ELSEVIER

Journal of Statistical Planning and Inference 71 (1998) 137–150

journal of statistical planning and inference

On the overall sensitivity of the posterior distribution to its inputs

Bertrand Clarke *, Paul Gustafson

Department of Statistics, University of British Columbia, 6356 Agricultural Road, Room 333, Vancouver, BC, Canada V6T 1Z2

Received 1 June 1996; revised 5 January 1998; accepted 6 January 1998

Abstract

In a parametric Bayesian analysis, the posterior distribution of the parameter is determined by three inputs: the prior distribution of the parameter, the model distribution of the data given the parameter, and the data themselves. Working in the framework of two particular families of parametric models with conjugate priors, we develop a method for quantifying the local sensitivity of the posterior to simultaneous perturbations of all three inputs. The method uses relative entropy to measure discrepancies between pairs of posterior distributions, model distributions, and prior distributions. It also requires a measure of discrepancy between pairs of data sets. The fundamental sensitivity measure is taken to be the maximum discrepancy between a baseline posterior and a perturbed posterior, given a constraint on the size of the discrepancy between the baseline set of inputs and the perturbed inputs. We also examine the perturbed inputs which attain this maximum sensitivity, to see how influential the prior, model, and data are relative to one another. An empirical study highlights some interesting connections between sensitivity and the extent to which the data conflict with both the prior and the model.

Keywords: 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

* Corresponding author.

0378-3758/98/$19.00
PII: S0378-3758(98)00014-7

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 (1994), Ruggeri and Wasserman (1993), Sivaganesan (1993), and Gustafson (1996) are a few of the many references. Sensitivity of inferences to the choice of model has been examined by White (1982), Gould and Lawless (1988), Neuhaus et al. (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 (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 et al. (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 *overall sensitivity*. Specifically, 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. Our primary interest lies in the maximal rate of change in the posterior relative to change in the inputs. Further, we examine the *relative influence* of the three inputs; that is, we assess how much of the 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 interpretive 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. In this regard our method extends McCulloch's (1989) method for examining prior robustness.

There are several aspects of this formulation that require comment. First, we use relative entropy as measure of discrepancy between the baseline and perturbed posterior distributions. Whether or not an asymmetric measure is appropriate is moot, since the quadratic approximation symmetrizes the discrepancy measure. As well, we choose a measure of discrepancy between two sets of inputs based on summing the relative entropies between the priors, the relative entropies between the models, and a measure of discrepancy between the data sets.

Second, our goal is to quantify and better understand how the posterior distribution is sensitive to all its inputs. It is for this reason that we entertain perturbations to the data, as well as to the model and prior. In particular, our interest in data sets near the observed or baseline data set should not be construed as having frequentist connotations. We are taking a mathematical view of sensitivity, and asking how sensitive the output of the inferential procedure is to small changes in all of the inputs.

Third, we note that the relative influence of the three inputs can indicate the presence of data-prior or data-model conflict. Conceptually, we think of data-prior conflict as arising when an estimate of the parameter falls in a low prior probability region of the parameter space. Provided that the parameter has an interpretable meaning as a population quantity, the degree to which the data and prior conflict can be assessed without regard to the choice of model. Data-model conflict arises when the model is a poor fit to the data. A numerical example suggests that when neither conflict is present, the relative influence of the data is high compared to the model and the prior. On the other hand, an elevated relative influence for the prior or model may indicate the presence of a conflict. The caveat is that in some situations both conflicts are present but operate in opposite directions on the posterior. In such cases, the data can still have high relative influence compared to the prior and model.

Both our examples involve the simple setting of estimating the mean of a continuous distribution on the positive reals. In both cases the mean parameter operates as a scale parameter, while the model index controls shape in the first example and tail behavior in the second example. We restrict ourselves to conjugate priors, in order to simplify the sensitivity calculations.

2. An illustration of the method

We find it clearer to describe our methodology in the context of a simple example, instead of delineating it in broad generality. In particular, 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 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 $\mathrm{IG}(a, b)$ denote the inverse gamma density proportional to $z^{-(a+1)}e^{-b/z}$. A prior distribution $\theta \sim \mathrm{IG}(\alpha_1, \alpha_2)$ leads to a posterior distribution of the form $\theta|X = x \sim \mathrm{IG}(\alpha_1^*, \alpha_2^*)$, 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 index α, the model index λ, and the data x. To study the effect of simultaneous small changes to the three inputs we compare the baseline posterior arising from $\omega = (\alpha, \lambda, x)$ to the posterior based on a nearby set of inputs $\tilde{\omega} = (\tilde{\alpha}, \tilde{\lambda}, \tilde{x})$. We measure the discrepancy between these two posteriors by the relative entropy, denoted as

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

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

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

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

and the discrepancy between two models to be

$$d_{\mathrm{M}}(\lambda,\tilde{\lambda}) = D(G(\lambda,\theta/\lambda)||G(\tilde{\lambda},\theta/\tilde{\lambda})). \tag{3}$$

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

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

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

chiefly for convenience but also because it is compatible with our choice of the relative entropy in several senses. First, Eq. (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 $d_{\mathrm{D}}(x,\tilde{x})$ does not depend on n, like $d_{\mathrm{PR}}(\alpha,\tilde{\alpha})$ and $d_{\mathrm{M}}(\lambda,\tilde{\lambda})$ which do not depend on n at all. That is, $d_{\mathrm{D}}(x,\tilde{x})$ tends to a finite, non-zero limit as n tends to infinity. 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 Eq. (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 given baseline inputs ω, our measure of overall sensitivity is based on maximizing $d_{\mathrm{PS}}(\omega,\tilde{\omega})$ subject to a constraint on how far $\tilde{\omega}$ can deviate from ω. Formally, let

$$d_{\mathrm{I}}(\omega,\tilde{\omega}) = d_{\mathrm{PR}}(\alpha,\tilde{\alpha}) + d_{\mathrm{M}}(\lambda,\tilde{\lambda}) + d_{\mathrm{D}}(x,\tilde{x}) \tag{5}$$

be the discrepancy between the two sets of inputs. On one level, we have defined this input discrepancy as the sum of constituent discrepancies simply as a convenient way to introduce neighbourhoods in the input space. But beyond that, the prior, model, and data are three disjoint pieces of information that go into an analysis. Provided that the parameter has an interpretable meaning as a population quantity, it is possible to specify each of the inputs individually, without regard to the others. Therefore, it seems natural to quantify distance between two sets of inputs using the additive form (5).

In a global approach to robustness, the maximum of Eq. (1) as a function of $\tilde{\omega}$, subject to an upper bound on Eq. (5), would be a basic measure of overall posterior sensitivity. Instead of doing this, we find that a computational and conceptual simplification results from 'localizing' the problem as follows. Expand expressions (1) and (5) about ω to get $d_{\mathrm{I}}(\omega,\tilde{\omega}) \approx d^*_{\mathrm{I}}(\omega,\tilde{\omega})$ and $d_{\mathrm{PS}}(\omega,\tilde{\omega}) \approx d^*_{\mathrm{PS}}(\omega,\tilde{\omega})$, where

$$d^*_{\mathrm{I}}(\omega,\tilde{\omega}) = \tfrac{1}{2}(\tilde{\omega}-\omega)^{\mathrm{T}} A_{\mathrm{I}}(\omega)(\tilde{\omega}-\omega), \tag{6}$$

$$d^*_{\mathrm{PS}}(\omega,\tilde{\omega}) = \tfrac{1}{2}(\tilde{\omega}-\omega)^{\mathrm{T}} A_{\mathrm{PS}}(\omega)(\tilde{\omega}-\omega). \tag{7}$$

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 Eq. (5) yields

$$A_{\mathrm{I}}(\omega) = \begin{pmatrix} A_{\mathrm{PR}}(\alpha) & 0 & 0 \\ 0 & A_{\mathrm{M}}(\lambda) & 0 \\ 0 & 0 & A_{\mathrm{D}}(x) \end{pmatrix}, \tag{8}$$

where A_{PR}, A_{M}, and A_{D} are second derivatives arising from Eqs. (2)–(4), respectively. In the cases of the prior and model these second derivatives can be interpreted as Fisher information matrices. In the present example, $A_{\mathrm{PR}}(\alpha)$ is the Fisher information matrix for the $\mathrm{IG}(\alpha_1, \alpha_2)$ family, which evaluates to

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

where ψ' is the trigamma function. Similarly $A_{\mathrm{M}}(\lambda)$ is the Fisher information matrix for the $G(\lambda, \theta/\lambda)$ family, when θ is known, which evaluates to $A_{\mathrm{M}}(\lambda) = \Psi'(\lambda) - \lambda^{-1}$. Finally, we have that $A_{\mathrm{D}}(x) = 2(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{-1} I_n$, where I_n is the $n \times n$ identity matrix.

Analogously, $A_{\mathrm{PS}}(\omega)$ 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 'updated' hyperparameter vector α^* which determines the posterior is a function of the inputs ω. Letting B denote this function, the Fisher information for the posterior distribution is

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

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

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

with derivative

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

Now as a local measure of overall sensitivity we seek the maximum value of Eq. (7) as a function of $\tilde{\omega}$, subject to the constraint that Eq. (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 $k\varepsilon^2$, where k is the largest eigenvalue of $[A_{\mathrm{I}}(\omega)]^{-1} A_{\mathrm{PS}}(\omega)$. This maximum is attained by taking $\tilde{\omega} = \omega + cv$, where v is the eigenvector corresponding to eigenvalue k, and c is a constant chosen so that

Eq. (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 Eq. (5) is present.

The approximations d_{I}^* and d_{PS}^* to d_{I} and d_{PS} 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 overall sensitivity. Note that this definition of overall sensitivity 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

$$v^{\mathrm{T}} A_{\mathrm{I}}(\omega) v = v_{\mathrm{PR}}^{\mathrm{T}} A_{\mathrm{PR}}(\alpha) v_{\mathrm{PR}} + v_{\mathrm{M}}^{\mathrm{T}} A_{\mathrm{M}}(\lambda) v_{\mathrm{M}} + v_{\mathrm{D}}^{\mathrm{T}} A_{\mathrm{D}}(x) v_{\mathrm{D}}, \tag{10}$$

where $v = (v_{\mathrm{PR}}, v_{\mathrm{M}}, v_{\mathrm{D}})$ is the partition of the maximal eigenvector into components corresponding to the prior, model, and data respectively. So, the ratio of $v_{\mathrm{PR}}^{\mathrm{T}} A_{\mathrm{PR}}(\alpha) v_{\mathrm{PR}}$ to $v^{\mathrm{T}} A_{\mathrm{I}}(\omega) v$ is the relative influence of prior uncertainty on the overall sensitivity. The relative influences of the model and data can be reported similarly.

As a numerical illustration with the current example, 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 overall sensitivity of the posterior based on only the first five observations is $k = 1.80$, with relative influence of $(0.13, 0.15, 0.73)$ from the prior, model, and data, respectively. If the first ten observations are used, the overall sensitivity is 4.92, with relative influence of $(0.02, 0.01, 0.96)$. If all twenty observations are considered, the overall sensitivity is 12.42, with relative influence $(0.01, 0.00, 0.99)$. (The reported relative influence entries do not necessarily sum exactly to one because of rounding.) These results show what we can obtain with our sensitivity analysis. Much more detailed numerical results are discussed for our second example in the next section.

3. Example

3.1. A family governing tail behavior

Again, consider estimating the mean θ of a distribution on $(0, \infty)$ which gives rise to independent and identically distributed observations $X = (X_1, \ldots, X_n)$. In this example, suppose the model index λ governs the right tail behavior of the distribution, via 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

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

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

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

where $Z_0 \sim G(1/\lambda, 1)$. We denote the parametric family (11) as $\mathrm{PG}(\lambda, \theta)$ (the P stands for power), and note we have ensured that θ has the same interpretation for all λ, i.e. $E_{\lambda,\theta} Z = \theta$, for all λ.

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

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

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

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

where $\theta_0 \sim \mathrm{IG}(\alpha_1, \alpha_2)$. Let $\mathrm{PIG}(\alpha_1, \alpha_2, \lambda)$ denote the parametric family (13). Then Bayesian updating proceeds as follows. If $X_1, \ldots, X_n$ are independent and identically distributed as $\mathrm{PG}(\lambda, \theta)$, and $\theta \sim \mathrm{PIG}(\alpha_1, \alpha_2, \lambda)$, then $\theta|X = x \sim \mathrm{PIG}(\alpha_1^*, \alpha_2^*, \lambda)$, 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 2, we need the Fisher information matrices for the $\mathrm{PIG}(\alpha_1, \alpha_2, \lambda)$ family, and the $\mathrm{PG}(\lambda, \theta)$ family when θ is known. These quantities are derived in the Appendix. The discrepancy between data sets is again measured using Eq. (4).

The fact that the class of conjugate priors depends on the model index λ necessitates slight changes in the methodology of Section 2. 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 $d_{\mathrm{PR}}(\alpha, \tilde{\alpha})$ by $d_{\mathrm{PR}}(\gamma, \tilde{\gamma})$. This in turn causes a modification to Eq. (8), in that $A_{\mathrm{PR}}(\gamma)$ and $A_{\mathrm{M}}(\lambda)$ will overlap. That is, both the prior discrepancy and the model discrepancy contribute additively to the λ block of A_1. The relationship (9) is still valid, provided that B is considered to map (α, λ, x) to (α^*, λ).

3.2. Computational results

Provided one has reliable data, there are two sorts of modeling errors a Bayesian can make. The prior may conflict with the data, in the sense that nonparametric estimates of the parameter fall in low prior probability regions of the parameter space.

Or, the data may conflict with the model in the sense that the model exhibits lack of fit. Consequently, we characterize four scenarios: no conflicts, data-prior conflict only, data-model conflict only, and both conflicts. Our goal is to investigate the overall sensitivity of the posterior and the relative influence of the inputs in these four scenarios.

Our empirical study proceeds as follows. 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 v_1 and v_2 be the prior mean and standard deviation. Thus v 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 $v_2 = 0.2$ and $v_2 = 0.9$, respectively. Now the degree of prior–data conflict can be summarized in the other hyperparameter v_1. We choose $v_1 = \theta^*$ for prior–data agreement, since x is a vector of quantiles under θ^*. For prior–data conflict we set $v_1 = \theta^* + 2v_2$ so that the true parameter lies two prior standard deviations away from the prior mean.

We take each of the true and assumed model indices, λ^* and λ, to be either 1 or 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 overall sensitivity under the given conditions, and one triple representing the relative influence of the prior, model, and data on the overall sensitivity.

When examining Tables 1–4, it is meaningful to compare the overall sensitivity 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 1–4 have several anticipated properties. First, in each column of each table the overall sensitivity increases with sample size. This is consistent with work of Gustafson and Wasserman (1995) showing that for fixed data and model, the norm of the mapping from prior to posterior increases with sample size. Indeed, regardless of the

Table 1
Overall sensitivity and relative influence 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 v_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	1.47	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 2
Overall sensitivity and relative influence 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 v_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	1.24	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 3
Overall sensitivity and relative influence 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 v_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	1.34	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)

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 influence of the prior on the overall sensitivity, as given by the first entry in each triple, always decreases with sample size. Third, the highest relative influences of the data on the overall sensitivity 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 influence of the prior on the overall sensitivity tends to be larger in the

Table 4
Overall sensitivity and relative influence 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 v_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	1.17 (0.75, 0.07, 0.18)	1.20 (0.71, 0.20, 0.09)	1.64 (0.91, 0.07, 0.02)	1.44 (0.83, 0.15, 0.02)
10	1.69 (0.40, 0.00, 0.60)	1.77 (0.41, 0.30, 0.29)	2.34 (0.78, 0.13, 0.08)	2.19 (0.59, 0.33, 0.08)
20	4.27 (0.11, 0.07, 0.82)	3.61 (0.16, 0.30, 0.53)	3.83 (0.61, 0.13, 0.26)	4.36 (0.34, 0.48, 0.18)

presence of data–prior conflict and the relative influence 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 overall sensitivity 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. In tandem, the two conflicts reinforce and give a larger overall sensitivity. 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 overall sensitivity values in column 2 of Tables 1 and 2 are larger than their counterparts in Tables 3 and 4, respectively.

4. 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 overall sensitivity so that the relative influence of these three inputs can be identified.

Our definitions of overall sensitivity and relative influence of the inputs are quite general. In principle they could be applied to any parameterized collection of parametric models, with any parametric family of priors, at least in the context of independent and identically distributed data. In practice, some work would be needed to compute $A_{\mathrm{PS}}(\omega)$ when non-conjugate priors are involved. However, the computational burden may not be much greater than the basic burden of computing posterior quantities under the baseline inputs. For instance, if the log-likelihood function is $l(\theta|\omega)$ and the log-prior density is $r(\theta|\omega)$ then the Fisher information for the family of posterior distributions indexed by the input vector ω can be expressed as

$$[A_{\mathrm{PS}}(\omega)]_{ij} = \mathrm{Cov}\left(\frac{\partial}{\partial\omega_i}[l(\theta|\omega)+r(\theta|\omega)], \frac{\partial}{\partial\omega_j}[l(\theta|\omega)+r(\theta|\omega)]\right), \tag{15}$$

where the covariance is with respect to the baseline posterior distribution of θ. Thus if Markov chain Monte Carlo methods are used to construct a baseline posterior sample, Eq. (15) can be estimated by a sample covariance. It may be worthwhile to pursue this approach.

The example in Section 3 suggests that for valid inferences the relative influence of the data on the overall sensitivity should be as high as possible, as 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 overall sensitivity may not be high due to a cancellation effect. So we cannot make a non-trivial statement about detecting this case by looking only at the relative influences of the prior and model. Nevertheless, the partitioning of the overall sensitivity as we have defined it here can be regarded as a partial check for model fit and good prior information. In interpreting relative influence, however, it is important to remember that it pertains only to the change in inputs which produces the maximum change in the posterior. It may be possible to obtain a near maximal change in the posterior with an alternate input perturbation that has quite different relative influences for the prior, model, and data.

We caution against extrapolating the qualitative conclusions from Section 3 to other classes of models and priors. For instance, consider the normal distribution with mean θ as the parameter and standard deviation λ as the model index. Under conjugate normal priors, the posterior mean for θ is a weighted average of the prior mean and sample mean, where the weights are determined by λ and the prior standard deviation. As a referee has pointed out, the effect of a change in the prior mean on the posterior mean will be the same regardless of the degree of conflict between the prior mean and the sample mean. Consequently, the interplay of sensitivity and conflict in this example may be quite different from that of Section 3.2, where the model index controls tail behavior.

We also note that 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 example 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 A.

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 $\Psi(s) = \partial/\partial s \log\Gamma(s)$ is the digamma function. The trigamma function is denoted $\Psi'(s) = \partial/\partial s \Psi(s)$.

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

$$d_{\mathrm{PR}}(\gamma,\tilde{\gamma}) = \log\left(\frac{\lambda\alpha_2^{\alpha_1}\Gamma(\tilde{\alpha}_1)}{\tilde{\lambda}\tilde{\alpha}_2^{\tilde{\alpha}_1}\Gamma(\alpha_1)}\right) + (\tilde{\alpha}_1\tilde{\lambda} - \alpha_1\lambda)E_{\alpha,\lambda}(\log\theta) + \tilde{\alpha}_2 E_{\alpha,\lambda}(\theta^{-\tilde{\lambda}}) - \alpha_2 E_{\alpha,\lambda}(\theta^{-\lambda}). \tag{A.1}$$

The expectations are easily evaluated by substituting the right-hand side of Eq. (14) into Eq. (A.1), and then applying the above-mentioned facts. The resulting expression is

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

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

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

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 Eqs. (11) and (12) we see that

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

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

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

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

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

The Bayesian updating takes the form

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

This leads to a derivative

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

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 (v_1, v_2), note that under Eq. (13), $E_{\alpha,\lambda}(\theta)$ is given by

$$v_1 = \alpha_2^{1/\lambda} \frac{\Gamma(\alpha_1 - 1/\lambda)}{\Gamma(\alpha_1)},$$

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

$$v_2^2 + v_1^2 = \alpha_2^{2/\lambda} \frac{\Gamma(\alpha_1 - 2/\lambda)}{\Gamma(\alpha_1)},$$

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

$$\frac{\Gamma(\alpha_1 - 2/\lambda)\Gamma(\alpha_1)}{(\Gamma(\alpha_1 - 1/\lambda))^2} = 1 + \frac{v_2^2}{v_1^2},$$

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

$$\alpha_2 = \left(\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1 - 1/\lambda)} v_1 \right)^{\lambda}.$$

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.

Cook, R.D., 1986. Assessment of local influence (with discussion). J. Roy. Statist. Soc. B 48, 133–169.

Dey, D.K., Birmiwal, L.R., 1994. Robust Bayesian analysis using entropy and divergence measures. Statist. Probab. Lett. 20, 287–294.

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

Gustafson, P., 1996. Local sensitivity of inferences to prior marginals. J. Amer. Statist. Assoc. 91, 774–781.

Gustafson, P., Wasserman, L., 1995. Local sensitivity diagnostics for Bayesian Inference. Ann. Statist. 23, 2153–2167.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., 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., Kadane, J.B., 1989. Approximate methods for assessing influence and sensitivity in Bayesian analysis. Biometrika 76, 663–674.

Lavine, M., 1991. Sensitivity in Bayesian statistics: the prior and the likelihood. J. Amer. Statist. Assoc. 86, 396–399.

McCulloch, R.E., 1989. Local model influence. J. Amer. Statist. Assoc. 84, 473–478.

Neuhaus, Kalbfleisch, Hauck, 1992. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika 79, 755–762.

Peng, F., Dey, D.K., 1995. Bayesian analysis of outlier problems using divergence measures. Can. J. Statist. 23, 199–213.

Ruggeri, F., Wasserman, L., 1993. Infinitesimal sensitivity of posterior distributions. Can. J. Statist. 21, 195–203.

Sivaganesan, S., 1993. Robust Bayesian diagnostics. J. Statist. Plann. Inference 35, 171–188.

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

Tsou, T.S., Royall, R.M., 1995. Robust likelihoods. J. Amer. Statist. Assoc. 90, 316–320.

Weiss, R.E., 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.