Journal of Statistical Planning and Inference 40 (1994) 205–220
North-Holland

Concentration functions and Bayesian robustness

Sandra Fortini and Fabrizio Ruggeri

CNR-IAMI, Via A.M. Ampère 56, I-20131 Milano, Italy

Received 1 June 1992; revised manuscript received 10 May 1993

Abstract

The concentration function, extending the classical notion of Lorenz curve, is well suited for comparing probability measures. Such a feature can be useful in different issues in Bayesian robustness, when a probability measure is deemed a baseline to be compared with other measures by means of their functional forms. Neighbourhood classes Γ of probability measures, including well-known ones, can be defined through the concentration function and both prior and posterior expectations of given functions of the unknown parameter are studied. The ranges of such expectations over Γ can be found, restricting the search among the extremal measures in Γ. The concentration function can be also used as a criterion to assess posterior robustness, when considering sensitivity to changes in the likelihood and the prior.

AMS Subject Classification: Primary 62F15; secondry 62F35.

Key words: Concentration function; Bayesian robustness; mixtures of probability measures; extremal probability measures.

1. Introduction

Cifarelli and Regazzini (1987) defined the concentration function of a probability measure P with respect to another one, say P_0, extending the classical notion of the Lorenz–Gini curve. By the concentration function, the discrepancy between two measures defined on the same probability space is studied, comparing the different concentrations of probability determined by the measures. As suggested by Regazzini (1992), the concentration function could be a valuable tool in robust Bayesian inference to analyse posterior probability measures under uncertainty about either prior measure or likelihood. Fortini and Ruggeri (1990) implemented such ideas and applied them to study the sensitivity of the posterior measures as the priors vary in an ε-contamination class. Such a sensitivity analysis, along with other issues in Bayesian robustness, has received much attention; see Berger (1984, 1985, 1990), Wasserman (1992) and the references contained therein.

In this paper, the role of the concentration function in Bayesian robustness is stressed, through applications to different issues. In Section 2, the use of the concentration function in Bayesian robustness is justified and a class Γ of priors is defined through a given class of concentration functions with respect to a base prior. In Section 3, some results in Fortini and Ruggeri (1992) are presented and applied to robust Bayesian inferences, so that the extremal points of Γ are identified and maximisation of prior and posterior quantities of interest can be performed over them, rather than all over Γ. In Section 4, some classes Γ, including some well-known ones, are defined through the concentration function and robustness analyses are performed over them. Some concluding remarks are contained in Section 5.

2. Comparison of probability measures

Bayesian robustness problems require the specification of criteria to analyse probability measures, for example to define neighbourhoods of an elicited prior or to compare the posterior probability measures as the prior varies in a given class. Suppose therefore that we are interested in comparing probability measures, say P and P_0, on the same measurable space $(\Theta, \mathscr{F})$, Θ being a Polish space and $\mathscr{F}$ its Borel σ-field. Many criteria have been proposed, such as probabilities of sets, means, etc. Such rules are often satisfactory but they usually say nothing about the functional form of the probability measures. Although rather neglected in Bayesian robustness, some criteria exist to compare functional forms, such as the variational distance $d_V(P, P_0)=\sup_{A\in\mathscr{F}} |P(A)-P_0(A)|$ and the Prohorov distance $d_P(P, P_0)=\inf\{\varepsilon>0: P(A)\leqslant P_0(A^\varepsilon)+\varepsilon \ \forall A\in\mathscr{F}\}$, where $A^\varepsilon=\{\theta\in\Theta: d(\theta, A)\leqslant\varepsilon\}$ and d is a metric on Θ.

However, such rules do not seem sufficiently sensitive on the sets with small probability under P_0. For example, if the variational metric is considered, then an ε-neighbourhood of P_0 contains all the probability measures P such that, for any $A\in\mathscr{F}$, $|P(A)-P_0(A)|\leqslant\varepsilon$. Consider a new set E such that $P_0(E)=\varepsilon/10$. Given P in the ε-neighbourhood of P_0, it follows that $P(E)\leqslant 11/10\varepsilon$ is the only restriction about P on E; i.e. P is considered to be close to P_0 even if its value on E is 11 times greater than $P_0(E)$. A similar reasoning holds for the ε-contamination class of priors, described in Section 4, which contains all the probability measures P such that, for any $A\in\mathscr{F}$, $(1-\varepsilon)P_0(A)\leqslant P(A)\leqslant(1-\varepsilon)P_0(A)+\varepsilon$. When such a consequence is deemed inconvenient, then different bounds on $P(A)$ could be considered and the concentration function (c.f.) is a flexible tool to get them. As an example, require, for any $A\in\mathscr{F}$, either $|P_0(A)-P(A)|\leqslant\varepsilon\min\{P_0(A), 1-P_0(A)\}$ or $|P_0(A)-P(A)|\leqslant P_0(A)(1-P_0(A))$, so that more stringent bounds are found on $P(E)$; in the former case, we have $(\varepsilon/10)(1-\varepsilon)\leqslant P(E)\leqslant(\varepsilon/10)(1+\varepsilon)$, while the latter implies $(\varepsilon^2/100)\leqslant P(E)\leqslant(\varepsilon/10)(2-\varepsilon/10)$, i.e. $P(E)$ does not exceed twice $P_0(E)$.

It is sometimes worth comparing functional forms of probability measures, both a priori and a posteriori. In the former case, it is reasonable to choose measures which

are 'functionally close' to an elicited prior P_0, allowing small changes in the concentration of the probability, due to errors in the elicitation process. In the latter case, the functional closeness could be worth comparing when we are interested in the posterior measures themselves, e.g. when we accept the point of view of 'some Bayesians (who) maintain that inference should ideally consist of simply reporting the *entire* posterior distribution ...' (Berger, 1985, p. 133).

Comparison of functional forms has received little attention in Bayesian robustness, mainly because the classes considered are not easy to work with. We think that the c.f. could simplify such task, by finding workable classes. Classes K_g of probability measures can be defined through the c.f.'s, as neighbourhoods around a base measure P_0, which is assumed nonatomic in this paper.

Definition 1. If $g:[0,1]\to[0,1]$ is a continuous, convex, monotone nondecreasing function with $g(0)=0$, then the set

$$K_g=\{P: P(A)\geqslant g(P_0(A)) \ \forall A\in\mathscr{F}\} \tag{1}$$

will be said to be a g-neighbourhood of P_0.

Observe that, if $P\in K_g$, then $g(P_0(A))\leqslant P(A)\leqslant 1-g(1-P_0(A))$. As proved in Fortini and Ruggeri (1992), $\{K_g\}$ generates a topology in which it becomes a fundamental system of neighbourhoods of P_0, when g belongs to an adequate class G of continuous, convex, monotone nondecreasing functions.

The requirement $g(0)=0$ is needed to avoid $P(\Theta)\leqslant 1-g(0)<1$, while monotonicity, continuity and convexity are thoroughly discussed in Fortini and Ruggeri (1992), as quite natural requirements from the definition of probability measure on a σ-field.

The definition of g-neighbourhood can be reformulated by means of the concentration function, which generalises the Lorenz curve, described in Marshall and Olkin (1979, p. 5). The classical definition of concentration refers to the discrepancy between a discrete probability P and a uniform one, say P_0. Cifarelli and Regazzini (1987) defined the c.f. of P with respect to (w.r.t.) P_0, where P and P_0 are two probability measures on the same measurable space $(\Theta, \mathscr{F})$. According to the Radon–Nikodym theorem, there is a unique partition $\{N, N^C\}\subset\mathscr{F}$ of Θ and a nonnegative function h on N^C such that

$$P(E)=\int_{E\cap N^C} h(\theta)P_0(\mathrm{d}\theta)+P_s(E\cap N), \quad \forall E\in\mathscr{F},$$

$$P_0(N)=0, \qquad P_s(N)=P_s(\Theta),$$

where

$$P_a(\cdot)=\int_{\cdot\cap N^C} h(\theta)P_0(\mathrm{d}\theta)$$

and P_s denote the absolutely continuous and the singular part of P w.r.t. P_0, respectively. Set $h(\theta)=\infty$ all over N and define

$$H(y)=P_0(\{\theta\in\Theta: h(\theta)\leqslant y\}), \qquad c(x)=\inf\{y\in\Re: H(y)\geqslant x\}.$$

Finally, let

$$L(x)=\{\theta\in\Theta: h(\theta)\leqslant c(x)\} \quad \text{and} \quad L^-(x)=\{\theta\in\Theta: h(\theta)<c(x)\}.$$

Definition 2. The function $\varphi:[0,1]\to[0,1]$ is said to be the concentration function of P w.r.t. P_0 if $\varphi(x)=P(L^-(x))+c(x)\{x-H(c(x)^-)\}$ for $x\in(0,1)$, $\varphi(0)=0$ and $\varphi(1)=P_a(\Theta)$.

When the dependence on P is to be emphasized, the c.f. will be denotd $\varphi_P(x)$. As proved in Cifarelli and Regazzini (1987), $\varphi(x)$ is a nondecreasing, continuous and convex function such that

$$\varphi(x)=\int_0^{c(x)}\{x-H(t)\,\mathrm{d}t=\int_0^x c(t)\,\mathrm{d}t$$

and

$$\varphi(x)\equiv 0 \Leftrightarrow P\perp P_0, \qquad \varphi(x)=x\ \forall x\in[0,1] \Leftrightarrow P=P_0.$$

As an example, the c.f. $\varphi(x)$ of $P\sim\mathcal{G}(2,2)$ w.r.t. $P_0\sim\mathcal{E}(1)$ is plotted in Figure 1 and it is shown that [0.216, 0.559] is the range spanned by the probability, under P, of the sets A with $P_0(A)=0.4$. Such a range follows from Theorem 2.2 in Cifarelli and Regazzini (1987), which provides an interesting interpretation of the c.f., especially for its use in Bayesian robustness; in fact, given *any* $x\in[0,1]$, then the probability, under P, of any A with P_0-measure x, is such that $\varphi(x)\leqslant P(A)\leqslant 1-\varphi(1-x)$. Under P_0 nonatomic, such a theorem can be expressed as follows.

Theorem 1. *Given any* $x\in[0,1]$, *then* B_x *exists such that* $P_0(B_x)=x$ *and* $\varphi(x)=P_a(B_x)=\min\{P(A): A\in\mathcal{F}$ *and* $P_0(A)\geqslant x\}$.

Theorem 1 allows g-neighbourhoods to be expressed by means of c.f.s.

Proposition 1. *The set* $K_g=\{P: \varphi_P(x)\geqslant g(x), \forall x\in[0,1]\}$ *is g-neighbourhood of* P_0 *as defined in* (1).

Definition 3. A function $g:[0,1]\to[0,1]$ is said to be compatible if g is a monotone nondecreasing, continuous, convex function, with $g(0)=0$.

Fortini and Ruggeri (1992) proved that any compatible g is a c.f.

Theorem 2. *Given a function* $g:[0,1]\to[0,1]$, *there exists at least one measure* P *such that* g *is the c.f. of* P *w.r.t.* P_0 *if and only if* g *is compatible.*

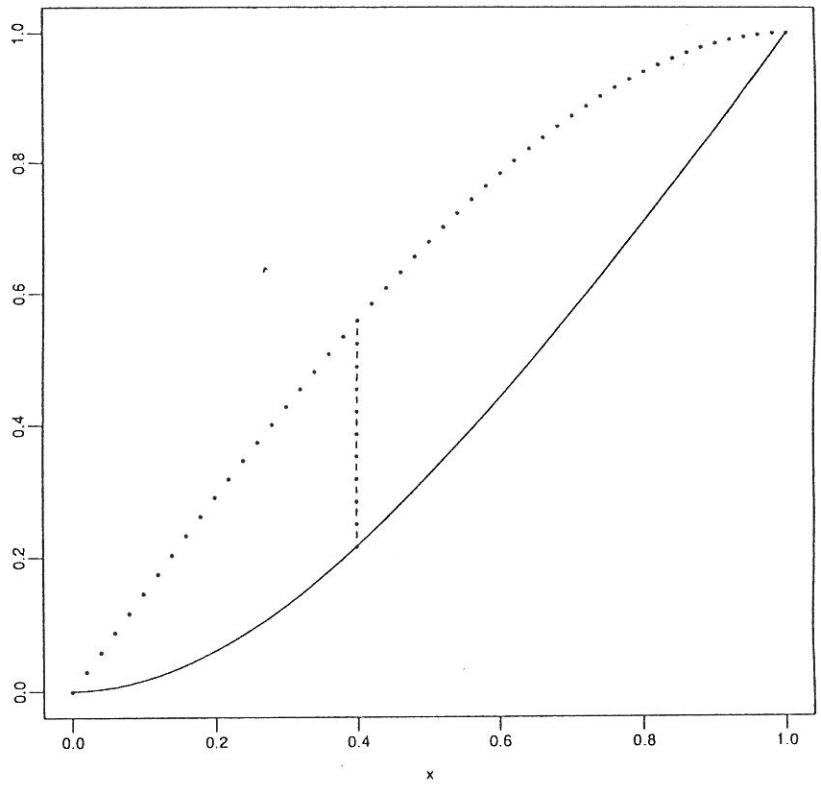

Fig. 1. Concentration functions $\varphi(x)$ (——) and $1-\varphi(1-x)$ ($\cdots$) of $P\sim\mathcal{G}(2,2)$ w.r.t. $P_0\sim\mathcal{E}(1)$.

3. Representation theorems

Consider the space $\mathcal{P}$ of all probability measures on Θ endowed with the weak topology. It is well known that $\mathcal{P}$ can be metrised as a complete separable metric space. Consider the set of extremal points of K_g, that is the probability measures $P\in K_g$ such that

$$P=\alpha P_1+(1-\alpha)P_2, \quad P_1\in K_g, P_2\in K_g, 0<\alpha<1 \Leftrightarrow P=P_1=P_2.$$

The following results were proved by Fortini and Ruggeri (1992).

Proposition 2. *The set of all the extremal points of* K_g *is contained in* E_g, *where* $E_g=\{P: \varphi_P(x)=g(x), \forall x\in[0,1]\}$. *If* $g(1)=1$, *then it coincides with* E_g.

Furthermore, every probability measure whose c.f. is greater than g can be represented as a mixture of probability measures having g as c.f., applying the Choquet theorem (Phelps, 1966).

Theorem 3. *Let the function $g:[0,1]\to[0,1]$ be compatible. Then for any probability measure $\bar{P}\in K_g$, there exists a probability measure $\mu_{\bar{P}}$ on $\mathscr{P}$ such that $\mu_{\bar{P}}(F_g)=1$ and $\bar{P}=\int_{\mathscr{P}} P\,\mu_{\bar{P}}(\mathrm{d}P)$, where $F_g\subseteq E_g$ is the set of the extremal points of K_g.*

In Section 4, it will be shown that F_g can be a proper subset of E_g, as in the cases of ε-contaminated and total variation neighbourhoods. The next theorem is proved similarly to Lemma A.1 in Sivaganesan and Berger (1989).

Theorem 4. *Let f and g be real-valued functions on Θ such that $\int_\Theta |f(\theta)|P(\mathrm{d}\theta)<\infty$ and $0<\int_\Theta g(\theta)P(\mathrm{d}\theta)<\infty$ for any $P\in K$. Then*

$$\sup_{P\in K_*}\frac{\int_\Theta f(\theta)P(\mathrm{d}\theta)}{\int_\Theta g(\theta)P(\mathrm{d}\theta)}=\sup_{P\in E_*}\frac{\int_\Theta f(\theta)P(\mathrm{d}\theta)}{\int_\Theta g(\theta)P(\mathrm{d}\theta)}.$$

The same result holds with 'sup' replaced by 'inf'.

Computations of bounds on prior expectations are simplified by the next theorem.

Theorem 5. *Let*

$$H_f(y)=P_0(\{\theta\in\Theta: f(\theta)\leqslant y\}),\qquad c_f(x)=\inf\{y: H_f(y)\geqslant x\}.$$

Then

$$\sup_{P\in K_*}\int_\Theta f(\theta)P(\mathrm{d}\theta)=\int_0^1 c_f(x)c(x)\,\mathrm{d}x,$$

where $c(x)=g'(x)$ a.e.

Such a result can be applied to find bounds on posterior expectations, too, using the linearisation technique presented by Lavine (1988).

4. Bayesian robustness

The results in Section 3 are used in building a class of prior measures in a neighbourhood of a given one and checking if inferences lead to posterior measures close to a base one. In the former case, a class of prior measures K_g is determined such that their c.f. with respect to a nonatomic base one, say P_0, is pointwise not smaller than a specified compatible function g. Such a function gives the maximum concentration of a measure w.r.t. a base one which is deemed compatible with our knowledge. It will be shown that some well-known neighbourhood classes can be described in such a way. Both prior and posterior expectations of any function $f(\theta)$, $\theta\in\Theta$, say $E(f)$ and $E(f|x)$ respectively, are maximised (or minimised) all over K_g by applying Theorems 4 and 5. Let $\bar{E}(f)$ and $\bar{E}(f|x)$ be such maximum values over K_g. Finally, the latter case specifies a bounding function again, denoting the maximum

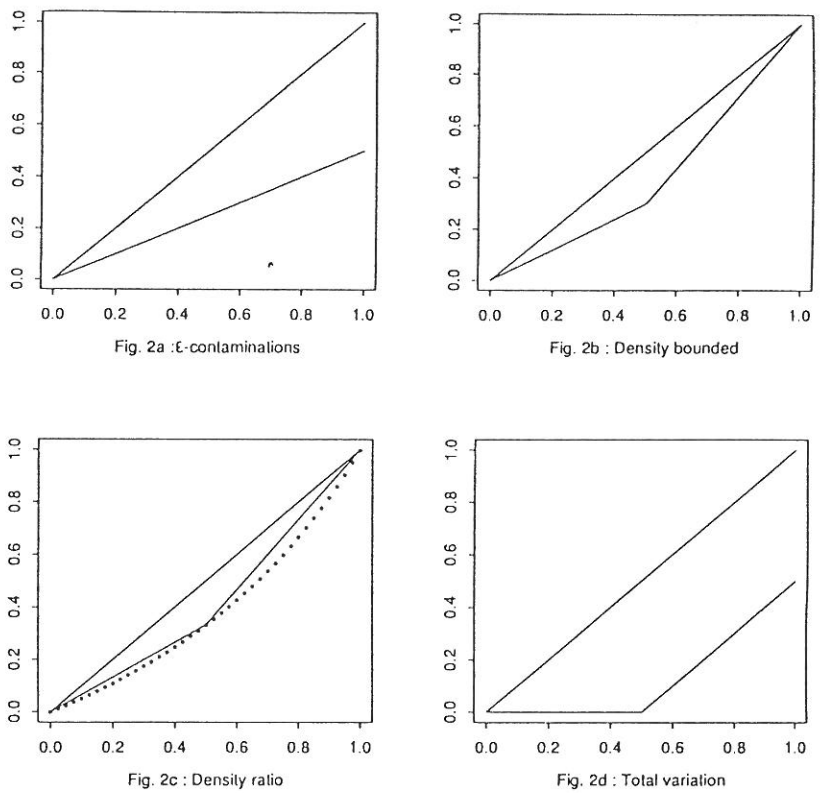

Fig. 2. Examples of classes defined through the concentration function.

concentration allowed so that a posterior measure is deemed acceptable. Since P_0 is nonatomic, the discrete measures can or cannot be ruled out by choosing $g(1)=1$ or <1, respectively.

4.1. Classes of prior measures

Some well-known classes are presented by means of the corresponding classes of c.f.'s, defined by the functions $g(x)$ plotted in Figure 2.

4.1.1. ε-Contaminations

Given a probability measures P_0 and $\varepsilon\in[0,1]$, the class $\Gamma_\varepsilon=\{P_Q=(1-\varepsilon)P_0+\varepsilon Q, Q\in\mathscr{Q}\}$, where $\mathscr{Q}\subseteq\mathscr{P}$, is said to be an ε-contamination class of priors. It was proved, by Fortini and Ruggeri (1990), that the c.f. $\varphi(x)$ of P_Q w.r.t. P_0 is such that $\varphi(x)=(1-\varepsilon)x+\varepsilon\varphi_0(x)$, where $\varphi_0(x)$ is the c.f. of Q w.r.t. P_0. Considering $g(x)=(1-\varepsilon)x$, $\forall x\in[0,1]$, it can be easily shown that $\Gamma_\varepsilon=K_g$ when $\mathscr{Q}=\mathscr{P}$, while E_g and F_g are obtained, respectively, for singular w.r.t. P_0 and Dirac contaminating

measures Q's, observing that any probability measure is a mixture of Dirac measures and the c.f. of a Dirac measure w.r.t. to a nonatomic measure is $\varphi_0 \equiv 0$. As shown in Berger (1990), $E(f)$ and $E(f|x)$ are maximised by contaminating Dirac measures, i.e. over F_g.

Taking $\varepsilon = 1$, the set $\mathscr{P}$ of all the probability measures can be considered as a special case of ε-contamination class of priors, so that $g(x) = 0, \forall x \in [0, 1]$, and any measure is a mixture of Dirac measures and $E(f)$ and $E(f|x)$ are maximised by one of them (see Sivaganesan and Berger, 1989).

4.1.2. Density-bounded class

The density-bounded classes $\Gamma^{\mathrm{B}}_{L,U}$ were firstly defined by Lavine (1991), while a special case, emphasising their role as a neighbourhood class, has been considered by some authors, e.g. Ruggeri and Wasserman (1991). Lavine defined $\Gamma^{\mathrm{B}}_{L,U}$ to be the set of probability measures P that satisfy $L(A) \leqslant P(A) \leqslant U(A)$ for all measurable A where L and U are measures such that $L(\Theta) \leqslant 1 \leqslant U(\Theta)$. When all the probability measures P have a density $p(\theta)$ w.r.t. some dominating measure λ, the class $\Gamma^{\mathrm{B}}_{L,U}$ is such that $l(\theta) \leqslant p(\theta) \leqslant u(\theta)$ a.e., where $l(\theta)$ and $u(\theta)$ are the densities of L and U, respectively. Given a probability measure P_0, consider the special case, studied by Ruggeri and Wasserman (1991), where $L = (1/k)P_0$, $U = kP_0$, $k \geqslant 1$, which will be denoted by Γ^{B}_k. From a point of view of the c.f., such a class Γ^{B}_k can be seen as a special case of K_g, where

$$g(x) = \max\left\{\frac{\beta}{\alpha}x, \frac{1-\beta}{1-\alpha}(x-1)+1\right\}.$$

It follows that $P \in E_g$ if and only if $K \in \mathscr{F}$ exists such that $P_0(K) = \alpha$ and it has density

$$p(\theta) = \frac{\beta}{\alpha}p_0(\theta)I_K(\theta) + \frac{1-\beta}{1-\alpha}p_0(\theta)I_{K^c}(\theta),$$

where I_A is the indicator function of the subset A.

The class Γ^{B}_k is obtained by taking $\alpha = k/(1+k)$ and $\beta = 1/(1+k)$ so that $g(x) = \max\{x/k, k(x-1)+1\}$. It can be easily shown that $\Gamma^{\mathrm{B}}_k = K_g$ and the maximising priors in E_g confirm known results, such as $\bar{E}(f)$ over $\Gamma^{\mathrm{B}}_{L,U}$ which was computed by Lavine (1991), while Ruggeri and Wasserman (1991) considered Γ^{B}_k and gave, in addition, bounds on the maximum of posterior expectations. By means of the c.f.s, the maximising measures are at least identified, even if actual computations remain a hard task.

Another case is obtained by taking $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}(1-\varepsilon)$ so that $g(x) = x - \varepsilon \min\{x, 1-x\}$. Such a case corresponds to the requirement, discussed in Section 2, that

$$\sup_{A \in \mathscr{F}: P_0(A) = x} |P_0(A) - P(A)| \leqslant \varepsilon \min\{P_0(A), 1 - P_0(A)\}.$$

4.1.3. Density ratio class

The density ratio classes were firstly defined by DeRobertis and Hartigan (1981), while a special case, emphasising their role as a neighbourhood class, has been considered by some authors, e.g. Ruggeri and Wasserman (1991). Given a probability measure P_0 with density $p_0(\theta)$, Ruggeri and Wasserman considered the density ratio neighbourhood around P_0, Γ^{DR}_k, defined as the set of all the probability measures whose densities $p(\theta)$ are such that there exists a $c > 0$ so that $p_0(\theta) \leqslant cp(\theta) \leqslant kp_0(\theta)$ for almost all θ. It can be proved that Γ^{DR}_k is also the class of the probability measures such that their c.f.'s w.r.t. P_0 are inside a triangle with vertices $(0,0)$, $(1,1)$ and any point on the curve $g(x) = x/[k-(k-1)x]$, $0 < x < 1$. Since g is compatible, it follows that $\Gamma^{\mathrm{DR}}_k \subset K_g$. Bounds on expectations in such classes are compared in Wasserman and Kadane (1992). Observe that it can be proved that

$$K_g = \left\{P: \frac{P(A)}{P(A^{\mathrm{C}})} \leqslant k\frac{P_0(A)}{P_0(A^{\mathrm{C}})}, \forall A \in \mathscr{F}\right\},$$

which resembles an equivalent definition of Γ^{DR}_k given by

$$\Gamma^{\mathrm{DR}}_k = \left\{P: \frac{p(\theta)}{p(\theta')} \leqslant k\frac{p_0(\theta)}{p_0(\theta')}, \quad \forall \theta, \theta' \in \Theta \text{ a.c.}\right\}.$$

Computations of $\bar{E}(f)$ and $\bar{E}(f|x)$ in Γ^{DR}_k are made possible by the results due to DeRobertis and Hartigan (1981) and Ruggeri and Wasserman (1991).

4.1.4. Total variation neighbourhood

A class Γ^{T} is said to be a total variation neighbourhood of a probability measure P_0 if it contains all the probability measures P that satisfy $\sup_{A \in \mathscr{F}} |P(A) - P_0(A)| \leqslant \varepsilon$, given a fixed $\varepsilon \in [0, 1]$. Since P_0 is nonatomic, then the measures in E_g coincide with P_0 over a subset B_1 such that $P_0(B_1) = 1 - \varepsilon$, give a total mass ε to B_2 where $P_0(B_2) = 0$ and vanish elsewhere. F_g is obtained by considering the measures in E_g giving the mass ε to a unique point.

Computations of $\bar{E}(f)$ are made possible by the results due to Wasserman and Kadane (1990). They also found bounds on posterior probabilities, while here the measures maximising $E(f|x)$ are identified.

4.1.5. Neighbourhood of the uniform distribution

The behaviour of the inferences when considering a neighbourhood of the uniform distribution has been considered in some recent works (e.g. Wasserman and Kadane, 1992). In the next example, we consider a neighbourhood which is given by the c.f.

Example 1. Suppose that a coin is flipped twice and that θ is the probability of getting the 'head' in a flip. A uniform distribution P_0 is a possible choice as prior measure

on θ. Such a choice can be modified, considering measures which slightly differ from it, by choosing P such that

$$\sup_{A\in\mathcal{F}:\,P_0(A)=x} |P_0(A)-P(A)| \leqslant P_0(A)\,(1-P_0(A)).$$

Hence, we consider the random variable $X\sim \mathrm{Bin}(2,\theta)$, with density $f(x|\theta)$, and the class $K_g=\{P:\ \varphi_P(x)\geqslant x^2,\ \forall\, x\in[0,1]\}$.

Since $g(1)=1$, maximisations of $E(f)$ and $E(f|x)$ are made over E_g, which contains all the probability measures such that $c(x)=2x$, $x\in[0,1]$. In particular, the next computations are made by applying Theorem 5.

As pointed out in Berger (1985, Sections 3.5.1, 4.7.2), marginal distributions are sometimes relevant, e.g. to construct the prior or to check assumptions on it and the model. A well-known method to select priors is based on the ML-II approach (see Berger, 1985, Section 3.5.4). Given a class Γ of probability measures and the observed data x, then $\hat{P}\in\Gamma$ is said to be an ML-II prior if it maximises, over all Γ, the marginal density $m(x|P)=\int_\Theta f(x|\theta)P(\mathrm{d}\theta)$.

When $x=0$ or $x=2$, then the maximum value of the marginal density is $\frac{1}{2}$, corresponding, respectively, to the ML-II priors $\hat{P}_0$ and $\hat{P}_2$ having densities $\hat{p}_0(\theta)=2(1-\theta)$ and $\hat{p}_2(\theta)=2\theta$, $\theta\in[0,1]$. When $x=1$, then the maximum marginal density is $\frac{5}{12}$, corresponding to the ML-II prior $\hat{P}_1$ having density $\hat{p}_1(\theta)=4\,\min\{\theta,1-\theta\}$, $\theta\in[0,1]$.

Given $x=1$, we compute now the upper and lower bounds, $\bar{\rho}$ and $\underline{\rho}$ respectively, on the posterior probability $P(A|x)$ of the subset $A=[0,\frac{1}{2}]$, when the prior measure varies in K_g. Hence, $\bar{\rho}=\frac{13}{16}$ is found taking $\hat{P}$ with density $\hat{p}(\theta)=2\theta+I_{[0,1/2]}(\theta)-I_{(1/2,1]}(\theta)$. Computing $\underline{\rho}$ similarly, it results that $\frac{3}{16}\leqslant P(A|1)\leqslant\frac{13}{16}$.

4.2. *Sensitivity to the prior*

The next example will show how the c.f. is used to assess robustness w.r.t. changes in the prior, when a function $g(x)$ is assumed as maximum tolerable concentration of a measure w.r.t. to a base one.

Example 2. Consider the random variable $X\sim\mathrm{Bin}(1,\theta)$ with density $f(x|\theta)$. Take $P_0\sim\mathcal{U}(0,1)$ and suppose that two experts elicit Beta priors with different parameters, i.e. $P_{2,1}\sim\mathrm{Be}(2,1)$ and $P_{2,2}\sim\mathrm{Be}(2,2)$. Let $\varphi_{2,1}$ and $\varphi_{2,2}$ denote the respective c.f.'s w.r.t. P_0. Such priors are deemed compatible with the prior knowledge if they do not differ 'too much' from P_0 and here we measure such a difference by means of their c.f.'s which should be not smaller than $g(x)=x^2$ (as in Example 1). Actually, the c.f.'s satisfy such a requirement, because it follows that $\varphi_{2,1}(x)=g(x)$ and $\varphi_{2,2}(x)=(\frac{3}{2})x^2-\frac{1}{2}x^3\geqslant g(x)$, for all $x\in[0,1]$.

Consider now a sample $\tilde{x}$ from X, so that the likelihood function becomes $l_{\tilde{x}}(\theta)=\theta^{\tilde{x}}(1-\theta)^{1-\tilde{x}}$ while the posterior measures become $P_0^*\sim\mathrm{Be}(1+\tilde{x},2-\tilde{x})$, $P_{2,1}^*\sim\mathrm{Be}(2+\tilde{x},2-\tilde{x})$ and $P_{2,2}^*\sim\mathrm{Be}(2+\tilde{x},3-\tilde{x})$. Robustness is checked by comparing $g(x)=x^2$ with the c.f.'s of $P_{2,1}^*$ and $P_{2,2}^*$ w.r.t. P_0^*, denoted, respectively, by $\varphi_{2,1}^*$ and $\varphi_{2,2}^*$. Given $\tilde{x}=1$, then it follows that $\varphi_{2,1}^*(x)=x^{3/2}\geqslant g(x)$ and $\varphi_{2,2}^*(x)=\frac{3}{2}x^2-\frac{1}{2}x^3\geqslant g(x)$, for all $x\in[0,1]$. Given $\tilde{x}=0$, it follows that $\varphi_{2,1}^*(x)=3x-2+2(1-x)^{3/2}\leqslant g(x)$ and $\varphi_{2,2}^*(x)=\frac{3}{2}x^2-\frac{1}{2}x^3\geqslant g(x)$, for all $x\in[0,1]$. It is clear that the robustness is achieved if and only if the sample $\tilde{x}=1$ is obtained.

5. Discussion

The c.f. is useful in comparing functional forms of the probability measures; such a comparison is well justified in Bayesian inferences (as it is also in robust ones) when entire posteriors are reported, rather than some of their features (mean, HPD, etc.). As discussed in Berger (1985, p. 144), reporting of the entire postrior measure is preferred by many Bayesians, as opposed to an HPD, which is not necessarily invariant under reparameterisation.

Given the probability measures in a g-neighbourhood of P_0, it would be interesting to check if, or when, the corresponding posterior measures form a g^*-neighbourhood of P_0^* (the posterior form P_0) or a proper subset of it. Proper inclusion holds for the ε-contamination class of priors, described in Section 4.1, when the function $g(x)=(1-\varepsilon)x$ is transformed into $g^*(x)=Cg(x)$, where the constant C is computed in Fortini and Ruggeri (1990). Such a C, or other indices like the Gini's area of concentration $2\int_0^1\{x-g(x)\}\,\mathrm{d}x$, could measure the effect of the data on the distribution of the parameter.

Finally, it should be observed that g-neighbourhoods are also special capacities, as defined in Buja (1986) and Bednarski (1981); furthermore, they are symmetric upper probabilities, as in Wasserman and Kadane (1992), where g is the lower distribution function and the convexity is equivalent to the two-alternating condition.

Acknowledgement

The authors would like to thank Eugenio Regazzini for inspiring their work.

References

Bednarski, T. (1981). On solutions of minimax test problems for special capacities. *Z. Wahrsch. Verw. Gebiete* **58**, 397–405.

Berger, J. (1984). The robust Bayesian viewpoint (with discussion). In: J. Kadane, Ed., *Robustness of Bayesian Analyses*. North-Holland, Amsterdam.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

Berger, J. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference* **25**, 303–328.

Buja, A. (1986). On the Huber–Strassen theorem. *Probab. Theory Relat. Fields* **73**, 367–384.

Cifarelli, D.M. and E. Regazzini (1987). On a general definition of concentration function. *Sankhya B* **49**, 307–319.
DeRobertis, L. and J. Hartigan (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235–244.
Fortini, S. and F. Ruggeri (1990). Concentration function in a robust Bayesian framework. Quaderno IAMI 90.6. CNR-IAMI, Milano.
Fortini, S. and F. Ruggeri (1992). On defining neighbourhoods of measures through the concentration function. *Sankhya A* (to appear).
Lavine, M. (1988). Prior influence in Bayesian statistics. Discussion paper 88-06. Institute of Statistics and Decision Sciences, Duke University, Durham.
Lavine, M. (1991). An approach to robust Bayesian analysis for multidimensional spaces. *J. Am. Statist. Assoc.* **86**, 400–403.
Marshall, A.W. and I. Olkin (1979). *Inequalities: Theory of Majorization and its Applications.* Academic Press, New York.
Phelps, R.R. (1966). *Lectures on Choquet's Theorem.* Van Nostrand, Princeton, NJ.
Regazzini, E. (1992). Concentration comparisons between probability measures. *Sankhya B* (to appear).
Ruggeri, F. and L. Wasserman (1991). Density based classes of priors: infinitesimal properties and approximations. Technical Report 528, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
Sivaganesan, S. and J. Berger (1989). Ranges of posterior measures for priors with unimodal contaminations. *Ann. Statist.* **17**, 868–889.
Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Eds., *Bayesian Statistics, Vol. 4.* Oxford University Press, Oxford.
Wasserman, L. and J. Kadane (1990). Bayes' theorem for Choquet capacities. *Ann. Statist.* **18**, 1328–1339.
Wasserman, L. and J. Kadane (1992). Symmetric upper probabilities. *Ann. Statist.* **20**, 1720–1736.

Comments on 'Concentration functions and Bayesian robustness' by Sandra Fortini and Fabrizio Ruggeri

Thomas Sellke

Purdue University, West Lafayette, IN, USA

The paper by Fortini and Ruggeri presents some interesting connections between concentration functions and Bayesian robustness. I especially liked the observations in Section 4 that certain standard class of priors, including the ε-contamination, density-bounded, and density ratio classes, can be described in terms of concentration functions.

The first section below considers the 'large sample theory' of concentration function classes of priors. The second section makes a somewhat related (but much less important) observation about classes of posteriors. The third section points out a fairly obvious weakness in the use of concentration function neighborhoods as classes of prior distributions.

1. Large sample theory

A fundamental issue concerning the use of concentration function neighborhoods K_g as classes of priors is the behavior of the corresponding classes of posteriors when the amount of information in the data is large. For example, the well-known *principle of stable estimation* (see Edwards et al. (1963), or Berger (1985)) says that the posterior density $\pi(\theta|y)$ will be approximately proportional to the likelihood $l(\theta|y)$ for data y if the likelihood is concentrated on a set of θ's where the prior density $\pi(\theta)$ is approximately constant. Thus, all reasonably smooth prior densities will yield approximately the same posterior in the presence of enough data. Is there an analogous 'large sample' result in the present context? Under mild conditions, the answer is 'yes', as is explained below.

Suppose that $g(1)=1$ and that $0<g'(0)\leqslant g'(1)<\infty$, where $g'(0)$ is of course the right-hand derivative of g at 0 and $g'(1)$ is the left-hand derivative at 1. Write $\Gamma^{\mathrm{DR}}(P_0;k)$ for the density ratio class Γ_k^{DR} defined in Section 4 of Fortini and Ruggeri to emphasize the dependence on P_0. Set $k_0=g'(1)/g'(0)$ and $k_\varepsilon=[1-g(1-\varepsilon)]/g(\varepsilon)$, $0<\varepsilon<1$. Note that $k_\varepsilon\uparrow k_0$ as $\varepsilon\downarrow 0$.

For data y, let $K_g|y$ be the class of posterior distributions corresponding to priors in K_g. It is easy to show that

$$K_g\subset\Gamma^{\mathrm{DR}}(P_0;k_0), \tag{1}$$

so that

$$K_g|y\subset\Gamma^{\mathrm{DR}}(P_0;k_0)|y. \tag{2}$$

Under the assumptions above that $g(1)=1$ and $0<g'(0)\leqslant g'(1)<\infty$, it turns out that (2) is an approximate equality when $P_0|y$ (the posterior distribution corresponding to prior P_0 and data y) is concentrated on a set B of small P_0-probability.

To give a more exact statement of the claim that (2) is an approximate equality for large samples, let us define a distance function between classes of distributions. For probability distributions P_1 and P_2 on the parameter space Θ, let

$$d(P_1,P_2)=\sup_{A\subset\Theta}|P_1(A)-P_2(A)| \tag{3}$$

be the total variation distance between them. For a class K of probability distributions on Θ, define the total variation distance from P_1 to K by

$$d(P_1,K)=\inf_{P_2\in K} d(P_1,P_2). \tag{4}$$

Finally, for two classes K_1 and K_2, define the total variation distance between them by

$$d[K_1\,\|\,K_2]=\max\left\{\sup_{P_1\in K_1} d(P_1,K_2),\ \sup_{P_2\in K_2} d(P_2,K_1)\right\}. \tag{5}$$

Theorem. *For $B\subset\Theta$, let $P_0(B)=\varepsilon$ be the prior probability of B under P_0, and let $P_0(B|y)=1-\delta$ be the posterior probability of B under P_0, given data y. Then*

$$d[K_g|y \,\|\, \Gamma^{DR}(P_0;k_0)|y] \leqslant 2\delta k_0 + (k_0 - k_\varepsilon). \tag{6}$$

Thus, if $y_1, y_2, \ldots$ is a sequence of increasingly informative data in the sense that there are sets $B_n\subset\Theta$ with $P_0(B_n)\downarrow 0$ and $P_0(B_n|y_n)\uparrow 1$, then the total variation distance between $K_g|y$ and $\Gamma^{DR}(P_0;k_0)|y$ will converge to 0.

Here is a sketch of the proof of the theorem. It is easy to check that

$$\Gamma^{DR}(P_0;k)|y = \Gamma^{DR}(P_0|y;k). \tag{7}$$

It is easy to check that

$$\Gamma^{DR}(P_0;k)|B = \Gamma^{DR}(P_0|B;k). \tag{8}$$

(Indeed, (8) is a special case of (7).) For $B\subset\Theta$ with $P_0(B)<\varepsilon$, a straightforward argument shows that

$$\Gamma^{DR}(P_0|B;k_\varepsilon)\subset K_g|B\subset\Gamma^{DR}(P_0|B;k_0). \tag{9}$$

By (7) and (9),

$$\Gamma^{DR}(P_0|y,B;k_\varepsilon)\subset K_g|y, \qquad B\subset\Gamma^{DR}(P_0|y,B;k_0). \tag{10}$$

Now note that for any P, and $k_2>k_1\geqslant 1$,

$$d[\Gamma^{DR}(P;k_1)\,\|\,\Gamma^{DR}(P;k_2)]\leqslant k_2-k_1. \tag{11}$$

Applying this to (10) yields

$$\begin{aligned} d[K_g|y,B\,\|\,\Gamma^{DR}(P_0|y,B;k_0] &\leqslant d[\Gamma^{DR}(P_0|y,B;k_\varepsilon]\,\|\,\Gamma^{DR}(P_0|y,B;k_0)] \\ &\leqslant k_0-k_\varepsilon. \end{aligned} \tag{12}$$

Also, for any class K_1 of probability distributions on Θ and any $B\subset\Theta$,

$$d[K_1|B\,\|\,K_1]\leqslant \sup_{P_1\in K_1} P_1(B^c). \tag{13}$$

Thus, by (8) and (13),

$$d[\Gamma^{DR}(P_0|y,B;k_0)\,\|\,\Gamma^{DR}(P_0|y;k_0)]\leqslant k_0 P_0(B^c|y)=k_0\delta. \tag{14}$$

Likewise,

$$d[K_g|y,B\,\|\,K_g|y]\leqslant k_0\delta. \tag{15}$$

The total variation distance (5) satisfies the triangle inequality, so (12), (14), and (15)

2. 'Weak dilation' of concentration functions

Again, let K_g be a concentration function neighborhood of P_0, and let $K_g|y$ be the corresponding posterior distributions for data y. When P_0 is nonatomic, the class $K_g|y$ will contain posterior distributions whose concentration functions with respect to $P_0|y$ are as close to g as desired. Under mild conditions (e.g. multidimensional Θ and a smooth likelihood function), it will even be true that $K_g|y$ contains distributions whose concentration functions with respect to $P_0|y$ are exactly g.

The gist of what is going on here may be understood by considering the case of a likelihood function which takes on only finitely many values. Let $B_1,\ldots,B_k$ be the subsets of Θ of positive P_0-probability upon which the likelihood function takes on its k possible values. Define a probability P_1 so that the concentration of $P_1|B_i$ relative to $P_0|B_i$ is g for each B_i. (That such a P_1 exists follows from a theorem in Fortini and Ruggeri (1992).) Then the concentration function of $P_1|y$ with respect to $P_0|y$ will be exactly g.

In general, if one can find a P_1 with concentration function g w.r.t. P_0 for which the Radon–Nikodym derivative $dP_1/d(P_0+P_1)$ is independent of the likelihood function under both P_0 and P_1, then the concentration function of $P_1|y$ w.r.t. $P_0|y$ will be g.

It follows from the above observations that, under mild conditions, any concentration function neighborhood centered at $P_0|y$ which contains $K_g|y$ must contain the g-neighborhood of $P_0|y$.

3. Quibbles

The papers by Fortini and Ruggeri as well as the cited papers by Cifarelli and Regazzini show that concentration functions are objects of basic mathematical interest and that concentration functions provide a unifying and enlightening perspective on some statistical issues.

However, a major weakness in the use of concentration function neighborhoods to study Bayesian robustness is that concentration function neighborhoods ignore the topology of the parameter space. It will typically be the case that the prior distributions of real interest have smooth densities (or at least some local smoothness properties) so that the principle of stable estimation will usually apply when the amount of information is large. Concentration function neighborhoods contain distributions whose densities are extremely irregular. Of course, it is often awkward to specify classes of smooth priors and to describe the corresponding classes of posteriors. Thus, it may sometimes be worthwhile to investigate aspects of Bayesian robustness using classes of priors which are artificial but tractable. For one thing, theoretical investigations are obviously easier for situations permitting nice, clean formulation and calculation. Ideally, some of the insights gained from looking at artificial examples should carry

Bayesian robustness can be established for a class of priors much larger than the class of real interest, then Bayesian robustness holds at least as well for the priors of real interest. However, one should not lose sight of the fact that classes of priors (such as those given by concentration function neighborhoods) which are chosen for reasons of tractability and mathematical elegance will seldom be more than crude surrogates for more appropriate classes. I wish to thank Larry Wasserman for very helpful conversations about concentration functions.

Additional reference

Edwards, W., H. Lindman and L.J. Savage (1963), Bayesian statistical inference for psychological research, *Psychol. Rev.* 70, 193–242; reprinted in J. Kadane, Ed. (1984). *Robustness of Bayesian Analyses.* North-Holland, Amsterdam.

Rejoinder to Thomas Sellke's comment

S. Fortini and F. Ruggeri

We wish to thank Professor Sellke for his stimulating results and comments. We like the idea of studying what happens to a neighbourhood K_g of a prior measure when considering the corresponding posterior measures. Sellke gives a condition under which the same g defines both prior and posterior neighbourhoods. We have analysed some classes in order to see what happens; in particular we mention that the ε-contaminations, when the contaminating class contains all the probability measures, are defined by $g(x)=(1-\varepsilon)x$, $\forall x\in[0,1]$, while a posteriori, the lowest c.f. is given by

$$g(x)=\frac{(1-\varepsilon)D_0}{(1-\varepsilon)D_0+\varepsilon l_x(0_0)}\,x,$$

where $l_x(0_0)$ is the likelihood at its mode and $D_0=\int_\Theta l_x(0)P_0(\mathrm{d}0)$ (see Fortini and Ruggeri, 1990). Such a g cannot be above the g a priori, confirming, as Sellke pointed out, that 'any concentration function of neighborhood centered at $P_0|y$ which contains $K_g|y$ must contain the g-neighborhood of $P_0|y$'.

Finally, we think that the concentration function neighbourhood is interesting not only for 'reasons of tractability and mathematical elegance' but also because it arises quite naturally when bounding the probability of any measurable subset, as pointed in Section 2.