ASPECTS OF BAYESIAN ROBUSTNESS IN HIERARCHICAL MODELS

By Paul Gustafson[1]
University of British Columbia

This article examines the sensitivity of inferences to perturbations at various stages of a hierarchically specified prior. For the most part, a local method of assessing sensitivity is adopted. First, the general behaviour of sensitivity across levels of the hierarchy is studied. Then quantitative measures of sensitivity are investigated; these are easily computed using Markov Chain Monte Carlo methods. Finally, asymptotic sensitivity of posterior based inference is considered, under misspecification of the first-stage prior distribution. Throughout, the role of influence functions is emphasized.

1. Introduction. Hierarchical Bayes models have been popular since the fundamental paper of Lindley and Smith (1972), though their origins can be traced back further; see Good (1980) and the references cited therein. The basic idea is to specify a joint distribution for data and parameters hierarchically, through a succession of conditional distributions. Specifically, the conditional distribution of a data vector θ_0 given a parameter vector θ_1 is specified, followed by the distribution of θ_1 given a second parameter vector θ_2, and so on. At some point the specification terminates, with the distribution of θ_{k+1} taken to be degenerate. That is the conditional for $\theta_k|\theta_{k+1}$ is specified, where θ_{k+1} is a known hyperparameter. After collecting the data, all statistical inference is based on the posterior distribution of the entire parameter $(\theta_1, \ldots, \theta_k)$ given the data θ_0 and the hyperparameter θ_{k+1}. Recent developments in Markov chain Monte Carlo methods facilitate exploration of the posterior distribution in most hierarchical models.

One can delineate at least four broad approaches to robustness which are applicable to hierarchical models. The present paper focusses on an essentially diagnostic approach. A single hierarchical model is specified, and parameter estimation is carried out. Various measures of sensitivity to assumptions are computed, without explicit modelling of alternate specifications. The second approach, investigated by Albert and Chib (1994), involves specification of a small number of competing hierarchical specifications. Bayes factors are used to determine which specifications are supported by the data. A third approach involves making part of the specification nonparametric. For instance, Escobar (1994) replaces a normal prior specification with a Dirichlet process on all possible prior distributions. Finally, Angers and Berger (1991) investigate the inherent robustness to outlying

[1] Research supported by a Postdoctoral Fellowship from the Natural Sciences and Engineering Research Council of Canada

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

Key words and phrases. Hierarchical models; Influence functions; Robustness.

random effects that can be attained by specifying thick-tailed random effect distributions.

The focus in this paper is on assessing the change in posterior–based inference induced by a change in prior specification. In multiparameter models generally, and hierarchical models specifically, it can be fruitful to assess the change in components of the inference induced by changes in components of the prior specification. For instance, for each (i, j) pair one can assess the sensitivity of the posterior marginal for $\theta_j | \theta_0, \theta_{k+1}$ to the prior specification for $\theta_i | \theta_{i+1}$. Since the prior is specified stage by stage, it may be important to know the relative influence of different prior stages. As well, it can be much easier to implement a sensitivity analysis if only one-dimensional prior distributions are perturbed.

Several authors have applied global Bayesian robustness analyses to hierarchical models. In particular, Cano (1993) and Polasek and Pötzelberger (1988) assess the range of inference produced when the last-stage prior (the prior for θ_k in the above notation) varies in a class of distributions. The work of Sivaganesan (1993) is also related to the present study.

The organization of the paper is as follows. Section 2 investigates the qualitative behaviour of sensitivity, focussing on the relationship between sensitivity and the respective levels of inference and perturbation within the hierarchy. Section 3 then reviews Bayesian local sensitivity techniques, which are based on infinitesimal perturbations to the prior distribution. These techniques are easily implemented in hierarchical models; an example is given in Section 4. Section 5 details an application of local sensitivity techniques in the large–sample limit.

2. Qualitative aspects. Recall the general hierarchical set-up of the previous section, whereby θ_0 is a vector of observables, the prior is specified hierarchically in terms of $\{\theta_i | \theta_{i+1}\}_{i=1}^{k}$, and θ_{k+1} is a known hyperparameter. Goel and DeGroot (1981) and Goel (1983) show quite generally that the amount of learning about θ_i from the data diminishes as i increases. More specifically, they show that the divergence between prior and posterior marginals for θ_i decreases as i increases, for many divergence measures. In a related vein, Haitovsky and Zidek (1986) truncate the true hierarchical prior by placing an improper prior on θ_i, and show that the resulting approximation to the true posterior density for θ_1 is increasingly accurate as i increases. One might like to know if similar orderings apply to sensitivity.

Let s_{ij} be a measure of sensitivity of the θ_j posterior marginal to uncertainty about the prior specification for the $\theta_i | \theta_{i+1}$ prior conditional. One might postulate that s_{ij} decreases as $|i - j|$ increases; that is sensitivity will be lower when there are many stages interceding between the stage of uncertain specification and the stage of inference. This turns out to be only partly true.

Consider two possible hierarchical specifications, $P^{(1)}$ and $P^{(2)}$, which differ only in terms of $\theta_i|\theta_{i+1}$. The discrepancy in inference about θ_j will be measured in terms of the total variation distance between the resulting posterior marginals, $P^{(1)}(\theta_j|\theta_0)$ and $P^{(2)}(\theta_j|\theta_0)$. Recall that $d_{TV}(P,Q) = \sup_A |P(A) - Q(A)|$. The total variation metric is attractive for use in robustness problems because of its straightforward interpretation in terms of probability. It also has the following intuitive and easily verified property.

LEMMA 1. *Assume that* (X_1,X_2) *and* (Y_1,Y_2) *each have a joint density on the sample space* $(\mathcal{X}_1,\mathcal{X}_2)$*, with respect to some dominating measure. If* $X_1|X_2 \overset{D}{\equiv} Y_1|Y_2$*, then* $d_{TV}(X_1,Y_1) \le d_{TV}(X_2,Y_2)$*, with equality if and only if* $X_2 \overset{D}{\equiv} Y_2$.

Succinctly, the distance between two different mixtures of the same conditional distribution is less than the distance between the mixing distributions. The lemma permits investigation of the behaviour of s_{ij}, as i is fixed and j varies.

RESULT 1. *Let* $d(j) = d_{TV}(P^{(1)}(\theta_j|\theta_0,\theta_{k+1}), P^{(2)}(\theta_j|\theta_0,\theta_{k+1}))$*, for fixed* i *and* θ_0*. Then* $d(j)$ *is a strictly unimodal function, which is maximized at either* $j = i$ *or* $j = i + 1$.

PROOF. First consider the case $1 \le j < i$. The aim is to show that $d_j < d_{j+1}$. Since the distribution of $(\theta_0,\ldots,\theta_j|\theta_{j+1})$ is completely determined by $\{P(\theta_m|\theta_{m+1})\}_{m=0}^{j}$, it follows that $P^{(1)}(\theta_j|\theta_{j+1},\theta_0,\theta_{k+1})$ and $P^{(2)}(\theta_j|\theta_{j+1},\theta_0,\theta_{k+1})$ are equal (neither distribution actually depends on θ_{k+1}, due to the conditioning on θ_{j+1}). Applying Lemma 1 gives the result. Similarly, consider $i < j < k$. Specifying $\{P(\theta_m|\theta_{m+1})\}_{m=j}^{k}$ completely determines the distribution of $\theta_{j+1}|\theta_j,\theta_{k+1}$, which is equivalent to $\theta_{j+1}|\theta_0,\theta_j,\theta_{k+1}$. Thus $P^{(1)}(\theta_{j+1}|\theta_0,\theta_j,\theta_{k+1})$ and $P^{(2)}(\theta_{j+1}|\theta_0,\theta_j,\theta_{k+1})$ are equivalent in distribution, and application of Lemma 1 shows that $d_{j+1} < d_j$.□

Hence for perturbations at a particular stage of the hierarchy, sensitivity falls off as the level of inference moves away from the level of perturbation, in either direction. This is in agreement with the postulated behaviour. The result is a global sensitivity result, in contrast to the rest of the paper which focusses on local sensitivity.

It is much harder to make general statements when the stage of inference is fixed and the stage of perturbation varies. It is possible to make some progress in the special case of a normal location model with known variances. In particular, let $\theta_i|\theta_{i+1} = \theta_{i+1} + \epsilon_i$, where $\epsilon_0, \ldots, \epsilon_k$ are independently normally distributed, with respective variances $\sigma_0^2, \ldots, \sigma_k^2$ (here θ_{k+1} is known, and $\epsilon_{k+1} = 0$). A local measure of the sensitivity of the posterior mean of θ_j to perturbation of the prior on ϵ_i is considered. That is, the location structure is left intact under perturbation; only the noise distribution is perturbed. In Gustafson (1996a) it is shown that when j is fixed, and i varies over $\{1, \ldots, j-1\}$, the sensitivity measure and σ_i^2 share the same ordering. The same is true as i varies over $\{j, \ldots, k\}$. Thus in this case the separation $|i-j|$ does not play a role in the sensitivity ordering. Rather, the stages of higher nominal prior variances are more influential on inference at a particular stage. This result indicates that inference will be sensitive to improper priors, a finding that agrees with Pericchi and Nazaret (1988).

3. Review of local sensitivity. There are many recent papers discussing local sensitivity measures based on differentiation of posterior quantities. The method presented here combines elements from Gustafson (1996b), Ruggeri and Wasserman (1993), and Sivaganesan (1993).

Consider a partition of the k-dimensional parameter θ, as $\theta = (\phi, \psi)$. Let $P(\phi, \psi)$ be the nominal joint prior on θ. In terms of perturbations to the prior, the marginal $P(\phi)$ is allowed to vary, while the conditional $P(\psi|\phi)$ remains fixed. Sometimes the ϕ prior marginal is denoted as P_ϕ. Typically ϕ will be one–dimensional; that is sensitivity to each one–dimensional prior marginal is assessed, in turn. For a likelihood function $L(\theta)$ and a function of interest $g(\theta)$, define

$$T(Q) = \frac{\int g(\theta)L(\theta)dP(\psi|\phi)dQ(\phi)}{\int L(\theta)dP(\psi|\phi)dQ(\phi)}.$$

Thus T maps the prior marginal for ϕ to the posterior mean of $g(\theta)$. Perhaps the simplest local measure of sensitivity is the directional derivative of T at P_ϕ in direction Q, which exists under weak conditions, and can be expressed in an influence function representation (Hampel, Ronchetti, Rousseeuw and Stahel, 1986) as follows:

$$(1) \qquad \frac{\partial}{\partial\epsilon} T((1-\epsilon)P_\phi + \epsilon Q)\Big|_{\epsilon=0} = \int IF_P(z)d[Q - P_\phi](z).$$

From Gustafson (1996a), the influence function IF_P is given by

$$(2) \qquad IF_P(z) = E^x(g(\theta) - \rho_g|\phi = z)\left[\frac{dP_\phi^x}{dP_\phi}(z)\right].$$

In the above expression, and throughout the remainder of the paper, P^x and E^x denote posterior distributions and expectations under the nominal prior P, while $\rho_g = E^x(g(\theta))$.

The influence function representation is useful since IF_P does not depend on the direction Q, and so encapsulates the local sensitivity to perturbations in all directions. Equation (1) only defines the influence function up to an additive constant. It is convenient to standardize by requiring $\int IF_P(z)dP_\phi(z) = 0$. By taking $Q = \delta_z$, it then follows that $\epsilon IF_P(z)$ is a first-order approximation to the change in T that arises upon ϵ-contamination of P_ϕ by a point mass at z. This gives a graphical interpretation for the influence function. Note that (2) is already in standardized form.

The Fréchet derivative is considered next. Temporarily extend the domain of T to all signed measures having total mass one. Let Γ be all signed measures U having total mass zero, and consider a norm $||U||$ on Γ. If there is a linear functional $\dot{T}(P_\phi)$ such that

$$(3) \qquad T(P_\phi + U) \;=\; T(P_\phi) + \dot{T}(P_\phi)U + o(||U||)$$

uniformly on U in bounded (in norm) subsets of Γ, then $\dot{T}(P_\phi)$ is the Fréchet derivative of T at P_ϕ. This is a stronger notion of derivative than (1). If the Fréchet derivative exists, however, it must coincide with the influence function. That is $\dot{T}(P_\phi)U = \int IF_P(z)dU(z)$. To ensure that $P_\phi + U$ is a probability measure, let Γ_R be the class of all directions of the form $U = \epsilon(Q - P_\phi)$, for some probability measure Q and some $\epsilon \in [0,1]$. Then the restricted norm of the derivative,

$$||\dot{T}(P_\phi)|| \;=\; \sup_{U\in\Gamma_R} \frac{|\dot{T}(P_\phi)U|}{||U||},$$

reflects the maximum change in the posterior expectation relative to change in the prior marginal.

A useful specification of a norm on Γ is

$$||U|| \;=\; \left(\int \left[\frac{dU}{dP_\phi}\right]^2 dP_\phi\right)^{1/2},$$

the L_2 norm, with respect to P_ϕ, of the 'density' dU/dP_ϕ. (Since P_ϕ is always a fixed nominal measure, there is no harm in letting the norm depend on P_ϕ.) In this case, the restricted norm of the derivative is simply

$$(4) \qquad ||\dot{T}(P_\phi)|| \;=\; \left(\int (IF_P(z))^2 dP_\phi(z)\right)^{1/2},$$

which would be reported as a single number summary of the sensitivity of the posterior mean of $g(\theta)$ to perturbation of the ϕ prior marginal.

Table 1: *Topical cream data. Successes / Number of cases.*

Clinic	1	2	3	4	5	6	7	8
Active	11/36	16/20	14/19	2 /16	6 /17	1 /11	1 /5	4 /6
Control	10/37	22/32	7/19	1/17	0/12	0/10	1/9	6/7

The advantage of the Fréchet derivative over the weaker directional derivative is that the former yields an approximation to a global robustness quantity:

$$\sup_{Q \in \Gamma_{R,\delta}} T(Q) \quad = \quad T(P_\phi) + \delta \|\dot{T}(P_\phi)\| + o(\delta),$$

where $\Gamma_{R,\delta} = \{P_\phi + U : U \in \Gamma_R, \|U\| \leq \delta\}$ is the class of all ϵ-contaminations of P_ϕ within a δ radius of P under χ^2 distance. Without the uniformity in (3), such an approximation is not valid. The above–specified norm on U yields the χ^2 distance. It is a compromise between the extreme richness of total variation neighbourhoods (based on L_1 distance) and lack of richness of density ratio neighbourhoods (based on L_∞ distance), and yields asymptotically sensible sensitivity measures (Gustafson 1996b).

The above discussion focusses on perturbations to prior marginals, even though hierarchical models are specified in terms of conditional distributions. The motivation for this is twofold. First, it is much easier to think about perturbations to a single distribution (marginal) than to a family of distributions (conditional). Second, it is often desirable to retain the conditional structure and perturb only the noise distribution. For instance, if the nominal specification is $\theta_i|\theta_{i+1} \sim N(\theta_{i+1}, \sigma^2)$, it may make more sense to write $\theta_i|\theta_{i+1} = \theta_{i+1} + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$, and then perturb the marginal prior distribution of ϵ_i in order to assess sensitivity to the conditional specification for $\theta_i|\theta_{i+1}$.

4. Quantitative aspects. In this section, influence functions and derivative norms are computed in a practical hierarchical model, to show how local sensitivity measures can be obtained from Markov chain Monte Carlo output. The example is a mixed effects probit model, applied to a data set from Beitler and Landis (1985) (given here in Table 1). The data arose from a multicentre clinical trial comparing the efficacies of two topical creams, one active and one control, in curing infection.

Let Y_{ijk} represent a patient's binary outcome (1=success, 0=failure), with $i = 1, 2$ indexing treatment (1=active, 2=control), $j = 1, \ldots, 8$ indexing centre, and $k = 1, \ldots, n_{ij}$ indexing patient within treatment/centre. Conditioned on parameters μ, α, and τ, all patient responses are assumed independent, with $P(Y_{ijk} = 1) = \Phi(\mu_i + \tau\alpha_j)$, where Φ is the standard normal distribution function. A priori, μ, α and τ are assumed independent. The components of μ are taken to be *iid* $N(\lambda, \sigma^2)$, reflecting a prior belief in the exchangeability of treatments. The components of α are taken to be *iid* $N(0,1)$. Finally, as (the square root of) a variance component, τ (which is assumed to be nonnegative) is assigned a half-Cauchy prior density with scale parameter τ_0. Thus τ has its prior mode at 0, which seems reasonable for a variance component. For the sake of illustration, hyperparameter values $\lambda = 0$, $\sigma^2 = 4$, and $\tau_0 = 0.375$ are selected. To roughly interpret these hyperparameters, the average ($\alpha_i = 0$) success probabilities approximately have prior quartiles 0.1 and 0.9, while when $\mu_i = 0$, the prior quartiles for the random effects are approximately 0.4 and 0.6, when translated to the probability scale.

More typically the model might be expressed hierarchically, by first specifying the success probability conditional on the random effect $\tau\alpha_j$, then specifying the normal random effect distribution with variance component τ^2. The parameterization given above is useful because it is expressed in terms of the quantities to which we would like to assess sensitivity. That is we can separately assess sensitivity to the prior marginal for α which gives the distributional form of the random effects, as well as to the prior marginal for τ which represents belief about the magnitude of the random effects.

The posterior distribution of parameters is explored via Gibbs sampling. Following Carlin and Polson (1992), latent variables are introduced to facilitate Gibbs sampling. Consequently all required conditional distributions have standard forms, except for τ which can easily be sampled by the rejection method with draws from a normal distribution. All inferences and sensitivity calculations are based on a sample of size 5000 from the Gibbs sampler.

Straightforwardly, the influence for any one-dimensional prior marginal and function of interest is given by (2). As discussed in Gustafson (1996a), there are identities which are useful for calculating influence functions from Markov Chain Monte Carlo output. Generically, consider a partioned parameter, $\theta = (\phi, \psi)$. To compute an influence function for sensitivity to perturbation of the ϕ prior marginal, a function of the form

$$a(z) = E^x(h(\theta)|\phi = z)p^x_\phi(z) \tag{5}$$

is required. (Here p^x_ϕ denotes the posterior marginal density of ϕ.) The following identity can be helpful in computing a:

$$a(z) \;=\; E^x\left[h(z,\psi)p^x_{\phi|\psi}(z \mid \psi)\right]. \tag{6}$$

In particular, given a sample from the posterior distribution of parameters, as can be approximately produced by a Markov chain Monte Carlo scheme, the

Table 2: *Posterior means and standard deviations of the functions of interest.*

quantity of interest	g_1	g_2	g_3	g_4
posterior mean	0.41	0.26	0.41	0.35
posterior standard deviation	0.12	0.10	0.11	0.11

sample average of the quantity in square brackets is an unbiased estimator of $a(z)$. In the special case of $h \equiv 1$, this reduces to the usual "Rao-Blackwell" estimator of a marginal density. The above is helpful when the posterior conditional of $\phi|\psi$ has a closed form. If this is not the case, one can resort to using

$$a(z) \quad = \quad E^x\left[h(z,\psi)\frac{p^x_{\phi|\psi}(z \mid \psi)}{p^x_{\phi|\psi}(\phi \mid \psi)}f(\phi|\psi)\right],$$

where f is any family of densities indexed by ψ. This is a generalization of an identity advocated by Chen (1994). In practice f can be taken to be the generating density, if ϕ is updated either by a Gibbs sampling step using the rejection method or by a Metropolis step.

In the present context, four inferential quantities are considered: $g_1(\theta) = \Phi(\mu_1)$, $g_2(\theta) = \Phi(\mu_2)$, $g_3(\theta) = \Phi(\mu_1 + \tau z_0) - \Phi(\mu_1 - \tau z_0)$, and $g_4(\theta) = \Phi(\mu_2 + \tau z_0) - \Phi(\mu_2 - \tau z_0)$. Here $z_0 = \Phi^{-1}(.75) \approx 0.674$. Note that all posterior quantities are on the probability scale: g_1 and g_2 are success probabilities for an average patient ($\alpha_i = 0$) on active and control treatments respectively, while g_3 and g_4 are within–treatment interquartile ranges for the success probabilities. The estimated posterior means and posterior standard deviations for these quantities of interest are given in Table 2.

Influence functions of the form (2) can be computed for all combinations of prior marginals and functions of interest. For the sake of brevity only the influence functions for the μ_1, μ_2, and τ prior marginals are plotted in Figure 1. (The last row of plots should be ignored for now.) Relative sensitivity statements can be made directly from the plots. For example, the priors on μ_1 and μ_2 have more influence on the average success probabilities (g_1 and g_2) than on the interquartile ranges (g_3 and g_4). Conversely, the prior on τ has much more influence on the interquartile ranges than on the average success probabilities, as is to be expected. To make comparisons on a numerical scale, the derivative norms (4) are computed and reported in Table 3. (Again, ignore the last row for now.)

Sometimes it is necessary to be careful about the meaning of parameters when distributions are perturbed. For instance, in the present model

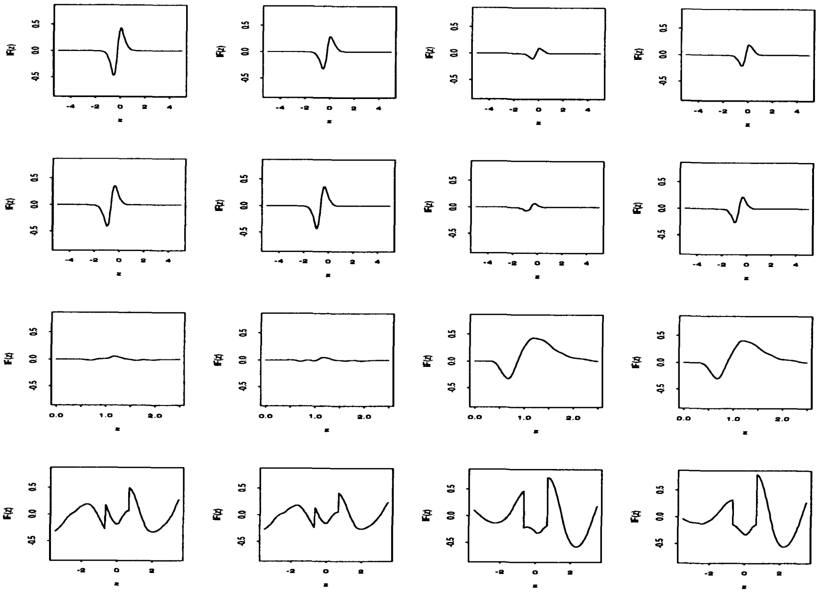

Figure 1: *Influence functions in the topical cream data example. The rows correspond to perturbed prior marginals for μ_1, μ_2, τ, and α. That is the last row corresponds to simultaneous perturbation of all random effect marginals in the same direction. The columns correspond to functions of interest g_1, g_2, g_3, and g_4.*

τ^2 can be interpreted as the variance of the random effects acting on the probit scale. But when the prior on α_i is perturbed, this is no longer true in general. Thus it may be desirable to constrain certain prior summaries under perturbation. This can be accomplished via the usual influence function, used only in conjunction with contaminations that preserve the desired summaries. However, direct graphical interpretation of the influence function is more difficult in this situation. In fact constraints can be incorporated in the influence function directly. This is done by forming the ϵ-contamination

Table 3: *Derivative norms for clinical trial data set. The rows index the perturbed prior marginal, while the columns index the quantity of interest.*

	g_1	g_2	g_3	g_4
μ_1	0.248	0.165	0.052	0.108
μ_2	0.188	0.192	0.037	0.117
τ	0.010	0.010	0.151	0.138
α	0.500	0.400	0.608	0.677

of a nominal distribution P by Q as usual, and then rescaling the resulting distribution to ensure it satisfies the constraints. For instance, if two constraints are placed on the prior density p, then the ϵ-contamination of p by q is taken to be:

$$(7) \qquad q_\epsilon(\cdot) \;=\; b(\epsilon)\left[(1-\epsilon)p\{b(\epsilon)(\cdot - a(\epsilon))\} + \epsilon q\{b(\epsilon)(\cdot - a(\epsilon))\}\right],$$

where $a(\epsilon)$ and $(b(\epsilon)$ are implicitly defined to ensure that Q_ϵ satisfies the desired constraints. Using the above, the directional derivative is given as follows.

RESULT 2. *Let the ϕ marginal of a prior with density $p(\phi,\psi)$ be perturbed according to (7). Then the derivative of $T(Q_\epsilon)$ with respect to ϵ, evaluated at $\epsilon = 0$, has an influence function representation of the form*

$$(8) \qquad IF(z) + Cov^x(g(\theta), p^*(\phi))a^*(z) + Cov^x(g(\theta), \phi p^*(\phi))b^*(z),$$

where $IF(\cdot)$ is the unconstrained influence function, and a^* satisfies $a'(0) = \int a^*(z)d[Q-P](z)$, with the analogous relation holding for $b'(0)$ and b^*. Also, $p^*(\phi) = -p'(\phi)/p(\phi)$.

As an example, say the constraints are $Q_\epsilon((-\infty,k_i)) = P((-\infty,k_i))$ for $i = 1,2$, with $k_1 < k_2$. This implicitly determines $a(\epsilon)$ and $b(\epsilon)$, and leads to

$$\begin{aligned} a^*(z) &= \frac{1}{k_2-k_1}\left[\frac{k_2 I_{(-\infty,k_1)}(z)}{p(k_1)} - \frac{k_1 I_{(-\infty,k_2)}(z)}{p(k_2)}\right], \\ b^*(z) &= \frac{1}{k_2-k_1}\left[\frac{I_{(-\infty,k_1)}(z)}{p(k_1)} - \frac{I_{(-\infty,k_2)}(z)}{p(k_2)}\right]. \end{aligned}$$

The analogous expressions for moment constraints are similarly determined.

In our example, constraints are introduced to preserve the quartiles of the prior on α, so as to preserve the interpretation of g_3 and g_4 as interquartile ranges. Thus $k_1 = \Phi^{-1}(0.25)$ and $k_2 = \Phi^{-1}(0.75)$. More precisely, the aim is to simultaneously perturb the prior on each α_i in the same direction, while preserving the quartiles. This addresses sensitivity to different specifications of the random effect distribution rather than sensitivity to outlying random effects. In turns out that in this case the influence functions add; that is the influence to misspecification is the sum of the influence to perturbations of the prior components individually. In the unconstrained setting, this is noted by Sivaganesan (1993), and made rigorous in a Fréchet derivative sense by Gustafson (1996a). To be clear, in the present example the influence function for perturbation of the random effects distribution is given by

$$IF_\alpha(z) \;=\; \left(\sum_i IF_{\alpha_i}(z)\right) + Cov^x\left(g(\theta), \sum_i \alpha_i\right) a^*(z) +$$

$$Cov^x\left(g(\theta), \sum_i \alpha_i^2\right) b^*(z), \tag{9}$$

where IF_{α_i} is the (unconstrained) influence function for perturbation of the α_i prior marginal only. These "overall" influence functions are plotted in the last row of Figure 1. The jaggedness observed in the plots is due to the sum of contributions from the IF_{α_i}, as well as the discontinuities in the integrands of the expressions for $a'(0)$ and $b'(0)$. The derivative norms in the last row of Table 3 are large compared to the individual marginal counterparts, suggesting that potentially misspecification is more influential than a single outlying random effect.

5. Asymptotic aspects. The local approach to sensitivity in hierarchical models can be applied in the large–sample limit. The hope is that this can shed some light on the asymptotic behaviour of the posterior distribution when the first stage of the prior is misspecified. Neuhaus, Hauck and Kalbfleisch (1992) and Neuhaus, Kalbfleisch and Hauck (1994) investigate this issue in some specific models, using different methods.

Consider modelling data $\{X_i\}_{i=1}^n$ as independent given the first–stage parameters $\{\lambda_i\}_{i=1}^n$, with densities $\{p(x_i|\lambda_i)\}_{i=1}^n$. This arises naturally when λ_i is a random effect acting on the distribution of observable X_i. In turn the random effects are postulated to be independent and identically distributed from a distribution belonging to the parametric family F_γ. Thus upon integrating out the random effects, the distribution of the data depends only on the parameter vector γ. The components of $\gamma = (\mu, \nu)$ are assumed to be the mean and variance of the random effect or mixing distribution.

In the large n case, what happens when in reality the random effect distribution does not belong to F_γ? This question is addressed in the special case of nominal first–stage prior that is conjugate to the model specification, with the true first–stage prior obtained as an ϵ-contamination of the nominal prior.

Let $\gamma^* = (\mu^*, \nu^*)$ be the true mean and variance of the mixing distribution. Assume that rather than arising from F_{γ^*}, the random effects are independent and identically distributed with distribution function

$$G_\epsilon(\cdot) = (1-\epsilon)F_{\gamma^*}(b(\epsilon)(\cdot - a(\epsilon))) + \epsilon G(b(\epsilon)(\cdot - a(\epsilon))),$$

where $a(\epsilon)$ and $b(\epsilon)$ are chosen to yield mean μ^* and variance ν^*, along the lines of the previous section. Now let $\gamma(\epsilon)$ be the value of γ which minimizes the Kullback-Leibler divergence between the assumed and underlying models for a single X_i, given by:

$$\int \log\left\{\frac{\int p(x|\lambda)dG_\epsilon(\lambda)}{\int p(x|\lambda)dF_\gamma(\lambda)}\right\} \int p(x|\lambda)dG_\epsilon(\lambda)\, dx. \tag{10}$$

Under weak conditions the posterior distribution on γ, as well as the maximum likelihood estimator for γ, will converge to $\gamma(\epsilon)$ as n increases to infinity (White, 1982). Note that $\gamma(0) = \gamma^*$. It is typically hard to compute $\gamma(\epsilon)$ since the random effects cannot be integrated out analytically, unless $\epsilon = 0$. However, it is feasible to compute the directional derivative of $\gamma(\epsilon)$ evaluated at $\epsilon = 0$, and give an influence function representation.

First, the derivative of (10) with respect to γ evaluated at γ^* is equated to zero. Then differentiation with respect to ϵ leads to the following:

RESULT 3.

$$\gamma_i'(0) = \int \left[I^{-1}(\gamma^*)b(\gamma^*, z)\right]_i d[G - F_{\gamma^*}](z), \tag{11}$$

where

$$\begin{aligned} b_i(\gamma^*, z) = {} & E\left[l_i(\gamma^*; X) \mid \lambda = z\right] + \\ & E_{\gamma^*}\left[l_i(\gamma^*; X) f^*_{\gamma^*}(\lambda)\right] \left(\frac{\mu^*}{2\nu^*}z^2 - \frac{\nu^* + (\mu^*)^2}{\nu^*}z\right) - \\ & E_{\gamma^*}\left[l_i(\gamma^*; X)\lambda f^*_{\gamma^*}(\lambda)\right] \left(\frac{1}{2\nu^*}z^2 - \frac{\mu^*}{\nu^*}z\right). \end{aligned} \tag{12}$$

In the above $l(\gamma; X) = \int p(X|\lambda) dF_\gamma(\lambda)$ is the nominal log likelihood function for a single X_i, with l_i and l_{ij} indicating partial first and second derivatives of l with respect to the components of γ. Furthermore, I is the usual information matrix, with $I_{ij}(\gamma) = -E_\gamma(l_{ij}(\gamma; X))$, and, as before, $f^*(\lambda) = -f'(\lambda)/f(\lambda)$. The integrand in (11) can be viewed as an asymptotic influence function which measures the robustness of the limiting inference to the misspecification of the random effect distribution.

A variant of this asymptotic influence function, without mean and variance constraints, is investigated for some common models in Gustafson (1996c). In particular, in the normal-normal, beta-binomial, and gamma-Poisson cases, the effect of misspecification is found to be small in general. As a further example, consider the case of X_{ij} being *iid* exponential variates ($j = 1, \ldots, m$) with common hazard λ_i. Nominally the random effects $\lambda_1, \lambda_2, \ldots$ are assumed to arise from a gamma distribution with mean μ^* and variance ν^*. With some effort the asymptotic influence function can be computed in this case (numerical integration is required for the first term in (12), but all other required quantities can be evaluated analytically).

Asymptotic influence functions are plotted for $\mu^* = 10$ and $\nu^* = 5$, $\nu^* = 10$, and $\nu^* = 20$ in Figure 2, for each of $m = 5$, $m = 25$, and $m = 125$. The three values of ν^* correspond to random effects which explain 5%, 10%,

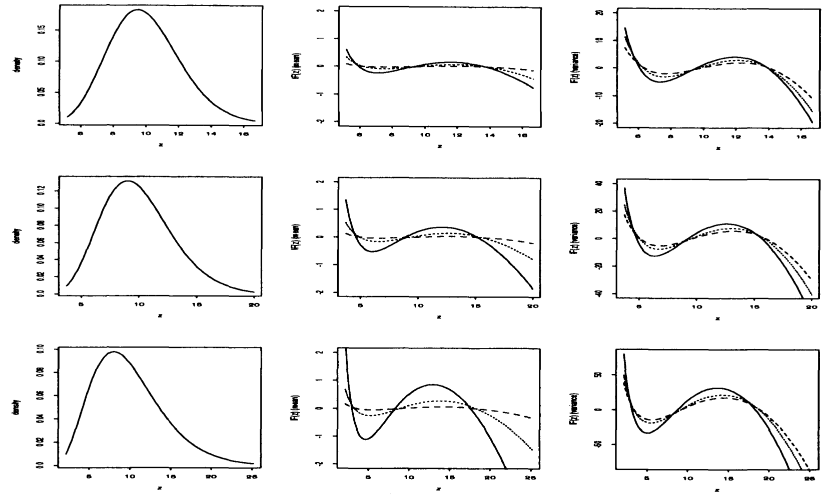

Figure 2: *Asymptotic influence functions in the gamma-exponential example. The rows correspond to true parameter values* $(\mu, \nu) = (10, 5)$, $(\mu, \nu) = (10, 10)$, *and* $(\mu, \nu) = (10, 20)$. *The first column gives the nominal random effect density. The second and third columns gives influence functions for the mean and variance respectively. The solid line corresponds to* $m = 5$, *the dotted line to* $m = 25$, *and the dashed line to* $m = 125$.

and 20% of the variance of an observation respectively. Note that the scale of the plots is fixed across values of ν^* for the influence on inference about μ, but scales with ν^* for the influence on inference about ν. Thus the plots suggest that inference about the random effect variance is equally robust across values of the true variance, as measured by the relative change in the limiting posterior distribution. In contrast inference about the mean becomes slightly less robust as the true variance increases. Clearly inferences become more robust to misspecification as the within cluster sample size m increases. This is more noticeable for the mean than the variance.

6. Discussion. Infinitesimal perturbations at various stages of a hierarchical prior are useful in assessing sensitivity quantitatively, qualitatively, and asymptotically. This amounts to a diagnostic approach to robustness in hierarchical models. There would be less need for diagnostics if inherently robust procedures could be employed. As mentioned earlier, Escobar (1994) pursues this approach by using a Dirichlet process as the first-stage prior distribution for a collection of normal means. If smoother priors are desired, one might consider using a mixture distribution, with the mixing distribution taken to be a Dirichlet process. This would mimic the approach to Bayesian robustness advocated by Bose (1994).

A potential shortcoming of the diagnostic approach is that the degree to which perturbed specifications are supported by the data is not taken into account. This is desirable for the subjective part of a hierarchical prior, but not for the structural part. Sivaganesan and Berger (1993) cite this as a reason why the range of inference over a class of first-stage prior distributions may not be of interest. Also, Weiss (1996) emphasizes this point in discussing non-hierarchical influence diagnostics. Unfortunately it seems difficult to incorporate the degree of data support into local diagnostics while retaining the simplicity of implementation. In particular, thought is required as to how to extend the higher–stage prior to reflect the plausibility of various first–stage perturbations.

REFERENCES

ALBERT, J., AND CHIB, S. (1994). Bayesian model diagnostics in hierarchical models, Preprint.

ANGERS, J.F., AND BERGER, J.O. (1991). Robust hierarchical Bayes estimation of exchangeable means. *Canad. J. Statist.* **19** 39-56.

BEITLER, P.J., AND LANDIS, J.R. (1985). A mixed-effects model for categorical data. *Biometrics* **41** 991-1000.

BOSE, S. (1994). Bayesian robustness with mixture classes of priors. *Ann. Statist.* **22** 652-667.

CANO, J.A. (1993). Robustness of the posterior mean in normal hierarchical models. *Comm. Stat. A* **22** 1999-2014.

CARLIN, B.P., AND POLSON, N.G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) 577-586. Oxford Univ. Press, London.

CHEN, M.H. (1994). Importance weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89** 818-824.

ESCOBAR, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268-277.

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

GOEL, P.K. AND DEGROOT, M.H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76** 140-147.

GOOD, I.J. (1980). Some history of the hierarchical Bayesian methodology (with discussion). in *Bayesian Statistics 2* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds.) 489-504. Oxford Univ. Press, London.

GUSTAFSON, P. (1996a). Local sensitivity of inferences to prior marginals. *J. Amer. Statist. Assoc.* **91** 774-781.

GUSTAFSON, P. (1996b). Local sensitivity of posterior expectations. *Ann. Statist.* **24** 174-195.

GUSTAFSON, P. (1996c). The effect of mixing distribution misspecification in conjugate mixture models. *Canad. J. Statist.* To appear.

HAITOVSKY, Y. AND ZIDEK, J.V. (1986). Approximating hierarchical normal priors using a vague component. *J. Multivariate Analysis* **19** 48-66.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. AND STAHEL, W.A. (1986). *Robust statistics: the approach based on influence functions.* Wiley, New York.

LINDLEY, D.V. AND SMITH, A.F.M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc.*, **34** 1-41 (with discussion).

NEUHAUS, J.M., HAUCK, W.W. AND KALBFLEISCH, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed–effects logistic models. *Biometrika* **79** 755-762.

NEUHAUS, J.M., KALBFLEISCH, J.D. AND HAUCK, W.W. (1994). Conditions for consistent estimation in mixed–effects models for binary matched–pairs data. *Canad. J. Statist.* **22** 139-148.

PERICCHI, L.R. AND NAZARET, W.A. (1988). On being imprecise at the higher levels of a hierarchical linear model. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith eds.) 361-375. Oxford Univ. Press, London.

POLASEK, W. AND PÖTZELBERGER, K. (1988). Robust Bayesian analysis in hierarchical models. In *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith eds.) 377-394. Oxford Univ. Press, Oxford.

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.

SIVAGANESAN, S. AND BERGER, J. (1993). Robust Bayesian analysis of the binomial empirical Bayes problem. *Canad. J. Statist.* **21** 107-119.

WEISS, R. (1996). An approach to Bayesian sensitivity analysis. *J. Roy. Statist. Soc. B.* To appear.

WHITE, H.A. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1-25.

UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS
2021 WEST MALL
VANCOUVER, BC V6T 1Z2 CANADA

Aspects Of Bayesian Robustness In Hierarchical Models

discussion by
SUDIP BOSE
The George Washington University

First of all, let me congratulate Professor Gustafson on an excellent paper. He has examined very thoroughly, Bayesian robustness of several features of hierarchical models. He has, for the most part, focused on local sensitivity. Before discussing specific sections of the paper, let me make a few remarks about local sensitivity vs. global sensitivity (robustness). The existing work on global robustness has mostly concentrated on determining ranges of posterior expectations of real parametric functions as the prior is allowed to vary over a neighborhood class. Special techniques have been presented, each typically pertaining to a specific type of neighborhood class. In some sense, there is no universal method - the linearization algorithm is probably the closest there is. On the other hand, the local sensitivity approach is almost by definition, a universal approach. Another advantage of the local sensitivity approach is that it can better deal with additional imposed constraints - something that is often very difficult with global robustness. Global robustness appears to have the advantage of greater ease of interpretation - the range of posterior expectations may be precisely the set of possible answers to the inference question, or the possible answers may be derived very simply from it. With local sensitivity, following Hampel, Ronchetti, Rousseeuw and Stahel (1986), one can interpret the value $IF_P(z)$ as the effect of an infinitesimal contamination at the point z on the posterior expectation, standardized by the mass of the contamination. Also, the gross-error sensitivity

$$\gamma = \sup_z |IF_P(z)|$$

measures the worst possible influence of a small amount of contamination of fixed size. Even when one has determined these quantities, whether they are 'large' or 'small' requires extra effort to judge. This is not meant as a criticism of local sensitivity and the influence function; rather to point out that we probably need to provide guidelines to practitioners to help them interpret, and appreciate the worth of, our sensitivity calculations.

In the rest of the discussion, the title of a section precedes the comments pertaining to it.

Qualitative aspects

It is reassuring to see that if one holds the stage of perturbation, i fixed, the effect (of the same perturbation) increases as the stage of inference, j moves away from i in either direction. It is worth noting that this result is different from the other results in that the effect is not just a local phenomenon.

For the case of j, the inference stage fixed, and i, the perturbation stage allowed to vary, the result about the normal hierarchical model is a bit of a surprise to me. The fact that $||\dot{T}(P)||$ should depend on the stage of perturbation only through the variance σ_i^2 seems somewhat counter-intuitive. One possible explanation is that the usual intuition regarding hierarchical models does not apply since all the parameters θ_i are one-dimensional. This is in contrast to the usual case where the dimension of θ_i decreases as i increases, or in other words one has fewer parameters at the higher stages of the hierarchy. It is also noted that $||\dot{T}(P)||$ increases with σ_i^2 - it makes me wonder whether this is true under other forms of perturbation. In particular it may be instructive to see what happens in the following two cases.

(i) Let $\epsilon_i \sim \sigma_i^{-1} f(\frac{\epsilon_i}{\sigma_i})$ and allow f to vary in the neighborhood of $N(0,1)$.

(ii) Let the distribution of ϵ_i vary in a neighborhood of $N(0, \sigma_i^2)$ with the mean and variance constrained to be 0 and σ_i^2, respectively.

Quantitative aspects

The author uses the representation

$$q_\epsilon(.) = b(\epsilon)[(1-\epsilon)p\{b(\epsilon)(. - a(\epsilon))\} + \epsilon q\{b(\epsilon)(. - a(\epsilon))\}]$$

in expression (7), where $a(\epsilon)$ and $b(\epsilon)$ are implicitly defined so as to ensure that the resulting measures satisfy the constraints. What is the motivation for choosing this particular representation? Is it somehow natural? Also how does the author propose to extend this to deal with more than two constraints?

In the multicentre clinical trial example that the author has analyzed, it is interesting to see that the influence functions of the α_i-s add. I wonder, just how general is this additivity of influence functions?

REJOINDER

Paul Gustafson

I would like to thank Professor Bose for his comments. I agree that there is a need to calibrate local sensitivity measures. McCulloch (1989) considered calibration of one particular local sensitivity measure, via a coin–tossing analogy. As well, the calibration issue was addressed by S. Sivaganesan in a talk at this conference.

With regard to the Qualitative aspects, the point about dimensionality is well taken. It would be of some interest to analyze the sensitivity over a simple "tree–like" normal hierarchy, where the number of parameters decreases with the stage of the hierarchy. In the one–dimensional case, the class (i) suggested by Professor Bose will yield exactly the same results as in the paper. This is because the underlying metric used is invariant to scale transformations. That is starting from a fixed distribution, we can either make a scale transformation and then form a class of size ϵ under the metric, or we can form the class first, and make the scale transformation to each member of the class. Either way, the resulting class is the same. It is not known if the constrained class (ii) gives the same sensitivity orderings as the unconstrained class. While the sensitivity with respect to (ii) can be computed numerically, it is hard to determine if orderings apply to all possible data sets and all possible nominal variances.

The location–scale transformation is used to adjust for constraints because it is simple, and because it is common to have two constraints on a prior distribution. There is no obvious way to proceed with an influence function representation if three or more constraints are present.

The additivity of influence functions is quite general. It manifests itself whenever a density giving rise to *iid* parameters or observations is perturbed.

Additional References

McCulloch, R.E. (1989). Local model influence. *J. Amer. Statist. Assoc.* **84**, 473-478.