underset{n\to\infty}{\longrightarrow} 0 \quad P\text{-a.s.} \tag{9}$$

but this is not necessarily true for the total variation norm. If, for example, P is equal to the standard normal distribution, we have

$$\|\mathbb{P}_x^{(n)} - P\| = 2 \quad \forall x \in \mathscr{X}^n \quad \forall n \in \mathbb{N}$$

which is the worst possible violation of (9). However, Theorem 3.1 states that this annoying difficulty totally disappears in the imprecise probability setup summarized in Section 2.1. If we replace P by a regular, finitely generated coherent upper prevision $\overline{P}$, we get

$$\|\mathbb{P}^{(n)} - \overline{P}\| \underset{n\to\infty}{\longrightarrow} 0 \quad P\text{-a.s.}^* \tag{10}$$

for every probability measure P in the credal set $\mathscr{M}$ of $\overline{P}$. In (10), writing a.s.* instead of a.s. indicates that there may be some problems concerning measurability because, in general, measurability of the map

$$x \mapsto \|\mathbb{P}_x^{(n)} - \overline{P}\| = \inf_{P\in\mathscr{M}} \sup_{f\in\mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{|\mathbb{P}_x^{(n)}[f] - P[f]|}{\|f\|}$$

will not be available. In order to stay mathematically rigorous, we have to use the more elaborated setup based on random variables and image measures. In doing so, (10) more precisely, means that

$$\|\mathbb{P}_{\omega}^{(n)} - \overline{P}\| \underset{n\to\infty}{\longrightarrow} 0 \quad U_0(d\omega)\text{-a.s.}^* \tag{11}$$

if $X_i(U_0) = P_0$ is a member of the credal set $\mathscr{M}$ of $\overline{P}$. Here, $\omega \mapsto \|\mathbb{P}_{\omega}^{(n)} - \overline{P}\|$ denotes the map

$$\Omega \to \mathbb{R}, \quad \omega \mapsto \inf_{P\in\mathscr{M}} \sup_{f\in\mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i(\omega)) - P[f]\right|}{\|f\|}.$$

Taking the measurability issues indicated by the asterisk in a.e.* into account, (11) precisely means: there is a sequence of $\mathscr{A}/\mathbb{R}$-measurable random variables

$$\Delta_n : \Omega \to \mathbb{R}, \quad \omega \mapsto \Delta_n(\omega)$$

such that

$$\|\mathbb{P}_{\omega}^{(n)} - \overline{P}\| \le \Delta_n(\omega) \quad \forall\, \omega \in \Omega \quad \forall\, n \in \mathbb{N}$$

and

$$\Delta_n \underset{n\to\infty}{\longrightarrow} 0 \quad U_0\text{-a.s.}$$

Confer e.g. van der Vaart (1998, Section 18) or van der Vaart and Wellner (1996, Section 1.9) for this definition of almost sure convergence of unmeasurable maps.

Now, the already pronounced theorem can be formulated:

Theorem 3.1. *Let $(\Omega, \mathscr{A}, U_0)$ be a probability space, $(\mathscr{X}, \mathscr{B})$ a measurable space, and let*

$$X_i : (\Omega, \mathscr{A}, U_0) \longrightarrow (\mathscr{X}, \mathscr{B}), \quad i \in \{1, \ldots, n\}$$

be random variables which are independent identically distributed according to $P_0 = X_i(U_0)$. Let $\overline{P}$ be a regular, finitely generated coherent upper prevision with credal set $\mathscr{M}$ such that

$$P_0 \in \mathscr{M}.$$

Then,

$$\|\mathbb{P}_{\omega}^{(n)} - \overline{P}\| \underset{n\to\infty}{\longrightarrow} 0 \quad U_0(d\omega)\text{-a.s.}^*$$

Theorem 3.1 may be regarded as a first Glivenko–Cantelli theorem for imprecise probabilities. The proof of Theorem 3.1 needs some preparations:

Lemma 3.2. *Let $\overline{Q}$ be a coherent upper prevision on $\mathscr{L}_\infty(\mathscr{Y}, \mathscr{B})$ with corresponding credal set $\mathscr{N}$ on $(\mathscr{Y}, \mathscr{B})$ and let Q_0 be a probability charge on $(\mathscr{Y}, \mathscr{B})$. Let G be a subset of $\mathscr{L}_\infty(\mathscr{Y}, \mathscr{B})$ such that*

- *G is convex,*
- *$g \in G \Rightarrow -g \in G$,*
- *G is bounded*: $\sup_{g\in G}\|g\| < \infty$.

Then,

$$\inf_{Q\in\mathscr{N}} \sup_{g\in G} |Q_0[g] - Q[g]| = \sup_{g\in G} Q_0[g] - \overline{Q}[g]. \tag{12}$$

In particular,

$$\|Q_0 - \overline{Q}\| = \sup_{g\in\mathscr{L}_\infty(\mathscr{Y},\mathscr{B})} \frac{Q_0[g] - \overline{Q}[g]}{\|g\|}.$$

That is, the distance $\|Q_0 - \overline{Q}\|$ exactly coincides with the operator norm if we consider

$$Q_0 - \overline{Q} : \mathscr{L}_\infty(\mathscr{Y}, \mathscr{B}) \to \mathbb{R}, \quad g \mapsto Q_0[g] - \overline{Q}[g]$$

as a (non-linear) operator.

Proof of Lemma 3.2. Eq. (12) obviously coincides with

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}|Q_0[g]-Q[g]| = \sup_{g\in G}\inf_{Q\in\mathcal{N}} Q_0[g]-Q[g]. \tag{13}$$

In (13) the inequality "$\geq$" is trivial and, therefore, it only remains to proof the inequality "$\leq$" in (13).

In order to prove this, firstly, note that property ($g \in G \Rightarrow -g \in G$) implies

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}|Q_0[g]-Q[g]| = \inf_{Q\in\mathcal{N}}\sup_{g\in G} Q_0[g]-Q[g]. \tag{14}$$

In order to show that "inf" and "sup" may be interchanged at the right side of Eq. (14), a minimax theorem is applied for

$$\Gamma : \mathcal{N}\times G \to \mathbb{R}, \quad (Q,g)\mapsto Q_0[g]-Q[g].$$

To this end, note that $\mathcal{N}$ is $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$-compact (cf. Section 1.2) and, for every $g \in G$, $Q \mapsto \Gamma(Q,g)$ is convex and $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$-continuous. In addition, $g \mapsto \Gamma(Q,g)$ is concave for every $Q \in \mathcal{N}$. So, it follows from the minimax theorem (Fan, 1953, Theorem 2) that

$$\inf_{Q\in\mathcal{N}}\sup_{g\in G}\Gamma(Q,g) = \sup_{g\in G}\inf_{Q\in\mathcal{N}}\Gamma(Q,g).$$

Together with (14), this implies (13) and (12). The last statement follows from (12) for $G=\{g\in\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})\,|\,\|g\|\leq 1\}$. $\square$

Proposition 3.3. *Let $\overline{Q}$ be a coherent upper prevision on $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$ with credal set*

$$\mathcal{N} = \left\{ Q \in \mathrm{ba}_1^+(\mathcal{Y},\mathcal{B}) \,\middle|\, Q[g]\leq\overline{Q}[g] \quad \forall g\in\mathcal{G}\right\}, \tag{15}$$

where $\mathcal{G}=\{g_1,\dots,g_s\}$ is a finite subset of $\mathcal{L}_\infty(\mathcal{Y},\mathcal{B})$. Assume that

$$\overline{Q}[g_i]-\underline{Q}[g_i]>0 \quad \forall i\in\{1,\dots,s\}. \tag{16}$$

Then, for every probability charge $Q_0\in\mathrm{ba}_1^+(\mathcal{Y},\mathcal{B})$,

$$\|Q_0-\overline{Q}\| \leq 2\cdot\sum_{i=1}^{s}\frac{(Q_0[g_i]-\overline{Q}[g_i])^+}{\overline{Q}[g_i]-\underline{Q}[g_i]}. \tag{17}$$

Proof of Proposition 3.3. If there is any $i\in\{1,\dots,s\}$ such that $Q_0[g_i]-\overline{Q}[g_i]\geq\overline{Q}[g_i]-\underline{Q}[g_i]$ then (17) is trivially fulfilled and nothing remains to be proven. Therefore, it can be assumed that

$$Q_0[g_i]-\overline{Q}[g_i]<\overline{Q}[g_i]-\underline{Q}[g_i] \quad \forall i\in\{1,\dots,s\}. \tag{18}$$

Without loss of generality, we may assume that the elements of $\mathcal{G}$ are indexed in such a way that there is an $r\in\{0,\dots,s\}$ such that

$$Q_0[g_i]>\overline{Q}[g_i] \quad \forall i\leq r \quad \text{and} \quad Q_0[g_i]\leq\overline{Q}[g_i] \quad \forall i>r.$$

Putting

$$\varepsilon_i := \frac{(Q_0[g_i]-\overline{Q}[g_i])^+}{\overline{Q}[g_i]-\underline{Q}[g_i]} \quad \forall i\leq r, \tag{19}$$

(18) implies $0<\varepsilon_i<1$ for every $i\in\{1,\dots,r\}$.

Let $\overline{Q}_0$ be the coherent upper prevision with credal set

$$\mathcal{N}_0 = \left\{ Q\in\mathrm{ba}_1^+(\mathcal{Y},\mathcal{B}) \,\middle|\, Q[g]\leq \max\left\{\overline{Q}[g],Q_0[g]\right\} \ \forall g\in\mathcal{G}\right\}.$$

Then, it follows from $Q_0\in\mathcal{N}_0$ and $\mathcal{N}\subset\mathcal{N}_0$ that

$$\overline{Q}_0[g]=\max\{\overline{Q}[g],Q_0[g]\} \quad \forall g\in\mathcal{G}$$

and, together with (19), this implies

$$\overline{Q}[g_i]\leq\overline{Q}_0[g_i]=\overline{Q}[g_i]+\varepsilon_i(\overline{Q}[g_i]-\underline{Q}[g_i]) \quad \forall i\leq r$$

and

$$\overline{Q}[g_i]=\overline{Q}_0[g_i] \quad \forall i>r.$$

Since $\overline{Q}$ and $\overline{Q}_0$ may be considered as natural extensions of coherent upper previsions on $\mathscr{G}$, (Hable, 2009b, Theorem 2.2) is applicable and yields

$$\overline{Q}[g] \le \overline{Q}_0[g] \le \overline{Q}[g] + \varepsilon(\sup g - \inf g) \le \overline{Q}[g] + 2\varepsilon\|g\| \quad \forall g \in \mathscr{L}_\infty(\mathscr{Y},\mathscr{B})$$

for $\varepsilon = \varepsilon_1 + \cdots + \varepsilon_r > 0$.

Put $G := \{g \in \mathscr{L}_\infty(\mathscr{Y},\mathscr{B}) |\, \|g\| \le 1\}$; then, Lemma 3.2 implies

$$\begin{aligned}\|Q_0 - \overline{Q}\| &= \sup_{g\in G} Q_0[g] - \overline{Q}[g] \le \sup_{g\in G} \overline{Q}_0[g] - \overline{Q}[g] \\ &\le \sup_{g\in G} \overline{Q}[g] + 2\varepsilon\|g\| - \overline{Q}[g] \le \sup_{g\in G} 2\varepsilon\|g\| \\ &= 2\cdot\sum_{i=1}^{r}\varepsilon_i = 2\cdot\sum_{i=1}^{s}\frac{(Q_0[g_i] - \overline{Q}[g_i])^+}{\overline{Q}[g_i] - \underline{Q}[g_i]}. \qquad \square\end{aligned}$$

Of course, (15) is, in general, a very rough bound. However, if Q_0 is equal to the empirical measure $\mathbb{P}^{(n)}$ and the true distribution lies in the credal set of $\overline{Q} = \overline{P}$, then the law of large numbers yields

$$(Q_0[g_i] - \overline{Q}[g_i])^+ \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore, bound (15) provides valuable information for increasing numbers of observations. Since Theorem 3.1 is about the asymptotic behavior of the distance $\|\mathbb{P}^{(n)} - \overline{P}\|$, bound (15) serves as the cornerstone of the proof. Theorem 3.1 states that the distance between the empirical measure and the coherent upper prevision converges to 0. However, Proposition 3.3 can also be used to prove the following stronger theorem which also makes some assertions about the rate of convergence:

Theorem 3.4. *Under the assumptions of Theorem* 3.1, *it follows that*

(a) $\|\mathbb{P}^{(n)}_\omega - \overline{P}\| = O\left(\sqrt{\frac{\ln\ln n}{n}}\right) \quad U_0(d\omega)$*-a.s.**

(b) *In addition, assume that*

$$\underline{P}_0[f_i] < \overline{P}_0[f_i] \quad \forall i \in \{1,\dots,s\}. \tag{20}$$

Then,

$$\lim_{n\to\infty} U_0^*(\{\omega \in \Omega |\, \|\mathbb{P}^{(n)}_\omega - \overline{P}\| > 0\}) = 0.$$

In part (b), the outer measure U_0^* is used instead of U_0 because the set

$$\{\omega \in \Omega |\, \|\mathbb{P}^{(n)}_\omega - \overline{P}\| > 0\}$$

is not assumed to be measurable.

Proof of Theorem 3.4. For every $i \in \{1,\dots,s\}$, put

$$h_i := \frac{f_i}{\overline{P}[f_i] - \underline{P}[f_i]}.$$

According to Proposition 3.3,

$$0 \le \|\mathbb{P}^{(n)}_\omega - \overline{P}\| \le 2\cdot\sum_{i=1}^{s}(\mathbb{P}^{(n)}_\omega[h_i] - \overline{P}[h_i])^+ \quad \forall \omega \in \Omega. \tag{21}$$

In particular, $\Omega \to \mathbb{R}$, $\omega \mapsto 2\sum_{i=1}^{s}(\mathbb{P}^{(n)}_\omega[h_i] - \overline{P}[h_i])^+$ is an $\mathscr{A}/\mathbb{B}$-measurable map which dominates the possibly unmeasurable map $\omega \mapsto \|\mathbb{P}^{(n)}_\omega - \overline{P}\|$.

(a) Note that, for every $i \in \{1,\dots,s\}$ and for every $\omega \in \Omega$,

$$(\mathbb{P}^{(n)}_\omega[h_i] - \overline{P}[h_i])^+ \le (\mathbb{P}^{(n)}_\omega[h_i] - P_0[h_i])^+ \le |\mathbb{P}^{(n)}_\omega[h_i] - P_0[h_i]| = \left|\frac{1}{n}\sum_{j=1}^{n}(h_i\circ X_j(\omega) - P_0[h_i])\right|.$$

Since $\int_\Omega h_i \circ X_j(\omega) U_0(d\omega) = P_0[h_i]$, the law of the iterated logarithm (Hoffmann-Jørgensen, 1994b, Section 10.25) yields

$$\left| \frac{1}{n} \sum_{j=1}^{n} (h_i \circ X_j(\omega) - P_0[h_i]) \right| = O\left(\sqrt{\frac{\ln \ln n}{n}} \right) \quad U_0(d\omega)\text{-a.s.}$$

and, therefore,

$$(\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}[h_i])^+ = O\left(\sqrt{\frac{\ln \ln n}{n}} \right) \quad U_0(d\omega)\text{-a.s.}$$

Together with (21), this implies the validity of part (a).

(b) For every $n \in \mathbb{N}$, put

$$A_n^{(1)} = \{\omega \in \Omega | \|\mathbb{P}_\omega^{(n)} - \overline{P}\| > 0\},$$
$$A_n^{(2)} = \left\{ \omega \in \Omega \left| \sum_{i=1}^{s} (\mathbb{P}_\omega^{(n)}[h_i] - \overline{P}[h_i])^+ > 0 \right. \right\}$$

and, for $i \in \{1, \dots, s\}$,

$$B_{i,n} = \{\omega \in \Omega | \mathbb{P}_\omega^{(n)}[h_i] > \overline{P}[h_i]\} = \{\omega \in \Omega | \mathbb{P}_\omega^{(n)}[h_i] - P_0[h_i] > \overline{P}[h_i] - P_0[h_i]\}.$$

Then, we have

$$A_n^{(1)} \overset{(21)}{\subset} A_n^{(2)} \subset \bigcup_{i=1}^{s} B_{i,n},$$

where $A_n^{(2)}, B_{1,n}, \dots, B_{s,n} \in \mathscr{A}$. Finally,

$$U_0^*(A_n^{(1)}) \le \sum_{i=1}^{s} U_0(B_{i,n}) \underset{n\to\infty}{\longrightarrow} 0$$

because $U_0(B_{i,n}) \underset{n\to\infty}{\longrightarrow} 0 \; \forall i \in \{1, \dots, s\}$ follows from the strong law of large numbers (Hoffmann-Jørgensen, 1994a, Section 4.12), assumption (20) and the fact that almost sure convergence implies convergence in probability; cf. Hoffmann-Jørgensen (1994a), Section 3.25. □

Proof of Theorem 3.1. This is a direct consequence of Theorem 3.4. □

Now, let us turn over to consistency of the minimum distance estimator. The true $P_{\theta_0} \in \mathscr{M}_{\theta_0}$ is totally unknown and we only want to estimate θ_0. A true parameter θ_0 is any $\theta_0 \in \Theta$ such that

$$P_{\theta_0} \in \mathscr{M}_{\theta_0}.$$

In case of a finite set of parameters Θ, a sensible estimator $\hat{\theta}_n$ should—at least for large sizes of n—lead to small error probabilities

$$U_0^*(P_0 \notin \mathscr{M}_{\hat{\theta}_n}).$$

The minimum distance estimator fulfills this requirement as stated in the following theorem. Again, using the outer measure U_0^* instead of U_0 is necessary because we do not assume $\{\omega \in \Omega | P_0 \notin \mathscr{M}_{\hat{\theta}_n(\omega)}\}$ to be measurable.

Theorem 3.5. *In addition to the fixings and assumptions given in* Section 2.1, *let the index set Θ be finite. Then,*

$$U_0^*(P_0 \notin \mathscr{M}_{\hat{\theta}_n}) \underset{n\to\infty}{\longrightarrow} 0. \tag{22}$$

Proof. According to the assumptions, there is some $\theta_0 \in \Theta$ such that $P_0 \in \mathscr{M}_{\theta_0}$. Firstly, fix any $n \in \mathbb{N}$. For every $\theta \in \Theta$, put

$$A_\theta^{(n)} := \left\{ \omega \in \Omega \,\middle|\, \|\mathbb{P}_\omega^{(n)} - \overline{P}_\theta\| \le \|\mathbb{P}_\omega^{(n)} - \overline{P}_{\theta_0}\| \right\}.$$

Note that the definition of $\hat{\theta}_n$ implies

$$\omega \in A_{\hat{\theta}_n(\omega)}^{(n)} \quad \forall \omega \in \Omega. \tag{23}$$

Put $\Theta_0 := \{\theta \in \Theta | P_0 \in \mathscr{M}_\theta\}$; in particular, $\theta_0 \in \Theta_0$. Then, the following relations are valid for every $\omega \in \Omega$:

$$P_0 \notin \mathscr{M}_{\hat{\theta}_n(\omega)} \Rightarrow \hat{\theta}_n(\omega) \in \Theta \backslash \Theta_0 \overset{(23)}{\Rightarrow} \omega \in \bigcup_{\theta \in \Theta \backslash \Theta_0} A_\theta^{(n)}.$$

Therefore,

$$U_0^*(P_{\theta_0} \notin \mathscr{M}_{\hat{\theta}_n}) \leq \sum_{\theta \in \Theta \backslash \Theta_0} U_0^*(A_\theta^{(n)}). \tag{24}$$

For every $\theta \in \Theta \backslash \Theta_0$, it follows from $P_{\theta_0} \notin \mathscr{M}_\theta$ that there is a $\varepsilon_\theta > 0$ such that

$$\sup_{i \in \{1,\ldots,s\}} (P_{\theta_0}[f_i] - \overline{P}_\theta[f_i]) \cdot \|f_i\|^{-1} > \varepsilon_\theta. \tag{25}$$

Then, for every $\omega \in \Omega$ and for every $\theta \in \Theta \backslash \Theta_0$,

$$\begin{aligned}
\|\mathbb{P}_\omega^{(n)} - \overline{P}_\theta\| &\geq \inf_{P_\theta \in \mathscr{M}_\theta} \sup_{i \in \{1,\ldots,s\}} |\mathbb{P}_\omega^{(n)}[f_i] - P_\theta[f_i]| \cdot \|f_i\|^{-1} \\
&\geq \sup_{i \in \{1,\ldots,s\}} \inf_{P_\theta \in \mathscr{M}_\theta} (\mathbb{P}_\omega^{(n)}[f_i] - P_\theta[f_i]) \cdot \|f_i\|^{-1} \\
&= \sup_{i \in \{1,\ldots,s\}} (\mathbb{P}_\omega^{(n)}[f_i] - P_0[f_i] + P_0[f_i] - \overline{P}_\theta[f_i]) \cdot \|f_i\|^{-1} \\
&\geq \inf_{i \in \{1,\ldots,s\}} (\mathbb{P}_\omega^{(n)}[f_i] - P_0[f_i]) \cdot \|f_i\|^{-1} + \sup_{i \in \{1,\ldots,s\}} (P_0[f_i] - \overline{P}_\theta[f_i]) \cdot \|f_i\|^{-1} \\
&\overset{(25)}{>} \inf_{i \in \{1,\ldots,s\}} (\mathbb{P}_\omega^{(n)}[f_i] - P_0[f_i]) \cdot \|f_i\|^{-1} + \varepsilon_\theta.
\end{aligned}$$

Put

$$Z_n(\omega) = \|\mathbb{P}_\omega^{(n)} - \overline{P}_{\theta_0}\| - \inf_{i \in \{1,\ldots,s\}} (\mathbb{P}_\omega^{(n)}[f_i] - P_0[f_i]) \cdot \|f_i\|^{-1} \quad \forall \omega \in \Omega$$

and note that, for every $\theta \in \Theta \backslash \Theta_0$,

$$Z_n(\omega) > \varepsilon_\theta \quad \forall \omega \in A_\theta^{(n)}.$$

Therefore, it follows from (24) that

$$U_0^*(P_0 \notin \mathscr{M}_{\hat{\theta}_n}) \leq \sum_{\theta \in \Theta \backslash \Theta_0} U_0^*(Z_n > \varepsilon_\theta). \tag{26}$$

Next, Theorem 3.1 and the strong law of large numbers (Hoffmann-Jørgensen, 1994a, Section 4.12) yield

$$Z \underset{n \to \infty}{\longrightarrow} 0 \quad U_0\text{-a.s.}^*$$

According to van der Vaart and Wellner (1996, Lemma 1.9.2), U_0-a.s.*-convergence implies convergence in U_0^*-probability. Hence,

$$U_0^*(P_0 \notin \mathscr{M}_{\hat{\theta}_n}) \overset{(26)}{\leq} \sum_{\theta \in \Theta \backslash \Theta_0} U_0^*(Z_n > \varepsilon_\theta) \underset{n \to \infty}{\longrightarrow} 0. \qquad \square$$

4. Calculation of the estimator

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. 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. In case of precise probabilities, discretization would even be part of the definition of the minimum distance estimator.

Recall the fixings and definitions given in Section 2.1. In order to calculate the minimum distance estimator, we have to calculate

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

Though $\mathscr{M}_\theta$ is a large subset of $\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})$ (which is hardly dominated by a σ-finite measure), 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 $(\mathscr{X},\mathscr{B})$ may be discretized as follows: for $\theta\in\Theta$, put

$$\varepsilon_\theta^{(j)} := \frac{\overline{P}_\theta[f_j]-\underline{P}_\theta[f_j]}{2s}\cdot\varepsilon \quad \forall j\in\{1,\dots,s\}$$

and choose simple functions $h_\theta^{(1)},\dots,h_\theta^{(s)}$ such that

$$f_j \le h_\theta^{(j)} \le f_j+\varepsilon_\theta^{(j)} \quad \forall j\in\{1,\dots,s\}.$$

Then, let $\mathscr{C}_\theta$ be the smallest σ-algebra on $\mathscr{X}$ such that the simple functions $h_\theta^{(1)},\dots,h_\theta^{(s)}$ are $\mathscr{C}_\theta/\mathbb{B}$-measurable. Note that $\mathscr{C}_\theta$ is a finite subset of $\mathscr{B}$ and that there is a finite partition $\{C_\theta^{(1)},\dots,C_\theta^{(r)}\}$ of $\mathscr{X}$ such that

$$\mathscr{C}_\theta=\left\{\bigcup_{j\in\mathscr{J}} C_\theta^{(j)}\,|\,\mathscr{J}\subset\{1,\dots,r\}\right\}.$$

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

$$\mathscr{N}_\theta=\left\{Q_\theta\in\mathrm{ba}_1^+(\mathscr{X},\mathscr{C}_\theta)\;\middle|\;Q_\theta[h_j]\le\overline{P}_\theta[f_j]+\varepsilon_\theta^{(j)}\quad\forall j\in\{1,\dots,s\}\right\}.$$

The following theorem states that

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

is approximately equal to $\|\mathbb{P}_x^{(n)}-\overline{P}_\theta\|$. In this way, the above discretization of the sample space guarantees that the distances are calculated with accuracy ε. However, the sample space may also be discretized in other ways as the result of the estimator only depends on the accuracy of the calculated distances and not on the discretization.

Theorem 4.1. *In the setup of the present subsection*,

$$\|\mathbb{P}_x^{(n)}-\overline{Q}_\theta\|\le\|\mathbb{P}_x^{(n)}-\overline{P}_\theta\|\le\|\mathbb{P}_x^{(n)}-\overline{Q}_\theta\|+\varepsilon\quad\forall\theta\in\Theta$$

for every $x\in\mathscr{X}^n$.

Proof. Fix any $x=(x_1,\dots,x_n)\in\mathscr{X}^n$ and any $\theta\in\Theta$. Let $\overline{Q}'_\theta$ be the natural extension of $\overline{Q}_\theta$ on $\mathscr{L}_\infty(\mathscr{X},\mathscr{B})$; i.e. $\overline{Q}'_\theta$ is a coherent upper prevision with credal set

$$\mathscr{N}'_\theta=\left\{Q'_\theta\in\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})\;\middle|\;Q'_\theta[h_j]\le\overline{P}_\theta[f_j]+\varepsilon_\theta^{(j)}\quad\forall j\in\{1,\dots,s\}\right\}.$$

Then,

$$\|\mathbb{P}_x^{(n)}-\overline{Q}'_\theta\|=\inf_{Q'_\theta\in\mathscr{N}'_\theta}\sup_{f\in\mathscr{L}_\infty(\mathscr{X},\mathscr{B})}\frac{|\mathbb{P}_x^{(n)}[f]-Q'_\theta[f]|}{\|f\|}.$$

Let $\overline{P}'_\theta$ be the coherent upper prevision on $(\mathscr{X},\mathscr{B})$ with credal set

$$\mathscr{M}'_\theta=\left\{P'_\theta\in\mathrm{ba}_1^+(\mathscr{X},\mathscr{B})\;\middle|\;P'_\theta[f_j]\le\overline{Q}'_\theta[f_j]\quad\forall j\in\{1,\dots,s\}\right\}.$$

It is an easy consequence of the definitions that $\mathscr{M}_\theta\subset\mathscr{N}'_\theta\subset\mathscr{M}'_\theta$,

$$\overline{P}_\theta[f]\le\overline{Q}'_\theta[f]\le\overline{P}'_\theta[f]\quad\forall f\in\mathscr{L}_\infty(\mathscr{X},\mathscr{B}) \tag{27}$$

and

$$\overline{P}'_\theta[f_j]=\overline{Q}'_\theta[f_j]\quad\forall j\in\{1,\dots,s\}. \tag{28}$$

Hence, we have

$$\overline{P}_\theta[f_j]\overset{(27)}{\le}\overline{P}'_\theta[f_j]\overset{(28)}{=}\overline{Q}'_\theta[f_j]\le\overline{P}_\theta[f_j]+\frac{\varepsilon}{2s}(\overline{P}_\theta[f_j]-\underline{P}_\theta[f_j])$$

according to the definition of $\varepsilon_\theta^{(j)}$. Therefore, an application of Hable (2009b, Theorem 2.2) yields

$$\overline{P}_\theta[f] \le \overline{P}'_\theta[f] \le \overline{P}_\theta[f] + \frac{\varepsilon}{2}(\sup f - \inf f) \quad \forall f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}).$$

Together with (27), this implies

$$\overline{P}_\theta[f] \le \overline{Q}'_\theta[f] \le \overline{P}_\theta[f] + \frac{\varepsilon}{2}(\sup f - \inf f) \quad \forall f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}). \tag{29}$$

A twofold application of Lemma 3.2 implies

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

and

$$\|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\| = \sup_{f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}'_\theta[f]}{\|f\|}.$$

Hence, (29) implies

$$\begin{aligned}
\|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\| \le \|\mathbb{P}_x^{(n)} - \overline{P}_\theta\| \\
&\le \sup_{f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{\mathbb{P}_x^{(n)}[f] - \left(\overline{Q}'_\theta[f] - \frac{\varepsilon}{2}(\sup f - \inf f)\right)}{\|f\|} \\
&\le \|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\| + \frac{\varepsilon}{2} \cdot \sup_{f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{\sup f - \inf f}{\|f\|} \\
&\le \|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\| + \varepsilon.
\end{aligned}$$

Therefore, it only remains to prove

$$\|\mathbb{P}_x^{(n)} - \overline{Q}'_\theta\| = \|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\|. \tag{30}$$

The inequality " $\ge$ " is trivially fulfilled in (30). In order to prove the inequality " $\le$ " it is enough to show

$$\sup_{f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{B})} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}'_\theta[f]}{\|f\|} \le \sup_{f \in \mathscr{L}_\infty(\mathscr{X},\mathscr{C}_\theta)} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|} \tag{31}$$

according to Lemma 3.2. That is, it only remains to prove (31) in the following:

To this end, 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 $\mathscr{X}$ which generates $\mathscr{C}_\theta$. Furthermore, put

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

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

$$\mathscr{J}_0 = \left\{ j \in \{1, \dots, r\} \mid n_j = 0 \right\} \quad \text{and} \quad \mathscr{J}_1 = \left\{ j \in \{1, \dots, r\} \mid n_j > 0 \right\}.$$

In particular, this means

$$\{x_1, \dots, x_n\} \cap C_\theta^{(j)} = \emptyset \quad \forall j \in \mathscr{J}_0 \tag{32}$$

and

$$\sum_{k=1}^{n} I_{C_\theta^{(j)}}(x_k) = n_j \quad \forall j \in \mathscr{J}_1. \tag{33}$$

Applying these settings, we can define the map

$$\xi : \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}) \to \mathscr{L}_\infty(\mathscr{X}, \mathscr{C}_\theta), \quad f \mapsto \xi(f),$$

where

$$\xi(f) = \sum_{j \in \mathscr{J}_0} f(c_j) I_{C_\theta^{(j)}} + \sum_{j \in \mathscr{J}_1} \left(\frac{1}{n_j} \sum_{i \in N_j} f(x_i) \right) I_{C_\theta^{(j)}}. \tag{34}$$

Note that this map is defined well and that $\mathscr{J}_0 \cap \mathscr{J}_1 = \emptyset$. Then, it is an immediate consequence of the definitions that ξ is linear, positive ($\xi(f) \geq 0 \, \forall f \geq 0$) and normalized ($\xi(I_{\mathscr{X}}) = I_{\mathscr{X}}$). Therefore, ξ defines a map

$$\rho : \mathrm{ba}_1^+(\mathscr{X}, \mathscr{C}_\theta) \to \mathrm{ba}_1^+(\mathscr{X}, \mathscr{B}), \quad Q \mapsto \rho(Q)$$

via $\rho(Q)[f] = Q[\xi(f)] \, \forall Q \in \mathrm{ba}_1^+(\mathscr{X}, \mathscr{C}_\theta)$, $\forall f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B})$. Since

$$\xi(f) = f \quad \forall f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{C}_\theta) \subset \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}).$$

it is easy to see that

$$\rho(Q_\theta) \in \mathscr{N}'_\theta \quad \forall Q_\theta \in \mathscr{N}_\theta. \tag{35}$$

In addition,

$$\begin{aligned}
\mathbb{P}_x^{(n)}[\xi(f)] &= \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{j\in\mathscr{J}_0} f(c_j) I_{C_\theta^{(j)}}(x_k) + \sum_{j\in\mathscr{J}_1}\left(\frac{1}{n_j}\sum_{i\in N_j} f(x_i)\right) I_{C_\theta^{(j)}}(x_k)\right) \\
&\overset{(32)}{=} \frac{1}{n}\sum_{j\in\mathscr{J}_1}\left(\left(\frac{1}{n_j}\sum_{i\in N_j} f(x_i)\right)\sum_{k=1}^{n} I_{C_\theta^{(j)}}(x_k)\right) \\
&\overset{(33)}{=} \frac{1}{n}\sum_{j\in\mathscr{J}_1}\left(\left(\frac{1}{n_j}\sum_{i\in N_j} f(x_i)\right)\cdot n_j\right) = \frac{1}{n}\sum_{j\in\mathscr{J}_1}\sum_{i\in N_j} f(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} f(x_i) = \mathbb{P}_n^{(n)}[f] \quad \forall f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B})
\end{aligned}$$

shows that

$$\rho(\mathbb{P}_x^{(n)}) = \mathbb{P}_x^{(n)}. \tag{36}$$

Finally, fix any $f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}) \setminus \{0\}$. Then,

$$\|\xi(f)\| \leq \|f\| \tag{37}$$

and

$$\begin{aligned}
\frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}'_\theta[f]}{\|f\|} &\overset{(36)}{=} \inf_{Q'_\theta \in \mathscr{N}'_\theta} \frac{\rho(\mathbb{P}_x^{(n)})[f] - Q'_\theta[f]}{\|f\|} \\
&\overset{(35)}{\leq} \inf_{Q_\theta \in \mathscr{N}_\theta} \frac{\rho(\mathbb{P}_x^{(n)})[f] - \rho(Q_\theta)[f]}{\|f\|} = \inf_{Q_\theta \in \mathscr{N}_\theta} \frac{\mathbb{P}_x^{(n)}[\xi(f)] - Q_\theta[\xi(f)]}{\|f\|}.
\end{aligned}$$

If $\xi(f) = 0$, this implies

$$\frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}'_\theta[f]}{\|f\|} \leq 0 = \frac{\mathbb{P}_x^{(n)}[I_{\mathscr{X}}] - \overline{Q}_\theta[I_{\mathscr{X}}]}{\|I_{\mathscr{X}}\|} \leq \sup_{f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{C}_\theta)} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|}$$

and, if $\xi(f) \neq 0$, this implies

$$\begin{aligned}
\frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}'_\theta[f]}{\|f\|} &\leq \inf_{Q_\theta \in \mathscr{N}_\theta} \frac{\mathbb{P}_x^{(n)}[\xi(f)] - Q_\theta[\xi(f)]}{\|f\|} \\
&\overset{(37)}{\leq} \inf_{Q_\theta \in \mathscr{N}_\theta} \frac{\mathbb{P}_x^{(n)}[\xi(f)] - Q_\theta[\xi(f)]}{\|\xi(f)\|} = \frac{\mathbb{P}_x^{(n)}[\xi(f)] - \overline{Q}_\theta[\xi(f)]}{\|\xi(f)\|} \\
&\leq \sup_{f \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{C}_\theta)} \frac{\mathbb{P}_x^{(n)}[f] - \overline{Q}_\theta[f]}{\|f\|}
\end{aligned}$$

again. Therefore, (31) follows. $\square$

That is, in order to approximately calculate $\|\mathbb{P}_x^{(n)} - \overline{P}_\theta\|$, it is possible to calculate

$$\|\mathbb{P}_x^{(n)} - \overline{Q}_\theta\| = \inf_{Q_\theta \in \mathscr{N}_\theta} \sup_{f \in \mathscr{L}(\mathscr{X}, \mathscr{A})} \frac{|\mathbb{P}_x^{(n)}[f] - Q_\theta[f]|}{\|f\|}, \tag{38}$$

where $\overline{Q}_\theta$ is a coherent upper prevision on the finite space $(\mathscr{X},\mathscr{C}_\theta)$. This value can be calculated by solving the following linear program. The notation is taken from the proof of Theorem 4.1.

$$\sum_{j\in\mathscr{J}_1} q_j-\gamma_j \longrightarrow \max_{q,\gamma}! \tag{39}$$

where

$$\sum_{j=1}^{r} q_j = 1 \tag{40}$$

and

$$\sum_{j=1}^{r} q_j h_k(c_j) \le \overline{P}_\theta[f_k]+\varepsilon_\theta^{(k)} \quad \forall k\in\{1,\dots,s\} \tag{41}$$

and

$$q_j-\gamma_j \le \frac{n_j}{n} \quad \forall j\in\mathscr{J}_1 \tag{42}$$

for the variables

$$q=(q_1,\dots,q_r)\in\mathbb{R}^r, \quad q_j\ge 0 \quad \forall j\in\{1,\dots,r\} \tag{43}$$

and

$$\gamma=(\gamma_j)_{j\in\mathscr{J}_1}\subset\mathbb{R}, \quad \gamma_j\ge 0 \quad \forall j\in\mathscr{J}_1. \tag{44}$$

Let β_θ be the optimal value of the above linear program. Then, we have

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

See Hable (2009a, Section 6.5) and Hable (2009c) for the derivation of this linear programm and more details about computational aspects.

5. Applications

5.1. Simulation study

Many statistical evaluations are based on the assumption that the data stem from a normal distribution. Though it is not possible to statistically assure the validity of this assumption, it is often tried to do this by a chi-square test. In order to do this, the sample space $\mathscr{X}=\mathbb{R}$ is divided into segments. Since the chi-square test only takes the probabilities of such segments into account, it is far away from covering all aspects of the normal distribution. Therefore, this situation does not cope with the strict assumption of a precise normal distribution but exactly corresponds to a finitely generated coherent upper prevision. This motivates the following definition of an imprecise model.

The sample space $(\mathscr{X},\mathscr{B})$ is equal to $(\mathbb{R},\mathfrak{B})$. It is divided by

$$a_0=-10, \quad a_1=-9.5, \quad a_2=-9, \quad \dots, \quad a_j=-10+j\cdot 0.5,\dots,a_{40}=10$$

into the segments $(-\infty,a_0),(a_0,a_1],(a_1,a_2],\dots,(a_{39},a_{40}],(a_{40},\infty)$. That is, we consider the set of functions $\mathscr{K}=\{f_1,f_1,f_2,\dots,f_{80},f_{81}\}$ where

$$f_0=I_{(-\infty,a_0]}, \quad f_{81}=I_{(a_{40},\infty)}$$

and

$$f_j=I_{(a_{j-1},a_j]}, \quad f_{40+j}=1-I_{(a_{j-1},a_j]} \quad \forall j\in\{1,\dots,40\}.$$

According to the above motivation, we want to deal with a family of normal distributions on $(\mathbb{R},\mathfrak{B})$:

$$P_\theta=\mathscr{N}(\mu,\sigma^2) \quad \text{where } \theta=(\mu,\sigma) \quad \text{for } -5\le\mu\le 5,\ 0.4\le\sigma\le 2.$$

That is, the index set is $\Theta=\Theta^{(1)}\times\Theta^{(2)}=[-5,5]\times[0.4,2]$. For the definition of the imprecise model, Θ is discretized:

$$\Theta_0=\left\{(\mu_0,\sigma_0)\in\Theta \,\middle|\, \begin{matrix}\mu_0=-5+0.1\cdot j_1,\ j_1\in\{0,1,\dots,100\},\\ \sigma_0=0.4+0.1\cdot j_2,\ j_2\in\{0,1,\dots,16\}\end{matrix}\right\}.$$

Table 1
Empirical mean squared errors calculated over the estimations for sample size $n=100$—obtained in 500 runs for each part of the simulation study.

	Ideal situation		Real situation	
	MinDistance	MaxLikelihood	MinDistance	MaxLikelihood
μ	0.05517	0.02056	0.05233	1.04284
σ	0.04309	0.01049	0.05007	0.23216

So, (μ_0,σ_0) corresponds to the rectangle $(\mu_0-0.05,\mu_0+0.05]\times(\sigma_0-0.05,\sigma_0+0.05]$ with center (μ_0,σ_0). Based on this discretization and the normal distributions, we can define the following upper previsions for the segments of the sample space:

$$\overline{P}_{\theta_0}[f_j]=(1-0.02)\cdot\sup_{\substack{\mu\in(\mu_0-0.05,\mu_0+0.05]\\ \sigma\in(\sigma_0-0.05,\sigma_0+0.05]}}\int_{\mathbb{R}} f_j\,d\mathcal{N}(\mu,\sigma^2)+0.02$$

for every $j\in\{0,\dots,81\}$ and $(\mu_0,\sigma_0)\in\Theta_0$. The value 0.02 leads to more imprecision in the imprecise model—that is, to a more cautious proceeding. Roughly speaking, 0.02 can be interpreted as the probability that the data stem from any distribution which can be totally different from normal distributions. This proceeding is very similar to the use of contamination neighborhoods in robust statistics. The credal sets of the coherent upper previsions are given by

$$\mathcal{M}_{\theta_0}=\left\{Q\;\middle|\;Q[f_j]\le\overline{P}_{\theta_0}[f_j]\;\forall j\in\{0,\dots,81\}\right\}\quad\forall\theta_0\in\Theta_0.\tag{45}$$

Note that these credal sets are much larger than contamination neighborhoods of normal distributions with radius 0.02 because, in (45), only the probabilities of the above segments are associated with normal distributions. For every $j\in\{1,\dots,40\}$, the value $\overline{P}_{\theta_0}[f_j]$ is an upper bound on the probability of the segment $(a_{j-1},a_j]$ and the value $\overline{P}_{\theta_0}[f_{40+j}]$ imposes a lower bound on the probability of the segment $(a_{j-1},a_j]$ because

$$\inf_{Q\in\mathcal{M}_{\theta_0}}Q((a_{j-1},a_j])=1-\sup_{Q\in\mathcal{M}_{\theta_0}}Q[1-f_j]=1-\overline{P}_{\theta_0}[f_{40+j}]$$

implies

$$Q_0[f_{40+j}]\le\overline{P}_{\theta_0}[f_{40+j}]\quad\Leftrightarrow\quad Q_0((a_{j-1},a_j])\ge\inf_{Q\in\mathcal{M}_{\theta_0}}Q((a_{j-1},a_j])$$

for any probability measure Q_0.

The simulation study has two parts where each part consists of 500 runs with sample size $n=100$. In the first part, the data $x_1,\dots,x_n$ are independent, identically distributed according to the ideal probability measure $P_\theta=\mathcal{N}(\mu,\sigma^2)$ with $\theta=(\mu,\sigma)=(-4,1.4)$. In the second part, the data $x_1,\dots,x_n$ are independent, identically distributed according to the probability measure $P_0=0.85\cdot\mathcal{N}(\mu,\sigma^2)+0.15\cdot\text{Cauchy}(-4,1)$ where again $\theta=(\mu,\sigma)=(-4,1.4)$ is the true parameter. Numerical calculations show that, in fact,

$$P_0\in\mathcal{M}_{(-4,1.4)}\tag{46}$$

even though P_0 consists of a quite strong—15 percent—contamination with a Cauchy-distribution. On the one hand, this demonstrates that the credal sets $\mathcal{M}_\theta$ are quite large. On the other hand, this demonstrates that the use of such a (strongly contaminated) distribution P_0 is not unreasonable because (46) implies that the Cauchy-distribution is extremely similar to the normal distribution—at least with respect to the probabilities of the above segments; and such probabilities are the only aspects of the normal distribution which are often taken into account e.g. by chi-square tests.

In the simulation study, our minimum distance estimator is compared to the classical maximum likelihood estimators for normal distributions. Since the parameter set Θ is restricted, estimations which exceed the bounds of Θ are not reasonable. Therefore, the classical estimators are truncated by the bounds of Θ.

Table 1 shows the (empirical) mean squared errors for both estimators in part 1 ("ideal situation") and part 2 ("real situation") of the simulation study. These values demonstrate that our minimum distance estimator behaves reasonable well in both cases while the classical estimators are quite perfect in the "ideal situation" but lead to unreliable results in real situations. This is also made visible by the boxplots shown in Fig. 1. In particular, it can be seen that the classical estimation of the standard deviation σ breaks down in the "real situation"—only the bounded parameter set and the implemented truncation prevent it from exploding estimations. According to the results shown in Table 1 and Fig. 1, the performance of the minimum distance estimator is nearly the same in both parts of the simulation study—that is, the performance is nearly the same whether or not the data are contaminated by a Cauchy distribution. Therefore, it seems that, here, the contamination with a Cauchy distribution does not have an essential influence on our minimum distance estimator. On the one hand, this is not surprising because the minimum distance estimator does not really keep an eye on the normal distributions. It is only concerned with the probabilities of some segments and, with respect to these probabilities, normal distributions and Cauchy distributions are nearly the same. On the other hand, this means

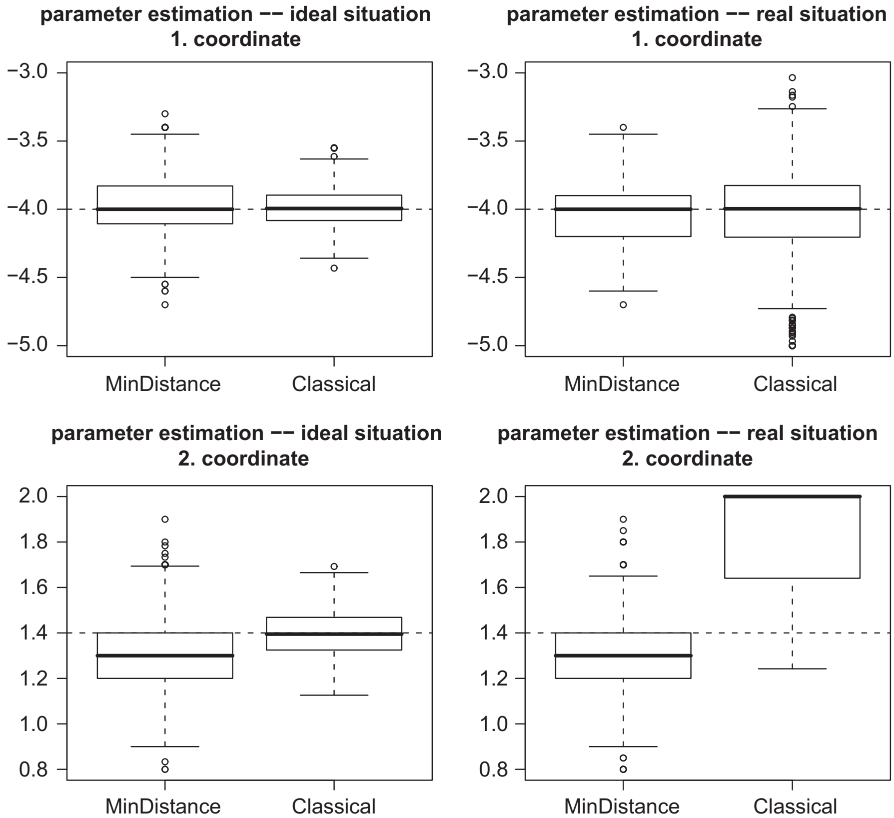

Fig. 1. Boxplots of the estimations of both parameters for sample size $n=100$—obtained in 500 runs for each part of the simulation study.

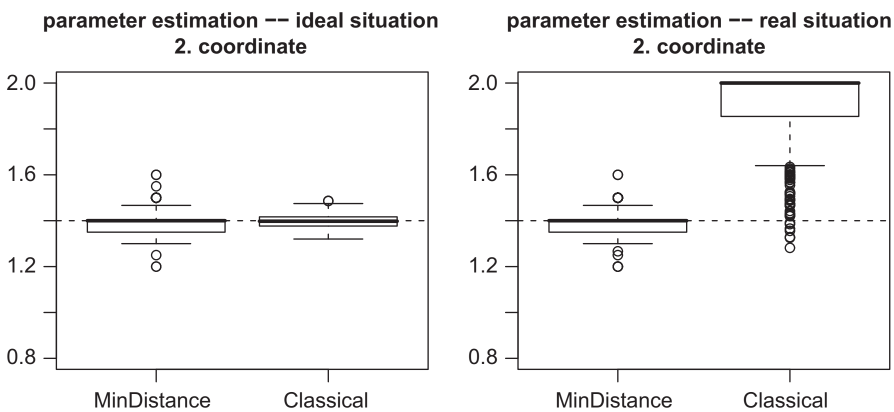

Fig. 2. Boxplots of the estimations of the standard deviation σ for sample size $n=1000$ obtained in 500 runs for each part of the simulation study.

that—as the minimum distance estimation of the standard deviation σ seems to have a negative bias in the "ideal situation"—the minimum distance estimation of the standard deviation σ also has the same negative bias in the "real situation". This negative bias is surprising here because one would rather expect that a contamination with a Cauchy distribution would lead to a positive bias (as in case of the classical maximum likelihood estimator). However, this bias of the minimum distance estimator is only a finite sample effect which disappears with growing sample sizes as follows from the asymptotic properties of the minimum distance estimator (see Section 3). In order to illustrate this, the simulation study has also been repeated for sample size $n=1000$. Fig. 2 shows the resulting boxplots for the estimation of the standard deviation σ in case of this larger sample size.

Table 2
Results of the estimations for the real data from NHANES.

	MinDistance	MaxLikelihood
μ	14.5	17.3
σ	19.0	19.2

5.2. *Real data*

Finally, the estimator is applied to a real data set. The data set consists of 200 data

$$(y_i, z_i) \in [0, \infty) \times [160, \infty), \quad i \in \{1, \dots, 200\}$$

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. 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. 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). (The original data set contains many additional variables which have been omitted here. The 200 persons whose data are analyzed here have been randomly picked out of the data from the National Health and Nutrition Examination Survey.) The following relation is assumed

$$y_i = b(z_i) + x_i, \quad i \in \{1, \dots, 200\}$$

for persons with a height of at least 160 cm. Accordingly, only persons have been considered who fulfill this condition. The function b corresponds to the medium normal weight given by the body mass index; i.e. $b : [0, \infty) \to [0, \infty)$, $x \mapsto 21.75 \cdot (x/100)^2$. The deviations $x_i = y_i - b(z_i)$ from the normal weight are assumed to be independent identically distributed according to an "approximate normal distribution". Accordingly, the following imprecise model is similar to the one of the simulation study in Section 5.1:

The sample space $(\mathscr{X}, \mathscr{B})$ is equal to $(\mathbb{R}, \mathbb{B})$ and is divided by

$$a_0 = -20, \quad a_1 = -15, \quad \dots, \quad a_k = -20 + k \cdot 5, \dots, a_{12} = 40$$

into segments. That is, we consider the set of functions $\mathscr{K} = \{f_0, \dots, f_{25}\}$ where

$$f_0 = I_{(-\infty, a_0]}, \quad f_{25} = I_{(a_{12}, \infty)}$$

and

$$f_k = I_{(a_{k-1}, a_k]} \quad \text{and} \quad f_{k+12} = 1 - I_{(a_{k-1}, a_k]} \quad \forall k \in \{1, \dots, 12\}.$$

The parameter set is $\Theta = [-15, 30] \times [1, 100]$ and the discretized parameter set is

$$\Theta_0 = \{-15, -14.5, -14, -13.5, \dots, 29.5, 30\} \times \{1, 2, 3, \dots, 100\}.$$

That is, $(\mu_0, \sigma_0) \in \Theta_0$ corresponds to the rectangle $(\mu_0 - 0.25, \mu_0 + 0.25] \times (\sigma_0 - 0.5, \sigma_0 + 0.5]$ with center (μ_0, σ_0). Based on this discretization and the normal distributions, we can define the following upper previsions for the segments of the sample space:

$$\overline{P}_{\theta_0}[f_j] = (1 - 0.01) \cdot \sup_{\substack{\mu \in (\mu_0 - 0.25, \mu_0 + 0.25] \\ \sigma \in (\sigma_0 - 0.5, \sigma_0 + 0.5]}} \int_{\mathbb{R}} f_j \, d\mathscr{N}(\mu, \sigma^2) + 0.01$$

for every $j \in \{0, \dots, 25\}$ and $(\mu_0, \sigma_0) \in \Theta_0$. The credal sets of the coherent upper previsions are given by

$$\mathscr{M}_{\theta_0} = \left\{ P \,\middle|\, P[f_j] \leq \overline{P}_{\theta_0}[f_j] \, \forall j \in \{0, \dots, 25\} \right\} \quad \forall \theta_0 \in \Theta_0. \tag{47}$$

In order to estimate the parameters μ and σ, our minimum distance estimator and the classical maximum likelihood estimator are applied; the results are given in Table 2. Fig. 3 shows the real data (y_i, z_i), the curve b (given by the body mass index) and the shifted curves $b_{\hat{\mu}} = b + \hat{\mu}$ for the estimates $\hat{\mu} = \hat{\mu}_{\text{MinDistance}}$ and $\hat{\mu} = \hat{\mu}_{\text{MaxLikelihood}}$, respectively. For the minimizing $\theta = \hat{\theta}_{\text{MinDistance}}$, the distance $\|\mathbb{P}^{(n)} - \overline{P}_\theta\|$ is only 0.03 which indicates that the data fit quite well to the assumed imprecise model.

6. Conclusions

The present article considers the estimation of a parameter in an imprecise probability model—a topic which has hardly been considered explicitly within the theory of coherent upper previsions so far. Since we are not yet able to calculate optimal estimators within this setup, a minimum distance estimator is developed which is proven to have some good asymptotic

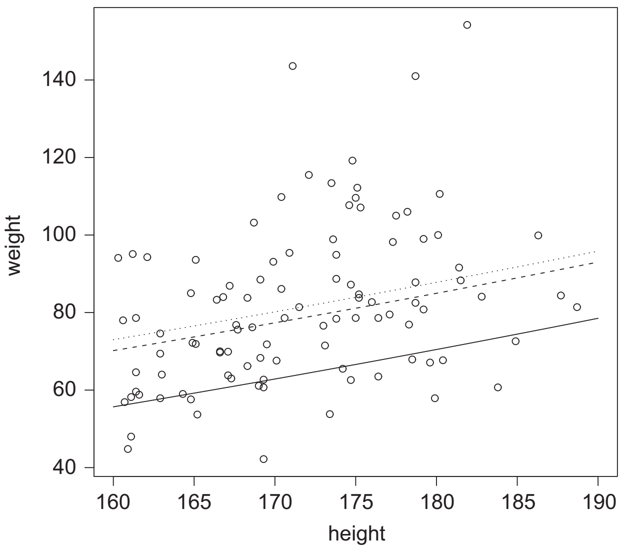

Fig. 3. Real data from NHANES and the curve of the body mass index (solid line) together with the shifted curves of the body mass index where the shift is equal to the estimated mean $\hat{\mu}_{\mathrm{MinDistance}}$ (dashed line) and $\hat{\mu}_{\mathrm{MaxLikelihood}}$ (dotted line).

properties. It is shown that the involved distances can be approximately calculated by discretizing the sample space so that approximate minimum distances can be calculated by finite space methods (such as linear programming). The applicability of the estimator is verified by a simulation study and on a real data set. However, it would have been out of the scope to investigate more advanced applications in involved statistical models. In order to encourage this, the estimator is programmed in R and publicly available as (open source) R package "imprProbEst"; cf. Hable (2008).

Acknowledgements

I would like to thank Thomas Augustin for valuable suggestions and Matthias Kohl for his help with programming in R. Furthermore, I thank the reviewer for helpful comments.

References

Augustin, T., 1998. Optimale Tests bei Intervallwahrscheinlichkeit. Vandenhoeck & Ruprecht, Göttingen.
Augustin, T., 2002. 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.
Bhaskara Rao, K.P.S., Bhaskara Rao, M., 1983. Theory of Charges. Academic Press Inc., New York.
Bickis, M., Bickis, U., 2007. Predicting the next pandemic: an exercise in imprecise hazards. In: de Cooman, G., Vejnarová, J., Zaffalon, M. (Eds.), Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications. SIPTA, Prague, pp. 41–46.
Buja, A., 1984. Simultaneously least favorable experiments. I. Upper standard functionals and sufficiency. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65 (3), 367–384.
De Cooman, G., Miranda, E., 2008. Weak and strong laws of large numbers for coherent lower previsions. Journal of Statistical Planning and Inference 138 (8), 2409–2432.
Donoho, D., Liu, R., 1988. The "automatic" robustness of minimum distance functionals. The Annals of Statistics 16 (2), 552–586.
Dunford, N., Schwartz, J., 1958. Linear Operators. I. General Theory. Wiley-Interscience Publishers, New York.
Fan, K., 1953. Minimax theorems. Proceedings of the National Academy of Sciences of the United States of America 39, 42–47.
Hable, R., 2008. imprProbEst: minimum distance estimation in an imprecise probability model. Contributed R-Package on CRAN, Version 1.0, 2008-10-23; maintainer Hable, R.
Hable, R., 2009a. Data-based decisions under complex uncertainty. Ph.D. Thesis, Ludwig-Maximilians-Universität (LMU) Munich. URL ⟨http://edoc.ub.uni-muenchen.de/9874/⟩.
Hable, R., 2009b. Discretization of sample spaces in data-based decision theory under imprecise probabilities. International Journal of Approximate Reasoning 50 (7), 1115–1128.
Hable, R., 2009c. A minimum distance estimator in an imprecise probability model—computational aspects and applications. In: 6th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'09). URL ⟨http://www.sipta.org/isipta09/proceedings/005.html⟩.
Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986. Robust Statistics. Wiley, New York.
Hoffmann-Jørgensen, J., 1994a. Probability with a View Toward Statistics, vol. I. Chapman & Hall, New York.
Hoffmann-Jørgensen, J., 1994b. Probability with a View Toward Statistics, vol. II. Chapman & Hall, New York.
Huber, P., 1965. A robust version of the probability ratio test. Annals of Mathematical Statistics 36, 1753–1758.
Huber, P., 1981. Robust Statistics. Wiley, New York.
Huber, P., 1997. Robust statistical procedures. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 68, second ed. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA.
Huber, P., Strassen, V., 1973. Minimax tests and the Neyman–Pearson lemma for capacities. The Annals of Statistics 1, 251–263.
Hutter, M., 2009. Practical robust estimators for the imprecise Dirichlet model. International Journal of Approximate Reasoning 50 (2), 231–242.
Kohl, M., 2005. Numerical contributions to the asymptotic theory of robustness. Ph.D. Thesis, Universität Bayreuth.

Kriegler, E., Held, H., 2003. Climate projections for the 21st century using random sets. In: Bernard, J., Seidenfeld, T., Zaffalon, M. (Eds.), ISIPTA'03, Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, Lugano. Carleton Scientific, Waterloo, pp. 345–360.

Marazzi, A., 1993. Algorithms, routines, and S functions for robust statistics. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. The FORTRAN library ROBETH with an interface to S-PLUS. With the collaboration of Johann Joss and Alex Randriamiharisoa. With a separately available computer disk.

Millar, P., 1981. Robust estimation via minimum distance methods. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 55 (1), 73–89.

Öztürk, Ö., Hettmansperger, T., 1998. Simultaneous robust estimation of location and scale parameters: a minimum-distance approach. The Canadian Journal of Statistics 26 (2), 217–229.

Parr, W., Schucany, W., 1980. Minimum distance and robust estimation. Journal of the American Statistical Association 75 (371), 616–624.

Quaeghebeur, E., de Cooman, G., 2005. Imprecise probability models for inference in exponential families. In: Cozman, F., Nau, R., Seidenfeld, T. (Eds.), ISIPTA'05, Proceedings of the Fourth International Symposium on Imprecise Probabilities and their Applications, Pittsburg. SIPTA, Manno, pp. 287–296.

Rieder, H., 1994. Robust Asymptotic Statistics. Springer, New York.

Tukey, J., 1960. A survey of sampling from contaminated distributions. In: Contributions to Probability and Statistics.Stanford University Press, Stanford, CA, pp. 448–485.

van der Vaart, A., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.

van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes. With Applications to Statistics. Springer, New York.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London.

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

Walter, G., Augustin, T., 2009. Imprecision and prior-data conflict in generalized Bayesian inference. Journal of Statistical Theory and Practice 3, 255–271.

Weichselberger, K., 2001. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept. Physica, Heidelberg.