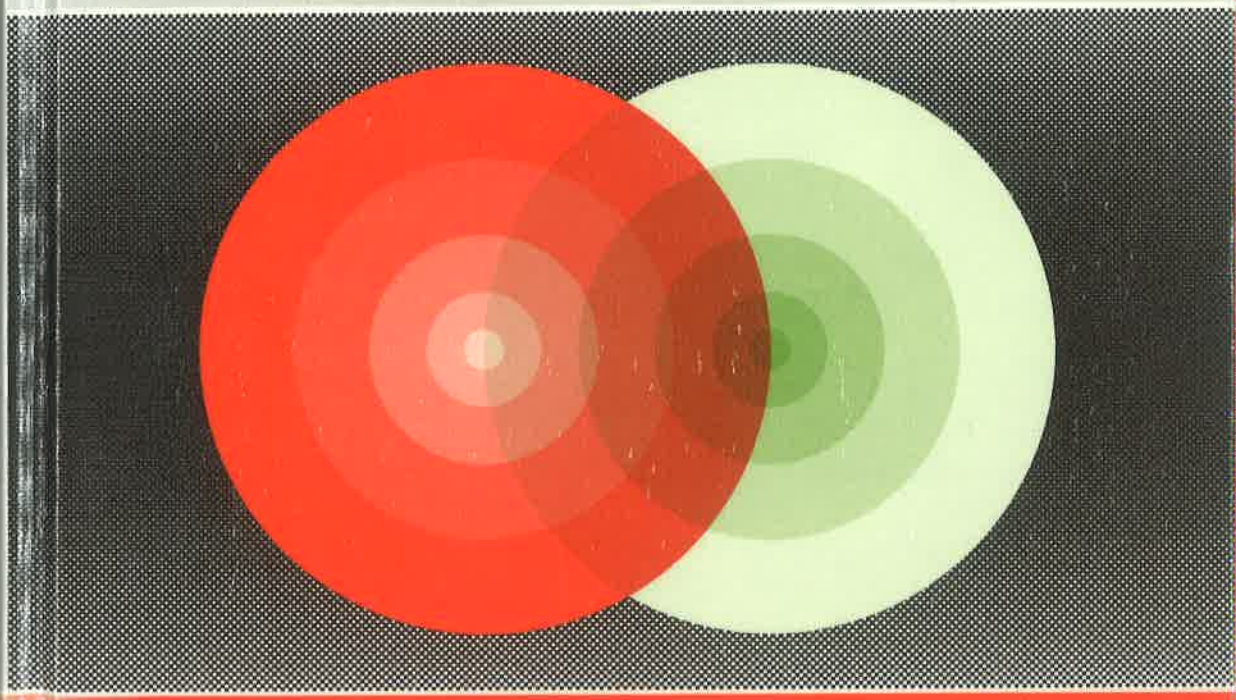
Maximum Entropy and Bayesian Methods

Santa Barbara, California, U.S.A., 1993

Edited by

Glenn R. Heidbreder

Kluwer Academic Publishers



Fundamental Theories of Physics

LOCAL POSTERIOR ROBUSTNESS WITH PARAMETRIC PRIORS : MAXIMUM AND AVERAGE SENSITIVITY

Sanjib Basu
Department of Mathematical Sciences
University of Arkansas, Fayetteville, AR 72701

Sreenivasa Rao Jammalamadaka *
Department of Statistics and Applied Probability
University of California, Santa Barbara, CA 93106

and Wei Liu
Ciba Geigy, Summit, NJ 07901

ABSTRACT. The local sensitivity of a posterior quantity $\rho(P)$ to the choice of the prior P is considered. When the prior P_λ is indexed by parameter λ, a natural measure is the total derivative of $\rho(P_\lambda)$ w.r.t. λ. Total derivative, however, is direction specific. To measure the local sensitivity of $\rho(P_\lambda)$ to specification of λ, one may either use the norm (maximum over all directions) of the total derivative or alternatively, the average sensitivity which evaluates the average of this total derivative over all directions. Simple expressions are given for the maximum and average sensitivity which make their evaluations very easy. Discussion and several examples illustrate implications of these ideas.

1. Introduction

Bayesian paradigm requires one to specify two parametric models; the sampling density $f(X|\theta)$ and the prior $P(\theta)$. However, in practice, knowledge about these models are never accurate, and such specifications are only approximations or guesses at best. Hence, sensitivity of the final action to deviations of these various inputs from their idealized models is of much concern. As Tukey (1960) writes, "A tacit hope in ignoring deviations from ideal models was that they would not matter; that statistical procedures which are optimal under the strict model would still be approximately optimal under the approximate model. Unfortunately, it turned out that this hope was often drastically wrong; even mild deviations often have much larger effects than were anticipated by most statisticians".

Robustness studies, in both Classical and Bayesian statistics, can broadly be divided into two subgroups; global sensitivity analysis and local or infinitesimal approach. The former examines the effect of misspecification, when the true model may or may not be close to the idealized one. In the Bayesian context, global sensitivity to misspecification of the prior has been expounded by many, see Berger (1993), Wasserman (1992), Basu and DasGupta (1992), Rivier et al. (1990), and the references therein. In contrast, local sensitivity studies explore the effect of infinitesimal perturbations from the idealized model. Recent advances in this area include Rodríguez (1994), Ruggeri and Wasserman (1993), and

*Research supported in part by ONR Grant number N00014-93-1-0174.

G. R. Heidbreder (ed.), Maximum Entropy and Bayesian Methods, 97–106.

Skilling (1990). Our efforts in this article will be directed towards studying local sensitivity of Bayesian analysis to the choice of the prior.

Formally, we observe data X from the sampling density $f(x|\theta)$. The observed likelihood function $f(X|\theta)$ will be denoted by $\ell(\theta)$ (with conditioning X understood), and $P(\cdot)$ will denote the prior distribution on θ. Let $m(P) = \int_\Theta \ell(\theta)\, dP(\theta)$ denote the marginal w.r.t. prior P. Given the likelihood $\ell(\cdot)$ and the prior $P(\cdot)$, the posterior probability distribution, defined as $P(A|X) = \frac{1}{m(P)} \int_A \ell(\theta)\, dP(\theta)$ for any set A, will be denoted by $P(\cdot|X)$ (with dependence on $\ell(\theta)$ understood). Similarly, $\pi(\cdot)$ and $\pi(\cdot|X)$ will respectively denote the prior and the posterior densities (whenever appropriate). We will use $\rho(P)$ or ρ_P to denote a posterior quantity (such as the posterior mean) corresponding to the prior P.

As we mentioned before, prior specification is typically imprecise. Thus, in reality, we have a multiplicity of P as possible choices of the prior, from which we choose a single P_0 as our idealized prior. We will use $\mathcal{P}$ to denote the class of all plausible priors. Sometimes, the prior class $\mathcal{P}$ is indexed by a parameter. For example, we may decide to use $P = N(\mu, \tau^2)$, but are not sure about any specific values of μ and τ^2, thus leading to the class $\{\, N(\mu, \tau^2) : (\mu, \tau^2)^T \in (-\infty, \infty) \otimes (0, \infty) \,\}$. Such parametric classes will be denoted by $\mathcal{P}_\Lambda = \{P_\lambda : \lambda \in \Lambda\}$. We will often assume that the indexing set $\Lambda \subseteq \Re^k$. In other situations, when any particular parametric form for the prior is not apparent, one uses a nonparametric class, such as an ε-contamination class $\mathcal{P}^\varepsilon$. An ε-contamination class arises when one is $100(1-\varepsilon)\%$ certain about the idealized P_0 as the choice of the prior, and $100\,\varepsilon\%$ uncertain ($0 \le \varepsilon < 1$), thus resulting in the class $\mathcal{P}^\varepsilon = \{P : P = (1-\varepsilon)P_0 + \varepsilon Q\}$ where Q is any arbitrary prior distribution.

When we have a class $\mathcal{P}$ of plausible priors, and an idealized prior P_0, the first question that comes to mind is : "if the true prior Q in $\mathcal{P}$ is close to the idealized P_0, is it guaranteed that $\rho(Q)$ will be close to $\rho(P_0)$?" In a limiting sense, this amounts to continuity of $\rho(P)$ (as a function of P) at $P = P_0$, and in the terminology of classical robustness literature, this corresponds to Hampel's (1971) definition of *qualitative robustness*. Note that we are posing the question in terms of $\rho(P)$, however, an exactly similar question can be posed in terms of the posterior distribution $P(\cdot|X)$. If *qualitative robustness* is achieved, a second natural question to ask would be : "is the change in $\rho(P)$ bounded by the change in P ?". To formalize this question, suppose $d(\cdot,\cdot)$ is a metric on the space of priors, and $\nu(\cdot,\cdot)$ is a metric on the space of the posterior quantities $\rho(P)$. Then, we can pose our question as follows: "does $\exists$ an $\alpha > 0$ such that $\nu(\rho(P), \rho(P_0)) \le M\,[d(P(\cdot), P_0(\cdot)]^\alpha$ for some $M > 0$?". Mathematically, this is a Lipschitz condition of order α. Basu, Jammalamadaka and Liu (1993) termed this second notion as *stability*, and studied the *qualitative robustness* and *stability* of $\rho(P)$ and $P(\cdot|X)$.

Qualitative robustness and *stability* are very necessary but rather weak characterizations of robustness. A local sensitivity study should also explore the rate of change of $\rho(P)$ as P deviates infinitesimally from the idealized P_0. If the prior class $\mathcal{P} = \mathcal{P}_\Lambda$ is parametric and $\Lambda \subseteq \Re$, this is easy. For $P_0 = P_{\lambda_0}$ and $\rho(P_\lambda) = \rho(\lambda)$, one simply computes the derivative $\rho'(\lambda) = \frac{d}{d\lambda}\rho(\lambda)$ at $\lambda = \lambda_0$. If $\rho'(\lambda_0)$ is small, it suggests that $\rho(\lambda)$ is not sensitive to mild perturbations of P_λ around $\lambda = \lambda_0$. The situation gets complicated when $\Lambda \subseteq \Re^k$. We consider a more complex setup when $\boldsymbol{\rho}$ is also multidimensional, i.e., $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]^T$. A proper concept of derivative in such multivariate situations is the total derivative. In section

2.1., we establish sufficient conditions for total differentiability of a posterior quantity $\boldsymbol{\rho}(\boldsymbol{\lambda})$. However, total derivative is direction specific, its value depends on the direction of deviation $\boldsymbol{\lambda}$ from the idealized value $\boldsymbol{\lambda}_0$. We thus evaluate the norm of the total derivative, or its maximum value over all directions. Theorem 2 supplies an easy formula for evaluation of this norm. An alternative viewpoint would suggest computing the average of the total derivative over all directions. This leads us to average sensitivity. Section 2.3. discusses this issue and again supplies simple expressions for ease of computation. Several univariate and multivariate applications are explored in section 3.. Finally, section 4. briefly discusses the issue of quantification of local sensitivity over nonparametric prior classes.

Use of derivatives to quantify the sensitivity of a posterior quantity is not new. Diaconis and Freedman (1986), and Ruggeri and Wasserman (1993) evaluated norm of Fréchet derivatives over the class of all signed measures and/or its appropriate nonparametric subclasses. Rodríguez (1994) used the concept of Lie derivative to quantify the intrinsic robustness of a hypothesis space. To our knowledge, such explorations over parametric classes have not been explicitly considered before.

2. Parametric prior classes

2.1. Total derivative

Mathematical and numerical convenience often attracts one to use a prior of a special parametric form (this is more true in multivariate situations). For example, in a linear model setup : $\underset{\sim}{Y} \sim N(X\underset{\sim}{\beta}, \sigma^2 I)$ with $\sigma^2 > 0$ known, it is common to use a $N(\mu, \Gamma)$ prior for $\underset{\sim}{\beta}$. Even if such a formulation is justified, specification of the prior hyperparameters poses a secondary problem, which is often handled through Empirical Bayes and/or Hierarchical Bayes methods, or the hyperparameters are specified as inputs by the user. Again, these inputs are never exactly accurate, so that local sensitivity to a particular choice of the hyperparameters is of concern.

Let $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k]^T$ denote a generic element of Λ, and let $\mathcal{P}_\Lambda = \{P_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \Lambda\}$ be the class of all plausible parametric priors from which we choose $P_{\boldsymbol{\lambda}_0}$ as an idealized prior. We will assume that Λ is an open subset in $\Re^k$ so that for each $\boldsymbol{\lambda}_0 \in \Lambda$, $\exists$ a neighborhood N_0 of $\boldsymbol{\lambda}_0$ such that $\boldsymbol{\lambda}_0 \in N_0 \subseteq \Lambda$. Let $\boldsymbol{\rho}(P_{\boldsymbol{\lambda}}) = \boldsymbol{\rho}(\boldsymbol{\lambda})$ be the posterior quantity of interest. $\boldsymbol{\rho}(\boldsymbol{\lambda})$ may be univariate (a single posterior quantity), or multivariate (a vector of such quantities); in general, we will assume that $\boldsymbol{\rho}$ is n-dimensional and $\boldsymbol{\lambda}$ is k-dimensional, i.e., $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]^T : \Lambda \subseteq \Re^k \mapsto \Re^n$. Often, we will focus on ratio-linear posterior quantities, i.e., $\boldsymbol{\rho}(\boldsymbol{\lambda}) = [\rho_1(\boldsymbol{\lambda}), \ldots, \rho_n(\boldsymbol{\lambda})]^T = [\frac{1}{m(P_{\boldsymbol{\lambda}})} \int h_i(\theta)\, \ell(\theta)\, dP_{\boldsymbol{\lambda}}(\theta)]_{i=1}^n$. Such quantities will be denoted by $\boldsymbol{\rho}^{\boldsymbol{h}}(\boldsymbol{\lambda})$.

Our concern is the local sensitivity of the posterior quantity $\boldsymbol{\rho}(\boldsymbol{\lambda})$ to the particular choice of the parameter $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$. The weaker local sensitivity properties of $\boldsymbol{\rho}(\boldsymbol{\lambda})$, namely, *qualitative robustness* and *stability*, are explored in Basu, Jammalamadaka and Liu (1993). Here, we focus on measuring the rate of change of $\boldsymbol{\rho}(\boldsymbol{\lambda})$ to small perturbations in $\boldsymbol{\lambda}$, in other words, the derivative of $\boldsymbol{\rho}(\boldsymbol{\lambda})$ w.r.t. $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$. Let $\bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_*) = [[\frac{\delta\rho_i(\boldsymbol{\lambda}_*)}{\delta\lambda_j}]_{j=1}^k]_{i=1}^n$ denote the matrix of partial derivatives of $\boldsymbol{\rho}$ w.r.t. $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}_*$. However, the appropriate derivative in multivariate calculus is not the partial derivative, but rather, the *total derivative* $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_*}$. The function $\boldsymbol{\rho} : \Lambda \subseteq \Re^k \mapsto \Re^n$ is called (totally) differentiable at $\boldsymbol{\lambda}_* \in \Lambda$ if $\exists$ a linear

function $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_*} : \Re^k \mapsto \Re^n$ such that $\frac{\|\boldsymbol{\rho}(\boldsymbol{\lambda}_*+\boldsymbol{v}) - \boldsymbol{\rho}(\boldsymbol{\lambda}_*) - T\boldsymbol{\rho}_{\boldsymbol{\lambda}_*}(\boldsymbol{v})\|_n}{\|\boldsymbol{v}\|_k} \to 0$ as $\|\boldsymbol{v}\|_k \to 0$ ($\|\boldsymbol{v}\|_k = \sqrt{v_1^2 + \ldots + v_k^2}$ denotes the standard Euclidean norm on the k-dimensional space $\Re^k$). Note that each $\boldsymbol{\lambda}_* \in \Lambda$ gives rise to a distinct linear transformation $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_*}$.

The existence of the total derivative $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_*}$, however, is easier to prove through the existence and continuity of the partial derivatives $\frac{\delta\rho_i(\boldsymbol{\lambda}_*)}{\delta\lambda_j}$. A well known result in differential calculus states that the total derivative $T\boldsymbol{\rho}$ exists over a neighborhood N_0 of $\boldsymbol{\lambda}_0$ and is continuous on the space $\mathcal{L}(\Re^k, \Re^n)$ of linear transformations from $\Re^k \mapsto \Re^n$ iff the partial derivatives $\frac{\delta\rho_i}{\delta\lambda_j}$ exist and are continuous on N_0 $\forall\, 1 \le i \le n,\ 1 \le j \le k$ (Rudin (1976), p 219). We use this result to investigate differentiability of the ratio-linear posterior quantity $\boldsymbol{\rho}^{\boldsymbol{h}}(\boldsymbol{\lambda})$ in Theorem 1. It is easier to state the result in terms of densities, thus we will assume that each $P_{\boldsymbol{\lambda}} \in \mathcal{P}_\Lambda$ has a density $\pi_{\boldsymbol{\lambda}}(\theta) = \pi(\theta, \boldsymbol{\lambda})$.

Theorem 1 *Let N_0 be a neighborhood of $\boldsymbol{\lambda} \in \Lambda$. Assume $|\ell(\theta)| \le M_0$, and for all $1 \le i \le n$, $|h_i(\theta)\,\ell(\theta)| \le M_i$ $\forall\, \theta \in \Theta$. We further assume the following :*

(i) For each $1 \le j \le k$, the partial derivative $\frac{\delta}{\delta\lambda_j}\pi(\theta, \boldsymbol{\lambda})$ exists $\forall\ (\theta, \boldsymbol{\lambda}) \in \Theta \otimes N_0$, and is continuous as a function of $\boldsymbol{\lambda}$ for every $\theta \in \Theta$.

(ii) For every $1 \le j \le k$, $\exists$ a function $g_j(\theta)$ on Θ such that (a) $g_j(\theta) \ge 0$ $\forall\, \theta \in \Theta$, (b) $\int_\Theta g_j(\theta\, d\mu(\theta) \le L_j < \infty$, and (c) $|\frac{\delta}{\delta\lambda_j}\pi(\theta, \underset{\sim}{\lambda})| \le g_j(\theta)$ $\forall\ (\theta, \underset{\sim}{\lambda}) \in \Theta \otimes N_0$.

Then the total derivative $T\boldsymbol{\rho}^{\boldsymbol{h}}_{\boldsymbol{\lambda}}$ of the posterior quantity $\boldsymbol{\rho}^{\boldsymbol{h}}(\boldsymbol{\lambda})$ exists for $\boldsymbol{\lambda} \in N_0$ and $T\boldsymbol{\rho}^{\boldsymbol{h}}$ is continuous on $\mathcal{L}(\Re^k, \Re^n)$.

Proof : Let $N_i(\boldsymbol{\lambda}) = \int h_i(\theta)\,\ell(\theta)\,\pi(d\theta, \boldsymbol{\lambda})$, thus $\rho_i(\boldsymbol{\lambda}) = \frac{N_i(\boldsymbol{\lambda})}{m(\pi_{\boldsymbol{\lambda}})}$, $1 \le i \le n$. The conditions of the theorem ensure that for $1 \le i \le n$, $1 \le j \le k$, and $\forall\, \boldsymbol{\lambda} \in N_0$, the partial derivative $\frac{\delta}{\delta\lambda_j}N_i(\boldsymbol{\lambda})$ exists and $= \int h_i(\theta)\,\ell(\theta)\,\frac{\delta}{\delta\lambda_j}\pi(\theta, \boldsymbol{\lambda})$ by the Dominated Convergence theorem. Continuity of $\xi_j(\boldsymbol{\lambda}) = \frac{\delta}{\delta\lambda_j}\pi(\theta, \boldsymbol{\lambda})$ and another application of D.C.T. prove that $\frac{\delta}{\delta\lambda_j}N_i(\boldsymbol{\lambda})$ is continuous in $\boldsymbol{\lambda} \in N_0$. Similarly, $\frac{\delta}{\delta\lambda_j}m(\pi_{\boldsymbol{\lambda}})$, and hence $\frac{\delta}{\delta\lambda_j}\rho_i^{\boldsymbol{h}}(\boldsymbol{\lambda})$ exist and are continuous in $\boldsymbol{\lambda}$ for every $1 \le i \le n$, $1 \le j \le k$. The proof of the theorem follows ∎

2.2. Maximum sensitivity

Our interest lies in measuring the rate of change of $\boldsymbol{\rho}(\boldsymbol{\lambda})$ as $\boldsymbol{\lambda}$ deviates from $\boldsymbol{\lambda}_0$. In particular, since we are not sure about any specific direction of deviation, we would like to find the maximum rate of change of $\boldsymbol{\rho}(\boldsymbol{\lambda})$ over all directions. However, note that the total derivative $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}$ is a linear function of $\boldsymbol{v} \in \Re^k$, i.e., even if we fix a direction $\boldsymbol{v}$, $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(c\cdot\boldsymbol{v}) = c\cdot T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})$ for any $c > 0$. Hence, $\sup_{\text{all } \boldsymbol{v} \neq \boldsymbol{0}} \|T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})\|_n$ is clearly infinite. What we need is the concept of the *norm* of a linear functional, defined by $\|T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}\| = \sup_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{1}{\|\boldsymbol{v}\|_k}\|T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})\|_n = \sup_{\|\boldsymbol{v}\|_k = c} \frac{1}{c}\|T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})\|_n$. Here $c > 0$ can be chosen arbitrarily small to make sure that $\boldsymbol{\rho}(\boldsymbol{\lambda}_0 + \boldsymbol{v})$ is well defined for all $\{\boldsymbol{v} : \|\boldsymbol{v}\|_k = c\}$ (see definition of $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}$). Also, note that

if $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are univariate, i.e., $k = n = 1$, then $\frac{1}{||v||_1}||T\rho_{\lambda_0}(v)||_1 = |\frac{d\rho_\lambda(\lambda_0)}{d\lambda}|$. Thus, $\frac{1}{||v||_k}||T\rho_{\boldsymbol{\lambda}_0}(\boldsymbol{v})||_n$ has an intuitive interpretation as the rate of change of $\boldsymbol{\rho}(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda}_0$ in the direction of $\boldsymbol{\lambda}_0 + \boldsymbol{v}$, and we are trying to find the maximum rate over all such directions $\boldsymbol{v}$.

Direct evaluations of the total derivative $T\boldsymbol{\rho}_{\boldsymbol{\lambda}}$ and its norm, however, are hard. The next theorem expresses $||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}||$ as a function of the partial derivatives $\frac{\delta\rho_i(\boldsymbol{\lambda})}{\delta\lambda_j}$, which are much easier to calculate.

Theorem 2 *Let Λ, $\boldsymbol{\rho}(\boldsymbol{\lambda})$, and $\bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda})$ be as defined before. Assume $\boldsymbol{\rho}(\cdot) : \Lambda \mapsto \Re^n$ is totally differentiable at an interior point $\boldsymbol{\lambda}_0$ of Λ. Then $||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}||^2$ = maximum eigenvalue of the $k \times k$ nonnegative definite matrix $\bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)$.*

Proof : It is well known that the total derivative $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}$ is a linear combination of the partial derivatives, i.e., $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v}) = \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)\,\boldsymbol{v}$ (Rudin (1976), pp. 215). Hence, $||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}||^2 = \sup_{\boldsymbol{v}^T\boldsymbol{v}\neq 0} \frac{1}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v}^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)\,\boldsymbol{v}$ = maximum eigenvalue of $\bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)$ (see, for instance, Rao, C.R. (1973), p 62) ■

Corollary 1 *Suppose we consider a single posterior quantity, i.e., $\rho(\cdot) : \Lambda \subseteq \Re^k \mapsto \Re$. Then $||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}|| = \sqrt{\sum_{i=1}^{k} [\frac{\delta}{\delta\lambda_i}\rho(\boldsymbol{\lambda}_0)]^2}$.*

Proof : Immediate from Theorem 2 ■

2.3. Average sensitivity

It should be mentioned that the norm of the total derivative, or the maximum sensitivity, is a very conservative estimate in the sense that it tries to guard against large changes in $\boldsymbol{\rho}(\boldsymbol{\lambda})$ by computing the fastest rate of change over all possible directions. Another less conservative concept would be to average the rate of change over all directions. Mathematically, this amounts to evaluating $\int_{\{||\boldsymbol{v}||_k=1\}} ||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})||_n\, d\boldsymbol{v}$. But since this integral is hard to compute, we square the integrand and evaluate $\int_{\{||\boldsymbol{v}||_k=1\}} ||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})||_n^2\, d\boldsymbol{v}$ instead. The choice of "1" as the radius of the hypersphere is completely arbitrary here; any other radius leads to an equivalent definition (through the linear structure of $T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})$).

Definition 1 *Assume $\boldsymbol{\rho}(\cdot) : \Lambda \subseteq \Re^k \mapsto \Re^n$ is totally differentiable at an interior point $\boldsymbol{\lambda}_0$ of Λ. Then the* **average sensitivity** *of the posterior quantity $\boldsymbol{\rho}(\boldsymbol{\lambda}) = \boldsymbol{\rho}(P_{\boldsymbol{\lambda}})$ w.r.t. the choice of the prior parameter $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ is defined to be $\overline{T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}} = \frac{1}{r^3}\int_{\{||\boldsymbol{v}||_k=r\}} ||T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}(\boldsymbol{v})||_n^2\, d\boldsymbol{v}$. Here $r > 0$ is arbitrary (the definition is independent of the choice of r).*

The next theorem shows how to evaluate $\overline{T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}}$ for a totally differentiable posterior quantity $\boldsymbol{\rho}(\boldsymbol{\lambda})$.

Theorem 3 *Assume the setup of Theorem 2 with $\Lambda \subseteq \Re^k$. Then $\overline{T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}} = \frac{w_k}{k} \times$ {sum of eigenvalues of the $k \times k$ matrix $\bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)$}, where w_k = surface area of the hypersphere $\{\boldsymbol{v} : ||\boldsymbol{v}||_k = 1\} = \frac{2\pi^{k/2}}{\Gamma(k/2)}$.*

Proof : Since $T\rho_{\boldsymbol{\lambda}_0}(\boldsymbol{v}) = \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)\,\boldsymbol{v}$, $||T\rho_{\boldsymbol{\lambda}_0}(\boldsymbol{v})||^2 = \boldsymbol{v}^T A\,\boldsymbol{v}$ with $A_{(k\times k)} = \bigtriangledown\rho(\lambda_0)^T \bigtriangledown\boldsymbol{\rho}(\boldsymbol{\lambda}_0)$. Let $\beta_1, \ldots, \beta_k$ be the eigenvalues of A, i.e., $A = P^T D\, P$ where $D_{(k\times k)} = \text{diag}\{\beta_1, \ldots, \beta_k\}$ and P is an orthogonal matrix. Let $\boldsymbol{u} = (u_1, \ldots, u_k)^T = P\,\boldsymbol{v}$. Then $\overline{T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}} = \int\limits_{\{||\boldsymbol{v}||_k=1\}} (\boldsymbol{v}^T A\,\boldsymbol{v})\,d\boldsymbol{v} = \sum\limits_{i=1}^{k}\{\beta_i \int\limits_{\{||\boldsymbol{u}||_k=1\}} u_i^2\,d\boldsymbol{u}\}$. Clearly, $S = \int\limits_{\{||\boldsymbol{u}||_k=1\}} u_i^2\,d\boldsymbol{u}$ is independent of "i", and $k\,S = \int\limits_{\{||\boldsymbol{u}||_k=1\}} \{\sum\limits_{i=1}^{k} u_i^2\}d\boldsymbol{u} = w_k$. This completes the proof of the theorem ■

Corollary 2 *Suppose $\rho(\lambda)$ is univariate, i.e, $\rho(\cdot) : \Lambda \subseteq \Re^k \mapsto \Re$. Then*
$\frac{k}{w_k}\,\overline{T\boldsymbol{\rho}_{\boldsymbol{\lambda}_0}} = \sum\limits_{i=1}^{k} [\frac{\delta}{\delta\,\lambda_i}\,\rho(\boldsymbol{\lambda}_0)]^2 = ||T\rho_{\boldsymbol{\lambda}_0}||^2$.

Proof : Follows trivially from Theorem 3 ■
Remark : It is clear that for $n = 1$, these two concepts of maximum and average sensitivity are equivalent (see Corollaries 1 and 2) □

3. Examples

We look at several applications of Theorem 2 and Theorem 3 in this section. The first three examples evaluate the maximum sensitivity of posterior quantities, while Example 4 examines average sensitivity.

Example 1 : Suppose we observe X from $N(\theta, \sigma^2)$, where $\sigma^2 > 0$ is known, and decide to use a $N(\mu, \tau^2)$ prior for θ. Thus, $P_{\boldsymbol{\lambda}} = N(\mu, \tau^2)$ with $\boldsymbol{\lambda} = (\mu, \tau^2)^T \in \Re \otimes (0, \infty)$. Our interest is the Bayes estimate of θ under squared-error loss, i.e., $\rho(\mu, \tau) = E_{\boldsymbol{\lambda}}(\theta \mid X) = \frac{\tau^2 X + \sigma^2 \mu}{\tau^2 + \sigma^2}$. To evaluate local sensitivity of $\rho(\mu, \tau)$ w.r.t. a particular choice of the prior location parameter μ and scale parameter τ, we evaluate the total derivative of ρ. Clearly, $\frac{\delta}{\delta\,\mu}\rho = \frac{\sigma^2}{\tau^2 + \sigma^2}$ and $\frac{\delta}{\delta\,\tau}\rho = \frac{2\,\tau\,\sigma^2\,(X-\mu)}{(\tau^2+\sigma^2)^2}$, thus, by Corollary 1, $||T\rho_{(\mu,\tau)}|| = \frac{\sigma^2}{\tau^2+\sigma^2}\sqrt{1 + \frac{4\tau^2\,(X-\mu)^2}{(\tau^2+\sigma^2)^2}}$. Notice that the local sensitivity index $||T\rho_{(\mu,\tau)}||$ decreases as $|X - \mu|$ decreases and/or as τ increases (subject to $\tau \geq \sigma$). Thus, for this particular example, our evaluation of $||T\rho_{(\mu,\tau)}||$ mathematically justifies the popular belief that if the center of the prior matches with that of the likelihood and/or if the prior has a flat tail, then (generally) posterior robustness (w.r.t. the prior) is achieved □

Example 2 : Let X be observed from $N(\theta, 1)$, and the user or a finite elicitation process specifies the prior median and quartiles of θ at 0 and ± 1 respectively. Several distributions satisfy these requirements (see Basu and DasGupta (1992)). For comparison, we only consider the sharp tailed $\pi^n(\mu, \tau^2) = N(\mu, \tau^2)$ with $\mu = 0, \tau = 1.48$, and the flat tailed $\pi^c(\mu, \tau^2) = \text{Cauchy}(\mu, \tau^2)$ with $\mu = 0, \tau = 1$. However, the specifications of median $= 0$ and quartiles $= \pm 1$ often can not be taken as exactly accurate. We thus consider the local sensitivity of the specification $\mu = 0, \tau = 1.48$ in the class of all $N(\mu, \tau^2)$ priors, and compare it with the sensitivity of the specification $\mu = 0, \tau = 1$ in the class of all Cauchy(μ, τ^2) priors. Let $\rho^n(\mu, \tau^2)$ and $\rho^c(\mu, \tau^2)$ denote the posterior means w.r.t. $\pi^n(\mu, \tau^2)$ and $\pi^c(\mu, \tau^2)$ respectively. The local sensitivity in the Normal class, i.e., $||T\rho^n_{(\mu=0,\tau=1.48)}||$, can be easily found from the calculations

done in Example 1. Let $\rho^c(\mu,\tau^2) = \frac{N^c(\mu,\tau^2)}{D^c(\mu,\tau^2)}$, where $N^c(\mu,\tau^2) = \int \theta\ell(\theta)\pi^c(\theta|\mu,\tau^2)\,d\theta$, $D^c(\mu,\tau^2) = \int \ell(\theta)\pi^c(\theta|\mu,\tau^2)\,d\theta$, and $\ell(\theta)$ is the appropriate likelihood. $N^c(\mu,\tau^2)$ is difficult to compute analytically. However, it is easy to check that the condition for interchange of derivative and integral is satisfied, i.e., $\frac{\delta}{\delta\mu}N^c(\mu,\tau^2) = \int \theta\ell(\theta)\,[\frac{\delta}{\delta\mu}\pi^c(\theta|\mu,\tau^2)]\,d\theta$. Similar result holds for $D^c(\mu,\tau^2)$. Now, $\frac{\delta}{\delta\mu}\rho^c(\mu,\tau^2) = \frac{1}{D^c(\mu,\tau^2)^2}\{D^c(\mu,\tau^2)\frac{\delta}{\delta\mu}N^c(\mu,\tau^2) - N^c(\mu,\tau^2)\frac{\delta}{\delta\mu}D^c(\mu,\tau^2)\}$, and each term in the above expression involves a simple numerical integration. Same is true for $\frac{\delta}{\delta\tau}\rho^c(\mu,\tau^2)$. Thus, $||T\rho^c_{(\mu=0,\tau=1)}||$ can be obtained with little numerical work.

Table 1: $||T\rho||$ for $N(0,2.19)$ and Cauchy$(0,1)$ priors

X	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
$\|\|T\rho^n\|\|$	0.346	0.428	0.537	0.661	0.792	0.927	1.065	1.205
$\|\|T\rho^c\|\|$	0.481	0.497	0.512	0.511	0.476	0.402	0.309	0.225

Table 1 shows the values of $||T\rho^n_{(\mu=1,\tau=1.48)}||$ and $||T\rho^c_{(\mu=0,\tau=1)}||$ for different values of X. As can be seen, the value of $||T\rho^n||$ increases with X, and is large for $X \geq 2.5$, whereas $||T\rho^c||$ fluctuates very little. Thus, misspecification of the prior parameters results in much less sensitivity for the heavy tailed Cauchy prior than for the sharp tailed Normal prior (especially when the the center of the prior and the likelihood do not match), which again agrees with prevalent beliefs □

Example 3 : Consider a standard linear model setup : $\underset{\sim}{Y} \sim N_m(X\beta, \Sigma)$. Here, $\underset{\sim}{Y}_{m\times 1}$ is an observed vector, $X_{m\times k}$ is a known design matrix, Σ is a known positive definite matrix, and $\beta_{k\times 1}$ is an unknown parameter vector. Under the Bayesian paradigm, we assume a $N_k(\mu,\Gamma)$ prior for β. It is well known that in this setup, the posterior mean for β is $\underset{\sim}{\beta}^* = [\Gamma^{-1} + X^T\Sigma^{-1}X]^{-1}[\Gamma^{-1}\mu + X^T\Sigma^{-1}X\underset{\sim}{b}]$, where $\underset{\sim}{b} = [X^T\Sigma^{-1}X]^{-1}X^T\Sigma^{-1}\underset{\sim}{Y}$ is the generalized least square estimate (or *mle*) of β. For notational simplicity, we denote $X^T\Sigma^{-1}X$ by A from now on. However, specification of the prior parameters μ and Γ is again of concern. First, we assume Γ is exactly known, and find the local sensitivity of $\underset{\sim}{\beta}^*$ w.r.t. misspecifications of μ. Clearly, $[\frac{\delta\underset{\sim}{\beta}^*}{\delta\mu}]_{k\times k} = [\Gamma^{-1}+A]^{-1}\Gamma^{-1}$, thus $||T\underset{\mu}{\beta}^*||^2 =$ maximum eigenvalue of $[\Gamma^{-1}+A]^{-1}\Gamma^{-1}\Gamma^{-1}[\Gamma^{-1}+A]^{-1}$. Surprisingly, this local sensitivity $||T\underset{\mu}{\beta}^*||$ does not depend on μ or on the observed value of $\underset{\sim}{Y}$.

We next assume that μ is correctly specified and examine the sensitivity of $\underset{\sim}{\beta}^*$ to misspecifications of Γ. In particular, we presume that Γ has a equicorrelated structure, i.e., $\Gamma = \sigma\{(1-r)I + r\,\underset{\sim}{1}\,\underset{\sim}{1}^T\}$, thus specification of Γ requires specifying the variance term σ and the correlation term r (The following calculations can also be done for a general positive definite Γ, but with increased complexity). For ease in calculations, we write $\Gamma = \tau\{I + \rho\,\underset{\sim}{1}\,\underset{\sim}{1}^T\}$, thus $\tau = \sigma(1-r)$, $\rho = \frac{r}{1-r}$, and $\Gamma^{-1} = \frac{1}{\tau}[I - \frac{\rho}{1+k\rho}\underset{\sim}{1}\,\underset{\sim}{1}^T]$. Calculation of $\frac{\delta\underset{\sim}{\beta}^*}{\delta\tau}$ and $\frac{\delta\underset{\sim}{\beta}^*}{\delta\rho}$, however, requires use of matrix derivatives. In particular, we need : (i) if V and W (both matrices) are functions of a matrix $U_{m\times n}$, then $\frac{d(VW)}{dU} = (\frac{dV}{dU})(W\otimes I_n) + (V\otimes I_m)(\frac{dW}{dU})$, and (ii) if V is invertible, then $\frac{d(V^{-1})}{dU} = -(V^{-1}\otimes I_m)(\frac{dV}{dU})(V^{-1}\otimes I_n)$. Here, $\otimes$ denotes a Kronecker product and, for $V_{p\times q}$, $U_{m\times n}$, $[\frac{dV}{dU}]_{mp\times nq} = V\otimes\frac{d}{dU}$ where $\frac{d}{dU}$ is a matrix of derivative oper-

ators $[\frac{\delta}{\delta u_{ij}}]_{m\times n}$ (see MacRae (1974), Polasek (1985) for more on matrix derivatives). Using these formulae, we find $\frac{\delta \underset{\sim}{\beta}^*}{\delta \tau} = \frac{1}{\tau}[\Gamma^{-1}+A]^{-1}\,\Gamma^{-1}\,\{[\Gamma^{-1}+A]^{-1}\,[\Gamma^{-1}\mu + A\,\underset{\sim}{b}] - \mu\}$ and $\frac{\delta \underset{\sim}{\beta}^*}{\delta \rho} = \frac{1}{(1+k\rho)^2\,\tau}\,[\Gamma^{-1}+A]^{-1}\,\underset{\sim}{1}\,\underset{\sim}{1}^T\,\{[\Gamma^{-1}+A]^{-1}\,[\Gamma^{-1}\mu + A\,\underset{\sim}{b}] - \mu\}$. Going back to our original parameters, we have $\bigtriangledown \underset{\sim}{\beta}^*(\sigma, r) = [\frac{\delta \underset{\sim}{\beta}^*}{\delta(\sigma,r)}]_{k\times 2} = [\frac{\delta \underset{\sim}{\beta}^*}{\delta(\tau,\rho)}]_{k\times 2}\,[\frac{\delta(\tau,\rho)}{\delta(\sigma,r)}]_{2\times 2} = [\frac{\delta \underset{\sim}{\beta}^*}{\delta \tau}, \frac{\delta \underset{\sim}{\beta}^*}{\delta \rho}]\begin{pmatrix} 1-r & -\sigma \\ 0 & 1/(1-r)^2 \end{pmatrix}$. Moreover, $||T\underset{\sim}{\beta}^*_{(\sigma,r)}||^2$ = maximum eigenvalue of $[\bigtriangledown \underset{\sim}{\beta}^*]^T\,[\bigtriangledown \underset{\sim}{\beta}^*]$. Notice that the matrix on the r.h.s is only 2×2, so that the maximum eigenvalue can be found easily.

For example, suppose we consider a simple linear regression model : $Y_i = \beta_0 + \beta_1\, x_i + \varepsilon_i$, $i = 1, \ldots, m$, where ε_i's are i.i.d. $N(0,1)$ and $|x_i| \leq 1$. Using an optimal design strategy, we take 10 observations at $x_i = 1$ and 10 at $x_i = -1$. Thus, $m = 20$, $k = 2$, $\Sigma = I$, and $X^T X = 20\, I$. In this setup, $||T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}|| = \frac{1}{1+20\sigma\,(1-|r|)}$. Notice, we did not require to specify either $\underset{\sim}{Y}$ or $\underset{\sim}{\mu}$ to evaluate $||T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}||$. Moreover, $||T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}||$ decreases, i.e., $\underset{\sim}{\beta}^*$ becomes less sensitive to specification of $\underset{\sim}{\mu}$ as the variance term σ increases and/or as the correlation r gets close to 0. Evaluation of $||T\underset{\sim}{\beta}^*_{(\sigma,r)}||$, however, requires us to know $\underset{\sim}{\mu}$, and $\underset{\sim}{Y}$, or equivalently, the least square estimate $\underset{\sim}{b}$. We specify $\underset{\sim}{\mu} = (0,0)^T$, and evaluate $||T\underset{\sim}{\beta}^*_{(\sigma,r)}||$ in Table 2 for three different values of $\underset{\sim}{b}$, namely, $\underset{\sim}{b} = (1,1)^T, (1,3)^T$, and $(3,3)^T$. From Table 2, we see that $\underset{\sim}{\beta}^*$ becomes less sensitive to specifications of (σ, r) as σ increases and/or r gets close to 0. However, positive and negative r values have different effects. Also, $\underset{\sim}{\beta}^*$ is less sensitive to (σ, r) for $\underset{\sim}{b} = (1,1)^T$ (which is close to the prior specification $\underset{\sim}{\mu} = (0,0)^T$) than for other values of $\underset{\sim}{b}$ □

Example 4 (Example 3 continued) : As before, consider a linear model setup : $\underset{\sim}{Y} \sim$

Table 2: $||T\underset{\sim}{\beta}^*_{(\sigma,r)}||$ for different values of $\underset{\sim}{b}$, σ and r

	$\sigma = 1$					$\sigma = 2$				
r	$-.75$	$-.5$	0	.5	.75	$-.75$	$-.5$	0	.5	.75
$\underset{\sim}{b} = (1,1)^T$	.22	.13	.09	.27	.30	.08	.04	.04	.14	.15
$\underset{\sim}{b} = (1,3)^T$	.44	.26	.18	.58	.99	.16	.09	.08	.30	.56
$\underset{\sim}{b} = (3,3)^T$	.66	.38	.27	.81	.90	.25	.13	.11	.41	.46

$N_m\,(X\underset{\sim}{\beta},\, \Sigma)$ with $\underset{\sim}{\beta}_{k\times 1} \sim N_k\,(\underset{\sim}{\mu},\, \Gamma)$. When Γ is known and we focus on the average sensitivity of $\underset{\sim}{\beta}^*$ (the posterior mean of $\underset{\sim}{\beta}$) to specification of $\underset{\sim}{\mu}$, we have : $\overline{T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}} = \frac{w_k}{k} \times$ {sum of eigenvalues of $[\Gamma^{-1} + A]^{-1}\,\Gamma^{-1}\,\Gamma^{-1}[\Gamma^{-1} + A]^{-1}$ }, where $A = X^T \Sigma^{-1} X$. If $\underset{\sim}{\mu}$ is correctly specified, and we want to evaluate the average sensitivity of $\underset{\sim}{\beta}^*$ w.r.t. σ and r (where $\Gamma = \sigma\{(1-r)I + r\,\underset{\sim}{1}\,\underset{\sim}{1}^T\}$), then $\overline{T\underset{\sim}{\beta}^*_{(\sigma,r)}} = \pi \times$ {sum of eigenvalues of $[\bigtriangledown \underset{\sim}{\beta}^*(\sigma,r)]^T\,[\bigtriangledown \underset{\sim}{\beta}^*(\sigma,r)]$ } (see Example 3).

In particular, if we consider the specific example : $m = 20$, $k = 2$, $\Sigma = I$, and $X^T X = 20\, I$, then $\overline{T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}} = \pi \times \{\frac{1}{[1+20\sigma\,(1+r)]^2} + \frac{1}{[1+20\sigma\,(1-r)]^2}\}$. Notice, $\overline{T\underset{\sim}{\beta}^*_{\underset{\sim}{\mu}}}$ increases as σ decreases. It also increases as $|r|$ increases. For $\underset{\sim}{\mu} = (0,0)^T$, we also evaluated $\overline{T\underset{\sim}{\beta}^*_{(\sigma,r)}}$, and plotted it against r for different values of $\underset{\sim}{b}$ (the least square estimate of $\underset{\sim}{\beta}$) and σ (plot not shown). These plots showed that the average sensitivity decreases with increase of σ. However, the effect of r was somewhat surprising, $\overline{T\underset{\sim}{\beta}^*_{(\sigma,r)}}$ (for fixed $\underset{\sim}{b}$ and σ) did not attain its

minimum at $r = 0$ as was expected □

4. Nonparametric classes

An important issue in prior elicitation is that a parametric functional form of the prior is generally hard to determine. Recent attention in robust Bayesian analysis is thus more focused towards nonparametric prior classes. Our technique of computing the total derivative to quantify the sensitivity of $\rho(P)$ fails here, since the relevant domain of $\rho(P)$ is no longer a Euclidean space, but a general polish space $\mathcal{M}$ of all probability measures on Θ. Thus, the notion of functional derivatives, in particular, Fréchet derivatives enters the picture. Diaconis and Freedman (1986), and Ruggeri and Wasserman (1990) quantified the local sensitivity of a posterior quantity $\rho(P)$ by computing the norm of its Fréchet derivative over the class of all signed measures or its appropriate subclasses. Srinivasan and Truszczynska (1990) used Fréchet derivatives to approximate ranges of posterior quantities.

Fréchet derivatives are defined on normed linear spaces, or more generally, on topological vector spaces. However, the posterior quantity $\rho(P)$ is defined on $\mathcal{M}$ which is convex, but not linear. Thus ρ has to be artificially extended to the linear space of all signed measures Δ before the notion of Fréchet differentiability could be applied to ρ.

A different line of attack was proposed by Huber (1981) and others who generalized the definition of Fréchet derivatives to encompass the case when ρ is defined only on $\mathcal{M}$. We find this approach more natural from a statistical viewpoint. This generalized definition, however, comes with a price since we can not use strong theorems which are available for Fréchet derivatives on vector spaces. In our current ongoing work, we have established (Huber's) Fréchet differentiability of ratio-linear posterior quantities. We have also argued that since $\mathcal{M}$ is only convex, a direct maximization of the Fréchet derivative is more intuitive rather than treating $\mathcal{M}$ as a subspace of the linear space Δ and computing the norm of the Fréchet derivative over $\mathcal{M}$. We are in the process of developing methods for computing this maximum over different subclasses of $\mathcal{M}$.

Acknowledgement : The authors thank Benny Cheng for suggesting an improvement in the proof of Theorem 3.

References

[1] Basu, S., Jammalamadaka, S.R., and Liu, W. (1993), "Qualitative robustness and stability of posterior distributions and posterior quantities", Technical Report, **238**, Department of Statistics and Applied Probability, University of California, Santa Barbara.

[2] Basu, S. and DasGupta, A. (1992), "Bayesian analysis with distribution bands : the role of the loss function", verbally accepted in *Statist. and Decisions*

[3] Berger, J. (1993), "An overview of robust Bayesian analysis", Technical Report, **93-53**, Department of Statistics, Purdue University.

[4] Diaconis, P. and Freedman, D. (1986), "On the consistency of Bayes estimates", *Ann. Statist.*, **14**, 1-67.

[5] Hampel, F.R. (1971), "A general qualitative definition of robustness", *Ann. Math. Statist.*, **42**, 1887-1896.

[6] Huber, P.J. (1981), *Robust Statistics*, John Wiley : New York.

[7] MacRae, E.C (1974), "Matrix derivatives with an application to an adaptive linear decision problem", *Ann. Statist.*, **2**, 337-346.

[8] Polasek, W. (1985), "A dual approach for matrix-derivatives", *Metrika*, **32**, 275-292.

[9] Rao, C.R. (1973), *Linear statistical inference and its applications*, Wiley.

[10] Rivier, N., Englman, R., and Levine, R. D. (1990), "Constructing priors in maximum entropy methods", In *Maximum Entropy and Bayesian Methods*, P.F. Fougére (Ed.), 233-242, Kluwer Academic Publishers.

[11] Rodríguez, C.C. (1994), "Bayesian robustness : a new look from geometry", to appear in *Maximum Entropy and Bayesian Statistics*, G. Heidbreder (Ed.), Kluwer Academic Publishers.

[12] Rudin, W. (1976), *Principles of mathematical analysis*, McGraw-Hill.

[13] Ruggeri, F. and Wasserman, L. (1993), "Infinitesimal sensitivity of posterior distributions", *Canad. J. Statist.*, **21**, 195-203.

[14] Skilling, J. (1990), "Quantified maximum entropy", In *Maximum Entropy and Bayesian Methods*, P.F. Fougére (Ed.), 341-350, Kluwer Academic Publishers.

[15] Srinivasan, C. and Truszczynska, H. (1990), "Approximation to the range of a ratio-linear posterior quantity based on Fréchet derivative", Technical Report, **289**, Department of Statistics, University of Kentucky.

[16] Tukey, J.W. (1960), "A survey of sampling from contaminated distributions", In *Contributions to Statistics and Probability*, I. Olkin et al (Ed.), 448-485, Stanford University Press, Stanford, California.

[17] Wasserman, L. (1992), "Recent Methodological advances in robust Bayesian inference", In *Bayesian Statistics 4*, J.M. Bernardo, et. al. (Eds.), Oxford University Press, Oxford.

[6] Huber, P.J. (1981), *Robust Statistics*, John Wiley : New York.

[7] MacRae, E.C (1974), "Matrix derivatives with an application to an adaptive linear decision problem", *Ann. Statist.*, **2**, 337-346.

[8] Polasek, W. (1985), "A dual approach for matrix-derivatives", *Metrika*, **32**, 275-292.

[9] Rao, C.R. (1973), *Linear statistical inference and its applications*, Wiley.

[10] Rivier, N., Englman, R., and Levine, R. D. (1990), "Constructing priors in maximum entropy methods", In *Maximum Entropy and Bayesian Methods*, P.F. Fougére (Ed.), 233-242, Kluwer Academic Publishers.

[11] Rodríguez, C.C. (1994), "Bayesian robustness : a new look from geometry", to appear in *Maximum Entropy and Bayesian Statistics*, G. Heidbreder (Ed.), Kluwer Academic Publishers.

[12] Rudin, W. (1976), *Principles of mathematical analysis*, McGraw-Hill.

[13] Ruggeri, F. and Wasserman, L. (1993), "Infinitesimal sensitivity of posterior distributions", *Canad. J. Statist.*, **21**, 195-203.

[14] Skilling, J. (1990), "Quantified maximum entropy", In *Maximum Entropy and Bayesian Methods*, P.F. Fougére (Ed.), 341-350, Kluwer Academic Publishers.

[15] Srinivasan, C. and Truszczynska, H. (1990), "Approximation to the range of a ratio-linear posterior quantity based on Fréchet derivative", Technical Report, **289**, Department of Statistics, University of Kentucky.

[16] Tukey, J.W. (1960), "A survey of sampling from contaminated distributions", In *Contributions to Statistics and Probability*, I. Olkin et al (Ed.), 448-485, Stanford University Press, Stanford, California.

[17] Wasserman, L. (1992), "Recent Methodological advances in robust Bayesian inference", In *Bayesian Statistics 4*, J.M. Bernardo, et. al. (Eds.), Oxford University Press, Oxford.