Ann. Inst. Statist. Math.
Vol. 51, No. 3, 499–513 (1999)

POSTERIOR SENSITIVITY TO THE SAMPLING DISTRIBUTION AND THE PRIOR: MORE THAN ONE OBSERVATION*

SANJIB BASU**

Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, U.S.A.

(Received April 28, 1995; revised March 13, 1998)

Abstract. Sensitivity of a posterior quantity $\rho(f, P)$ to the choice of the sampling distribution f and prior P is considered. Sensitivity is measured by the range of $\rho(f, P)$ when f and P vary in nonparametric classes Γ^f and Γ^P respectively. Direct and iterative methods are described which obtain the range of $\rho(f, P)$ over $f \in \Gamma^f$ when prior P is fixed, and also the overall range over $f \in \Gamma^f$ and $P \in \Gamma^P$. When multiple i.i.d. observations $X_1, \ldots, X_k$ are observed from f, the posterior quantity $\rho(f, P)$ is not a ratio-linear function of f. A method of steepest descent is proposed to obtain the range of $\rho(f, P)$. Several examples illustrate applications of these methods.

Key words and phrases: Bayesian robustness, Gâteaux derivative, model robustness, model selection, predictive distribution, prior robustness, steepest descent.

1. Introduction

Bayesian analysis requires specifications of two models; the sampling distribution $f(x \mid \theta)$ and the prior $P(\theta)$. When considering the effect of one's modeling assumptions, perturbations of both $f(x \mid \theta)$ and $P(\theta)$ should be taken into account. Robust Bayesian analysis has, so far, concentrated more on imprecisions of the prior $P(\theta)$. There are several reasons for this emphasis. See Pericchi and Pérez (1994) for an exposition. In this article, we focus on sensitivity of Bayesian analysis w.r.t. the sampling distribution and also joint perturbations of the prior and the sampling distribution.

The literature on sampling model determination and robustness in Bayesian analysis includes Smith (1983), Gelfand *et al.* (1992), Pericchi and Pérez (1994) and others. These works explored the methods and effects of choosing a particular sampling distribution from a finitely many available choices. However, in many situations, these finitely many choices, or even the functional forms of the sampling density are hard to determine. We focus on nonparametric classes of sampling

* Research supported by ASTA grant 94-B-32.

** Now at Division of Statistics, Northern Illinois University, DeKalb, IL 60115, U.S.A.

distributions, as considered in Lavine (1991), Bayarri and Berger (1993), Ghosh and Dey (1994), and Dey *et al.* (1997).

Formally, let Θ be the parameter space. We assume Θ is a connected subset of $\Re$. Let $f(x \mid \theta)$ denote the sampling density of an one-dimensional random variable x and $P(\theta)$ denote the prior distribution on θ. For observed data X from $f(\cdot \mid \theta)$, let $m(X \mid f, P) = \int_\Theta f(X \mid \theta) dP(\theta)$ be the marginal w.r.t. $f(X \mid \theta)$ and $P(\theta)$. The posterior measure $P(A \mid X, f, P) = \frac{1}{m(X|f,P)} \int_A f(X \mid \theta) dP(\theta)$ is denoted by $P(\cdot \mid X, f, P)$. Similarly, $\pi(\cdot)$, $\pi(\cdot \mid X, f, \pi)$ respectively denote the prior or posterior density functions (cdfs). We use $\rho(X, f, P)$ ($\rho(X, f, \pi)$) to denote a posterior quantity w.r.t. prior P (π) and sampling distribution $f(X \mid \theta)$.

In the following, we present our posterior sensitivity analysis under the assumption that the class of densities $f(x \mid \theta)$ have a location structure, i.e., $f(x \mid \theta) = f(x - (\theta - \theta_0) \mid \theta_0)$ (for a fixed $\theta_0 \in \Theta$) $= f(x - \theta)$ (where, w.l.g., we assume $\theta_0 = 0$ and with abuse of notation). The methods that we describe, however, are clearly applicable in wider generalities. We make this simplifying assumption to keep our presentation clear and our notations simple.

We reflect our uncertainty about the functional form of the sampling density by assuming that $f(\cdot)$ vary in a class Γ^f. Through the location structure, this generates Γ^f_θ classes for $f(x \mid \theta)$, $\Gamma^f_\theta = \{f(x \mid \theta) = f(x - \theta), f \in \Gamma^f\}$.

We should point out here that an alternative way to model the uncertainty about the functional form of the sampling density f will be to treat f as an infinite dimensional parameter and to put a prior on f. This latter approach will impose a nonparametric Bayesian structure on the problem. We follow the former route which retains the simplicity of the parametric Bayesian structure. The two approaches differ in the sense that whereas the former is a robustness investigation, the latter is more of a model elaboration or model extension.

We further consider sensitivity of posterior results to joint perturbations of the sampling density and the prior. Here, in addition to $f \in \Gamma^f$, we also model the uncertainty about the prior choice by letting the prior P vary in a class Γ^P. The robustness of a posterior quantity $\rho(f, P)$ is measured by finding its range over these classes.

Lavine (1991) also considered posterior sensitivity by varying the sampling distribution $f(X \mid \theta)$ and the prior $P(\theta)$ simultaneously. He allowed P to vary in a class Γ^P and for each fixed θ, $f(X \mid \theta)$ is allowed to vary in a class Γ^c. This, for example, may allow a single $f(X \mid \theta)$ to have a normal functional form for $\theta = \theta_1$ and a Cauchy functional form for a different θ, say $\theta = \theta_2$. In our model this is not allowed, f cannot have two different functional forms for two different θ values.

The important achievement of this article is that we consider multiple i.i.d. observations $X_1, \ldots, X_k$ from the density $f(x \mid \theta)$. Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ and let the joint density be $\boldsymbol{f}(\boldsymbol{X} \mid \theta) = f(X_1 \mid \theta) \cdots f(X_k \mid \theta)$. If we assume that each marginal $f(X_i \mid \theta)$ belongs to a nonparametric class, the resulting class for the joint density $\boldsymbol{f}(\boldsymbol{X} \mid \theta)$ becomes very difficult to work with and has not been extensively explored. One important exception is Bayarri and Berger (1993) where the sampling density $f(x \mid \theta)$ is assumed to be of the form $f_w(x \mid \theta) = w(x) v_w(\theta) g(x \mid \theta)$ and the weight function $w(x)$ is allowed to vary. This structure leads to marked simplicity in the analysis.

In our sensitivity study, we do not assume any specific functional forms for the sampling density except for the assumption that it has a location structure. We first focus on sampling model robustness alone assuming the prior $P(\theta)$ to be fixed. Let $\rho(\boldsymbol{X}, f, P)$ be our posterior quantity of interest based on the sampling density $f(x \mid \theta)$ and observed data $\boldsymbol{X} = (X_1, \ldots, X_k)$. Section 3 describes an iterative method for finding $\inf_{f \in \Gamma^f} \rho(\boldsymbol{X}, f, P)$ and $\sup_{f \in \Gamma^f} \rho(\boldsymbol{X}, f, P)$. This method is a modified steepest descent algorithm. It iterates between (i) finding the steepest direction of descent and (ii) doing a line minimization along that direction. The implementation of this algorithm is discussed in detail in three different examples. Example 1 considers a setup where the range of $\rho(\boldsymbol{X}, f, P)$ can be obtained by a known non-iterative method. We compare the performances of our steepest descent method with the non-iterative method. Examples 2 and 3 consider Darwin's data with 15 observations. We consider model selection among the normal scale mixture family by maximizing the marginal density of $\boldsymbol{X}$ (Example 2) or the "pseudo-marginal density" of $\boldsymbol{X}$ (Example 3).

Finally, in Section 4, we consider joint variations of the prior P and the sampling density f when multiple i.i.d. observations $\boldsymbol{X} = (X_1, \ldots, X_k)$ are observed from f. We briefly describe techniques for finding the range of the posterior quantity $\rho(\boldsymbol{X}, f, P)$ over $f \in \Gamma^f$ and $P \in \Gamma^P$.

2. Preliminaries and the case of one observation

So far, we have discussed about the sampling density class Γ^f and the prior class Γ^P in general terms. Some specific choices for these classes are:

(i) $\Gamma_1 = \{\text{all distributions on } \Re\}$.

(ii) $\Gamma_2 = \{\text{all distributions on } \Re \text{ which are symmetric about a point } M\}$.

(iii) $\Gamma_3 = \{\text{all distribution on } \Re \text{ unimodal about } M\}$.

(iv) $\Gamma_4 = \{\text{all distribution on } \Re \text{ symmetric and unimodal about } M\}$.

(v) $\Gamma_5 = \{\text{all normal scale mixture distributions on } \Re \text{ with cdfs } F(\theta) = \int_{[0,\infty)} \Phi(\frac{\theta - M}{s}) dG(s)\}$.

(vi) $\Gamma_6 = \{\varepsilon\text{-contamination class}\} = \{p = (1-\varepsilon)p_0 + \varepsilon q; q \in \mathcal{Q}\}$ where $\mathcal{Q}$ equals Γ_i of any of the previous (i)–(v).

(vii) $\Gamma_7 = \{\text{Density ratio class}\} = \{\text{density } p(\eta) : L(\eta) \leq \alpha p(\eta) \leq U(\eta) \text{ for some } \alpha > 0\}$.

(viii) $\Gamma_8 = \{\text{Distribution band}\} = \{\text{cdf } F(\eta) : F_L(\eta) \leq F(\eta) \leq F_U(\eta), F_L \text{ and } F_U \text{ are two fixed cdfs}\}$.

All these classes can be used as choices for Γ^P whereas Γ_i, $i = 3, \ldots, 7$ can be used as choices for Γ^f. In Bayesian robustness literature, each of $\Gamma_1, \ldots, \Gamma_8$ has been used as a choice for prior class Γ^P. See Berger (1994) for an extensive review. On the other hand, the literature on Bayesian robustness studies of the sampling density is not very extensive. Lavine (1991) used the class Γ_7 to investigate sensitivity of likelihood in Bayesian analysis. In frequentist robustness, however, the focus is specifically on the robustness of the sampling density. Many of the classes from the above list have been used in such studies. Tukey (1960) introduced the ε-contamination class Γ_6 in a frequentist sensitivity study. Also see Huber (1981) in this context. Basu and DasGupta (1995*b*) used Γ_3, Γ_4, Γ_5 and Γ_6 as choices for sampling density classes.

We have listed above a collection of eight classes. One may ask that in a specific robustness investigation, which classes from $\Gamma_1, \ldots, \Gamma_8$ one should choose as choices for Γ^f and Γ^P. In general, this question has no definite answer. The choices of Γ^f and Γ^P depend on the extent of knowledge of the user about the sampling density and the prior. The choice $\Gamma^p = \Gamma_1$ implies that the user has no knowledge about the prior. However, such a completely non-informative choice will generally result in excessively wide ranges of posterior quantities. If symmetry and unimodality of the prior or the likelihood are assured, one should use the classes Γ_4 or Γ_5. Among these two, Γ_5 only allows distributions which have heavier tails than normal, whereas Γ_4 allows both heavier as well as lighter tails. The class Γ_6 models a different scenario. It suggests that the user is only $100(1-\varepsilon)\%$ sure that p_0 is the correct choice and $100\varepsilon\%$ uncertain. Each of the above eight classes thus models a different degree of uncertainty. In a specific robustness investigation, the user needs to choose a specific Γ^f and a specific Γ^P based on the specific setup and the user's knowledge about the specific problem.

Remark. As Lavine (1991) pointed out, when we have only one observation X, the selection of either of $\Gamma_3, \ldots, \Gamma_6$ as a choice of Γ^f may lead to trivial answers. These classes do not bound densities away from ∞. Γ_3, Γ_4 and Γ_5 allow $f(\theta)$ to go to ∞ at $\theta = M$. These phenomena translate to Γ_6 when any of $\Gamma_3, \ldots, \Gamma_5$ is used as a choice of $\mathcal{Q}$. In some cases, it may lead to trivial posterior extrema, for example, the supremum of a posterior probability may equal 1 and so on.

It is easy to see that the listed classes Γ_i, $i = 1, \ldots, 8$ are convex. Moreover, for Γ_i, $i = 1, \ldots, 6$, their extreme point classes $\mathcal{E}_i$, $i = 1, \ldots, 6$ are also known and are listed below. Notice that each of the $\mathcal{E}_i$ classes below is driven by one real parameter.

(i) $\mathcal{E}_1 = \{\text{all degenerate distributions on } \Re\}$.

(ii) $\mathcal{E}_2 = \{P_v(\theta) = 0.5I_{\{M-v\}}(\theta) + 0.5I_{\{M+v\}}(\theta),\ v \geq 0\}$.

(iii) $\mathcal{E}_3 = \{U_z : z \in \Re\}$, where $U_z = \text{Uniform}[M, M+z]$ if $z > 0$, $= \text{Uniform}[M-z, M]$ if $z < 0$ and $U_0(\theta) = I_{\{M\}}(\theta)$.

(iv) $\mathcal{E}_4 = \{U_z^* : z \geq 0\}$, where $U_z^* = \text{Uniform}[M-z, M+z]$ if $z > 0$ and $U_0^*(\theta) = I_{\{M\}}(\theta)$.

(v) $\mathcal{E}_5 = \{F_s(\theta) = \Phi(\frac{\theta}{s}),\ s \geq 0\}$.

(vi) $\mathcal{E}_6 = \{p : p = (1-\varepsilon)p_0 + \varepsilon q; q \in \mathcal{Q}^*\}$ where $\mathcal{Q}^*$ is the corresponding $\mathcal{E}_i$ of (i)–(v).

Suppose our posterior quantity of interest is $\rho(X, f, P)$. We reflect our uncertainties about the sampling density $f(x \mid \theta)$ and the P by assuming $f \in \Gamma^f$ and $P \in \Gamma^P$. We assume Γ^P is one of the above Γ_i classes, $i = 1, \ldots, 8$ and Γ^f is one of Γ_i, $i = 3, \ldots, 7$. We measure the sensitivity of $\rho(X, f, P)$ by its range, namely $\overline{\rho} = \sup_{f \in \Gamma^f, P \in \Gamma^P} \rho(X, f, P)$ and $\underline{\rho} = \inf_{f \in \Gamma^f, P \in \Gamma^P} \rho(X, f, P)$.

In this section, we briefly outline the method for finding the range of $\rho(X, f, P)$ in the easy case when only one observation X is observed from the density $f(x \mid \theta)$. The discussion on one observation brings out the difficulty which arises in the multiple observations case. The multiple observations case is discussed in Sections 3 and 4.

We focus on finding $\sup_{f\in\Gamma^f, P\in\Gamma^P} \rho(X,f,P)$. The infimum case is similar. Suppose further that $\rho(X,f,P)$ is ratio-linear, i.e., $\rho(X,f,P) = \{\int_\Theta h(\theta)f(X-\theta)dP(\theta)\}/\{\int_\Theta f(X-\theta)dP(\theta)\}$ for real-valued function $h(\theta)$. Examples of such are posterior mean, posterior probability of a fixed set C and more.

If $\rho(X,f,P)$ is ratio-linear, the linearization technique (see, for example, Basu and DasGupta (1995*a*)) can often be used to reduce the evaluation $\sup_{f\in\Gamma^f, P\in\Gamma^P} \rho(X,f,P)$ to two simpler steps: (i) for a fixed $\lambda \in \Re$, find $\overline{\rho}(\lambda) = \sup_{P\in\Gamma^P}\sup_{f\in\Gamma^f}\int(h(\theta)-\lambda)f(X-\theta)dP(\theta)$, and (ii) do an univariate optimization over λ or solve an equation in λ. Step (ii) is easy to obtain (analytically and/or numerically).

For step (i), let $\rho(X,f,P,\lambda) = \int(h(\theta)-\lambda)f(X-\theta)dP(\theta)$. Note that $\rho(X,f,P,\lambda)$ is a linear function of $f(\cdot)$ and $P(\cdot)$. Hence, this is a linear optimization problem. It then can be argued that $\sup_{P\in\Gamma^P}\sup_{f\in\Gamma^f}\rho(X,f,P,\lambda) = \sup_{P\in\mathcal{E}^P}\sup_{f\in\mathcal{E}^f}\rho(X,f,P,\lambda)$ if $\Gamma^f = \Gamma_i$, $i = 3,\ldots,6$ and $\Gamma^P = \Gamma_j$, $j = 1,\ldots,6$ and $\mathcal{E}^f$ and $\mathcal{E}^P$ are the corresponding extreme point classes. This reduction happens due to the convex structure of the Γ_i classes. In fact, a similar reduction happens if $\Gamma^f = \Gamma_7$ or if $\Gamma^P = \Gamma_7$ or Γ_8. We refrain from the details of the reduction argument and refer the reader to Basu (1994).

We mentioned earlier that each of the extreme point classes is driven by at most one real parameter. Hence the final optimization over the extreme point classes $\mathcal{E}^f$ and $\mathcal{E}^P$ is a relatively easy job.

We should mention here that the underlying mechanism behind the reduction of the optimization problem above is the famous Krein-Milman theorem. For each of $\Gamma_1,\ldots,\Gamma_6$, the corresponding extreme point class $\mathcal{E}_i$ has the property that Γ_i is the closed convex hull of $\mathcal{E}_i$. This is proved in Basu (1996). The reduction of the optimization $\sup_{P\in\Gamma^P}\sup_{f\in\Gamma^f}\rho(X,f,P,\lambda) = \sup_{P\in\mathcal{E}^P}\sup_{f\in\mathcal{E}^f}\rho(X,f,P,\lambda)$ follows from this result and the fact that $\rho(X,f,P,\lambda)$ is a linear function of $f(\cdot)$ and $P(\cdot)$. This discussion also points out that the reduction of the optimization problem is not specific only to the classes we listed, but will also hold for any class Γ which is a closed convex hull of its extreme points. We listed those classes which are popular in robustness studies.

3. More than one observation: sensitivity to the sampling distribution

3.1 *The method*

The discussion of Section 2 focused on the case of a single observation X from the sampling density. In this section, we consider the more practical case when we observe multiple i.i.d. observations $\boldsymbol{X} = (X_1,\ldots,X_k)$ from the sampling distribution $f(x \mid \theta) = f(x-\theta)$ (having the location structure). This yields the joint density $\boldsymbol{f}(\boldsymbol{X} \mid \theta) = f(X_1-\theta)\cdots f(X_k-\theta)$. We reflect our uncertainty about f by assuming that $f(\cdot)$ is an arbitrary choice from the class Γ^f. In this section, we focus only on sampling model robustness and assume that prior $P(\cdot)$ is fixed.

Let $\rho(\boldsymbol{X},f,P)$ be a posterior quantity based on $\boldsymbol{f}(\boldsymbol{X} \mid \theta)$ and $P(\theta)$. Even if $\rho(\boldsymbol{X},f,P)$ is the posterior expectation of a function $h(\theta)$, i.e., $\rho(\boldsymbol{X},f,P) = \{\int h(\theta)f(X_1-\theta)\cdots f(X_k-\theta)dP(\theta)\}/\{\int f(X_1-\theta)\cdots f(X_k-\theta)dP(\theta)\}$, $\rho(\boldsymbol{X},f,P)$ is not ratio-linear in f; the linearization technique of Section 2 does not apply and new methods are needed. A new iterative method is described below.

We assume that $f(x \mid \theta) = f(x-\theta)$ and $f(\cdot) \in \Gamma^f$ where $\Gamma^f = \Gamma_i$, $i = 3, \ldots, 6$. In the following, we propose an iterative method for finding $\inf_{f\in\Gamma^f} \rho(\boldsymbol{X}, f, P)$ and $\sup_{f\in\Gamma^f} \rho(\boldsymbol{X}, f, P)$ which uses the Gâteaux derivative of the function $\rho(\boldsymbol{X}, f, P)$. This method is widely applicable in the sense that it is valid for ratio-linear as well as other posterior quantities $\rho(\boldsymbol{X}, f, P)$. Gâteaux derivative (in statistical context) is generally defined on the linear space Δ of all signed measures. However, since our focus is on a convex subset $\Gamma^f \subseteq \Delta$, we follow Huber (1981) to define Gâteaux derivative on a convex set $\Gamma \subseteq \Delta$ (see also Basu (1996)).

DEFINITION 1. The functional $\rho(\boldsymbol{X}, \cdot, P) : \Gamma \to \Re$ is called Gâteaux differentiable at $f \in \Gamma$ if $\exists$ a linear functional $G\rho_f : \Delta \to \Re$ such that $\frac{1}{t}|\rho(\boldsymbol{X}, (1-t)f + tg, P) - \rho(\boldsymbol{X}, f, P) - tG\rho_f(g-f)| \to 0$ as $t \to 0$ for all $g \in \Gamma$.

The Gâteaux derivative $G\rho_f(g-f)$ is thus simply $\frac{d}{dt}\rho(\boldsymbol{X}, (1-t)f+tg, P)|_{t=0}$. The next result connects the concepts of Gâteaux derivative and local extrema.

THEOREM 1. *Let Γ be a convex subset of Δ. Assume $\rho(\boldsymbol{X}, f, P)$, as a function of f, is well defined on $\bar{\Gamma}$, the closure of Γ in weak convergence topology. Suppose $\inf_{f\in\Gamma} \rho(\boldsymbol{X}, f, P)$ $(\sup_{f\in\Gamma} \rho(\boldsymbol{X}, f, P))$ is attained at $f_I \in \bar{\Gamma}$ $(f_S \in \bar{\Gamma})$. If $\rho(\boldsymbol{X}, \cdot, P)$ is weakly continuous and Gâteaux differentiable at f_I (at f_S) then $\inf_{g\in\Gamma} G\rho_{f_I}(g - f_I) = 0$ $(\sup_{g\in\Gamma} G\rho_{f_S}(g - f_S) = 0)$.*

Remark. This result also appears, though in different forms, in Luenberger (1968) and Srinivasan and Truszczynska (1990).

PROOF. Take $g \in \Gamma$. Since $\bar{\Gamma}$ is convex, $(1-t)f_I + tg \in \bar{\Gamma}$ and $\rho(\boldsymbol{X}, (1-t)f_I + tg, P)$, as a function of $0 \le t < 1$, has a minimum at $t = 0$. Hence $G\rho_{f_I}(g - f_I) = \lim_{t\downarrow 0} \frac{1}{t}\{\rho(\boldsymbol{X}, (1-t)f_I + tg, P) - \rho(\boldsymbol{X}, f_I, P)\} \ge 0$. Next, take a sequence $\{g_n\}_{n\ge 1} \subseteq \Gamma$ such that $g_n \to f_I$ weakly. Then $G\rho_{f_I}(g_n - f_I) \to 0$ by weak continuity of $\rho(\boldsymbol{X}, \cdot, P)$. This completes the proof. □

From now on, we will concentrate on the infimum problem. The supremum problem is similar. Theorem 1 gives us a necessary condition for a local infimum. The method to be described below tries to find a f_I where this condition is met. Thus, as with many other numerical methods, our method is not guaranteed to converge to the global infimum.

Suppose Γ^f is convex and closed. Notice that this is true if $\Gamma^f = \Gamma_i$, $i = 3, \ldots, 6$. Further, assume that $G\rho_f(\cdot)$ is weakly continuous $\forall f \in \Gamma^f$.

[0] We start from an initial guess f_*. Then, we iterate through the following steps.

[1] We find the direction in which $\rho(\cdot)$ has the fastest rate of decrease by finding $\inf_{g\in\Gamma^f} G\rho_{f_*}(g - f_*)$.

Notice that $G\rho_{f_*}(g - f_*)$ is linear in g, i.e., for $g_1, g_2 \in \Gamma^f$, $G\rho_{f_*}((1-\alpha)g_1 + \alpha g_2 - f_*) = (1-\alpha)G\rho_{f_*}(g_1 - f_*) + \alpha G\rho_{f_*}(g_2 - f_*)$. Suppose Γ^f has extreme points $\mathcal{E}^f$. Then $\inf_{g\in\Gamma^f} G\rho_{f_*}(g - f_*) = \inf_{g\in\mathcal{E}^f} G\rho_{f_*}(g - f_*)$ (see Theorem 3 of Basu (1996)). For $\Gamma^f = \Gamma_i$, $i = 3, \ldots, 6$, the extreme points class $\mathcal{E}^f$ is driven by one real parameter. Thus, minimization over $\mathcal{E}^f$ is often a relatively easy job.

[2] If $\inf_{g\in\Gamma^f} G\rho_{f_*}(g - f_*) = 0$, we stop iteration. $\rho(\boldsymbol{X}, \cdot, P)$ is then non-decreasing in every direction g from f_*. Thus, f_* is our final answer from this iteration run and is a candidate for a local minimum. We may go back to step [0], start from a different initial guess and finally compare the answers from different runs.

[3] Suppose $\inf_{g\in\Gamma^f} G\rho_{f_*}(g - f_*) = G\rho_{f_*}(g_* - f_*) < 0$ (g_* exists since Γ^f is closed and $G\rho_{f_*}(\cdot)$ is continuous). Then $\rho(\boldsymbol{X}, \cdot, P)$ has the fastest rate of decrease from f_* in the direction g_*.

[4] Now, we look in the direction g_* and do a line minimization. Let $f_t = (1-t)f_* + tg_*$, $0 \le t \le 1$. We find $\min_{0\le t\le 1} \rho(\boldsymbol{X}, f_t, P) = \rho(\boldsymbol{X}, f_{t_*}, P)$. This is a univariate minimization on a bounded set and can often be done easily.

[5] f_{t_*} is our new guess for the minimal f_I. We put $f_* = f_{t_*}$ and go back to step [1].

A huge literature exists on the general performance and convergence of the steepest descent method. It is known that this steepest descent method is not very efficient in the sense that it may take many small steps to converge to a minimum even in well behaved problems. In Euclidean or Hilbert spaces, many faster algorithms exist, such as the conjugate gradient method. Since our space is only convex, we cannot avail ourselves of such faster methods. In the next section, we look at the performance of our method in several examples.

3.2 *Applications*

In this section, we look at three examples. Example 1 deals with one observation from the sampling density and examines a situation where the range of the posterior quantity can be obtained by a known non-iterative method. We illustrate the steps of the steepest descent method in this example and compare its performance to the non-iterative method. Examples 2 and 3 consider Darwin's data where we have multiple observations and other methods fail to obtain the range.

Example 1. Suppose we observe one observation X from the sampling density $f(x \mid \theta) = f(x - \theta)$ and θ has a fixed prior $P(\theta)$. We model our uncertainty about f by assuming that $f \in \Gamma^f$ = the ε-contamination class $\Gamma_6 = \{f = (1-\varepsilon)f_0 + \varepsilon q : q \in \mathcal{Q}\}$. We assume f_0 is symmetric, unimodal about 0 and the contaminating class $\mathcal{Q} = \Gamma_5 = \{$all normal scale mixture distributions with median at $0\}$.

Our posterior quantity of interest is the posterior variance $V(X, f, P) = m_2(X \mid f, P)/m(X \mid f, P) - \{m_1(X \mid f, P)/m(X \mid f, P)\}^2$ where $m_i(X \mid f, P) = \int \theta^i f(X - \theta) dP(\theta)$ and $m(X \mid f, P) = m_0(X \mid f, P)$. $V(X, f, P)$ is nonlinear in f. However, if we fix the value of the posterior mean $\mu(X, f, P) = m_1(X \mid f, P)/m(X \mid f, P)$, then $V(X, f, P)$ is a linear function of f and the method of Sivaganesan and Berger (1989) can be applied to obtain the ranges of $V(X, f, P)$ for each fixed value of $\mu(X, f, P)$. Finally, we can vary $\mu(X, f, P)$ on its range to find the overall range of $V(X, f, P)$.

For example, let $f_0(\cdot) = \frac{1}{\sqrt{2.2}}\phi(\frac{\cdot}{\sqrt{2.2}})$ and $P(\cdot)$ be the $N(0,1)$ distribution. For these choices of f_0 and P, we evaluate $\inf_{f\in\Gamma^f} V(X, f, P)$ by the Sivaganesan and Berger (1989) method. These values are listed in Table 1.

Table 1. $\min_{f \in \Gamma f} V(X, f, P)$ and the minimal contaminating q_* in Example 1.

	non-iterative method				
X	1	2	3	4	5
q_*	$N(0,(0.35)^2)$	$N(0,(1.12)^2)$	$N(0,(1.31)^2)$	$N(0,(1.38)^2)$	$N(0,(1.42)^2)$
$V(X,(1-\varepsilon)f_0+\varepsilon q_*,P)$	0.6524	0.6812	0.6849	0.6861	0.6866

	steepest descent method				
X	1	2	3	4	5
q_*	$N(0,(0.36)^2)$	$N(0,(1.12)^2)$	$N(0,(1.30)^2)$	$N(0,(1.38)^2)$	$N(0,(1.42)^2)$
$V((1-\varepsilon)f_0+\varepsilon q_*)$	0.6524	0.6812	0.6849	0.6861	0.6866

On the other hand, we can treat $V(X, f, P)$ as a nonlinear function of f on the convex set Γ^f and apply our steepest descent method. The extreme point class of Γ^f is $\mathcal{E}^f = \{f(\cdot) = (1-\varepsilon)f_0(\cdot) + \varepsilon\frac{1}{\sigma}\phi(\frac{\cdot}{\sigma}) : \sigma > 0\}$.

We start by choosing $f_* = f_1 = (1-\varepsilon)f_0 + \varepsilon q_1$ where $q_1(\cdot) = \frac{1}{\sigma_1}\phi(\frac{\cdot}{\sigma_1})$ for some $\sigma_1 > 0$. After the end of the n-th iteration cycle, we have $f_*(\cdot) = f_n(\cdot) = (1-\varepsilon)f_0(\cdot) + \varepsilon\sum_{i=1}^n \alpha_i q_i(\cdot)$ where $\alpha_i > 0$, $\sum_{i=1}^n \alpha_i = 1$, and $q_i(\cdot) = \frac{1}{\sigma_i}\phi(\frac{\cdot}{\sigma_i})$, $i = 1, \ldots, n$. Let $V_0 = V(X, f_0, P)$ (the posterior variance under f_0) and $\mu_0 = \mu(X, f_0, P)$ (the posterior mean under f_0). Similarly, let $V_i = V(X, q_i, P)$, $\mu_i = \mu(X, q_i, P)$, $i \geq 1$.

We next proceed to the $(n+1)$-th iteration cycle. The Gâteaux derivative of the posterior variance, namely $GV(\cdot)$, is obtained in Basu (1996). Using this result, we have $GV_{f_n}(g - f_n) = \{m(X \mid f_n, P)^2 m_2(X \mid g, P) - 2m_1(X \mid f_n, P)m(X \mid f_n, P)m_1(X \mid g, P) + (2m_1(X \mid f_n, P)^2 - m_2(X \mid f_n, P)m(X \mid f_n, P))m(X \mid g, P)\}/m(X \mid f_n, P)^3$ for any $g \in \Gamma^f$. Let $a = m(X \mid f_n, P)^2$, $b = -2m_1(X \mid f_n, P)m(X \mid f_n, P)$, $c = 2m_1(X \mid f_n, P)^2 - m_2(X \mid f_n, P)m(X \mid f_n, P)$, and assume prior $P = N(0, 1)$. Since $g = (1-\varepsilon)f_0 + \varepsilon q_{n+1}$ with $q_{n+1}(\cdot) = \frac{1}{\tau}\phi(\frac{\cdot}{\tau})$, we have $m_i(X \mid g, P) = (1-\varepsilon)m_i(X \mid f_0, P) + \varepsilon m_i(X \mid q_{n+1}, P)$ and $m(X \mid f_n, P)^3 GV_{f_n}(g - f_n) = (1-\varepsilon)\text{constant} + \frac{\varepsilon}{\sqrt{\tau^2+1}}\phi(\frac{X}{\sqrt{\tau^2+1}})\{a(\frac{\tau^2}{\tau^2+1} + \frac{X^2}{(\tau^2+1)^2}) + \frac{bX}{\tau^2+1} + c\}$ which is a function of $\tau^2 \geq 0$. Finding $\inf_{g\in\mathcal{E}^f} GV_{f_n}(g - f_n)$ is now an easy job.

Suppose $\tau = \sigma_{n+1}$ is obtained as the infimum attaining choice and let the corresponding $g = (1-\varepsilon)f_0 + \varepsilon q_{n+1}$ (with abuse of notation). Our next job is to evaluate $V(X, (1-t)f_n + tg, P)$ and then minimize over $0 \leq t \leq 1$. Let $\beta_0 = 1-\varepsilon$, $\beta_i = (1-t)\varepsilon\alpha_i$, $i = 1, \ldots, n$, and $\beta_{n+1} = t\varepsilon$ be the weights of f_0, and $q_1, \ldots, q_{n+1}$ in $(1-t)f_n + tg$. In the posterior, these weights get updated to $\gamma_0 = \beta_0 m(X \mid f_0, P)/\{\beta_0 m(X \mid f_0, P) + \sum_{i=1}^{n+1}\beta_i m(X \mid q_i), P\}$ and $\gamma_i = \beta_i m(X \mid q_i, P)/\{\beta_0 m(X \mid f_0, P) + \sum_{i=1}^{n+1}\beta_i m(X \mid q_i, P)\}$, $i = 1, \ldots, n+1$, and the posterior variance $V(X, (1-t)f_n + tg, P) = \sum_{i=0}^{n+1}\gamma_i V_i + \sum_{i=0}^{n+1}\gamma_i\mu_i^2 - (\sum_{i=0}^{n+1}\gamma_i\mu_i)^2$. Thus, evaluating $V(X, (1-t)f_n + tg, P)$ and obtaining $\min_{0\leq t\leq 1} V(X, (1-t)f_n + tg, P) = V(X, (1-t_*)f_n + t_*g, P)$ are also not hard. Once t_* is obtained, the minimal f_* gets updated to $f_*(\cdot) = f_{n+1}(\cdot) = (1-\varepsilon)f_0(\cdot) + \varepsilon\sum_{i=1}^n (1-t_*)\alpha_i q_i(\cdot) + \varepsilon t_* q_{n+1}(\cdot)$. The algorithm then proceeds to the next $(n+2)$-th iteration cycle.

For $f_0(\cdot) = \frac{1}{\sqrt{2.2}}\phi(\frac{\cdot}{\sqrt{2.2}})$ and $P = N(0, 1)$, we also use the steepest descent method to obtain the infimum of the posterior variance, $\inf_{f\in\Gamma^f} V(X, f, P)$. Our initial guess is $f_1(\cdot) = (1-\varepsilon)f_0(\cdot) + \frac{\varepsilon}{\sigma_1}\phi(\frac{\cdot}{\sigma_1})$ with $\sigma_1 = 1$. In all the cases, the iterations converge in 1 or at most 2 steps. The values are listed in Table 1. Caution is needed in the implementation, since the end points $\sigma = 0$ and $\sigma = \infty$ are local extreme points and the descent method shows a tendency to converge to them.

As seen in Table 1, the non-iterative method and the steepest descent method give almost identical results. The implementation of the steepest descent method, however, is more involved. The real power of the descent method is revealed when we have multiple observations as in the next two examples.

Example 2. Here and in Example 3, we consider Darwin's data (see Table 2) on the differences (in eights of an inch) of heights of 15 pairs of cross- and self-

Table 2. Darwin's data on differences of heights.

X_i	−67	−48	6	8	14	16	23	24	28	29	41	49	56	60	75

fertilized plants (see Andrews and Herzberg (1985)). These data were analyzed by Box and Tiao (1973) and Pericchi and Pérez (1994). The latter authors assumed that the fifteen observations $X_1, \ldots, X_{15}$ are i.i.d. from a sampling density $f(\cdot \mid \theta, \sigma)$. They assumed a prior $\pi(\theta, \sigma) = 1/\sigma$ on (θ, σ) and studied the performances of five different choices of the sampling distribution f : $N(\theta, \sigma^2)$, Cauchy(θ, σ^2), Uniform$(\theta - \sqrt{3}\sigma, \theta + \sqrt{3}\sigma)$, Left Exponential, and Right Exponential. Their calculations showed that the marginal $m(\boldsymbol{X} \mid f, \pi) = \int \prod_{i=1}^{15} f(X_i \mid \theta, \sigma)\pi(\theta, \sigma)d\theta d\sigma$ is maximized (among the five choices) and equals 2.451×10^{-32} when $f = N(\theta, \sigma^2)$.

We reexamine this data and assume $X_1, \ldots, X_{15}$ are i.i.d. from a location density $f(x - \theta)$. We allow a much wider choice for f and assume that f is a generic member of the nonparametric class $\Gamma^f = \Gamma_5 = \{$all normal scale mixture distributions with median at $0\}$. This choice of Γ_5 is due to simplicity in calculations, we could have easily used any of Γ_3, Γ_4 or Γ_6 as a choice for Γ^f. Let $\pi(\theta)$ be the prior on θ. For $f \in \Gamma^f$, let $m(\boldsymbol{X} \mid f, \pi) = \int \prod_{i=1}^{k} f(X_i - \theta)d\pi(\theta)$ where $k = 15$. We maximize this marginal: choose $f_* \in \Gamma^f$ such that $m(\boldsymbol{X} \mid f_*, \pi) = \max_{f \in \Gamma^f} m(\boldsymbol{X} \mid f, \pi)$. Some statisticians use $m(\boldsymbol{X} \mid f_*, \pi)$ as a model selection criterion (for example, from a Bayes factor viewpoint).

To apply the steepest ascent method to the nonlinear function $m(\boldsymbol{X} \mid f, \pi)$ of $f \in \Gamma^f$, we start from an initial guess $f_1(\cdot) = \frac{1}{\sigma_1}\phi(\frac{\cdot}{\sigma_1}) \in \mathcal{E}^f$ for some $\sigma_1 > 0$. After the end of the n-th iteration, we have $f_*(\cdot) = f_n(\cdot) = \sum_{i=1}^{n} \frac{\alpha_i}{\sigma_i}\phi(\frac{\cdot}{\sigma_i})$ where $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$ and $\sigma_i \geq 0$. For the $(n+1)$-th iteration cycle, we first evaluate $Gm_{f_n}(g - f_n) = \frac{d}{dt}m(\boldsymbol{X} \mid (1-t)f_n + tg, \pi)|_{t=0} = \frac{d}{dt}[\int \prod_{i=1}^{k}\{(1-t)f_n(X_i - \theta) + tg(X_i - \theta)\}d\pi(\theta)]_{t=0} = \sum_{i=1}^{k} \int g(X_i - \theta) \prod_{j \neq i}^{k} f_n(X_i - \theta)d\pi(\theta) - km(f_n, \pi)$. For $g(\cdot) = \frac{1}{\tau}\phi(\frac{\cdot}{\tau}) \in \mathcal{E}^f$, $Gm_{f_n}(g - f_n)$ is a function of $\tau \geq 0$ and its maximum value can be obtained numerically. Once the maximum attaining $\tau = \sigma_{n+1}$ and the corresponding g is obtained, we go to step [4] of the ascent method. Now, $\frac{d}{dt}m(\boldsymbol{X} \mid (1-t)f_n + tg, \pi) = \sum_{i=1}^{k} \int (g(X_i - \theta) - f_n(X_i - \theta)) \prod_{j \neq i}^{k}\{(1-t)f_n(X_j - \theta) + tg(X_j - \theta)\}d\pi(\theta)$ which is a $(k-1)$ degree polynomial $p(t)$ in t. Thus, $\max_{0 \leq t \leq 1} m(\boldsymbol{X} \mid (1-t)f_n + tg, \pi) = m(\boldsymbol{X} \mid (1-t_*)f_n + t_*g, \pi)$ where $0 < t_* \leq 1$ is a root of $p(t)$ or $t_* = 1$. We now update f_n to $f_{n+1}(\cdot) = \sum_{i=1}^{n}(1-t_*)\frac{\alpha_i}{\sigma_i}\phi(\frac{\cdot}{\sigma_i}) + \frac{t_*}{\sigma_{n+1}}\phi(\frac{\cdot}{\sigma_{n+1}})$ and proceed to the $(n+2)$-th iteration cycle.

In the numerical example, we assume a noninformative prior $\pi(\theta) = 1$ and start our steepest ascent algorithm with the initial guess $\sigma_1 = 38$ (based on the data standard deviation 37.74). For our numerical work, we bound the domain of σ to $[0.01, \infty)$ ($\sigma \to 0$ causes several numerical problems). After 49 iterations, the ascent algorithm converges (Gâteaux derivative = 0) to the sampling distribution $f_*^{(1)}(x) = \int \phi(\frac{x}{\sigma})dG_1(\sigma)$ where an approximate description of the 50 point mixing distribution G_1 is given in Table 3. For this $f_*^{(1)}$, we obtain $m(\boldsymbol{X} \mid f_*^{(1)}, \pi) =$

Table 3. Mixing distribution G_1 in Example 2 (σ values within a 0.5 interval are grouped).

σ	0.01	4.71	10.53	12.33	13.12	13.73	14.23	(14.5, 15]	(15, 15.5]
Prob.	.003	.020	.032	.021	.021	.021	.021	.041	.076

σ	(15.5, 16]	38	(48.5, 49]	(49, 49.5]	49.92	51.71	54.73	55.23
Prob.	.167	.132	.179	.170	.028	.027	.015	.026

1.246×10^{-31}. The maximum $m(\boldsymbol{X} \mid f, \pi)$ value obtained by Pericchi and Pérez (1994) was 2.451×10^{-32}. Thus, in terms of Bayes factors, $f_*^{(1)}$ is preferred 5.1 times higher than the best model of Pericchi and Pérez (1994). Instead of the mixing prior $= 1/\sigma$ of Pericchi and Pérez (1994) which concentrates mass around 0, our analysis suggests that the data prefer a more dispersed mixing distribution G_1 "centered" around 38.0.

A referee pointed out that Pericchi and Pérez (1994) used prior on both θ and σ ($\pi(\theta, \sigma) = 1/\sigma$) whereas we only use prior $\pi(\theta) = 1$ on the location parameter θ. We redid the Pericchi and Pérez (1994) computations with prior only on the location parameter θ. Consider the Pericchi and Pérez (1994) case when the sampling density $f = f_N = \mathrm{N}(\theta, \sigma^2)$. For the prior $\pi(\theta) = 1$, we obtain $m(\boldsymbol{X} \mid \sigma, f_N, \pi) = \int \prod_{i=1}^{15} f_N(X_i \mid \theta, \sigma) d\pi(\theta)$. We then maximize this marginal over σ, i.e., $\bar{m}(\boldsymbol{X} \mid f_N, \pi) = \max_{\sigma > 0} m(\boldsymbol{X} \mid \sigma, f_N, \pi)$. These calculations are repeated for each of the five sampling distribution choices of Pericchi and Pérez (1994). The overall maximum over the five choices is again attained at $f = \mathrm{N}(\theta, \sigma^2)$ and this maximum value is $\bar{m}(\boldsymbol{X} \mid f_N, \pi) = 5.113 \times 10^{-32}$. In comparison, the maximum obtained by our method is $m(\boldsymbol{X} \mid f_*^{(1)}, \pi) = 1.246 \times 10^{-31}$ which is more than 2.4 times higher.

In Fig. 1 we show a dotplot of Darwin's data along with the initial, intermediate and the final normal scale mixture density obtained by our algorithm. All three densities are centered at the data average $\bar{X} = 20.933$. The solid curve shows the initial normal density with $\sigma_1 = 38$. The normal scale mixture density obtained after the 16th iteration is shown in the picture with small and thick dashes, whereas the large and light dashes represent the final density obtained at the 49th iteration. These latter two densities are completely indistinguishable at the resolution of the figure. The value of $m(\boldsymbol{X} \mid f, \pi)$ increased from 1.119×10^{-31} at the 16th iteration to 1.246×10^{-31} at the final iteration. The painful slowness and small steps of the steepest ascent algorithm also becomes visible in this example. From the 16th iteration onwards ($i = 16/2 = 8$), the algorithm keeps on iterating between $\sigma_{2i} \in (48.5, 49.5]$ and $\sigma_{2i+1} \in (15, 16]$ until it converges in the 50th iteration ($i = 25$). It is reasonable to guess that there are optimal $\tau_1 \in (48.5, 49.5]$ and $\tau_2 \in (15, 16]$ such that $\sigma_{16} = \tau_1$ and $\sigma_{17} = \tau_2$ would have led to a quick convergence, but the algorithm fails to take this quick step.

What is learned about Darwin's data from this analysis? At a first glance, an empirical Bayes analysis may plan to use a normal sampling model with standard

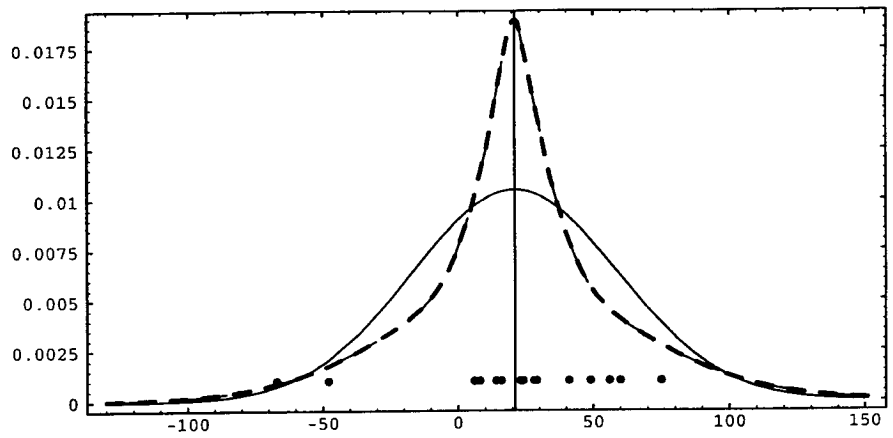

Fig. 1. Dotplot of Darwin's data and plots of the initial f_1, intermediate f_{16} and final $f_*^{(1)}$ sampling densities in Example 2. f_1 = solid curve, f_{16} = short thick dashes, $f_*^{(1)}$ = long light dashes. The latter two densities are indistinguishable.

deviation = 38 (based on the data standard deviation 37.74). However, in the Bayesian model we consider, our analysis and Fig. 1 show that the data favor a different sampling density which is sharper than the normal density in the center and almost similar to the normal density in the tail.

Example 3. Here we again consider Darwin's data. However, here we maximize the pseudo-marginal or the predictive marginal.

To distinguish between observed data and running variables, we use upper case letters (e.g. X) to denote observed realization of random variables (data), whereas lower case letters (e.g. x) are used to denote random variables or running variables. Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be the observed data vector ($k = 15$). Also, let $\boldsymbol{X}_{(i)}$ denote the $(k-1) \times 1$ data vector with i-th observation X_i deleted. Let $\boldsymbol{f}(\boldsymbol{X} - \theta) = \prod_{i=1}^k f(X_i - \theta)$ and $\boldsymbol{f}(\boldsymbol{X}_{(i)} - \theta) = \prod_{j \neq i} f(X_j - \theta)$ be the joint likelihood of $\boldsymbol{X}$ and $\boldsymbol{X}_{(i)}$ respectively. Let $m_k(\boldsymbol{X} \mid f, \pi) = \int \boldsymbol{f}(\boldsymbol{X} - \theta) d\pi(\theta)$ and $m_{k-1}(\boldsymbol{X}_{(i)} \mid f, \pi) = \int \boldsymbol{f}(\boldsymbol{X}_{(i)} - \theta) d\pi(\theta)$. With these notations, the posterior distribution of θ given only $\boldsymbol{X}_{(i)}$ is $d\pi(\theta \mid \boldsymbol{X}_{(i)}, f, \pi) = \boldsymbol{f}(\boldsymbol{X}_{(i)} - \theta) d\pi(\theta) / m_{k-1}(\boldsymbol{X}_{(i)} \mid f, \pi)$. The cross validated predictive density of the random variable x_i given the remaining $(k-1)$ observed data $\boldsymbol{X}_{(i)}$ is then $f(x_i \mid \boldsymbol{X}_{(i)}, f, \pi) = \int f(x_i - \theta) d\pi(\theta \mid \boldsymbol{X}_{(i)}, f, \pi) = m_k(\boldsymbol{Y} \mid f, \pi) / m_{k-1}(\boldsymbol{X}_{(i)} \mid f, \pi)$ where $\boldsymbol{Y} = (X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_k)^T$. A check of this predictive density $f(x_i \mid \boldsymbol{X}_{(i)}, f, \pi)$ with the observed X_i indicates how compatible X_i is to the rest of the data $\boldsymbol{X}_{(i)}$ (in view of the model). Several model selection criteria have been suggested based on this predictive distribution. We consider the "pseudo-marginal density" $D(f) = \prod_{i=1}^k f(X_i \mid \boldsymbol{X}_{(i)}, f, \pi)$ (a modification of the marginal density $m(\boldsymbol{X} \mid f, \pi)$ from the predictive viewpoint) which has been suggested as a model selection crietrion by Gelfand *et al.* (1992), and Geisser and Eddy (1979). We plan to select $f_* \in \Gamma^f$ for which $D(f_*) = \max_{f \in \Gamma^f} D(f)$. Our focus here is not comparing the relative merits of different

Table 4. Mixing distribution G_2 in Example 3.

σ	0.01	4.46	10.51	12.14	38	46.11	51.05	51.18	53.08	55.41
Prob.	0.01	0.06	0.15	0.07	0.44	0.02	0.10	0.01	0.04	0.10

model selection criteria but to see how our steepest ascent method works in a complicated setup.

Since $D(f) \geq 0$, maximizing $D(f)$ is equivalent to maximizing $\log D(f) = k \log m_k(\boldsymbol{X} \mid f, \pi) - \sum_{i=1}^{k} \log m_{k-1}(\boldsymbol{X}_{(i)} \mid f, \pi)$. The Gâteaux derivative equals $G \log D(g - f) = kGm_k(g - f, \boldsymbol{X})/m_k(\boldsymbol{X} \mid f, \pi) - \sum_{i=1}^{k} \{Gm_{k-1}(g - f, \boldsymbol{X}_{(i)})/ m_{k-1}(\boldsymbol{X}_{(i)} \mid f, \pi)\}$. The Gâteaux derivative $Gm_k(g-f, \boldsymbol{X})$ of the marginal $m_k(\boldsymbol{X} \mid f, \pi)$ is obtained in Example 2 whereas $Gm_{k-1}(g-f, \boldsymbol{X}_{(i)})$ is same as $Gm_k(g-f, \boldsymbol{X})$ with k replaced by $(k-1)$ and $\boldsymbol{X}$ replaced by $\boldsymbol{X}_{(i)}$. Thus, the implementation of steepest ascent method here only needs minor modifications to the algorithm already used in Example 2.

In the numerical example, we again assume prior $\pi(\theta) = 1$ and start our algorithm from the initial guess $\sigma_1 = 38$. After 9 iterations, the ascent algorithm converges to the normal scale mixture distribution $f_*^{(2)}(x) = \int \phi(\frac{x}{\sigma}) dG_2(\sigma)$ where the 10-point mixing distribution G_2 is described in Table 4. The value of the "predictive marginal" is $D(f_*^{(2)}) = 2.45 \times 10^{-33}$. Notice the similarity between $f_*^{(2)}$ and $f_*^{(1)}$ or between G_2 in Table 4 and G_1 in Table 3. Thus, maximizing the marginal and maximizing the "predictive marginal" obtain (somewhat) similar models for Darwin's data.

4. Joint perturbations of the prior and the sampling distribution

We briefly mention here the technique for considering joint perturbations of the prior P and the sampling distribution f when multiple observations $\boldsymbol{X} = (X_1, \ldots, X_k)$ are observed from $f(\cdot)$. The technique is basically a combination of the methods of Sections 2 and 3. Assume $f(x \mid \theta) = f(x - \theta)$ and f and P are generic members of the classes Γ^f and Γ^P. Let $\rho(\boldsymbol{X}, f, P)$ be our posterior quantity of interest and we want to find its range, i.e., $\inf_{f \in \Gamma^f, P \in \Gamma^P} \rho(\boldsymbol{X}, f, P)$ and $\sup_{f \in \Gamma^f, P \in \Gamma^P} \rho(\boldsymbol{X}, f, P)$. We only describe the supremum problem.

Suppose $\rho(\boldsymbol{X}, f, P) = \rho(h, \boldsymbol{X}, f, P) = E(h(\theta) \mid \boldsymbol{X}, f, P)$. Also, suppose $\Gamma^f = \Gamma_i$, $i = 3, \ldots, 6$ and $\Gamma^P = \Gamma_j$, $j = 1, \ldots, 7$. Then, for each fixed $f \in \Gamma^f$, finding $\sup_{P \in \Gamma^P} \rho(h, \boldsymbol{X}, f, P)$ can be reduced to maximization over $P \in \mathcal{E}^P$ by the techniques of Section 2 ($\mathcal{E}^P$ is also described in Section 2). $\mathcal{E}^P$ is at most one-dimensional. Next, for a fixed $P \in \mathcal{E}^P$, we can find $\sup_{f \in \Gamma^f} \rho(h, \boldsymbol{X}, f, P) = \overline{\rho}(h, \boldsymbol{X}, P)$ by the steepest ascent method described in Section 3. What remains is a maximization over $\mathcal{E}^P$, i.e., at most a one-dimensional maximization.

If $\rho(\boldsymbol{X}, f, P)$ is not ratio-linear, the situation is more difficult. However, we can always use the steepest ascent method whenever we know the extreme point

class $\mathcal{E}$ of Γ. Assume $\Gamma^f = \Gamma_i$, $i = 3, \ldots, 6$ and $\Gamma^P = \Gamma_j$, $j = 1, \ldots, 6$. A possible strategy in this case is the following:

(i) Start from an initial guess about the sampling distribution $f = f_*$. Use the steepest ascent method to obtain $\max_{P \in \Gamma^P} \rho(\boldsymbol{X}, f_*, P) = \rho(\boldsymbol{X}, f_*, P_S)$ (say).

(ii) Now, fix prior $P = P_S$, and use the steepest ascent method to obtain $\max_{f \in \Gamma^f} \rho(\boldsymbol{X}, f, P_S) = \rho(\boldsymbol{X}, f_S, P_S)$ (say).

(iii) Put $f_* = f_S$ and go back to step (i) until no significant increase of $\rho(\boldsymbol{X}, f, P)$ is obtained.

The implementations of these algorithms will be along the lines of Examples 1–3, though we have not tried them in any example.

5. Closing remarks

In this article, we hope to achieve two competing Bayesian robustness goals; a sampling model sensitivity study of posterior quantities by letting the sampling distribution vary in a nonparametric class and a joint sensitivity study w.r.t. the sampling distribution and the prior. Most previous works in this area dealt only with one observation X from the sampling distribution $f(x \mid \theta)$. We have been successful in dealing with multiple i.i.d. observations $X_1, \ldots, X_k$. The iterative steepest descent method that we develop is very powerful, its scope appears to go much beyond our specific goals. The use of this method in obtaining ranges of nonlinear posterior quantities is already illustrated in Example 1, further applications remain to be explored.

References

Andrews, D. F. and Herzberg, A. M. (1985). *Data: a Collection of Problems from Many Fields for the Student and Research Worker*, Springer, New York.

Basu, S. (1994). Posterior sensitivity to the sampling density and the prior: more than one observation, Tech. Report, No. 66, University of Arkansas.

Basu, S. (1996). Local sensitivity, functional derivatives and nonlinear posterior quantities, *Statist. Decisions*, **14**, 405–418.

Basu, S. and DasGupta, A. (1995*a*). Robust Bayesian analysis with distribution bands, *Statist. Decisions*, **13**, 333–349.

Basu, S. and DasGupta, A. (1995*b*). Robustness of standard confidence intervals for location parameters under departure from normality, *Ann. Statist.*, **23**, 1433–1442.

Bayarri, M. J. and Berger, J. (1993). Robust Bayesian analysis of selection models, Tech. Report, No. 93-96, Purdue University, Indiana.

Berger, J. (1994). An overview of robust Bayesian analysis, *Test*, **3**, 5–59.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, New York.

Dey, D. K., Ghosh, S. K. and Lou, K. (1997). *On Local Sensitivity Measures in Bayesian Analysis*, IMS Lecture Notes—Monograph Series, Vol. 29, 21–39, Hayward, California.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection, *J. Amer. Statist. Assoc.*, **74**, 153–160.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementations via sampling-based methods, *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 147–167, Oxford University Press, Oxford.

Ghosh, S. K. and Dey, D. K. (1994). Sensitivity diagnostics and robustness issues in Bayesian inference, Tech. Report, Nol. 94-30, University of Connecticut.

Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

Lavine, M. (1991). Sensitivity in Bayesian analysis: the prior and the likelihood, *J. Amer. Statist. Assoc.*, **86**, 396–399.

Luenberger, D. G. (1968). *Optimization by Vector Space Methods*, Wiley, New York.

Pericchi, L. R. and Pérez, M. E. (1994). Posterior robustness with more than one sampling model (with discussion), *J. Statist. Plann. Inference*, **40**, 279–294.

Sivaganesan, S. and Berger, J. (1989). Ranges of posterior measures for priors with unimodal contamination, *Ann. Statist.*, **17**, 868–889.

Smith, A. F. M. (1983). Bayesian approaches to outliers and robustness, *Specifying Statistical Models* (eds. J. P. Florens, M. Mouchart, J. P. Raoult, L. Simar and A. F. M. Smith), 13–35, Springer, New York.

Srinivasan, C. and Truszczynska, H. (1990). On the ranges of posterior quantities, Tech. Report, No. 294, University of Kentucky.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions, *Contributions to Statistics and Probability Essays in Honor of Harold Hotelling* (ed. I. Olkin), 418–485, Stanford University Press, California.