Cumulative Robustness of the Posterior in Exponential Scale Families

BERTRAND CLARKE and PAUL GUSTAFSON

For scale parameters in exponential families one can evaluate the relative entropy between a fixed posterior and a deviation from it as a cumulative measure of robustness. The deviated posterior is formed by perturbing the prior, model, and data set in the direction of maximal increase of the relative entropy. A large value for this relative entropy may indicate a problem with the modeling strategy. As we see in an example, the converse need not hold.

In addition, the relative contributions of the prior, model, and data to the cumulative robustness can be identified. Computational results suggest that for good modeling most of the sensitivity of the posterior should be attributable to the data. However, this must be qualified by an examination of the directions in which the prior, model, and data deviate individually along the overall direction of most rapid change.

We demonstrate our methods in the context of two parametric models in exponential form, each with its own class of conjugate priors. In each case, one parameter is treated as a model index, so we do not estimate it. In the first example, the model index controls shape, in the second it controls tail behavior. The other parameter, which we do estimate, has a common interpretation across models.

KEY WORDS: Bayesian Robustness; Relative Entropy.

1 Introduction

It is not clear how robust inferences should be to the information that was used to form them. Too much robustness reflects a failure to model key features of a phenomenon and too little robustness means that inferences will not generalize adequately. As a consequence, there is a substantial body of work examining diverse aspects of robustness in various contexts.

Robustness of inferences has been examined in both the Bayesian and frequentist contexts. Sensitivity of inferences to the choice of prior has been extensively investigated; for a review see Berger (1994). Also, Lavine (1991) considers sensitivity of the posterior to the prior and model jointly. Much recent work has focussed on local sensitivity, where infinitesimal changes in the prior are studied. McCulloch (1989), Dey and Birmiwal (1993), Ruggeri and Wasserman (1993), Sivaganesan (1993), and Gustafson (?????) are a few of the many references. Sensitivity of inferences to the choice of model has been examined by White (1982), Gould and Lawless (1987), Neuhaus, Kalbfleisch, and Hauck (1992), Basu (1994), Tsou and Royall (1995), and others, from a variety of viewpoints. Sensitivity to the data, in terms of the problem of outliers or unreliable measurements in a data set has also been examined in terms of local influence, see Cook (1986). Diverse methods for reducing influence appropriately have been proposed. For reviews, see Huber (1981) and Hampel et. al (1986) amongst others. From a Bayesian point of view, many authors have investigated the effect of outliers, including Kass, Tierney and Kadane (1989), Weiss and Cook (1992), and Peng and Dey (1995).

Restricting to the Bayesian context, a posterior distribution is determined by a prior distribution for unknown parameters, a model for the conditional distribution of data given these parameters, and the observed data themselves. The novelty in our approach is that we examine the robustness of the posterior distribution to all of these inputs simultaneously. We call this *cumulative* robustness. That is, we permit the prior, model, and data to vary so as to obtain a perturbation of the baseline posterior. The relative entropy between the baseline posterior and its perturbation is compared to a measure of distance between the baseline inputs and the perturbed inputs. The distance used on the input triples is a sum of three distances for the three inputs: prior, model, and data. Our interest lies in the maximal rate of change in the posterior relative to change in the inputs. Further, we can examine how much of this maximal change is due to change in the prior, how much is due to change in the model, and how much is due to change in the data.

For computational and interpretational simplicity, we work locally. That is, we examine the effects of small changes in the inputs by examining second–order Taylor series approximations to both the relative entropy between posteriors and the input distance. Our method is an extension of McCulloch's (1989) method for examining prior robustness. Specifically, we examine the cumulative robustness in two examples. Both are exponential families with a scale parameter admitting sufficient statistics and conjugate classes of priors. However, these are not restrictions necessitated by the method we propose; rather, they are motivated by the desire to provide examples involving quantities that are easily computed when they cannot be obtained in closed form.

There are several aspects of this formulation that require comment. First, we must choose a measure of distance between the baseline posterior and a perturbation of it. Here we choose the relative entropy and note that its asymmetry is appropriate: In effect we are assuming the baseline posterior is 'true' and assessing divergence from it. This is consistent with the use of the relative entropy in seeking codes with minimal redundancy. Also, we choose a measure of distance for prior, model, data triples. We use the sum of the relative entropy on the priors (baseline and deviated), the relative entropy on the models and a third measure of distance on data sets. We specify the latter in the forthcoming example and note that it may be the most important for detecting incompatibility of the data with a given prior and model.

***Second, we have perturbed the data, and done so in a Bayesian context. ***

Third, in an effort to use robustness to detect insufficient goodness of fit, we recognize two sources of conflict, either of which may be present in a particular application. These are data-prior conflict and data-model conflict. Prior-model conflict does not occur because the prior and model represent disjoint sources of information which are chosen by the experimenter. The idea of data-model conflict is essentially goodness of fit. Either the data are representative of some member of the class of conditional densities defined by the model or they are not. The idea of data-prior conflict is that the prior represents information about where the parameter lies. The data might not give a estimate of the parameter which is within the region where the prior density is highest.

We propose that examination of the cumulative robustness can, in some cases, be used to detect the presence of data-model and data-prior conflict. Specifically, our examples below suggest that in a good data analysis, one finds maximal sensitivity to the data and comparatively little sensitivity to the prior and model. Consequently, an elevated sensitivity to either the prior or model may indicate the presence of one of the types of conflict identified. There is a caveat in this: If the directions of the deviation from the data to the prior and from the data to the model are opposite, the sensitivities may cancel giving the illusion of high sensitivity to the data when in fact the sensitivity to the prior and/or model is higher.

The setting in which we demonstrate our proposed technique is that of scale parameters in exponential families with conjugate priors. It will be seen that this is the easiest setting for the practical implementation of our technique. In a scale family, the relative entropy between models depends only on the model parameter, and not on the estimand. This leads to an interpretable term for the model in the input distance. Using conjugate families of priors makes it easier to find the direction of maximal rate of change of the relative entropy between the baseline posterior and its perturbation. In addition, when estimating the same parameter using different models, it is important to ensure that the parameter has the same interpretation in all of the models. In our examples we parametrize so that the parameter of interest is the mean, for all models.

To fix ideas, consider the following example. Let $X = (X_1, \ldots, X_n)$ be independent and identically distributed observations from a gamma distribution. Suppose that the mean of this distribution is to be estimated from the observed data $X = x$, while the shape parameter is a model index determined from physical modeling, or other external considerations. In particular, let $G(a, b)$ denote the gamma distribution with density proportional to $z^{a-1}e^{-z/b}$. Then the data are modeled as arising from the $G(\lambda, \theta/\lambda)$ distribution, where θ is the unknown mean parameter and λ is the known shape parameter. Inverse gamma distributions are conjugate priors for θ. Let $IG(a, b)$ denote the inverse gamma distribution with density proportional to $z^{-(a+1)}e^{-b/z}$. A prior distribution $\theta \sim IG(\alpha_1, \alpha_2)$ leads to a posterior distribution of the form $\theta|X = x \sim IG(\alpha_1^*, \alpha_2^*)$, where $\alpha_1^* = \alpha_1 + n\lambda$ and $\alpha_2^* = \alpha_2 + n\lambda\bar{x}$, with $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ being the sample mean.

Now, for sample size n, the posterior distribution is determined by the prior indexed by α, the model index λ, and the data x. To study the effect on the posterior distribution of simultaneous small changes to the three inputs we compare the baseline posterior given by $\omega = (\alpha, \lambda, x)$ to the posterior based on a nearby set of inputs $\tilde{\omega} = (\tilde{\alpha}, \tilde{\lambda}, \tilde{x})$ and measure the discrepancy between these two posteriors by the relative entropy. We denote this by

$$d_{PS}(\omega, \tilde{\omega}) \quad = \quad D(IG(\alpha_1^*, \alpha_2^*)||IG(\tilde{\alpha}_1^*, \tilde{\alpha}_2^*)), \tag{1}$$

where $D(p||q) = \int p(x)\log(p(x)/q(x))dx$ for arbitrary densities p and q with respect to Lebesgue measure.

Analogously, we take the discrepancy between the two prior distributions to be

$$d_{PR}(\alpha, \tilde{\alpha}) \quad = \quad D(IG(\alpha_1, \alpha_2)||IG(\tilde{\alpha}_1, \tilde{\alpha}_2)), \tag{2}$$

and the discrepancy between two models to be

$$d_M(\lambda, \tilde{\lambda}) \quad = \quad D(G(\lambda, \theta/\lambda)||G(\tilde{\lambda}, \theta/\tilde{\lambda})). \tag{3}$$

Since relative entropy is invariant under transformation of the sample space, the value of d_M does not depend on the scale parameter θ, only on the model indices λ and $\tilde{\lambda}$.

Finally, we must specify a measure of discrepancy on data sets. We choose

$$d_D(x,\tilde{x}) = \frac{\sum_{i=1}^{n}(\tilde{x}_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{4}$$

chiefly for convenience but also because it is compatible with our choice of the relative entropy in several senses. First, (4) is invariant under a common affine transformation of x and $\tilde{x}$, which mimics the invariance of relative entropy under transformation of the sample space. Second, to first order $d_D(x,\tilde{x})$ does not depend on n, like $d_{PR}(\alpha,\tilde{\alpha})$ and $d_M(\lambda,\tilde{\lambda})$ which do not depend on n at all. Thus it is possible to isolate the effect of sample size on the posterior from the effects of changes in the prior, model, and data. Note that (4) is based on changes in the individual data points, and so is sensible when the differences of the form $|\tilde{x}_i - x_i|$ are small compared to the spacings between the order statistics of $\tilde{x}$ or x, as is the case when $\tilde{x}$ is a local perturbation of x.

For a given inputs ω, we maximize d_{PS} subject to a constraint on the difference between ω and $\tilde{\omega}$. Formally, let

$$d_I(\omega,\tilde{\omega}) = d_{PR}(\alpha,\tilde{\alpha}) + d_M(\lambda,\tilde{\lambda}) + d_D(x,\tilde{x}) \tag{5}$$

be the input distance. Then the maximum of (1) as a function of $\tilde{\omega}$, subject to an upper bound on (5), would be a basic measure of posterior sensitivity to all inputs jointly.

A useful simplification results from approximating the solution to the constrained maximization problem is approximated. Expand expressions (4) and (5) about ω to get $d_I(\omega,\tilde{\omega}) \approx d_I^*(\omega,\tilde{\omega})$ and $d_{PS}(\omega,\tilde{\omega}) \approx d_{PS}^*(\omega,\tilde{\omega})$, where

$$d_I^*(\omega,\tilde{\omega}) = \frac{1}{2}(\tilde{\omega}-\omega)^T A_I(\omega)(\tilde{\omega}-\omega), \tag{6}$$

and

$$d_{PS}^*(\omega,\tilde{\omega}) = \frac{1}{2}(\tilde{\omega}-\omega)^T A_{PS}(\omega)(\tilde{\omega}-\omega). \tag{7}$$

In each case, $A(\omega)$ is the second derivative of $d^*(\omega,\tilde{\omega})$ with respect to $\tilde{\omega}$, evaluated at $\tilde{\omega} = \omega$.

The additive form of (5) yields

$$A_I(\omega) \;=\; \begin{pmatrix} A_{PR}(\alpha) & 0 & 0 \\ 0 & A_M(\lambda) & 0 \\ 0 & 0 & A_D(x) \end{pmatrix} \tag{8}$$

where A_{PR}, A_M, and A_D are second derivatives arising from (2), (3), and (4) respectively. One can verify that $A_{PR}(\alpha)$ is the Fisher information matrix for the $IG(\alpha_1, \alpha_2)$ i.e.,

$$A_{PR}(\alpha) \;=\; \begin{pmatrix} \psi'(\alpha_1) & -1/\alpha_2 \\ -1/\alpha_2 & \alpha_1/\alpha_2^2 \end{pmatrix},$$

where ψ' is the trigamma function. Expression (2) gives that $A_M(\lambda)$ is the Fisher information matrix for the parametric family of model densities indexed by λ, under a fixed value of the θ. In the present case, A_M is the Fisher information for the $G(\lambda, \theta/\lambda)$ family, when θ is known and it does not depend on θ, i.e., $A_M(\lambda) = \Psi'(\lambda) - \lambda^{-1}$. Finally, we have that $A_D(x) = 2(\sum_{i=1} n(x_i - \bar{x})^2)^{-1} I_n$, where I_n is the $n \times n$ identity matrix.

Analogously, $A_{PS}(\omega)$ is the Fisher information matrix for the family of posterior distributions indexed by the input vector ω. This can be determined directly from the form of the posterior distribution but it is simpler to use conjugacy. The hyperparameter vector α is updated to α^*, which map from the input vector ω to the updated hyperparameter vector α^*. (***: not ω ?) Letting B denote this mapping, the Fisher information for the posterior distribution is

$$A_{PS}(\omega) \;=\; \{B'(\omega)\}^T A_{PR}(B(\omega))\{B'(\omega)\}, \tag{9}$$

where B' is the derivative of B. In the present case,

$$B\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \lambda \\ x \end{pmatrix} \;=\; \begin{pmatrix} \alpha_1 + n\lambda \\ \alpha_2 + n\lambda\bar{x} \end{pmatrix},$$

with derivative

$$B'\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \lambda \\ x \end{pmatrix} \;=\; \begin{pmatrix} 1 & 0 & n & 0 & \dots & 0 \\ 0 & 1 & n\bar{x} & \lambda & \dots & \lambda \end{pmatrix}.$$

Now we seek the maximum value of (7) as a function of $\tilde{\omega}$, subject to the constraint that (6) does not exceed some fixed value ϵ^2. A standard linear algebra result (see, for instance, Srivastava and Carter, 1983, Corollary 1.10.1) gives the maximum as $k\epsilon^2$, where k is the largest eigenvalue of $[A_I(\omega)]^{-1}A_{PS}(\omega)$. This maximum is attained by taking $\tilde{\omega} = \omega + c\nu$, where ν is the eigenvector corresponding to eigenvalue k, and c is a constant chosen so that (6) is equal to ϵ^2. This approach was first used by McCulloch (1989). to investigate sensitivity to the prior in which case only the first term in the right-hand side of (5) is present.

Here, the approximations d_I^* and d_{PR}^* to d_I and d_{PR} are better when $\tilde{\omega} - \omega$ is smaller. Consequently, k can be regarded as the locally maximal rate at which d_{PS} changes relative to d_I. We therefore define k to be the cumulative robustness. Note that this definition of cumulative robustness permits attribution of sensitivity to the model, data, and prior. Specifically, the discrepancy in inputs along the direction of maximal change can be partitioned as

$$\nu^T A_I(\omega)\nu \quad = \quad \nu_{PR}^T A_{PR}(\alpha)\nu_{PR} + \nu_M^T A_M(\lambda)\nu_M + \nu_D^T A_D(x)\nu_D, \tag{10}$$

where $\nu = (\nu_{PR}, \nu_M, \nu_D)$ is the partition of the maximal eigenvector into components corresponding to the prior, model, and data respectively. So, the ratio of $v_{PR}^T A_{PR}(\alpha) v_{PR}$ to $v^T A_I(\omega) v$ is the relative contribution of prior uncertainty to the cumulative robustness. The relative contributions of the model and data can be reported similarly.

As a numerical illustration let the baseline model specification be $\lambda = 2$ and let $\alpha = (3, 2)$. This makes both the prior mean and prior variance for θ equal to one. A data set of size 20 is simulated from the $Gamma(2, 1/2)$ distribution. Such a data set arises when the baseline model specification is correct and the true value of θ is equal to the prior mean for θ. The cumulative robustness of the posterior based on only the first five observations is $k = 1.80$, with relative contributions of $(.13, .15, .73)$ from the prior, model, and data respectively. If the first ten observations are used, the cumulative robustness is 4.92, with relative contributions of $(.02, .01, .96)$. If all twenty observations are considered, the cumulative robustness is 12.42, with relative contributions $(.01, .00, .99)$. (The relative contributions do not necessarily sum to exactly to one because of rounding.) These numerical results are consistent with the discussion in the following sections.

In Section 2, after formulating another example, we provide further numerical results to show how the cumulative robustness and its partition may be used to detect lack of agreement between the prior and the data or between the model and the data. In the discussion of Section 3 we give some methodological implications.

2 A Power Family of Gamma

2.1 A Parameter for Tail behavior

Again, consider estimating the mean θ of a distribution on $(0, \infty)$ which gives rise to independent and identically distributed observations $X_1, \ldots, X_n$. In this example, suppose the model index λ governs the right tail behavior of the distribution, and that the data are modeled by a density proportional to $\exp(-(x/\sigma)^\lambda)$, where λ is known and σ is unknown. That is, λ is presumed determined by a physical model. Note that $\lambda = 1$ yields an exponential model, and $\lambda = 2$ corresponds to a truncated-normal model.

Since the mean θ is the quantity of interest, we switch from the (λ, σ) parameterization to the (λ, θ) parameterization. This is accomplished by setting $\sigma = \theta / c_\lambda$, where $c_\lambda = \Gamma(2/\lambda)/\Gamma(1/\lambda)$. Under the desired parameterization, the density of a single observation is

$$p_\lambda(z|\theta) \quad = \quad \left(\frac{c_\lambda}{\theta}\right) \frac{\lambda}{\Gamma(1/\lambda)} \exp\left(-\left[\left(\frac{c_\lambda}{\theta}\right) z\right]^\lambda\right). \tag{11}$$

Alternatively, this distribution corresponds to the power of a gamma random variable. In particular, the parametric family can be represented as

$$Z \quad = \quad \left(\frac{\theta}{c_\lambda}\right) Z_0^{1/\lambda}, \tag{12}$$

where $Z_0 \sim G(1/\lambda, 1)$. We denote the parametric family (11) as $PG(\lambda, \theta)$ (the P stands for power).

This example differs from that of the previous section in that the family of conjugate priors for θ depends on the model index λ. Parametrizing by $\alpha = (\alpha_1, \alpha_2)$, the conjugate prior density has the form

$$p_{\lambda,\alpha}(\theta) = \lambda \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \frac{1}{\theta^{\alpha_1 \lambda + 1}} e^{-\alpha_2/\theta^\lambda}. \tag{13}$$

In parallel with (12), (13) can be expressed as a power of an inverse gamma variate,

$$\theta = \theta_0^{1/\lambda}, \tag{14}$$

where $\theta_0 \sim IG(\alpha_1, \alpha_2)$. It is straightforward to verify that for each λ and α, (13) is a unimodal density. Let $PIG(\alpha_1, \alpha_2, \lambda)$ denote the parametric family (13). Then Bayesian updating proceeds as follows. If $X_1, \ldots, X_n$ are independent and identically distributed as $PG(\lambda, \theta)$, and $\theta \sim PIG(\alpha_1, \alpha_2, \lambda)$, then $\theta | X = x \sim PIG(\alpha_1^*, \alpha_2^*, \lambda)$, where $\alpha_1^* = \alpha_1 + n/\lambda$ and $\alpha_2^* = \alpha_2 + c_\lambda^\lambda \sum_{i=1}^n x_i^\lambda$.

Following the method outlined in Section 1, we need the Fisher information matrix for the $PIG(\alpha_1, \alpha_2, \lambda)$ and the Fisherr information for the $PG(\lambda, \theta)$ family when θ is known. These quantities are derived in the Appendix. Discrepancy between data sets is again measured using (4).

The fact that the class of conjugate priors depends on the model index λ necessitates slight changes in the methodology of Section 1. In particular, the discrepancy between priors depends not just on α and $\tilde{\alpha}$, but also on λ and $\tilde{\lambda}$. Let $\gamma = (\alpha, \lambda)$ and replace $d_{PR}(\alpha, \tilde{\alpha})$ by $d_{PR}(\gamma, \tilde{\gamma})$. This in turn causes a modification to (8), in that $A_{PR}(\gamma)$ and $A_M(\lambda)$ will overlap. That is, both the prior discrepancy and the model discrepancy contribute additively to the λ block of A_I. The relationship (9) is still valid, provided that B is considered to map (α, λ, x) to (α^*, λ).

2.2 Computational Results

Provided one has reliable data, there are two sorts of modeling errors a Bayesian can make. The prior beliefs may be wildly wrong in the sense that the sample mean is far from the prior mean. Or, the data may conflict with the model in the sense that the model fails a goodness of fit test. Consequently it is desirable to characterize four cases. The first is the ideal case of no conflict. That is, the data reinforce the prior beliefs and the model fits the data. The second case is that of data–model conflict but no data–prior conflict. That is, the data reinforce the prior beliefs but the model fit is poor. The third case is the reverse of this: there is no data model conflict but the data contradict the prior beliefs. In the fourth case the model fit is poor (data–model conflict) and the prior beliefs are wrong (data–prior conflict).

To investigate these cases we compute the cumulative robustness and the relative contributions of the prior, model, and data to it. Given the sample size n, the "true" model index λ^*, and the true parameter value θ^*, we take the data vector x to be the $(1/(n+1), \ldots, n/(n+1))$ quantiles of $P_{\lambda^*}(\cdot|\theta^*)$. This ensures that the data set is representative of the true model and parameter values. Here, in fact, we set $\theta^* = 1$. This is without loss of generality, because θ is a scale parameter.

For interpretability, priors are specified by their moments. For a given λ, let ν_1 and ν_2 be the prior mean and standard deviation. Thus ν is a reparameterization of α; mathematical details are given in the Appendix. To compare relatively informative and noninformative priors, we take the prior standard deviation to be $\nu_2 = 0.2$ and $\nu_2 = 0.9$ respectively. Now the degree of prior–data conflict can be summarized in the other hyperparameter ν_1. We choose $\nu_1 = \theta^*$ for prior–data agreement, since x is a vector of quantiles under θ^*. For prior–data conflict we set $\nu_1 = \theta^* + 2\nu_2$ so that the true parameter lies two prior standard deviations away from the prior mean.

We take the true and assumed model indices, λ^* and λ, to be elements of the set $\{1, 2\}$. The presence or absence of data–model conflict is represented by taking $\lambda^* = \lambda$ or $\lambda^* \neq \lambda$ respectively. Thus two representations of conflict are possible: $(\lambda^*, \lambda) = (1, 2)$ and $(\lambda^*, \lambda) = (2, 1)$. These two possibilities correspond to using a model with a lighter tail when a heavier tail is appropriate, and using a model with a heavier tail when a lighter tail is appropriate. Thus the two conflicts are in opposite directions. We return to this point presently.

The results of our computations have been organized into four tables. Each table has three rows for sample sizes $n = 5$, 10, and 20, and four columns corresponding to the four cases described at the beginning of this subsection. For Tables 1 and 2, $\lambda^* = 1$; for Tables 3 and 4, $\lambda^* = 2$. Tables 1 and 2 differ in the informativity of the prior; Tables 3 and 4 differ in the same way. Each table entry consists of one number representing the cumulative robustness under the given conditions, and one triple representing the relative contributions of the prior, model, and data to that cumulative robustness.

When examining Tables 1 through 4, it is meaningful to compare the cumulative robustness values to each other, and to compare the values within one triple to the corresponding values in another triple. However, it may be less meaningful to compare values within a triple to each other. This is so because within a triple the divergence measure for the prior and model is a relative entropy between one dimensional distributions comparable to each other and to the divergence between posteriors. By contrast, the divergence measure on the data is somewhat ad hoc.

Tables 1 through 4 have several anticipated properties. First, in each column of each table the cumulative robustness increases with sample size. This is consistent with work of Gustafson and Wasserman (????) showing that for fixed data and model, the norm of the mapping from prior to posterior increases with sample size. Indeed, regardless of the actual value of x, the posterior concentrates as the sample size increases. Consequently, as the sample size increases, slight shifts in the data are magnified by the concentration. Second, the relative contribution of the prior to the cumulative robustness, as given by the first entry in each triple, always decreases with sample size. Third, the highest relative contributions of the data to the cumulative robustness occurs when the prior and model agree with the data. This is true in every case when $n = 20$, and in most cases for smaller sample sizes. This suggests that one wants a posterior which is relatively robust to deviations in the prior and model so that most of the sensitivity is to the data. Finally, the relative contribution of the prior to the cumulative robustness tends to be larger in the presence of data–prior conflict and the relative contribution of the model tends to be larger in the presence of data–model conflict.

The tables exhibit some unexpected properties as well. First note that in Tables 1 and 2 the sensitivity tends to be larger in the second and third columns than in the first and fourth columns. Heuristically, it is tempting to expect that the sensitivity should be greater in the presence of both conflicts than in the presence of either conflict alone. In fact, while this is plausible it is masked here because the directions of the two conflicts cancel each other. In the first two tables the data represent a thicker–tailed distribution ($\lambda^* = 1$). Adding data–model conflict by modeling with a thinner–tailed distribution ($\lambda = 2$) tends to bias estimates downward since the data are positive. However, the data–prior conflict we have used—centering the prior two standard deviations higher than θ^*—biases estimates upward. Thus these two sources of conflict tend to cancel leading to a cumulative robustness smaller than under either conflict alone, at least for the larger of the sample sizes we have used.

The reverse is seen in Tables 3 and 4. In these cases the data come from the thinner–tailed distribution, but under data–model conflict they are modeled with the thicker–tailed distribution. The wrong model tends to bias estimates to the right, as does the prior when data–prior conflict is present. TogetherIn tandem the two conflicts reinforce and give a larger cumulative robustness. The frequentist robustness literature suggests it is less damaging to use a thick–tailed distribution with thin–tailed data than to use a thin–tailed distribution with thick–tailed data. This is supported in the present context because the cumulative robustness values in column 2 of Tables 1 and 2 are larger than their counterparts in Tables 3 and 4 respectively.

Finally, we note that our results here are for a very special case: the data are univariate and positive, the parameter is a positive scalar, and the classes of priors are conjugate, indexed by a two–dimensional hyperparameter. Because of these simplifying features, it is possible to determine when data–prior and data–model conflicts cancel or reinforced each other.

3 Discussion

The main methodological novelty of the present work is twofold. First, we have proposed a comprehensive measure of a posterior's sensitivity to its three inputs: the prior, the model, and the data. Second, we have partitioned this cumulative robustness so that the relative contributions of these inputs can be identified. We suggest these relative contributions can be used to detect inappropriate prior densities (data–prior conflict) or ill–fitting models (data–model conflict).

Both of the examples here suggest that for valid inferences the relative contribution of the data to the cumulative robustness should be as high as possible, and that high contributions from the prior or model are associated with data–prior and data–model conflict respectively. When both sources of conflict are present the cumulative robustness may not be high due to a cancellation effect. So, as yet, we cannot make a non–trivial statement about detecting this case by looking only at the relative contributions of the prior and model to the cumulative robustness. Nevertheless, the partitioning of cumulative robustness as we have defined it here is a partial check for model fit and good prior information because high sensitivity to the prior or model may indicate a modeling problem.

Settings with data–model conflict constitute a general limitation on robustness methods. This is so because lack of fit is not always detectable through criteria reflecting robustness exclusively. In particular, an ill–fitting model may be highly robust. However, as in the examples here, it is often the case that lack of fit is associated with lack of robustness, because a slight change to an ill fitting model may yield substantially better inferences. Thus our use of robustness when the only source of conflict is between the data and the model gives results consistent with intuition. When there are two sources of conflict, the discrepancy between robustness and goodness–of–fit can be more pronounced, as is seen in our results. This is simply due to the fact that variations in the prior and model can either cancel or reinforce one another.

Appendix

Details for the example of Section 3 are given here. Some of the expressions were determined or verified using the MAPLE software package. Several facts are used repeatedly in what follows. First, note that if G is a standard gamma random variable with shape parameter s, then $E(G^a) = \Gamma(a+s)/\Gamma(s)$, provided $a > -s$. Furthermore, $E(\log G) = \Psi(s)$, where $\Psi(s) = \frac{\partial}{\partial s}\log\Gamma(s)$ is the digamma function. The trigamma function is denoted $\Psi'(s) = \frac{\partial}{\partial s}\Psi(s)$.

From (13), the relative entropy between two conjugate priors under different models is seen to be

$$\begin{aligned} d_{PR}(\gamma,\tilde{\gamma}) &= \log\left(\frac{\lambda\alpha_2^{\alpha_1}\Gamma(\tilde{\alpha}_1)}{\tilde{\lambda}\tilde{\alpha}_2^{\tilde{\alpha}_1}\Gamma(\alpha_1)}\right) + (\tilde{\alpha}_1\tilde{\lambda} - \alpha_1\lambda)E_{\alpha,\lambda}(\log\theta) + \\ & \quad \tilde{\alpha}_2 E_{\alpha,\lambda}(\theta^{-\tilde{\lambda}}) - \alpha_2 E_{\alpha,\lambda}(\theta^{-\lambda}). \qquad (15) \end{aligned}$$

The expectations are easily evaluated by substituting the right–hand side of (14) into (15), and then applying the above–mentioned facts. The resulting expression is

$$\begin{aligned} d_{PR}(\gamma,\tilde{\gamma}) &= \log\left(\frac{\lambda\alpha_2^{\alpha_1}\Gamma(\tilde{\alpha}_1)}{\tilde{\lambda}\tilde{\alpha}_2^{\tilde{\alpha}_1}\Gamma(\alpha_1)}\right) + (\tilde{\alpha}_1\tilde{\lambda} - \alpha_1\lambda)\left(\frac{\log\alpha_2 - \psi(\alpha_1)}{\lambda}\right) + \\ & \quad \frac{\tilde{\alpha}_2}{\alpha_2^{\tilde{\lambda}/\lambda}}\frac{\Gamma(\alpha_1 + \tilde{\lambda}/\lambda)}{\Gamma(\alpha_1)} - \alpha_1. \end{aligned}$$

Differentiating twice and setting $\tilde{\gamma} = \gamma$ gives the (symmetric) second derivative matrix:

$$A_{PR}(\gamma) = \begin{pmatrix} \psi'(\alpha_1) & \frac{-1}{\alpha_2} & -\frac{\log\alpha_2 - \psi(\alpha_1)}{\lambda} \\ & \frac{\alpha_1}{\alpha_2^2} & \frac{\alpha_1(\psi(\alpha_1) - \log\alpha_2) + 1}{\alpha_2\lambda} \\ & & \frac{\alpha_1\left[(\psi(\alpha_1) - \log\alpha_2 + (1/\alpha_1))^2 + \psi'(\alpha_1)\right] - (1/\alpha_1) + 1}{\lambda^2} \end{pmatrix}.$$

A very similar argument leads to an expression for the relative entropy between two sampling densities for a single observation X, under a common mean θ but different models λ and $\tilde{\lambda}$. By the invariance of relative entropy, we take $\theta = 1$ without loss of generality. From (11) and (12) we see that

$$d_M(\lambda, \tilde{\lambda}) \;=\; E_\lambda \left\{ \log\left(\frac{c_\lambda \lambda \Gamma(1/\tilde{\lambda})}{c_{\tilde{\lambda}} \tilde{\lambda} \Gamma(1/\lambda)} \right) + \left(\frac{c_{\tilde{\lambda}}}{c_\lambda} Y^{1/\lambda} \right)^{\tilde{\lambda}} - Y \right\},$$

where $Y = (c_\lambda X)^\lambda$. From (12) it is seen that $Y \sim \text{Gamma}(1/\lambda)$ under model λ. Thus the expectations can be calculated, leading to

$$d_M(\lambda, \tilde{\lambda}) \;=\; \log\left(\frac{c_\lambda \lambda \Gamma(1/\tilde{\lambda})}{c_{\tilde{\lambda}} \tilde{\lambda} \Gamma(1/\lambda)} \right) + \left(\frac{c_{\tilde{\lambda}}}{c_\lambda} \right)^{\tilde{\lambda}} \frac{\Gamma((\tilde{\lambda}+1)/\lambda)}{\Gamma(1/\lambda)} - \frac{1}{\lambda}.$$

Differentiating twice with respect to $\tilde{\lambda}$ yields:

$$\begin{aligned} A_M(\lambda) \;=\; & \left(\frac{1}{\lambda}\right)^4 \{\psi'(1/\lambda)\} + \\ & \left(\frac{1}{\lambda}\right)^3 \left\{\psi'(1/\lambda) + 4[\psi(1/\lambda) - \psi(2/\lambda)] + 4[\psi(1/\lambda) - \psi(2/\lambda)]^2\right\} + \\ & \left(\frac{1}{\lambda}\right)^2 \{1 + 4[\psi(1/\lambda) - \psi(2/\lambda)]\}. \end{aligned}$$

The Bayesian updating takes the form

$$B \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \lambda \\ x \end{pmatrix} \;=\; \begin{pmatrix} \alpha_1 + (n/\lambda) \\ \alpha_2 + c_\lambda^\lambda \sum_{i=1}^n x_i^\lambda \\ \lambda \end{pmatrix}.$$

This leads to a derivative

$$B'(\omega) \;=\; \begin{pmatrix} 1 & 0 & \frac{-n}{\lambda^2} & 0 & \dots & 0 \\ 0 & 1 & \frac{\partial}{\partial\lambda} c_\lambda^\lambda \sum x_i^\lambda & c_\lambda^\lambda \lambda x_1^{\lambda-1} & \dots & c_\lambda^\lambda \lambda x_n^{\lambda-1} \\ 0 & 0 & 1 & 0 & \dots & 0 \end{pmatrix},$$

where

$$\frac{\partial}{\partial\lambda} c_\lambda^\lambda \sum x_i^\lambda \;=\; c_\lambda^\lambda \left\{ \left[\sum_i x_i^\lambda \log x_i \right] + [\log c_\lambda + (1/\lambda)\psi(1/\lambda) - (2/\lambda)\psi(2/\lambda)] \left[\sum_i x_i^\lambda \right] \right\}.$$

To determine hyperparameters (α_1, α_2) in terms of the prior mean and standard deviation (ν_1, ν_2), note that under (13), $E_{\alpha,\lambda}(\theta)$ is given by

$$\nu_1 \;=\; \alpha_2^{1/\lambda} \frac{\Gamma(\alpha_1 - 1/\lambda)}{\Gamma(\alpha_1)},$$

(provided $\alpha_1 > 1/\lambda$), and $E_{\alpha,\lambda}(\theta^2)$ is

$$\nu_2^2 + \nu_1^2 = \alpha_2^{2/\lambda} \frac{\Gamma(\alpha_1 - 2/\lambda)}{\Gamma(\alpha_1)},$$

(provided $\alpha_1 > 2/\lambda$). Both these calculations are expectations of (negative) powers of gamma random variables. We can numerically solve

$$\frac{\Gamma(\alpha_1 - 2/\lambda)\Gamma(\alpha_1)}{(\Gamma(\alpha_1 - 1/\lambda))^2} = 1 + \frac{\nu_2^2}{\nu_1^2},$$

for α_1. A solution exists for $\alpha_1 \in (2/\lambda, \infty)$, since the left–hand side decreases from ∞ down to 1 over this range. This fact follows from the concavity of the digamma function. Subsequently, α_2 is determined as

$$\alpha_2 = \left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 - 1/\lambda)} \nu_1 \right)^{\lambda}.$$

References

Basu, S. (1994), "Posterior sensitivity to the sampling distribution and the prior: more than one observation," Technical Report 66, Department of Mathematical Sciences, University of Arkansas.

Berger, J.O. (1994), "An overview of robust Bayesian analysis," *Test,* 3, 5-58.

Clarke, B. (1989), "Asymptotics of entropy risk with applications," Ph.D. Thesis, Department of Statistics, University of Illinois.

Cook, R.D. (1986), "Assessment of local influence (with discussion)," *Journal of the Royal Statistical Society B*, 48, 133-169.

Dey, D.K. and Birmiwal, L.R. (1994), "Robust Bayesian analysis using entropy and divergence measures," *Statistics and Probability Letters*, 20, 287-294.

Gould, A. and Lawless, J.F. (1988), "Consistency and efficiency of regression coefficient estimates in location–scale models," *Biometrika*, 75, 535–540.

Gustafson, P. (in press), "Local sensitivity of inferences to prior marginals," *Journal of the American Statistical Association.*

Gustafson. P. and Wasserman, L. (in press), "Local sensitivity diagnostics for Bayesian Inference," *Annals of Statistics.*

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.

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

Kass, R.E., Tierney, L., and Kadane, J.B. (1989), "Approximate methods for assessing influence and sensitivity in Bayesian analysis," *Biometrika* 76, 663-74.

Lavine, M. (1991), "Sensitivity in Bayesian statistics: the prior and the likelihood," *Journal of the American Statistical Association* 86, 396-399.

McCulloch, R.E. (1989), "Local model influence," *Journal of the American Statistical Association,* 84, 473-478.

Neuhaus, Kalbfleisch, and Hauck (1992). "The effects of mixture distribution misspecification when fitting mixed–effects logistic models," *Biometrika*, 79, 755–62.

Peng, F. and Dey, D.K. (1995), "Bayesian analysis of outlier problems using divergence measures," *Canadian Journal of Statistics*, 23, 199-213.

Ruggeri, F. and Wasserman, L. (1993). "Infinitesimal sensitivity of posterior distributions," *Canadian Journal of Statistics*, 21, 195-203.

Sivaganesan, S. (1993). "Robust Bayesian diagnostics," *Journal of Statistical Planning and Inference*, 35, 171-188.

Srivastava, M.S. and Carter, E.M. (1983). *An Introduction to Applied Multivariate Statistics.* North-Holland, New York.

Tsou, T.S., and Royall, R.M. (1995), "Robust likelihoods," *Journal of the American Statistical Association*, 90, 316–320.

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

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

n	no conflicts	data-model conflict	data-prior conflict	both conflicts
5	1.47	2.01	2.91	2.99
	(0.28, 0.28, 0.44)	(0.20, 0.27, 0.53)	(0.80, 0.18, 0.02)	(0.99, 0.00, 0.01)
10	2.96	6.22	5.16	4.93
	(0.09, 0.07, 0.85)	(0.14, 0.33, 0.53)	(0.67, 0.26, 0.07)	(0.96, 0.01, 0.03)
20	6.97	15.39	8.87	8.44
	(0.02, 0.01, 0.96)	(0.13, 0.35, 0.52)	(0.54, 0.27, 0.19)	(0.79, 0.09, 0.12)

Table 1: Cumulative robustness and relative contributions of inputs. The data are representative of $\lambda^* = 1$ and $\theta^* = 1$. The prior standard deviation is 0.9 throughout. The model index λ is 1 (2) under absence (presence) of data–likelihood conflict. The prior mean ν_1 is 1.0 (2.8) under absence (presence) of data–prior conflict.

n	no conflicts	data-model conflict	data-prior conflict	both conflicts
5	1.24	1.22	1.49	1.65
	(0.74, 0.15, 0.11)	(0.81, 0.00, 0.19)	(0.86, 0.12, 0.02)	(0.94, 0.04, 0.02)
10	1.79	2.38	2.18	2.18
	(0.43, 0.16, 0.41)	(0.30, 0.16, 0.54)	(0.66, 0.24, 0.10)	(0.85, 0.03, 0.12)
20	3.86	9.39	4.10	3.58
	(0.15, 0.08, 0.76)	(0.08, 0.40, 0.52)	(0.41, 0.31, 0.28)	(0.59, 0.00, 0.40)

Table 2: Cumulative robustness and relative contributions of inputs. The data are representative of $\lambda^* = 1$ and $\theta^* = 1$. The prior standard deviation is 0.2 throughout. The model index λ is 1 (2) under absence (presence) of data–likelihood conflict. The prior mean ν_1 is 1.0 (1.4) under absence (presence) of data–prior conflict.

n	no conflicts	data-model conflict	data-prior conflict	both conflicts
5	1.34	1.49	2.92	2.89
	(0.20, 0.14, 0.66)	(0.31, 0.42, 0.27)	(0.99, 0.00, 0.01)	(0.74, 0.24, 0.02)
10	3.38	2.76	4.86	5.33
	(0.09, 0.16, 0.75)	(0.14, 0.32, 0.55)	(0.98, 0.00, 0.02)	(0.60, 0.35, 0.05)
20	7.58	5.81	7.98	9.54
	(0.06, 0.14, 0.80)	(0.06, 0.23, 0.71)	(0.89, 0.02, 0.09)	(0.47, 0.41, 0.12)

Table 3: Cumulative robustness and relative contributions of inputs. The data are representative of $\lambda^* = 2$ and $\theta^* = 1$. The prior standard deviation is 0.9 throughout. The model index λ is 2 (1) under absence (presence) of data–likelihood conflict. The prior mean ν_1 is 1.0 (2.8) under absence (presence) of data–prior conflict.

n	no conflicts	data-model conflict	data-prior conflict	both conflicts
5	1.17	1.20	1.64	1.44
	(0.75, 0.07, 0.18)	(0.71, 0.20, 0.09)	(0.91, 0.07, 0.02)	(0.83, 0.15, 0.02)
10	1.69	1.77	2.34	2.19
	(0.40, 0.00, 0.60)	(0.41, 0.30, 0.29)	(0.78, 0.13, 0.08)	(0.59, 0.33, 0.08)
20	4.27	3.61	3.83	4.36
	(0.11, 0.07, 0.82)	(0.16, 0.30, 0.53)	(0.61, 0.13, 0.26)	(0.34, 0.48, 0.18)

Table 4: Cumulative robustness and relative contributions of inputs. The data are representative of $\lambda^* = 2$ and $\theta^* = 1$. The prior standard deviation is 0.2 throughout. The model index λ is 2 (1) under absence (presence) of data–likelihood conflict. The prior mean ν_1 is 1.0 (1.4) under absence (presence) of data–prior conflict.