The Annals of Statistics
1995, Vol. 23, No. 6, 2153–2167

LOCAL SENSITIVITY DIAGNOSTICS FOR BAYESIAN INFERENCE

BY PAUL GUSTAFSON[1] AND LARRY WASSERMAN[2]

University of British Columbia and Carnegie Mellon University

We investigate diagnostics for quantifying the effect of small changes to the prior distribution over a k-dimensional parameter space. We show that several previously suggested diagnostics, such as the norm of the Fréchet derivative, diverge at rate $n^{k/2}$ if the base prior is an interior point in the class of priors, under the density ratio topology. Diagnostics based on ϕ-divergences exhibit similar asymptotic behavior. We show that better asymptotic behavior can be obtained by suitably restricting the classes of priors. We also extend the diagnostics to see how various marginals of the prior affect various marginals of the posterior.

1. Introduction. In Bayesian inference, the sensitivity of the posterior to the prior is always a concern. *Global sensitivity analysis* is sometimes used to assess this sensitivity [Berger (1984, 1990)]. Here, the prior P is embedded in a class of priors Γ and the degree to which the posterior changes as the prior varies over Γ is used to assess sensitivity to the prior. This approach is infeasible for complicated models. When there are many parameters we might want to ask questions like: "How sensitive is the posterior marginal density for one parameter to changes in the prior of another parameter?" There are a multitude of such questions and it is usually too time-consuming to answer them all using global analysis. Instead, we might use *local sensitivity analysis* in which we study the effects of small perturbations to the prior. A small but quickly growing literature on Bayesian local sensitivity analysis has developed lately; see Basu, Jammalamadaka and Liu (1996), Berger (1986), Cuevas and Sanz (1988), Delampady and Dey (1994), Dey and Birmiwal (1994), Diaconis and Freedman (1986), Gelfand and Dey (1991), Gustafson, Srinivasan and Wasserman (1995), Ruggeri and Wasserman (1993), Sivaganesan (1993) and Srinivasan and Truszczynska (1990).

One approach to local sensitivity is to use the norm of the Fréchet derivative of the posterior with respect to the prior. Diaconis and Freedman (1986) were apparently the first to suggest this, albeit in a different context. Under weak conditions, this norm turns out to be $L(\hat{\theta})/m$, where L is the likelihood function, $\hat{\theta}$ is the maximum likelihood estimator of the k-dimen-

Received July 1993; revised November 1994.

[1] Research supported by a postgraduate scholarship and postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada.

[2] Research supported by NSF Grant DMS-90-05858 and NIH Grant R01-CA54852-01.

AMS 1991 *subject classifications*. Primary 62F15; secondary 62F35.

Key words and phrases. Classes of probabilities, ϕ-divergence, Fréchet derivatives, Kullback–Leibler distance, Hellinger distance, robustness, total variation distance.

sional parameter θ and m is the marginal density of the data. Typically, the norm increases with sample size at rate $n^{k/2}$, leading to the counterintuitive conclusion that the posterior becomes increasingly sensitive to the prior as we accumulate more data; Krasker and Pratt (1986) noted this in their discussion of Diaconis and Freedman (1986). The asymptotic behavior of the norm of the Fréchet derivative makes it unsuitable as a diagnostic.

The norm is equal to the Fréchet derivative of the posterior evaluated in the direction of the prior $\delta_{\hat{\theta}}$ which is a point mass at $\hat{\theta}$. Obviously, $\delta_{\hat{\theta}}$ is an unreasonable prior and one might suspect that, by eliminating such priors from consideration, the norm will have better asymptotic behavior. We will show that eliminating point masses is neither necessary nor sufficient for correcting the behavior of the derivative. It is not necessary because under some distance functions our diagnostic is $o(1)$ even when the weak closure of the class of priors contains point masses. It is not sufficient in the sense that there are classes of uniformly bounded density functions (which therefore do not include point masses) that give norms of order $n^{k/2}$. We will show that diagnostics based on ϕ-divergences [Gelfand and Dey (1991), Dey and Birmiwal (1994) and Delampady and Dey (1994)] exhibit similar behavior.

Another problem with using the norm of the Fréchet derivative as a diagnostic is apparent if we consider multiparameter models. Suppose that $\theta = (\theta_1, \theta_2)$. Then the norm of the Fréchet derivative of the marginal posterior for θ_1 turns out to be the same as that for θ_2. Again, this is counterintuitive since we expect that inferences for one parameter may be more or less sensitive than inferences for another. None of this is meant as a criticism of Diaconis and Freedman (1986), whose results were intended for a different purpose.

We have two main goals in this paper. The first is to explore the behavior of sensitivity diagnostics. The second is to develop new diagnostics that avoid these problems. In Section 2 we develop some basic theory. We define a local sensitivity diagnostic and give a sufficient condition for the divergence of the diagnostic. This sets up the appropriate background for Section 3, where we propose modified forms of the diagnostic that have good asymptotic behavior. In Section 4 we extend the diagnostics to multiparameter models. In that section, we are particularly interested in assessing the sensitivity of a posterior marginal to perturbations in various marginals of the prior. In Section 5 we provide some concluding remarks.

2. Basic theory.

2.1. *Notation and definitions.* Let $X = X_1^n = (X_1, \dots, X_n)$ be n independent and identically distributed random variables, each with density $f(\cdot \mid \theta)$. Here $\theta \in \Theta \subset \mathbb{R}^k$ is the unknown parameter. Let $\mathscr{B}$ be a σ-algebra on the parameter space Θ, and let P be a prior probability measure on $(\Theta, \mathscr{B})$. The posterior probability measure $P^x = P_n^x$ is defined by $P_n^x(d\theta) = \{m_P(X_1^n)\}^{-1} L(\theta; x_1^n) P(d\theta)$, where $L(\theta; x_1^n) = \prod_{i=1}^n f(x_i \mid \theta)$ is the likelihood function, and $m_P(x_1^n) = \int_\Theta L(\theta; x_1^n) P(d\theta)$ is the marginal density of the data.

To quantify changes in priors and posteriors, we require a function d: $\mathscr{P} \times \mathscr{P} \to [0, \infty]$, where $\mathscr{P}$ is the class of all priors on $\mathscr{B}$. We want d to indicate "distance" between elements of $\mathscr{P}$, but we do not require that d be a metric. Define *the local sensitivity of* $\mathscr{P}$ *in the direction of* Q by

$$s(P, Q; x) = \lim_{\varepsilon \downarrow 0} \frac{d(P^x, Q_\varepsilon^x)}{d(P, Q_\varepsilon)}. \tag{1}$$

Here we define Q_ε, the perturbation of P in direction Q, in one of two ways. We can take the *linear perturbation* $Q_\varepsilon = (1 - \varepsilon)P + \varepsilon Q$ or the *geometric perturbation* $dQ_\varepsilon \propto [dQ/dP]^\varepsilon \, dP$, as suggested by Gelfand and Dey (1991). The latter may seem less familiar but turns out to be quite natural for some choices of d. The local sensitivity (1) is the rate at which the contaminated posterior Q_ε^x tends to the nominal posterior P^x, relative to the analogous rate for the prior. A natural measure of the overall sensitivity is $s(P, \Gamma; x) = \sup_{Q \in \Gamma} s(P, Q; x)$, for some class of priors $\Gamma \subset \mathscr{P}$.

When linear perturbations are used we can relate $s(P, \Gamma; x)$ to Fréchet derivatives. Let $d = d_{\mathrm{TV}}$ be total variation distance, which is defined by $d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathscr{B}} |P(A) - Q(A)|$. Define $\delta = Q - P$, $\|\delta\| = d_{\mathrm{TV}}(P, Q)$ and $T(P) = P^x$. The Gateaux differential of the posterior is defined by

$$\dot{T}_P(\delta) = \lim_{\varepsilon \downarrow 0} \frac{d_{\mathrm{TV}}(P^x, Q_\varepsilon^x)}{\varepsilon} = \frac{m_Q(x)}{m_P(x)} d_{\mathrm{TV}}(P^x, Q^x). \tag{2}$$

Equation (2) follows from the fact that $Q_\varepsilon^x = (1 - \lambda)P^x + \lambda Q^x$, where $\lambda = \varepsilon m_Q(x)[(1 - \varepsilon)m_P(x) + \varepsilon m_Q(x)]^{-1}$; $\dot{T}_P(\delta)$ can be regarded as the derivative of P^x evaluated at P in the direction of the signed measure $\delta = Q - P$. If the likelihood is bounded, then this quantity is also the Fréchet derivative [see Diaconis and Freedman (1986)], which means that $\dot{T}_P$ is a linear map on signed measures such that $T(P + \delta) = T(P) + \dot{T}_P(\delta) + o(\|\delta\|)$ as $\|\delta\| \to 0$, uniformly over all signed measures δ with mass 0. Now, for a class Γ, define the restricted norm of the derivative by $\|\dot{T}_P\| = \sup_{\delta = Q - P;\, Q \in \Gamma} \|\dot{T}_P\| / \|\delta\|$. It follows immediately that $\|\dot{T}_P\| = s(P, \Gamma; x)$. Usual practice in functional analysis would be to compute the norm over the linear space of all signed measures, but we are interested in smaller classes. The quantity $s(P, Q; x)$ has a more direct interpretation than $\dot{T}_P(\delta)$ since $s(P, Q; x)$ has been normalized by $d(P, Q)$. The relationship between $\dot{T}$ and $s(P, Q; x)$ breaks down for other distance functions, but $s(P, Q; x)$ retains a clear meaning.

Besides total variation distance, we will consider the ϕ-divergence [Csiszar (1977) and Goel (1983)] defined by $d_\phi(P, Q) = \int \phi(dP/dQ) \, dQ$, where ϕ is a smooth, convex function such that $\phi(1) = 0$. This contains Kullback–Leibler, Hellinger, directed divergence and chi-squared distances, among others, as special cases. Sensitivity diagnostics based on ϕ-divergences are studied by Delampady and Dey (1994) and Dey and Birmiwal (1994).

The effect of sample size on the local sensitivity diagnostic $s(P, Q, x_1^n)$ is of central interest. In particular, it seems reasonable to expect that $s(P, \Gamma; x_1^n)$ converges to 0 in probability as $n \to \infty$. Surprisingly, subject to weak conditions, the local sensitivity measure diverges to infinity with probability 1 unless Γ is appropriately restricted. A simple example illustrates this point.

EXAMPLE 1. Let $X_1, \dots, X_n \mid \theta \sim N(\theta, 1)$ and $\theta \sim N(0,1)$. If $d = d_{\mathrm{TV}}$ and $\bar{x}_n$ is the sample mean, then

$$s(P, \mathscr{P}; x_1^n) = \frac{L(\hat{\theta}; x_1^n)}{m_P(x)} = \sqrt{n+1}\,\exp\left\{\left(\frac{n}{n+1}\right)\frac{\bar{x}_n^2}{2}\right\} \approx \sqrt{n}\,\exp\left\{\frac{\bar{x}_n^2}{2}\right\},$$

which diverges at rate $\sqrt{n}$ for almost all sample paths. Obviously, similar behavior will result for any regular parametric family as long as $\Gamma = \mathscr{P}$.

The rest of Section 2 is devoted to answering the following question: is the behavior in Example 1 typical?

Throughout this section, we assume the following regularity conditions:

(R1) For all x_1^n, $L(\theta; x_1^n)$ is bounded and continuous and attains its maximum at some finite value $\hat{\theta}$.

(R2) The prior P has full support and has density p with respect to Lebesgue measure μ. Furthermore, p is bounded and continuous.

(R3) Let θ_* be the true value of θ. Then

$$n^{-k/2}\frac{L(\hat{\theta}; x_1^n)}{m_P(x_1^n)} \to (2\pi)^{-k/2}\frac{\sqrt{|I(\theta_*)|}}{p(\theta_*)} \quad \text{almost surely } P_{\theta_*},$$

where P_θ is the product measure on the sample space arising from $f(\cdot \mid \theta)$, and $I(\theta)$ is the Fisher information at θ.

Note that (R3) holds as long as the usual conditions that imply asymptotic normality hold. Conditions (R1)–(R3) are stronger than needed but they lead to simpler proofs. We shall say that $s(P, \Gamma; x_1^n)$ diverges if $s(P, \Gamma; x_1^n) \to \infty$ almost surely P_{θ_*} as $n \to \infty$. Let

$$\Lambda_\varepsilon(p) = \left\{q; \operatorname*{ess\,sup}_{\theta, \gamma \in \Theta}\left[\log\frac{p(\theta)q(\gamma)}{p(\gamma)q(\theta)}\right] \le \varepsilon\right\}$$

be the density ratio sphere of size ε around p [DeRobertis (1978), page 141]. Let Γ_N^k be the set of all k-dimensional normal distributions, and let $\phi(\cdot; \mu, \sigma)$ denote a Normal density with mean μ and variance σ^2; we write $\phi(\cdot)$ for $\phi(\cdot; 0, 1)$.

2.2. *The total variation metric.* Let d be the total variation metric. We break this section into two subsections focussing on linear and geometric perturbations, respectively.

2.2.1. *Linear perturbations.* Under linear perturbations we have the following lemma. The first two statements are immediate from results in Diaconis and Freedman (1986). The last statement follows from (R3).

LEMMA 1. (i) *For every* $Q \in \mathscr{P}$, $s(P, Q; x) = (m_Q(x)/m_P(x)) \times (d(P^x, Q^x)/d(P, Q))$; (ii) $s(P, \mathscr{P}; x_1^n) = L(\hat{\theta}; x_1^n)/m_P(x_1^n)$; *and* (iii) $s(P, \mathscr{P}; x_1^n)$ *diverges at rate* $n^{k/2}$.

We are now interested in exploring what types of restrictions on Γ will induce more reasonable asymptotic behavior on $s(P, \Gamma; x)$.

THEOREM 1. *Suppose that* $\Lambda_\varepsilon(p) \subset \Gamma$ *for some* $\varepsilon > 0$. *Then* $s(P, \Gamma; x_1^n)$ *diverges at rate* $n^{k/2}$.

PROOF. Lemma 1 implies that if $s(P, \Gamma; x_1^n)$ diverges, it cannot do so faster than $n^{k/2}$. Choose $a \in (1, e^\varepsilon)$, and let $N_b(\hat{\theta})$ be the size b Euclidean neighborhood of $\hat{\theta}$, where $b > 0$. Define Q_b by $Q_b(B) = c_b[aP(B \cap N_b(\hat{\theta})) + P(B \cap N_b(\hat{\theta})^c)]$, where $c_b^{-1} = \{\Delta P(N_b(\hat{\theta})) + 1\}$ and $\Delta = a - 1$. Q_b is contained in $\Lambda_\varepsilon(P) \subset \Gamma$. Now, $m_{Q_b}(x_1^n) = m_P(x_1^n)c_b/c_b^x$, where $\{c_b^x\}^{-1} = \{\Delta P_n^x(N_b(\hat{\theta})) + 1\}$. Thus, $Q_b^x(B) = c_b^x[aP_n^x(B \cap N_b(\hat{\theta})) + P_n^x(B \cap N_b(\hat{\theta})^c)]$ so that $d(P, Q_b) = \Delta c_b P(N_b(\hat{\theta}))(1 - P(N_b(\hat{\theta})))$ and $d(P_n^x, Q_b^x) = \Delta c_b^x P_n^x(N_b(\hat{\theta}))(1 - P_n^x(N_b(\hat{\theta})))$. Applying Lemma 1, we see that

$$s(P, \Gamma; x_1^n) \geq s(P, Q_b; x_1^n) = \frac{P_n^x\big(N_b(\hat{\theta})\big)\big(1 - P_n^x\big(N_b(\hat{\theta})\big)\big)}{P\big(N_b(\hat{\theta})\big)\big(1 - P\big(N_b(\hat{\theta})\big)\big)}.$$

The right-hand side tends to

$$\frac{p_n^x(\hat{\theta})}{p(\hat{\theta})} = \frac{L\big(\hat{\theta}; x_1^n\big)}{m_P(x_1^n)}$$

as $b \to 0$, and the result follows from (R3). □

REMARK 1. The density ratio sphere $\Lambda_\varepsilon(p)$ is, in a sense, very small in that it contains densities very similar to p [Wasserman (1992)]. Hence, divergence does not require that Γ be large.

REMARK 2. In Theorem 1, it is not necessary to let $N_b(\hat{\theta})$ shrink to $\hat{\theta}$. Choose a sequence of sets A_n such that $P^x(A_n) \to c \in (0, 1)$ and $P(A_n) \to 0$, and define Q_n as Q_b was defined. Then $s(P, Q_n; x_1^n)$ still diverges at rate $n^{k/2}$.

REMARK 3. Lavine (1991) defines a density bounded class $\Gamma = \{q; l(\theta) \leq q(\theta) \leq u(\theta) \text{ for almost all } \theta\}$, where l and u are such that $\int l(\theta)\mu(d\theta) < 1 < \int u(\theta)\mu(d\theta)$. By Theorem 1, $s(P, \Gamma; x_1^n)$ diverges for any P that is interior to Γ, assuming that $l < u$.

The sequence $\{Q_b\}$ used in Theorem 1 is not pathological. These priors have bounded densities that are as smooth as p except on a set of measure 0. Thus, it is not the point masses in $\mathscr{P}$ that cause the problem. In fact, it is possible to replace the sequence $\{Q_b\}$ in Theorem 1 with a sequence of probability measures having continuous, differentiable densities and obtain the same result. The next result shows that $s(P, \Gamma_N^k; x_1^n)$ diverges. (Recall that Γ_N^k is the set of all k-dimensional normals). This follows by taking Q to be normal with mean $\hat{\theta}$ and letting its variance tend to 0. Then $s(P, Q; x_1^n)$ behaves in the limit as if Q were a point mass. The details of the proof are omitted.

THEOREM 2. *The quantity* $s(P, \Gamma_N^k; x_1^n)$ *diverges at rate* $n^{k/2}$.

2.2.2. *Geometric perturbations.* Still using total variation distance, it is possible to replace the linear perturbations with geometric perturbations. Unfortunately, this does not produce a tractable expression for $s(P, Q, x_1^n)$ akin to statement (i) of Lemma 1. However, we can still show that divergence occurs.

THEOREM 3. *If* $\Lambda_\varepsilon(p) \subset \Gamma$ *for some* $\varepsilon > 0$, *then* $s(P, \Gamma; x_1^n)$ *diverges at least as fast as* $n^{k/2}$.

PROOF. Let Q_b be as in Theorem 1. Tedious calculation yields

$$s(P, Q_b; x_1^n) = \frac{P_n^x\big(N_b(\hat{\theta})\big)\big(1 - P_n^x\big(N_b(\hat{\theta})\big)\big)}{P\big(N_b(\hat{\theta})\big)\big(1 - P\big(N_b(\hat{\theta})\big)\big)}$$

so the proof of Theorem 1 applies here also. □

2.3. *ϕ-Divergences.* We now consider diagnostics based on the ϕ-divergence. Again both linear and geometric perturbations are considered; correspondingly, we break this section into two subsections.

2.3.1. *Linear perturbations.* Under linear perturbations, the form of $s(P, Q; x)$ is given by Theorem 4.1 in Dey and Birmiwal (1994), which we state here as a lemma.

LEMMA 2. *Suppose that* $\int q^2/p < \infty$. *Then* $s(P, Q; x) = V_{P^x}(dQ/dP)/V_P(dQ/dP)$, *where* $V_R(h) = \int h^2\, dR - (\int h\, dR)^2$.

THEOREM 4. *If* $\Lambda_\varepsilon(P) \subset \Gamma$ *for some* $\varepsilon > 0$, *then* $s(P, \Gamma; x_1^n)$ *diverges at rate* $n^{k/2}$.

PROOF. Let Q_b be as in Theorem 1. It turns out that

$$s(P, Q_b; x_1^n) = \frac{P_n^x\big(N_b(\hat{\theta})\big)\big(1 - P_n^x\big(N_b(\hat{\theta})\big)\big)}{P\big(N_b(\hat{\theta})\big)\big(1 - P\big(N_b(\hat{\theta})\big)\big)},$$

so the proof of Theorem 1 applies again. This establishes that $s(P, \Gamma, x_1^n)$ diverges at rate $n^{k/2}$ or faster. However, Delampady and Dey [(1994), Theorem 3] show that $s(P, \Gamma, x_1^n)$ is bounded above by $L(\hat{\theta}; x_1^n)/m_P(x_1^n)$, hence the result holds. □

Now we consider the behavior of $s(P, \Gamma_N; x_1^n)$.

THEOREM 5. *When $k = 1$, $s(P; \Gamma_N^1; x_1^n)$ diverges at rate $n^{1/2}$.*

PROOF. Let p_n^x be the posterior density function and let $q(\theta) = \phi(\theta; \mu, \tau^2)$. Then

$$\begin{aligned} s(P, Q; x_1^n) &= \frac{V_{P_n^x}(q/p)}{V_P(q/p)} \\ &= \left[\int \left(\tau^{-1}\phi((\theta-\mu)/\tau)/p(\theta)\right)^2 p_n^x(\theta)\,d\theta \right. \\ &\qquad \left. -\left(\int \tau^{-1}\phi((\theta-\mu)/\tau)/p(\theta)\, p_n^x(\theta)\,d\theta\right)^2\right] \\ &\quad \times\left[\int \left(\tau^{-1}\phi((\theta-\mu)/\tau)/p(\theta)\right)^2 p(\theta)\,d\theta - 1\right]^{-1} \\ &= \left[\int \left[p_n^x(\theta)/p^2(\theta)\right]\left(1/\tau\sqrt{2\pi}\right)\phi\left(\sqrt{2}\,(\theta-\mu)/\tau\right)d\theta \right. \\ &\qquad \left. -\tau\left(\int \left[p_n^x(\theta)/p(\theta)\right]\tau^{-1}\phi((\theta-\mu)\tau)\,d\theta\right)^2\right] \\ &\quad \times\left[\int [1/p(\theta)]\left(1/\tau\sqrt{2\pi}\right)\phi\left(\sqrt{2}\,(\theta-\mu)/\tau\right)d\theta - \tau\right]^{-1}, \end{aligned}$$

which converges to $p_n^x(\mu)/p(\mu) = L(\mu)/m_P(x_1^n)$ as $\tau \to 0$ since the measure with density $\sqrt{2}/\tau\phi(\sqrt{2}\,(\theta-\mu)/\tau)$ converges to a point mass at μ. Now take $\mu = \hat{\theta}$ and apply (R3). □

2.3.2. *Geometric perturbations.* Now we consider geometric perturbations. The next lemma corresponds to Theorem 4.2 in Dey and Birmiwal (1994).

LEMMA 3. *Suppose that $\int(\log(dQ/dP))^2\,dP < \infty$ and $\int(\log(dQ/dP))^2 \times(dQ/dP)^{\varepsilon}\,dP < \infty$, for some $\varepsilon > 0$. Then*

$$s(P, Q; x) = \frac{V_{P^x}(\log(dQ/dP))}{V_P(\log(dQ/dP))}.$$

REMARK 4. By examining terms in the first and second derivatives of $\phi(dQ_\varepsilon/dP)(dQ_\varepsilon/dP)$ with respect to ε, convergence of the above integrals justifies the interchanges of differentiation and integration required to establish Lemma 3.

THEOREM 6. *If* $\Lambda_\varepsilon(P) \subset \Gamma$ *for some* $\varepsilon > 0$, *then* $s(P, \Gamma; x_1^n)$ *diverges at rate* $n^{k/2}$.

PROOF. Again, if $s(P, \Gamma, x_1^n)$ diverges, it cannot do so faster than $n^{k/2}$ [Delampady and Dey (1994)]. Define Q_b as in Theorem 1. Then

$$s(P, Q; x_1^n) = \frac{P_n^x\big(N_b(\hat{\theta})\big)\big(1 - P_n^x\big(N_b(\hat{\theta})\big)\big)}{P\big(N_b(\hat{\theta})\big)\big(1 - P\big(N_b(\hat{\theta})\big)\big)},$$

so the proof of Theorem 1 applies. □

REMARK 5. In general, the form of $s(P, Q, x_1^n)$ will depend on the choice of distance and on whether linear or geometric perturbations are used. It is remarkable that $s(P, Q_b; x_1^n)$ has exactly the same form in Theorems 1, 3, 4 and 6. We have no explanation for this curious phenomenon.

2.4. *Summary of Section* 2. We considered, to various extents, the four diagnostics that arise by selecting either d_{TV} or d_ϕ, and using either linear or geometric perturbations. In all four cases, if P is interior to Γ with respect to the density ratio topology, then $s(P, \Gamma; x_1^n)$ diverges at rate $n^{k/2}$. The combination of geometric perturbations and total variation distance is generally intractable, and therefore not useful. We also considered taking Γ to be all normal distributions. For linear perturbations under both total variation distance and ϕ-divergence, this diagnostic diverges. In the next section we will consider the normal class under ϕ-divergence and geometric perturbations.

3. Parametric diagnostics.

3.1. *Introduction.* The results of the previous section suggest that we will have to consider choosing Γ to be a parametric family. In this section we assume a one-dimensional parameter space ($k = 1$). We will see that parametric perturbations produce diagnostics with good behavior. We discuss multiparameter versions in Section 4. In Section 3.2 we use the total variation metric with linear perturbations and we consider a normal prior and likelihood with normal perturbations having the same variance as the base prior. In Section 3.3 we show that by taking d to be ϕ-divergence and using geometric perturbations we can relax the assumptions of Section 3.2. In particular, we do not assume a normal prior or likelihood. Furthermore, we use a much larger class of perturbations, namely, the set of all normals.

3.2. *Total variation metric with normal perturbations.* Let $\theta \sim N(\mu, \tau^2)$ and assume that the data are i.i.d. $N(\theta, \sigma^2)$. A natural class to consider is the set $\Gamma = \{Q_{a,\tau^2}; a \in \mathbb{R}\}$. It is easy to show that $d(P, Q_{a,\tau^2}) = H(|\mu - a|/(2\tau))$, where $H(b) = \Phi(b) - \Phi(-b)$ and Φ is the cumulative distribution function for a standard normal. Thus,

$$s(P, Q_{a,\tau^2}; x_1^n) = \frac{\phi(\bar{x}; a, v^2)}{\phi(\bar{x}; \mu, v^2)} \frac{H\big(|\mu - a|\sigma/\big(2\tau\sqrt{n\tau^2 + \sigma^2}\big)\big)}{H(|\mu - a|/2\tau)}, \tag{3}$$

where $v^2 = (n\tau^2 + \sigma^2)/n$. Since $2b\phi(b) \le H(b) \le 2b\phi(0)$, it follows that

$$s(P, Q; x_1^n) \le \frac{\sigma}{\sqrt{n\tau^2 + \sigma^2}} \frac{\phi(\bar{x}; a, v^2)}{\phi(\bar{x}; \mu, v^2)} \frac{\phi(0)}{\phi(|\mu - a|/(2\tau))}.$$

The right-hand side is maximized at $\hat{a} = (n\tau^2\mu + \mu\sigma^2 - 4n\tau^2\bar{x})(\sigma^2 - 3n\tau^2)$, thus,

$$\begin{aligned} s(P, \Gamma; x_1^n) &= \sup_a s(P, Q_{a,\tau^2}; x_1^n) \\ &\le \frac{\sigma}{\sqrt{n\tau^2 + \sigma^2}} \exp\left\{ \frac{2n^2\tau^2(\bar{x} - \mu)^2}{(3n\tau^2 - \sigma^2)(n\tau^2 + \sigma^2)} \right\} \\ &= O\left(\frac{1}{\sqrt{n}} \right). \end{aligned}$$

A lower bound may be obtained in a similar way and we have

$$\begin{aligned} &\frac{\sigma}{\sqrt{n\tau^2 + \sigma^2}} \exp\left\{ \frac{(\bar{x} - \mu)^2}{2\tau^2} \left(1 + O\left(\frac{1}{n}\right)\right) \right\} \\ &\quad \le s(P, \Gamma; x_1^n) \\ &\quad \le \frac{\sigma}{\sqrt{n\tau^2 + \sigma^2}} \exp\left\{ \frac{(\bar{x} - \mu)^2}{1.5\tau^2} \left(1 + O\left(\frac{1}{n}\right)\right) \right\} \\ &\quad \sim \frac{\text{s.e.}(\hat{\theta})}{\tau} \exp\left\{ \frac{(\bar{x} - \mu)^2}{1.5\tau^2} \right\}, \end{aligned} \tag{4}$$

where $\text{s.e.}(\hat{\theta}) = \sigma/\sqrt{n}$. The latter formula may be interpreted as a Bayesian standard error: the usual standard error of the maximum likelihood estimate times a factor that measures the conflict between the prior and the data. We have found that $s(P, Q_{\bar{x},\tau^2}; x_1^n)$, which requires no maximizations, is an accurate approximation to $s(P, \Gamma; x_1^n)$. Basu, Jammalamadaka and Liu (1996) considered a different type of parametric derivative.

3.3. *ϕ-divergence with geometric contaminations.* The diagnostics in the previous section are useful but require that we restrict the perturbations to have the same variance as the base prior. In some problems we would like to permit a larger class of perturbations. Here we use the ϕ-divergence with geometric perturbations and we consider the class Γ_N^1 of all normal priors.

THEOREM 7. (i) *We have*

$$(5)\quad s(P, \Gamma_N^1, x_1^n) = \sup_{a,t} \frac{b_{11}^x a^2 - b_{12}^x a + d^x t^4 + c_2^x t^2 - 2c_1^x a t^2 + \frac{1}{4} b_{22}^x}{b_{11} a^2 - b_{12} a + d t^4 + c_2 t^2 - 2c_1 a t^2 + \frac{1}{4} b_{22}},$$

where $b_{kl} = \mathrm{Cov}_P(\theta^k, \theta^l)$, $c_k = \mathrm{Cov}_P(\theta^k, \log p(\theta))$, $d = \mathrm{Var}_P(\log p(\theta))$ *and* b_{kl}^x, c_k^x *and* d^x *are the analogous posterior covariances.*

(ii) *If* P *is normal, this simplifies to*

$$(6)\qquad s(P, \Gamma_N^1, x_1^n) = \sup_a \frac{b_{11}^x a^2 - b_{12}^x a + \frac{1}{4} b_{22}^x}{b_{11} a^2 - b_{12} a + \frac{1}{4} b_{22}}.$$

(iii) *If* P *is normal and the model is* $X_1, \dots, X_n \mid \theta \sim N(\theta, 1)$, *then*

$$s(P, \Gamma_N^1, x_1^n) = O_P\left(\frac{1}{n}\right).$$

PROOF. If $Q = N(a, t^2)$, then, by Lemma 3,

$$(7)\qquad s(P, Q, x) = \frac{V_{P_n^x}\left[(\theta - a)^2 + 2t^2 \log p(\theta)\right]}{V_P\left[(\theta - a)^2 + 2t^2 \log p(\theta)\right]}.$$

Manipulation of (7) yields (i). If $P = N(\mu, \tau^2)$, then (7) becomes

$$\frac{V_{P_n^x}\left[\theta^2 - 2\gamma(a, t^2)\theta\right]}{V_P\left[\theta^2 - 2\gamma(a, t^2)\theta\right]},$$

where $\gamma(a, t^2) = (1/\tau^2 - 1/t^2)^{-1}\,(\mu/\tau^2 - a/t^2)$. We can safely ignore the case $t^2 = \tau^2$. Since the image of γ is the whole real line, it suffices to maximize over γ. This is equivalent to setting $t^2 = 0$ and maximizing over a, giving (ii). When the model is also normal, tedious algebraic manipulations yield

$$s(P, Q; x_1^n) = \left(\frac{\tau_x^2}{\tau^2}\right)\left\{\frac{\tau_x^2 + 2(\mu_x - \gamma)^2}{\tau^2 + 2(\mu - \gamma)^2}\right\},$$

where μ_x and τ_x^2 are the posterior mean and variance of θ, and γ is defined as above. Letting $d = \mu - \mu_x$,

$$\begin{aligned} s(P, \Gamma_N^1; x_1^n) &\le \left(\frac{\tau_x^2}{\tau^2}\right)\left\{\left(\frac{\tau_x^2}{\tau^2}\right) + 2\sup_\gamma \frac{(\mu_x - \gamma)^2}{\tau^2 + 2(\mu - \gamma)^2}\right\} \\ &= \left(\frac{\tau_x^2}{\tau^2}\right)\left\{\left(\frac{\tau_x^2}{\tau^2}\right) + 1 + \frac{2d^2}{\tau^2}\right\} = O_P\left(\frac{1}{n}\right), \end{aligned}$$

since $\tau_x^2/\tau^2 = O_p(1/n)$ and $d = O_p(1)$. □

REMARK 6. We conjecture that statement (iii) holds for all regular models and priors.

REMARK 7. The two-dimensional maximization in (i) can be reduced to one-dimensional numerical maximization, since for a fixed value of t we can maximize over a analytically.

EXAMPLE 1 (Continued). In the setting of Example 1, the present diagnostic is computed for a variety of n and $\bar{X}$. The results are displayed in Table 1.

4. Multiparameter models.

4.1. *Introduction.* We now turn to multiparameter models. Our goal is to examine the sensitivity of a marginal of the posterior to different aspects of the prior. Let $\theta = (\theta_1, \dots, \theta_k) \in \Theta = \Theta_1 \times \cdots \times \Theta_k$, and let $P^{x,i}(d\theta_i)$ denote the posterior marginal for θ_i. We are interested in studying the sensitivity of $P^{x,i}$ to perturbation of the jth marginal $P(d\theta_j)$. We shall not deal with other cases such as perturbations to conditionals.

To measure the sensitivity of $P^{x,i}(d\theta_i)$ to a perturbation in the direction of Q, we define

$$s_i(P, Q; x) = \lim_{\varepsilon \downarrow 0} \frac{d\left(P^{x,i}, Q_\varepsilon^{x,i}\right)}{d(P, Q_\varepsilon)}. \tag{8}$$

Letting Γ be a set of one-dimensional priors and using $\theta_{(j)}$ to denote $(\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_k)$, we define

$$\Gamma_j = \left\{Q; Q(d\theta_1, \dots, d\theta_k) = P\left(d\theta_{(j)} \mid \theta_j\right) Q_j\left(d\theta_j\right), Q_j \in \mathrm{T}\right\}$$

as a class of priors with perturbed θ_j marginal. Finally, we define

$$s_i^j(P, \Gamma; x) = s_i\left(P, \Gamma_j; x\right), \qquad i, j = 1, \dots, k.$$

For example, s_1^2 measures the sensitivity of the θ_1 posterior marginal to a perturbation of the θ_2 prior marginal. The matrix of values $\{s_i^j\}$ provides a convenient summary of the sensitivity of different components of the poste-

TABLE 1
Diagnostic based on normal perturbations with ϕ-divergence, univariate case

n	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$
5	0.167	0.314	0.999	2.155	3.774
20	0.048	0.104	0.361	0.792	1.397
40	0.024	0.055	0.193	0.425	0.750
60	0.016	0.037	0.132	0.290	0.512
80	0.012	0.028	0.100	0.220	0.389
100	0.010	0.023	0.080	0.177	0.313

rior to different components of the prior. Analogous to Section 3, we now consider two cases.

4.2. *Linear perturbation and total variation metric.* Let $Q_{A,\tau}$ be a bivariate normal with mean vector A and covariance τ. Let $\Gamma_0 = \{Q_{A,\tau};\ A \in \mathbb{R}^2\}$, and let $\phi(\theta; c, D)$ denote a bivariate normal density with mean vector c and covariance matrix D. It can be shown that

$$(9)\quad s(P, Q_{A,\tau}; x_1^n) = \frac{\phi(\bar{x}; A, V)}{\phi(\bar{x}; \mu, V)} \frac{H\left(\frac{1}{2}\{(\mu^x - A^x)'\tau_x^{-1}(\mu^x - A^x)\}^{1/2}\right)}{H\left(\frac{1}{2}\{(\mu - A)'\tau^{-1}(\mu - A)\}^{1/2}\right)},$$

where $V = \tau + (1/n)\sigma$, $\tau_x = (\tau^{-1} + n\sigma^{-1})^{-1}$, $\mu^x = \tau_x(\tau^{-1}\mu + n\sigma^{-1}\bar{x})$, $A^x = \tau_x(\tau^{-1}A + n\sigma^{-1}\bar{x})$ and $\bar{x}$ is the sample average vector.

To perturb only the θ_1 margin of the prior, we hold the prior on θ_2 given θ_1 fixed and replace the marginal prior with a $N(a_1, \tau_{11})$ distribution. This implies that the joint perturbation prior is $N(A_1, \tau)$, where $A_1 = (a_1, \mu_2 + (\tau_{12}/\tau_{11})(a_1 - \mu_1))'$. By similar reasoning, a perturbation to the θ_2 marginal is represented by the prior $N(A_2, \tau)$, where $A_2 = (\mu_1 + (\tau_{12}/\tau_{22})(a_2 - \mu_2), a_2)'$. Hence,

$$(10)\qquad s_i^j = \frac{\phi(\bar{x}; A_j, V)}{\phi(\bar{x}; \mu, V)} \frac{H\left(|\mu^x(i) - A_j^x(i)|/\left(2\sqrt{\tau_x(i,i)}\right)\right)}{H\left(\frac{1}{2}\{(\mu - A_j)'\tau^{-1}(\mu - A_j)\}^{1/2}\right)},$$

where $b(i)$ is the ith component of the vector b and $B(i,j)$ is the (i,j)th component of the matrix B. Maximizing over the free parameter in these expressions can be numerically intensive. An alternative is to fix a direction by taking $a_1 = \bar{x}_1$ and $a_2 = \bar{x}_2$. Numerical calculations (not shown here) indicate that the diagnostic behaves well. As before, we shall now show that ϕ-divergences permit a larger class of perturbations.

4.3. *Geometric perturbation and ϕ-divergence.* As in Section 3, we will focus on geometric perturbations using ϕ-divergence and the class of all normal priors. However, first we establish the form of the local sensitivity of a posterior marginal.

THEOREM 8. *Suppose the regularity conditions of Lemma* 3 *hold. Then*

$$s_i(P, Q, x) = \frac{V_{P^x} E_{P^x}(\log(dQ/dP) \mid \theta_i)}{V_P(\log(dQ/dP))}.$$

The proof, which is similar to arguments in Dey and Birmiwal (1994), is omitted.

REMARK 8. The relation between conditional and unconditional variance immediately implies that $s_i(P, Q, x) < s(P, Q, x)$ for each i; that is, for any perturbation, the whole posterior is at least as sensitive as any marginal.

Theorem 7 extends to the multidimensional case.

THEOREM 9.

(i) *The quantity* $s_i^i(P, \Gamma_N^1; x)$ *is given by equation* (5), *where* $b_{kl} = \mathrm{Cov}_P(\theta_i^k, \theta_i^l)$, $c_k = \mathrm{Cov}_P(\theta_i^k, \log p(\theta_i))$, $d = \mathrm{Var}_P(\log p(\theta_i))$, *and* b_{kl}^x, c_k^x *and* d^x *are the analogous posterior covariances. If* $P(d\theta_i)$ *is normal, then* $s_i^i(P, \Gamma_N^1; x)$ *simplifies to equation* (6).

(ii) *When* $i \neq j$, $s_i^j(P, \Gamma_N^1; x)$ *is given by equation* (5) *with* $b_{kl} = \mathrm{Cov}_P(\theta_j^k, \theta_j^l)$, $c_k = \mathrm{Cov}_P(\theta_j^k, \log p(\theta_j))$, $d = \mathrm{Var}_P(\log p(\theta_j))$, $b_{kl}^x = \mathrm{Cov}_{P^x}(E_{P^x}[\theta_j^k \mid \theta_i], E_{P^x}[\theta_j^l \mid \theta_i])$, $c_k^x = \mathrm{Cov}_{P^x}(E_{P^x}[\theta_j^k \mid \theta_i], E_{P^x}[\log p(\theta_j) \mid \theta_i])$ *and* $d^x = \mathrm{Var}_{P^x}(E_{P^x}[\log p(\theta_j) \mid \theta_i])$. *If* $P(d\theta_j)$ *is normal, then* $s_i^j(P, \Gamma_N^1; x)$ *simplifies to equation* (6).

REMARK 9. Note that all the quantities appearing in the formula of Theorem 9 can be easily approximated by simulation from the posterior. Thus, when the likelihood is nonnormal, the diagnostic is still tractable. Also, a referee has pointed out that the eigenvalues of the sensitivity matrix can be used to summarize the sensitivity.

EXAMPLE 2. We compute the diagnostic when the model specifies (X_{1i}, X_{2i}) as i.i.d. bivariate normal with mean vector θ and covariance matrix Σ, and the prior on θ is bivariate normal with mean vector μ and covariance matrix T. For our numerical calculations we fix $\mu = (0,0)^T$ and

$$\Sigma = T = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

We consider a variety of sample sizes, and two data mean vectors: $\bar{X} = (\bar{X}_1, \bar{X}_2)^T = (1,1)^T$ and $\bar{X} = (1, -1)^T$. The results are displayed in Table 2. Note that only the upper row of the sensitivity matrix is given, as with these prior, model and data specifications, $s_2^1 = s_1^2$ and $s_2^2 = s_1^1$.

TABLE 2
Diagnostic based on normal perturbations with ϕ-divergence, bivariate case

	$\boldsymbol{X = (1,1)^T}$		$\boldsymbol{X = (1,-1)^T}$	
$\boldsymbol{n}$	$\boldsymbol{s_1^1}$	$\boldsymbol{s_1^2}$	$\boldsymbol{s_1^1}$	$\boldsymbol{s_1^2}$
5	0.314	0.077	0.314	0.077
20	0.104	0.026	0.104	0.026
40	0.055	0.014	0.055	0.014
60	0.037	0.009	0.037	0.009
80	0.028	0.007	0.028	0.007
100	0.023	0.006	0.023	0.006

5. Discussion. When studying sensitivity to prior distributions there is always a tradeoff: classes of priors that are too large have poor behavior and classes of priors that are too small may understate sensitivity. This is especially so for local diagnostics. In Sections 3 and 4 we presented two diagnostics that attempt to strike a balance between these extremes. Our conclusion is that the ϕ-divergence with all normal perturbations provides a useful diagnostic. This diagnostic can be calibrated by comparing the value of the diagnostic in a given problem with the value in a series of canonical problems. For example, let s_n be the value of the diagnostic defined by (5) when the model is $X_n \sim N(\theta, 1/n)$, the prior is $N(0,1)$ and $X_n = 0$. This corresponds to the first column in Table 1. Now for a given value of the diagnostic s we can find n such that $s_n \approx s$. This calibrates s in terms of sample size in the canonical normal problem. Of course, other calibration schemes are possible.

The most important application of the diagnostics is to multivariate problems. In these cases, sensitivity diagnostics help sort out the degree to which different parts of the prior affect a given marginal of the posterior. We do this by examining the relative values of the entries in the rows of the matrix $\{s_i^j\}$. For example, if a parameter of interest is highly sensitive to the prior on a nuisance parameter, then there is cause for concern since the prior on the nuisance parameter is usually less dependable than the prior on the parameter of interest.

Another approach to marginal sensitivity analysis, investigated in Gustafson (1994), is to measure the sensitivity of posterior expectations instead of studying sensitivity of the whole posterior. These diagnostics have good asymptotic behavior and can be simpler to compute.

An issue we did not address is the problem of defining sensitivity diagnostics for improper priors. The techniques we used do not go through in any obvious way for improper priors. Indeed, the probems are reminiscent of the asymptotic problems discussed in this paper. These delicate issues are compounded in situations where one parameter has a proper prior and another has an improper prior, which happens in some Bayesian testing problems. It is possible that some modification of our methods might lead to suitable diagnostics in these cases.

Acknowledgments. The authors thank an Associate Editor and two referees for helpful comments and suggestions.

REFERENCES

BASU, S., JAMMALAMADAKA, S. R. and LIU, W. (1996). Local posterior robustness with parametric priors: maximum and average sensitivity. In *Maximum Entropy and Bayesian Statistics* (G. Heidbreder, ed.). Kluwer, Dordrecht. To appear.

BERGER, J. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness in Bayesian Statistics* (J. Kadane, ed.). North-Holland, Amsterdam.

BERGER, J. (1986). Comment on "On the consistency of Bayes estimates" by P. Diaconis and D. Freedman. *Ann. Statist.* **14** 30–37.

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

CSISZAR, I. (1977). Information measures: a critical survey. In *Transactions of 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes 1974*. Reidel, Dordrecht.

CUEVAS, A. and SANZ, P. (1988). On differentiability properties of Bayes operators. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 569–577. Oxford Univ. Press.

DELAMPADY, M. and DEY, D. (1994). Bayesian robustness for multiparameter problems. *J. Statist. Plann. Inference* **40** 375–382.

DEROBERTIS, L. (1978). The use of partial prior knowledge in Bayesian inference. Ph.D. dissertation, Yale Univ.

DEY, D. and BIRMIWAL, L. R. (1994). Robust Bayesian analysis using entropy and divergence measures. *Statist. Probab. Lett.* **20** 287–294.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

GELFAND, A. and DEY, D. (1991). On Bayesian robustness of contaminated classes of priors. *Statist. Decisions* **9** 63–80.

GOEL, P. (1983). Information measures and Bayesian hierarchical models. *J. Amer. Statist. Assoc.* **78** 408–410.

GUSTAFSON, P. (1994). Local sensitivity of posterior expectations. Ph.D. dissertation, Dept. Statistics, Carnegie Mellon Univ.

GUSTAFSON, P., SRINIVASAN, C. and WASSERMAN, L. (1995). Local sensitivity analysis. In *Bayesian Statistics 5*. To appear.

KRASKER, W. S. and PRATT, J. W. (1986). Discussion of "On the consistency of Bayes estimates," by P. Diaconis and D. Freedman. *Ann. Statist.* **14** 55–58.

LAVINE, M. (1991). An approach to robust Bayesian analysis for multidimensional parameter spaces. *J. Amer. Statist. Assoc.* **86** 400–403.

RUGGERI, F. and WASSERMAN, L. (1993). Infinitesimal sensitivity of posterior distributions. *Canad. J. Statist.* **21** 195–203.

SIVAGANESAN, S. (1993). Robust Bayesian diagnostics. *J. Statist. Plann. Inference* **35** 171–188.

SRINIVASAN, C. and TRUSZCZYNSKA, H. (1990). On the ranges of posterior quantities. Technical Report 294, Dept. Statistics, Univ. Kentucky.

WASSERMAN, L. (1992). The conflict between improper priors and robustness. Technical Report 559, Dept. Statistics, Carnegie Mellon Univ.

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BRITISH COLUMBIA
CANADA V6T 1Z2

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213