Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

SUFFICIENCY AND INFLUENCE[1]

BY ROBERT WEISS
University of California, Los Angeles

Consider two models M_1 and M_2 proposed as models for the same data. Assume that the conclusions from both models are posteriors $p_1(\theta|Y)$ and $p_2(\theta|Y)$ of some inferential target θ given the data Y. The unknowns θ may be parameters with identical interpretations in both models. Under mild conditions, the difference between the two conclusions is reducible to a one dimensional summary $h(\theta)$ for any two models. The result has implications for Bayesian diagnostics and sensitivity analysis. Applications of influence sufficiency to case and prior influence are illustrated, with emphasis on the influence of different priors and calculation of Bayes factors.

1. Introduction. I start with a brief discussion of influential and outlying observations, influential and unsupported assumptions and Bayesian robustness.

An observation is influential if the conclusion changes in an important manner when the observation's likelihood contribution changes. In contrast, outliers are observations whose responses differ from what is predicted by the model. Outliers need not be influential (consider either a model with t errors or regression through the origin) and influential observations need not be outliers; consider an observation in linear regression with leverage $h_i \approx 1$. As the number of observations increases, with all other aspects of the model held fixed, we can usually expect the influence of individual observations to become small.

Observations are not the only things in an analysis that can be influential. Assumptions such as linearity, normality, constant variance or smoothness contribute strongly to the likelihood. An assumption is influential if the conclusion changes when the assumption is relaxed. An assumption is outlying if the data support a relaxation of the assumption; usually we call this an unsupported assumption.

Classically, robustness has meant that within a range around a particular model, the derivatives of an inference, usually narrowly defined as a point estimate, with respect to various inputs are 'small'. Usually only the response variables are considered as inputs. From a Bayesian perspective, the classical definition of robustness can be encorporated into an analysis through choice of robust likelihoods (for example, Ramsay and Novick 1980). This is done through a priori beliefs and a posteriori data selection amongst models for the sampling density and not a blind requirement that models should be robust in the classical sense.

[1]Research supported by NIH grant GM50011.
Key Words and Phrases: Bayesian Data Analysis, Bayes Factors, Case Deletion, Diagnostics, L_1 norm, Prior Sensitivity.

Bayesian robustness has historically studied the sensitivity of a point estimate, Bayes factor or posterior to the prior. We take sets of priors and investigate the possible ranges of the posterior or point estimate. This is beginning to change, as researchers (for example, see Lavine 1991a, 1991b) realize that the prior is often not the main source of information in a model, rather, the likelihood contributes substantially more to conclusions. Another problem with Bayesian robustness is that the perturbation sets are not necessarily chosen for their a priori plausibility and support by the data. If the most influential priors are not plausible a posteriori and usually they are not, then it is unclear if we should be interested in their influence a posteriori. Further, in regression and other complex models with available substantive information, prior influence has not been extensively studied. Two examples are Carlin, Kass, Lerch and Huguenard (1992) and Greenhouse and Wasserman (1996). The example in this paper illustrates aspects of prior influence in a linear regression setup.

A Bayesian jointly models data Y and unknown parameters θ by model M_1 which specifies $p_1(\theta, Y) = p_1(\theta)f_1(Y|\theta) = p_1(\theta|Y)f_1(Y)$. Alternatively, a competing model M_2 is proposed with joint prior $p_2(\theta, Y)$ and corresponding prior $p_2(\theta)$, sampling $f_2(Y|\theta)$, posterior $p_2(\theta|Y)$ and prior predictive $f_2(Y)$ distributions. Typically, models include covariates X, but dependence of Y on X is suppressed for convenience. For influence assessment, the parameter θ must have the same interpretation under both models. Here I assume that M_1 is the current model under consideration, while model M_2 is a modified version of M_1, the result of a change in assumptions. Which model should be preferred? Model M_1 will be used for inference if the sensitivity analysis identifies no problems with the current model. Several means of choosing between the two models are possible. Influence analysis advises on whether the conclusion of our analysis changes when the assumptions change. Model M_2 must be considered only if it leads to different conclusions from M_1, and if the data support M_2 over M_1 (Weiss 1996).

The next section describes an idea in influence analysis which is both pretty and powerful. A particular value of the idea is its use in developing procedures for assessing posterior influence and data support and in providing computational algorithms. Computational results are emphasized, and new methodology for calculating the Bayes factor and L_1 divergence between M_1 and M_2 is given.

2. Sufficient Perturbation Functions. While the base material for this section is Weiss (1996), different aspects are emphasized, and several new methods are presented, especially in section 2.4. Consider the joint priors $p_1(\theta, Y)$ and $p_2(\theta, Y)$, the assumptions of the analysis given models

M_1 and M_2 respectively. Define

$$\tau^*(\theta, Y) = \frac{p_2(\theta, Y)}{p_1(\theta, Y)} \tag{1}$$

The function $\tau^* = \tau^*(\theta, Y)$ mathematically embodies the change of assumptions in going from M_1 to M_2. Thinking of M_2 as a modification or perturbation of M_1, then τ^* is the perturbation function that changes $p_1(\theta, Y)$ into $p_2(\theta, Y)$. Kass, Tierney and Kadane (1989) introduced the idea of perturbation functions and Weiss (1996) has expanded on the idea.

A posteriori we have

$$\tau(\theta, Y) = \frac{\tau^*(\theta, Y)}{\mathrm{E}_1[\tau^*(\theta, Y)|Y]} = \frac{p_2(\theta|Y)}{p_1(\theta|Y)} \tag{2}$$

where $\mathrm{E}_j[g(\theta)|Y] = \int g(\theta) p_j(\theta|Y) d\theta$, for $j = 1, 2$, and

$$\mathrm{E}_1[\tau^*(\theta, Y)|Y] = B_{21} = \frac{f_2(Y)}{f_1(Y)} \tag{3}$$

is the Bayes factor in favor of M_2 against M_1, assuming that it is well defined. If $\theta = (\theta_1, \theta_2)$, $p_j(\theta) = p_j(\theta_1) p_j(\theta_2)$, and $p_j(\theta_1)$ are both proper, then if $p_1(\theta_2) = p_2(\theta_2)$, even if improper, the Bayes factor is well defined as the Bayes factor where $p_j(\theta_2)$ are equal and proper but vague in the sense that the posteriors $p_j(\theta|Y)$ do not change.

2.1. *Examples of perturbation functions.*

2.1.1. *Case deletion.* What happens when we delete case i from the sample? The perturbation function $\tau^*_{1i} = f_{2i}(y_i) * [f(y_i|\theta, x_i)]^{-1}$ is proportional to the inverse sampling distribution of y_i given the parameters θ and covariates x_i and where $f_{2i}(y_i)$ is a new sampling density for y_i which does not depend on θ.

2.1.2. *Prior Perturbation.* Changing the prior from $p_1(\theta)$ to $p_2(\theta)$ gives $\tau^*_p = p_2(\theta)/p_1(\theta)$.

2.1.3. *Likelihood perturbation.* Changing the sampling density from $f_1(Y|\theta)$ to $f_2(Y|\theta)$ gives $\tau^*_L = f_2(Y|\theta)/f_1(Y|\theta)$.

2.1.4. *Covariate and response perturbation.* We can also perturb responses y_i to $y_i + \delta_i$ or perturb covariates x_i or sets of responses or covariates. Changing y_i is useful when the observations are uncertain in ways not accounted for by the sampling density. This happens with rounding, if imputed values are substituted for actual observed values or if censoring times are substituted for unobserved responses. The presence of errors in variables can suggest that the x_i's should be perturbed.

2.1.5. *Combinations.* If a set of individual perturbations are of interest, it is probable that combinations of the perturbations are also of interest. Virtually no work has occurred in this area. It appears to be an area where the tools of experimental design could usefully be applied; as with multiple case influence, combinatorial explosion problems can occur.

2.2. *Sufficiency.* If M_1 and M_2 both provide joint distributions for data Y and parameters θ, then model M_2 can be arrived at as a perturbation of M_1, provided only that the ratio $\tau^*(\theta, Y)$ is finite everywhere. That is, the joint prior of M_2 should have a density relative to the joint prior of M_1.

The perturbation τ or τ^* contains in a real sense all of the influence of the change from M_1 to M_2. In particular, given τ, there is no further influence on θ due to switching from M_1 to M_2. To see this, change variables from θ to τ, ρ where ρ is chosen to make the change of variables one-to-one and measurable. Then (Weiss 1996)

$$p_1(\rho|\tau, Y) = p_2(\rho|\tau, Y) \tag{4}$$

The posterior of ρ given τ is the same under both models M_1 and M_2. The proof is

$$\tau = \frac{p_2(\theta|Y)}{p_1(\theta|Y)} = \frac{p_2(\tau|Y)p_2(\rho|\tau, Y)}{p_1(\tau|Y)p_1(\rho|\tau, Y)}. \tag{5}$$

Now in (5) multiply leftmost and rightmost formulae by $p_1(\rho|\tau, Y)$ and integrate with respect to ρ. This gives

$$\tau = \frac{p_2(\tau|Y)}{p_1(\tau|Y)}. \tag{6}$$

Dividing (5) by equation (6) gives (4).

The updating of $p_1(\tau|Y)$ to $p_2(\tau|Y)$ is simple, since by (6), the ratio is proportional to τ. Thus, given a histogram of samples from $p_1(\tau|Y)$, or a plot of the density, we have a substantial amount of information about the effects of the perturbation. The beauty of the perturbation function is that τ is a univariate function of θ: by investigating a univariate marginal $p_1(\tau|Y)$ of $p_1(\theta|Y)$, many of the consequences of perturbing M_1 to M_2 can be explored; in particular we need not explore the high dimensional posteriors $p_1(\theta|Y)$ and $p_2(\theta|Y)$, since all conditionals $p_j(\rho|\tau, Y)$ are equal.

The function $\tau = \tau(\theta, Y)$, generally a function of the data and the parameters, can be called a *sufficient perturbation function* due to the results above. Conditional on τ, there is no further influence due to changing from M_1 to M_2. Unconditionally, further influence of the perturbation on a function $\beta = \beta(\theta)$ of interest is due to the influence on $p_1(\tau|Y)$ and any posterior association between τ and β. If τ is a function of β, then β is also a sufficient perturbation function for the perturbation from M_1 to M_2. We distinguish

β from τ by calling τ, or any 1-1 measurable function of τ, a *minimally sufficient perturbation.*

2.3. *Summarizing Influence.* Two different approaches for summarizing influence are numerical and graphical. The numerical approach summarizes the differences between $p_1(\theta|Y)$ and $p_2(\theta|Y)$ by a numerical summary of the differences. Discussions of these summaries often revolve around which summary is best, but a scalar summary is not required, and multiple summaries should be considered. On the other hand, most analysts don't want to wade through tons of influence statistics.

One approach is to summarize the influence through the change in posterior expectation of some quantity of interest. Another popular approach for summarizing the difference between $p_1(\beta(\theta)|Y)$ and $p_2(\beta(\theta)|Y)$ is a divergence measure

$$D_{\beta(\theta)}(g) = \int g\left(\frac{p_2(\beta(\theta)|Y)}{p_1(\beta(\theta)|Y)}\right) p_1(\beta(\theta)|Y)d\theta,$$

where $g(a)$ is convex and $g(1) = 0$. (See Csiszár 1967, Weiss and Cook 1992, and Weiss 1996.) By (5) and (6),

$$D_\theta(g) = D_\tau(g), \tag{7}$$

and by convexity of g,

$$D_\theta(g) \geq D_{\beta(\theta)}(g) \geq 0 \tag{8}$$

(Weiss 1996). Various measures that have been proposed are the L_1 norm with $g_L(a) = .5|a - 1|$, the several Kullback divergences, for example K with $g_K(a) = -a\log(a)$; functions of Hellinger distance (Geisser 1993) with $g_p(a) = |a^{1/p} - 1|^p$, $p \geq 1$ and the χ^2 divergence with $g_{\chi^2}(a) = (a-1)^2$. In my experience, the choice of divergence does not matter for ranking different perturbations especially in regards to case deletion, however, some divergences are easier to interpret. ¿From Weiss (1996), χ^2 divergence is the square of the Kass, Tierney, and Kadane (1989) maximum standardized change (MSC) and the L_1 is the maximum difference between M_1 and M_2 in posterior probability content of any interval. Kullback divergence with $g_K(a)$ is often recommended because closed form computations are sometimes possible, and because of the optimality identified by Bernardo (1979; 1985).

A graphical approach to influence summarization is a compact way of displaying many numerical influence statistics. The primary goal of influence analysis is to understand the difference between $p_1(\theta|Y)$ and $p_2(\theta|Y)$. A graphical approach plots these two posteriors and inspects them directly. When θ is a scalar, this is straightforward. When $p_j(\theta|Y)$ is more than 1 or 2 dimensional, this is hard. One way of easing the problem is to inspect

marginal posteriors, but by (8) these underestimate global influence, often drastically. Define a posterior influence plot as a plot of $p_1(\theta|Y)$ and $p_2(\theta|Y)$ or of the marginals $p_j(\beta|Y)$ of β. The sufficient perturbation comes into play here. Consider inspecting a plot of $p_1(\tau|Y)$ and $p_2(\tau|Y)$. If we consider that influence is properly summarized by a divergence measure, then this plot loses nothing over inspecting $p_1(\theta|Y)$ and $p_2(\theta|Y)$ since the $D_\tau(g) = D_\theta(g)$. If β is a non-minimal sufficient perturbation function, we can also inspect $p_1(\beta|Y)$ and $p_2(\beta|Y)$ without missing any influence. If β is easier than τ to interpret, its posterior influence plot may be preferable to the posterior influence plot using τ. When a not necessarily sufficient parameter β is of particular interest in an analysis, then one should inspect $p_1(\beta|Y)$ and $p_2(\beta|Y)$ to investigate influence. This also obviates the need for chosing a summary influence statistic.

2.4. *Computations.* For a single sample $\theta^{(l)}$ with $l = 1, \ldots, L$ from $p_1(\theta|Y)$ one can use

$$\hat{B}_{21} = \hat{\mathrm{E}}_1[\tau^*(\theta, Y)] = L^{-1} \sum \tau^*(\theta^{(l)}, Y), \tag{9}$$

by (3) to estimate B_{21} (Weiss 1996) and to calculate $D_\theta(g)$,

$$\hat{D}_\theta(g) = L^{-1} \sum_{l=1}^{L} g\left(\frac{\tau^*(\theta^{(l)}, Y)}{\hat{B}_{21}} \right), \tag{10}$$

by (7) to estimate influence functions of the perturbation from M_1 to M_2 (Weiss 1996). To calculate the change in posterior expectation of β, one can use

$$\mathrm{E}_2[\beta|Y] - \mathrm{E}_1[\beta|Y] = \mathrm{Cov}_1[\beta, \tau|Y], \tag{11}$$

where $\mathrm{Cov}_1[\beta, \tau|Y]$ is M_1's posterior covariance between β and τ. In principle, one can produce a posterior influence plot of $p_1(\tau|Y)$ and $p_2(\tau|Y)$ by approximating a sample from $p_2(\tau|Y)$ by reweighting $\theta^{(l)}$ by $\tau(\theta^{(l)})$. This and the results (9), (10), and (11) are importance sampling-type calculations. Samples from one density ($p_1(\theta|Y)$) are used to learn about a second density ($p_2(\theta|Y)$). Importance sampling can be used in theory to explore any posterior with the same support as $p_1(\theta|Y)$, provided that one can integrate arbitrary functions of θ given M_1. In practice, importance sampling generally only works if the importance density $p_1(\theta|Y)$ is close to the alternative density, $p_2(\theta|Y)$ and if special conditions are met by the density ratio τ. Influential perturbations are unlikely to meet these special conditions and the caveat of integrating arbitrary functions is very strong and virtually never met in practice.

The above calculations are the direct result of sufficiency. An alternate application of sufficiency uses samples from both $p_j(\theta|Y)$ is to form

estimates $\hat{p}_j(\tau^*|Y)$ using kernel density or other form of semiparametric estimate. These two posteriors are one dimensional, and the curse of dimensionality should be avoidable. One can then compute influence diagnostics numerically using one dimensional Riemann integration $d\tau$ applying (7). Assuming proper priors one can compute B_{21} using

$$\hat{B}_{21} = \frac{\hat{p}_1(\tau^*|Y)\tau^*}{\hat{p}_2(\tau^*|Y)}. \tag{12}$$

Any value of τ^* can be used, provided that both densities are accurately estimated at that point. The log scale is often easier to work with. Since the Jacobian cancels, the posterior of any monotone transformation of τ^* could be used in place of $p_j(\tau^*|Y)$. The difference in posterior expectations $\mathrm{E}_2[\beta|Y] - \mathrm{E}_1[\beta|Y]$ is estimated by separately estimating the expectations using the two samples.

Inspection of the densities $\hat{p}_j(\tau^*|Y)$ can shed light on the accuracy of these various calculational formulae, as illustrated in the example.

3. Housing Data. This data set was collected to help predict the *COST* to rehabilitate housing in St. Paul, Minnesota, USA. It is desired to estimate the cost to rehabilitate housing in all of St. Paul, in individual census tracts, to compare different census tracts this year and to compare housing stock to many years ago. The prediction is to be based on the average ratings of external parts of the house, *EAVES*, *WINDOWS* and *YARD* by three sidewalk surveyors. Ratings of *EAVES* and *WINDOWS* are integer valued from 1 to 6, and *YARD* is rated either 2 or 5. Lower ratings indicate houses in better condition. The data are given in table 1. The *COST* in kilodollars is estimated by a building contractor who must enter the house. Generally this is problematic, as it involves getting permission of the homeowner and getting the contractor to the house when the homeowner is present. In contrast, sidewalk surveyors' work takes only a few minutes, their time is inexpensive, and the survey is nonintrusive. The data collection instrument is designed to deliver consistency in ratings, but differences do occur among raters, so multiple raters are sent to each house.

Let *COST* y_i be modeled $y_i = x_i^t\beta + \epsilon_i$ with x_i a 4−vector of covariates, a one followed by the average *EAVES*, *WINDOWS*, and *YARD* ratings from the three raters, coefficients $\beta = (\beta_j)$, $j = 0, 1, 2, 3$ and errors $\epsilon_i|\sigma^2 \sim N(0, \sigma^2)$ given σ^2 a priori independent and identically distributed (iid).

In the original modeling, a least squares approach was used after model selection supplemented by some case deletion diagnostics. The current model with three predictors contains fewer predictors than are available. The author desired to do model selection to select an even more parsimonious model; the clients did not understand model selection, and did not like the results,

Table 1: *Housing data; The COST is in 1000's of dollars. The EAVES and WINDOWS ratings are truncated at 2 decimal points, but all computations carried 5 digits.*

Point #	*COST*	*EAVES*	*WINDOWS*	*YARD*
1	15.783	3.00	2.00	2
2	12.570	1.66	2.33	3
3	19.600	3.33	2.33	2
4	8.206	1.66	1.66	2
5	15.333	2.33	2.33	5
6	14.955	5.00	3.00	2
7	13.710	4.33	3.00	2
8	11.388	2.33	2.33	3
9	4.802	1.33	1.66	2
10	12.547	3.00	2.66	2
11	13.677	3.00	3.33	2
12	9.683	1.33	2.33	2
13	16.798	2.66	3.00	4
14	25.615	3.00	3.33	4
15	15.734	3.00	3.00	2
16	13.510	3.00	3.00	2
17	13.855	3.33	3.00	2
18	3.986	2.33	1.66	2
19	5.997	2.33	2.00	2
20	9.778	2.00	2.66	2
21	18.108	1.00	1.00	2
22	10.152	2.00	3.00	2

since it resulted in very few distinct predicted costs to rehabilitate homes. It was also desired to use the prior information available from a previous survey seven years earlier. The first analysis did not use the previous data due to time and computing constraints. In the end, least squares was used to estimate the parameters of a regression equation.

Three priors are used with this example; each represents a different perspective based on the previous discussion; a model selection prior, a flat prior, and an informative prior. All three have $p(\beta, \sigma^2) = p(\beta)p(\sigma^2)$ with $p(\sigma^2) \propto \sigma^{-2}$. The first prior is a version of the hierarchical model selection prior of George and McCulloch (1993). Consider the prior $p(\beta_j|\delta_j) \sim N(0, V_0 W^{\delta_j})$, $j = 1, 2, 3$, $\delta_j \sim \text{Bernoulli}(\pi_0)$. For the housing data set I selected $\pi_0 = .7$, $V_0 = .25$, and $W = 16$. Because of the scaling of the X's it was felt that the same prior might reasonably be used for all three coefficients. The variances of the mixture components suggest that the coefficients are less than 1 with prior probability .95 for variance $V_0 = .25$ and less than 4 with prior probability .95 for prior variance $V_0 W = 4$. A coefficient of 4 would mean that, holding everything else fixed, changing a covariate from 1 to 6 would change a fitted value by 20,000 dollars, which was felt to be an enormous amount. The prior for the intercept assumed $\beta_0 \sim N(0, V_0 W)$. This prior is called the model selection (MS) prior.

The second prior, the flat (F) prior, used a noninformative prior $p(\beta, \sigma) \propto c\sigma^{-2}$. The choice of constant c is important for calculating Bayes' factors, and I used the volume of the smallest rectangular region with sides parallel to the coordinate axes that covered all 2000 samples from the posterior, with side lengths rounded up slightly. Since all three priors for σ^2 are improper, the volume for σ^2 was not included. The appropriate normalizing constant was $(137024)^{-1}$.

The third prior was a proper informative (I) prior for β based on data taken 7 years previously. The prior data has sample size 39, where one outlier was deleted from the earlier analysis. The costs were inflated by a factor so that the means of the earlier sample (without case deletion) and the current sample are the same. The necessary statistics are given in table 2. The prior is the posterior t distribution $p(\beta|Y_{\text{old}})$ based on an uninformative prior with $p(\beta, \sigma^2) \propto \sigma^{-2}$.

All calculations were based on samples of size 2000 from the three posteriors based on the three priors. The posteriors will be named MS, F, and I posteriors after the priors.

4. Case Influence Analysis. This section discusses case diagnostics and the influence of the three priors on the case diagnostics. A priori, I expected that use of a proper prior would reduce case influence, and increase outlier statistics over a noninformative prior.

The sufficient perturbation for case deletion is independent of the prior.

In normal linear regression it is $\tau_i(\theta) = \tau_i(x_i^t\beta, \sigma^2) = \text{CPO}_i / f(y_i|\theta) =$

$$(2\pi\sigma^2)^{1/2} \exp\left(.5\sigma^{-2}(y_i - x_i^t\beta)^2\right) \text{CPO}_i$$

where $\text{CPO}_i = (E[(f(y_i|\beta, \sigma^2)^{-1}|Y])^{-1}$ is the conditional predictive ordinate (Geisser 1993, p. 108).

Table 3 gives three case statistics for each of the three priors. Column 1 is the case number, columns 2-4 give CPO_i, columns 5-7 give the L_1 divergence case statistic $.5 \int |p(\theta|Y) - p(\theta|Y_{(i)})| d\theta$ where $Y_{(i)}$ denotes the case deleted sample; and columns 8-10 give the diagnostic $P(|\epsilon_i| > 2\sigma|Y)$, the posterior probability that $|\epsilon_i|$ is larger than $2 * \sigma$, proposed as an outlier statistic by Chaloner and Brant (CB) (1988). A value of .00 indicates a number less than .01 after rounding, but not originally equal to 0. The diagnostics based on the MS posterior are in columns 2, 5, and 8; based on the F posterior in columns 3, 6, and 9; and based on the I posterior in columns 4, 7, and 10. As expected, the L_1 case influence statistics are smaller for the proper priors. However, the CB outlier statistics are smaller with the MS and I priors in 9 out of the 10 cases with a non-zero value in the flat prior. The exception is the most outlying case. The conditional predictive ordinates are comparable among the three models, except for the noticeable changes for the two most outlying cases, which are less outlying with the MS and I posteriors.

5. Prior Influence. There are three priors, so influence can be assessed in the context of switching between any two of them. I discuss marginal influence followed by the global affects of switching between priors and the data support of the priors.

There is less influence on individual parameters than on the full posterior by (8). The marginals based on the MS prior are particularly deceptive as summaries of the full multivariate posterior $p(\beta|Y, MS)$, since the multivariate posterior has lumps of probability close to coordinate axes and subspaces with some $\beta_j = 0$. Figure 1 shows the marginals of the intercept,

Table 2: *Prior mean $\hat{\beta}_0$, row 1, and X^tX matrix, rows 2-5, The prior residual sum of squares is 1674.3726, and sample size is 39.*

	intercept	*EAVES*	*WINDOWS*	*YARD*
$\hat{\beta}_0$	-3.3697	0.9522	2.4878	3.5082
X^tX	39.000	92.8333	93.6666	81.6666
	92.8333	270.9166	255.3888	219.9722
	93.6666	255.3888	255.4444	217.7222
	81.6666	219.9722	217.7222	202.6111

Table 3: *Case diagnostics,* CPO, L_1 *influence statistic, and Chaloner and Brant outlier statistic* $P(|\epsilon_i| > 2\sigma)$. *Results for given for the three posteriors MS model selection; F flat; and I informative. Posteriors are nested within diagnostic. A .00 indicates a value greater than 0 but less than .01 after rounding.*

	CPO			L_1			CB		
Model Case #	MS	F	I	MS	F	I	MS	F	I
1	.056	.055	.054	.10	.14	.11	0	0.00	0
2	.081	.083	.080	.08	.08	.08	0	0	0
3	.023	.022	.025	.20	.27	.19	0.09	0.12	0.07
4	.079	.077	.080	.07	.09	.07	0	0	0
5	.054	.033	.035	.19	.37	.29	0.01	0.05	0.04
6	.070	.059	.067	.12	.19	.12	0	0.01	0.00
7	.074	.070	.071	.10	.12	.09	0	0.00	0
8	.073	.071	.068	.08	.08	.08	0	0	0
9	.051	.046	.056	.10	.18	.09	0.00	0.01	0.00
10	.083	.084	.082	.08	.09	.07	0	0	0
11	.080	.080	.080	.08	.09	.08	0	0	0
12	.081	.080	.081	.08	.09	.07	0	0	0
13	.079	.076	.071	.09	.10	.10	0	0	0
14	.014	.0097	.029	.30	.49	.25	0.26	0.25	0.06
15	.069	.070	.074	.08	.10	.07	0	0	0
16	.082	.083	.082	.08	.08	.07	0	0	0
17	.082	.083	.082	.08	.08	.07	0	0	0
18	.026	.019	.029	.19	.29	.19	0.07	0.17	0.05
19	.044	.040	.046	.09	.15	.09	0.00	0.01	0.00
20	.077	.078	.076	.08	.09	.08	0	0	0
21	.0030	.00061	.0013	.45	.74	.58	0.72	0.69	.78
22	.074	.073	.071	.09	.11	.09	0	0	0

Table 4: *Influence on parameter posteriors, estimates, and sd's; joint influence on the posterior. The last three lines give estimated Bayes factors in favor of the second model against the first model using equation* (12) *for BF and equation* (9) *for BF-j, j* = 1, 2.

	L_1			mean			sd		
model(s) parameter	MS/F	MS/I	F/I	MS	F	I	MS	F	I
intercept	.48	.50	.35	.85	.19	-2.09	1.784	5.44	2.73
EAVES	.20	.05	.17	1.41	1.65	1.46	1.04	1.60	1.11
WINDOWS	.28	.19	.24	1.31	1.09	1.91	1.34	2.43	1.62
YARD	.17	.26	.20	2.09	2.41	2.83	0.99	1.37	.97
sigma	.19	.04	.17	4.69	5.16	4.72	.91	1.17	.86
mean cost	.10	.16	.21	12.1	12.3	12.4	.96	1.18	.75
posterior	.65	.57	.49						
BF	9.3	21000	2500						
BF-1	13.	21000	2300						
BF-2	9.0	21000	2800						

coefficients of *EAVES*, *WINDOWS* and *YARD*, and σ. The solid curves are the MS marginals, the dashed curves are the posterior marginals based on the I prior and the dotted curves are those based on the F marginals. As might be expected, the F marginals are less peaked and have more variance, and the MS marginals have bumps near zero. The plot labeled *Mean Cost* is the posterior estimated cost to rehabilitate an average house with *EAVES*, *WINDOWS* and *YARD* coefficients of 2.47, 2.41, and 2.22 respectively; the average ratings of the $40+22$ houses in the prior and current samples. Table 4 summarizes these plots with the L_1 norm between the various marginals and the posterior means and standard deviations. Rather surprising is the amount of influence on the intercept, which approaches the influence on the joint posterior, given in the line labeled 'posterior'. Also surprisingly, plausible values of σ decrease slightly with the MS and I priors.

The global effect of switching priors is substantial. For switching between any two priors, the L_1 norm between posteriors are estimated to be .65, .57 and .49, all quite large. These were calculated using a 1-d numerical integration of kernel density estimates of $p_1(\log\tau|Y)$ and $p_2(\log\tau|Y)$. These densities are plotted in Figure 2's left hand column (LHC). The solid, dashed and dotted densities are from the MS, I and F posteriors respectively. The choice of kernel density estimator mattered by .02 in the second digit. The importance sampling equality (10) gives two additional calculations each, depending on which sample is used. Of these six calculations, two for each pair of priors, four roughly agreed with the table figures; the two comparing

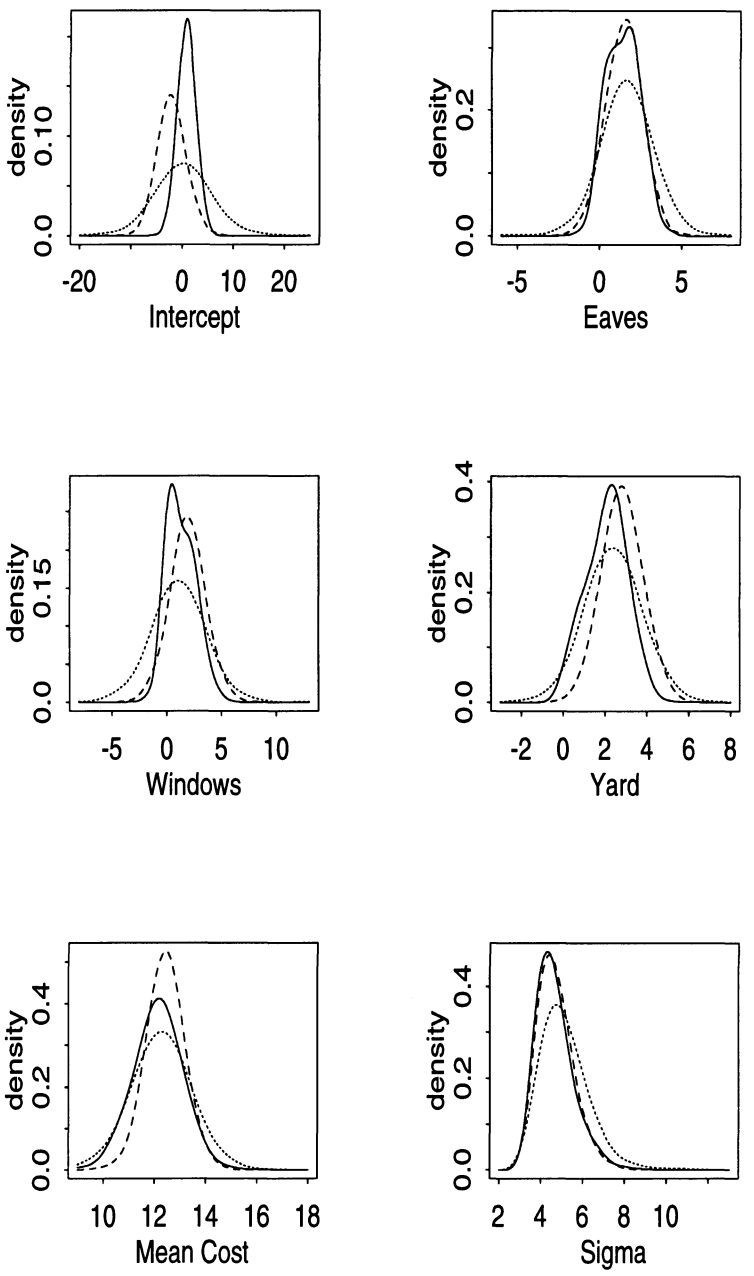

Figure 1: *Marginal Posteriors. Solid: MS prior; Dashed: I prior; Dotted: F prior.*

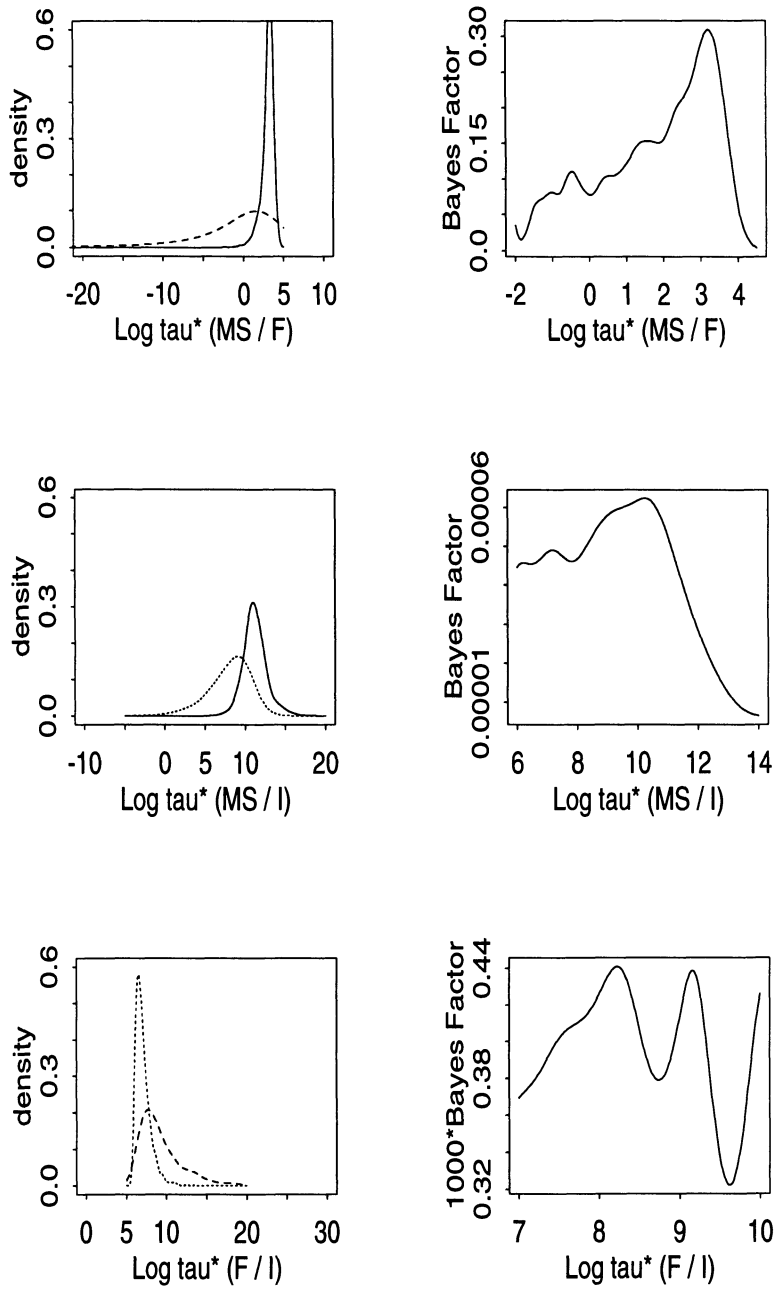

Figure 2: *Comparing MS, F and I Priors. Left column: posterior influence plots; Right: Bayes factors in favor of second prior using* (12)*, as a function of* $\log \tau^*$*. Row 1: MS and F; Row 2: MS and I; Row 3: F and I.*

the MS and F priors differed by .1 from table 4. From the LHC of Figure 2, we see that comparing MS and F posteriors, the F posterior might be reweighted to give the MS posterior but not vice-versa. The F posterior's reweighted L_1 was .55, while the MS posterior's reweighted posterior calculation gave .43, an underestimate. Inspecting the MS and I posteriors suggest that neither should work well for estimating under the other. Still, the L_1 calculations were .58 (MS) and .60 (I). Comparing the F and I priors gave L_1 values of .55 (I) and .49 (F), not too terrible, given that the plot suggests that the F posterior should work ok, while the I might not.

The data have a preference for the informative prior. The Bayes factor in favor of the I prior is around 2500 over the F prior, and around 21000 over the MS prior. These Bayes factors were calculated using formula (12). Table 4 gives median calculations using a range of τ^* values, plotted in the right hand column of Figure 2. If calculations were perfect, each of these figures should be a straight line across at the actual value of the Bayes factor. Clearly numerical problems still exist some of which may be due to bias from the kernel density estimates and to sampling variability. The last three rows of table 4 gives the Bayes factor calculations. The calculations range by a factor of roughly 1.5. The last row uses the second of the listed densities, the next to last uses the first of the listed posteriors. It is good that the data support the informative prior over the other two. The F prior's normalizing constant was chosen to favor the F prior. If other methods had been used to select a constant, chances are the Bayes factors would have favored the informative prior even less than it was already favored.

Acknowledgement. Thanks to Charlie Zhang for help with the plots and to two anonymous referees for helpful comments.

REFERENCES

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7** 686-690.

Bernardo, J. M. (1985). Discussion of Pettit and Smith (1985). In *Bayesian Statistics 2*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith. North Holland, Amsterdam 492-493.

Carlin, B. P., Kass, R. E., Lerch, F. J., and Huguenard, B. R. (1992). Predicting working memory failure: A subjective Bayesian approach to model selection. *J. Amer. Statist. Assoc.* **87** 319-327.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75** 651-659.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientarium Mathematicarum Hungarica* **2** 299-318.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881-889.

Greenhouse, J. and Wasserman, L. (1996). A practical, robust method for Bayesian model selection: a case study in the analysis of clinical trials. In *Bayesian Robustness*, edited by J.O. Berger, B. Betro, E. Moreno, L.R. Pericchi, F. Ruggeri, G. Salinetti and L. Wasserman, 41-58.

Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76** 663-674.

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

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

Ramsay, J. O. and Novick, M. R. (1980). PLU robust Bayesian decision theory: point estimation. *J. Amer. Statist. Assoc.* **75** 901-907.

Weiss, R. E. (1996). An approach to Bayesian sensitivity analysis. *J. Roy. Statist. Soc. Ser. B* **58** 739-750.

Weiss, R. E. and Cook, R. D. (1992). A graphical case statistic for assessing posterior influence. *Biometrika* **79** 51-55.

Department of Biostatistics
UCLA School of Public Health
Los Angeles CA 90095-1772 U.S.A.

Sufficiency and Influence

discussion by
JULIAN DE LA HORRA
Universidad Autónoma de Madrid

I would like to comment some possible extensions of the ideas in this interesting paper.

a) Let us consider the case in which our main interest is to predict the next observation x, given past observations y, where x and y are independent given θ. The ratio of posterior densities $p_2(\theta|y)$ and $p_1(\theta|y)$ (coming from M_2 and M_1, respectively) is given by

$$\frac{p_2(\theta|y)}{p_1(\theta|y)} = \frac{\tau^*(\theta,y)}{E_1[\tau^*(\theta,y)|y]} = \tau(\theta,y),$$

where $\tau^*(\theta,y)$ is the perturbation function that multiplies $p_1(\theta,y)$ to give $p_2(\theta,y)$ (case deletion, prior perturbation, ...), and the expectation is taken with respect to $p_1(\theta|y)$.

If $f_2(x|\theta) = f_1(x|\theta)$ (the sampling model for the next observation is the same under the two models), the ratio of posterior predictives is given by:

$$\begin{aligned}
\frac{f_2(x|y)}{f_1(x|y)} &= \frac{\int_\Theta f_2(x|\theta)p_2(\theta|y)d\theta}{\int_\Theta f_1(x|\theta)p_1(\theta|y)d\theta} \\
&= \frac{\int_\Theta f_1(x|\theta)\tau^*(\theta,y)p_1(\theta|y)d\theta / E_1[\tau^*(\theta,y)|y]}{\int_\Theta f_1(x|\theta)p_1(\theta|y)d\theta} \\
&= \frac{\int_\Theta \tau^*(\theta,y)p_1(\theta|x,y)d\theta}{E_1[\tau^*(\theta,y)|y]} \\
&= \frac{E_1[\tau^*(\theta,y)|x,y]}{E_1[\tau^*(\theta,y)|y]}.
\end{aligned}$$

This result is similar to that obtained in Weiss (1995) for a function of θ, $\nu(\theta)$, that captures the goals of the analysis:

$$\frac{p_2(\nu|y)}{p_1(\nu|y)} = \frac{E_1[\tau^*(\theta,y)|\nu,y]}{E_1[\tau^*(\theta,y)|y]}.$$

b) It could be interesting to study the influence of a perturbation on the posterior distribution, when the prior is of mixed type. This is the usual prior when we are interested in testing $\theta = \theta_0$ versus $\theta \neq \theta_0$.

For the model M_1, suppose that y is a set of observations from the density $f_1(y|\theta)$, and the prior distribution is given by a mass π_1 on θ_0 and a density $p_1(\theta|\theta \neq \theta_0)$ spreading the rest of the mass over $\Theta - \{\theta_0\}$. The posterior distribution is given by the mass on θ_0, $P_1(\theta_0|y)$, and by the density spreading the rest of the mass over $\Theta - \{\theta_0\}$, $p_1(\theta|y, \theta \neq \theta_0)$.

Consider now a perturbed model M_2 with elements $f_2(y|\theta)$, π_2 and $p_2(\theta|\theta \neq \theta_0)$. The posterior distribution is given by $P_2(\theta_0|y)$ and $p_2(\theta|y, \theta \neq \theta_0)$. The influence of this perturbation is usually measured by a function of the ratio $P_2(\theta_0|y)/P_1(\theta_0|y)$. This is suitable if we are interested in testing $\theta = \theta_0$ versus $\theta \neq \theta_0$. But, perhaps, we can sometimes need a measurement of the influence of the perturbation on the whole posterior distribution. This influence can be measured by the variation distance [this influence measure is considered in Weiss (1995) and is equivalent to L_1 distance between densities, when they exist]:

$$\begin{aligned} & \sup_{B \subset \Theta} [P_2(B|y) - P_1(B|y)] \\ = \; & \max\{P_2(\theta_0|y) - P_1(\theta_0|y) \\ + \; & \int_{B^*} ([1 - P_2(\theta_0|y)]p_2(\theta|y, \theta \neq \theta_0) - [1 - P_1(\theta_0|y)]p_1(\theta|y, \theta \neq \theta_0))d\theta, \\ & \int_{B^*} ([1 - P_2(\theta_0|y)]p_2(\theta|y, \theta \neq \theta_0) - [1 - P_1(\theta_0|y)]p_1(\theta|y, \theta \neq \theta_0))d\theta \Big\}, \end{aligned}$$

where

$$B^* = \{\theta \in \Theta : [1 - P_2(\theta_0|y)]p_2(\theta|y, \theta \neq \theta_0) - [1 - P_1(\theta_0|y)]p_1(\theta|y, \theta \neq \theta_0) > 0\}.$$

c) As the author points out, a popular approach for summarizing the difference between $p_1(\theta|y)$ and $p_2(\theta|y)$ is to use a divergence measure. In a recent paper, Dey and Birmiwal (1994) consider the local curvature for a divergence measure (small values of curvature indicating robustness). They obtain the local curvature of the divergence between the posterior coming from a baseline prior $p_1(\theta)$ and a posterior coming from either

$$\Gamma_a = \{p_2(\theta) : p_2(\theta) = (1 - \varepsilon)p_1(\theta) + \varepsilon q(\theta), q \in \mathcal{Q}\}$$

or

$$\Gamma_g = \{p_2(\theta) : p_2(\theta) = c(\varepsilon)p_1^{1-\varepsilon}(\theta)q^{\varepsilon}(\theta), q \in \mathcal{Q}\}.$$

It could be interesting to study the application of this approach to other types of perturbations of the model (other classes of priors, different sampling models, etc).

REFERENCES

DEY, D.K. AND BIRMIWAL, L.R. (1994). Robust Bayesian analysis using divergence measures. *Statist. Probab. Lett.* **20**, 287-294.

WEISS, R. (1995). An approach to Bayesian sensitivity analysis. *Preprint.*