Assessing Posterior Robustness Using Divergence Ratio

Mithat Gönen
Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10021
email: gonenm@mskcc.org

Elza Erkip
Polytechnic University, Department of Electrical and Computer Engineering, Six Metrotech Center, Brooklyn, NY 11201
email: elza@poly.edu

SUMMARY. Sensitivity to the choice of prior is a major concern in Bayesian analysis. In this article we propose a method based on the ratio of divergences between the prior and posterior densities to assess the sensitivity of the posterior to changes in the prior. Our method is designed to detect changes in the entire posterior distribution and can be used as a first check to determine whether further sensitivity analysis is needed. We provide an approximation that can be used with contamination classes of priors and can be easily calculated either with conjugacy or posterior simulation. We demonstrate the benefits of our method with published Bayesian analyses from the literature: a meta-analysis of environmental tobacco smoke and early stopping of a clinical trial for unacceptable toxicity.

KEY WORDS: robust Bayes, sensitivity analysis, Kullback-Leibler distance, $\epsilon$-contamination class
1 Introduction

One of the persistent criticisms raised against the use of Bayesian methods in the analysis of medical data is the possible sensitivity of conclusions to the selection of prior. Some applied statisticians regard Bayesian methods with skepticism, mostly because they think of it as a way of arriving at the desired conclusions by choosing an appropriate prior. Such objections can be addressed by providing a sensitivity analysis considering alternative priors and observing the changes in the posterior. If the changes in posterior are relatively small, then the analysis is said to be “posterior robust.” Berger (1985 Section 4.7, 1984, 1989 and 1994) summarizes several methods through which such a sensitivity analysis can be carried out. The central idea is to elicit a class of priors instead of a single distribution, evaluate a ‘target’ feature of the posterior on which the conclusion will be drawn, say a credible set, and then vary prior over the class identified by the elicitation process to see how the results are affected. Spiegelhalter, Freedman and Parmar (1993, 1994) gave examples from realistic clinical trial settings as to how such different priors can be necessary: a “clinical” prior representing the typically optimistic beliefs of the investigators, a “skeptical” prior that might be suitable for a regulatory agency approaching the experiment with a certain degree of suspicion and a “non-informative” prior representing lack of prior information (which usually falls somewhere in between the clinical and skeptical priors).

While a few priors can be selected using the approach of Spiegelhalter, Freedman and Parmar, there are more general ways of eliciting a larger class
which might contain those few priors of interest. One such class that has attracted attention in the literature is the contamination class. For general properties of contamination classes see Goldstein (1982) and for applications in robust Bayesian analysis, see, for example, Berger and Berliner (1986); Sivaganesan, Berliner and Berger (1993) or Greenhouse and Wasserman (1995).

In this article, we propose a method to evaluate robustness using divergence, also known as Kullback-Leibler distance, between the priors of choice and the corresponding posterior densities. Our method has the distinction of assessing the robustness of the entire posterior as opposed to a single feature. Hence it can be useful as a first check before any lengthy sensitivity analysis is undertaken. It can provide a degree of comfort, it might point to a serious problem or it might be used to justify and dictate the extent of further efforts to perform sensitivity analysis. We will demonstrate the uses of the proposed method on two examples from the literature: a random-effects meta-analysis of environmental tobacco smoke (Givens, Smith and Tweedie, 1997) and monitoring of clinical trials (Greenhouse and Wasserman, 1995).

Using divergence to perform sensitivity analysis in a Bayesian framework is an idea dating back to Johnson and Geisser (1983). Other works with a similar premise include McCulloch (1989), Gelfand and Dey (1991) and Dey and Birmiwal (1994). Our approach is mainly distinguished by the way we calibrate the posterior divergence which results in an easily-computed approximation via posterior simulation in the case of contamination classes. This can be beneficial for the practitioner since contamination classes are commonly used for robust Bayesian analysis in practice.
2 Using Divergence to Assess Sensitivity

We will first define divergence, also called Kullback-Leibler divergence or Kullback-Leibler distance. For any two probability distributions $f_1$ and $f_2$ (provided that the support of $f_1$ is a subset of the support of $f_2$) the divergence between $f_1$ and $f_2$ is given by (Cover and Thomas, 1991)

$$D(f_1||f_2) = \int \log \frac{f_1}{f_2} df_1.$$  \hspace{1cm} (1)

For applications of divergence in statistics, especially in relation to exponential families and maximum likelihood estimation, see Brown (1986). Divergence is always non-negative and is zero if and only if $f_1 = f_2$. It is not symmetric, however, and does not satisfy the triangle inequality either. Hence it is not a metric in strict terms. Nevertheless it is often useful in assessing how “far” two distributions are (Kullback, 1959).

We will use $x$ to denote the observables and $\theta$ the parameters in a generic Bayesian model. Let $\pi(\theta)$ be the prior, $L(x|\theta)$ the likelihood and $P(\theta|x)$ the posterior. These three are related to each other via the Bayes’ Theorem:

$$P(\theta|x) = \frac{L(x)\pi(\theta)}{m(x)}$$

where $m(x)$ is called the marginal (or predictive) density of the observables.

Consider two prior distributions, $\pi_1$ and $\pi_2$, both of which are deemed to be reasonable for the problem under consideration and let $P_1$ and $P_2$ be the corresponding posteriors for a given $x$. Let

$$D_\pi = D(\pi_1||\pi_2)$$

be the divergence between the priors and

$$D_P = D(P_1||P_2)$$
be the divergence between the posteriors. Note that $D_\pi$ does not depend on $x$, but $D_P$ does. $D_P$ contains some information regarding the sensitivity of the model; if it is small then the posteriors are close to each other and hence the model is not sensitive to the choice of $\pi_1$ or $\pi_2$. An immediate problem is the calibration of $D_P$; we don’t know what a “small” $D_P$ means. Here we suggest using “divergence ratio” which we define as

$$DR = \frac{D_P}{D_\pi}$$

(2)

For brevity, we will slightly abuse notation and suppress the dependence of $DR$ on $\pi_1$, $\pi_2$ and $x$ unless necessary to avoid confusion. Hence $DR$, $DR(x)$ and $DR(\pi_1, \pi_2)$ will all refer to the same quantity where $\pi_1$, $\pi_2$ and $x$ should be inferred from context when left unspecified.

If $D_P$ is much smaller than $D_\pi$ then our analysis is likely to be robust; the distance shrunk considerably from the baseline. On the other hand $D_P$ can be small but if $D_\pi$ was also small and $DR$ is calculated to be moderately large, then this can be considered an inconclusive evaluation: the two priors were so close to start with that we cannot claim robustness by the fact that $D_P$ is small. McCullach (1989), who also considered using $D_P$, suggested a calibration by a “mapping” from the set of all densities to the set of all Bernoulli densities. His method involves using $D_P$ only.

It is important to note that $E_{m_1(x)}DR(x) = \int DR(x)m_1(x) < 1$, which easily follows from Ahlswede and Gács (1976). Hence observing $DR < 1$ may not necessarily indicate robustness, but $DR > 1$ certainly raises a red flag. This is due to the fact that the information contained in $x$ will, on the average, bring the posteriors closer than priors. Therefore it is important
to familiarize ourselves with the scale of $DR$. We will start with a simple example.

**Example 1** (Berger, 1985, p. 195) We observe $X \sim N(\theta, 1)$. We also specify that $\pi(\theta)$ has median 0 and quartiles ±1. Either a Cauchy$(0,1)(\pi_C)$ or a $N(0,2.19)(\pi_N)$ density will match these prior specifications. A simple calculation gives us $D_\pi = D(\pi_C||\pi_N) = 0.1860$. Table 1 provides $D_P$ and $DR$ for different observed values of $X = x$.

[INSERT TABLE 1 HERE]

For lower values of $x$ (0, 1 and 2) we have excellent robustness. As $x$ gets larger however, the two posteriors diverge from one another, indicating sensitivity to the choice of prior. This conclusion agrees with that of Berger, where he analyzes sensitivity using posterior means. This example also serves to remind us that robustness in the Bayesian sense is conditional on $x$, in particular a model can be robust or sensitive to the choice of prior depending on the observed values of $x$.

We will next consider a sensitivity analysis performed in the context of a random-effects meta-analysis model using different priors.

**Example 2** Givens, Smith and Tweedie (1997, GST hereafter) present a Bayesian meta-analysis of the risks of environmental tobacco smoke (ETS) using a hierarchical model suggested by DuMouchel (1990). There are $J = 35$ published studies which report the relative risk (RR) of the exposed group. The random effect model they consider is

$$(\log RR)_j = \Delta + \beta_j + \epsilon_j, \quad j = 1, \ldots, 35$$

where $\Delta$ is an overall mean, $\beta_j \sim N(0, \tau^2)$ is a random effect introduced
to account for heterogeneity between studies, and $\epsilon_j \sim N(0, \sigma_j^2)$ represents the within-study variability. GST used inverse gamma priors for the variance components $\tau^2$ and $\sigma^2_j$ and a $N(0, 0.15^2)$ (call this $\pi_0$) prior for $\Delta$. To evaluate the sensitivity of their conclusions to the prior for $\Delta$, they considered three alternatives:

1. A more informative prior $\pi_1 = N(0, 0.10^2)$

2. An empirical prior $\pi_2 = N(0.1133, 0.11^2)$

3. A more uninformative prior $\pi_3 = N(0, 0.4^2)$

(labels “more informative,” “empirical” and “more uninformative” are introduced by GST). Realizing the need for a sensitivity analysis, they calculated the posterior mean and a 95% posterior interval for each of the three scenarios and argued that their analysis is robust to changes in the prior for their observed values of relative risks. It is interesting to note that all the discussants of the GST article pointed out robustness as a potential issue, leaving the readers with the feeling that they were not convinced by the sensitivity analysis presented.

Here we will use divergence ratio as defined in (2) to assess their claim of robustness. Table 2 shows values of $D_\pi = D(\pi_0||\pi_i)$, $D_P = D(P_0||P_i)$ and $DR = D_P/D_\pi$, $i = 1, 2, 3$ for their observed data.  

[INSERT TABLE 2 HERE]

The table supports the authors’ claim that their analysis is not sensitive to the choice of prior among $\pi_1$, $\pi_2$ or $\pi_3$ since $DR$ is reasonably small for all cases. The values of $DR$ observed here are in the same range of the robust situations ($x = 0, 1, 2$) from the previous example.
The original claims of Berger and GST in these examples were based on a single or a few selected features of the posterior. Our conclusions using divergence ratio agrees with these methods. Since $DR$ is an overall measure based on the entire posterior, we can conclude that, in both of these examples, the entire posterior distributions were close to each other.

In the next section we will consider an example where different features of the posterior might yield different conclusions regarding robustness and how $DR$ can be used to become aware of this situation in a relatively easy fashion.

3 Contamination Classes of Priors

The examples we considered in the previous section involved performing sensitivity analysis using a few different priors. It is more convincing to use a class with (possibly infinitely) many prior distributions. In this section we will investigate how $DR$ can be calculated when contamination classes are used as priors.

Let $\pi_0$ denote a 'base' prior selected for our problem. It can be the result of preliminary elicitation, or a conjugate prior chosen for its analytical convenience. Let $Q$ denote a class of priors each of which might also serve well for the problem under consideration, but either considered less suitable than $\pi_0$ or less convenient. Define

$$\Gamma_\epsilon = \{ \pi = (1 - \epsilon)\pi_0 + \epsilon q, \; q \in Q \},$$

which is called as an $\epsilon$-contamination class. As Berger (1994) argues, this class is easy to elicit, relatively easy to compute and amenable for use in
higher dimensions. It can be undesirably big, depending on the choice of \( Q \) and \( \epsilon \). This class has attracted some attention; for example, Greenhouse and Wasserman (1995) analyzed early stopping of a clinical trial via a contamination class. We will use their example to illustrate the use of \( DR \) with \( \epsilon \)-contamination classes.

When working with a contamination class of priors, a reasonable instrument to assess sensitivity is \( \sup_{\pi_1,\pi_2 \in \Gamma_\epsilon} DR(\pi_1,\pi_2) \). However, evaluation of divergences over the entire class in search of the supremum value of \( DR \) can be a difficult exercise. One way to circumvent this problem is to consider the case of \( \epsilon \to 0 \), which is called “local influence” or “infinitesimal robustness.” Local influence has attracted a good deal of attention in Bayesian robustness literature, see, for example McCulloch (1989) or Gustafson, Srinivasan and Wasserman (1996). Following lemma provides an approximation which can be used to assess local influence using \( DR \).

**Lemma 1** Let \( q_1 \) and \( q_2 \) be in \( Q \). Let \( D(\pi_1||\pi_2) \) be the distance between the two priors \( \pi_1 = (1-\epsilon)\pi_0 + \epsilon q_1 \) and \( \pi_2 = (1-\epsilon)\pi_0 + \epsilon q_2 \) and let \( D(P_1||P_2) \) be the distance between the corresponding posteriors. If we let \( u_i = q_i/\pi_0, \ i = 1,2 \) then, for small \( \epsilon \)

\[
DR = \frac{D(P_1||P_2)}{D(\pi_1||\pi_2)} \approx \frac{\text{Var}_{P_0}(u_1 - u_2)}{\text{Var}_{\pi_0}(u_1 - u_2)}.
\]

Therefore

\[
\sup_{\pi \in \Gamma_\epsilon} DR \approx \sup_{q_1, q_2 \in Q} \frac{\text{Var}_{P_0}(\frac{q_1-q_2}{\pi_0})}{\text{Var}_{\pi_0}(\frac{q_1-q_2}{\pi_0})}. \tag{5}
\]
In effect, this lemma tells us that $DR$ for any two members of the class can be approximated by the ratio of the variance of the difference of the two contaminating densities divided by the base prior (note that $u_1 - u_2 = [q_1 - q_2]/\pi_0$), where the variances are evaluated with respect to the base posterior or the base prior. Here division by $\pi_0$ plays the role of “normalization” with respect to the base prior. This result has intuitive appeal. If the normalized difference of the two contaminating densities (hence two members of the class) has large variation with respect to the base posterior, then it is likely that the posteriors under consideration will be far apart. A similar interpretation holds for the denominator as well. Note that $u_1 - u_2$ is not enough to assess the robustness, it is the variability of $u_1 - u_2$ with respect to the base prior (or posterior) that plays the major role.

The approximation presented in the lemma requires the calculation of only $P_0$ and not any of the posteriors resulting from the use of $\Gamma_e$. This has important implications for the practitioner; first of all it is sufficient to deal only with one density and still make conclusions about the entire class. In addition, the contaminated densities are usually not easy to work with, but $P_0$ is. If $\pi_0$ is chosen for convenience (such as conjugacy) then $P_0$ will be readily available. Even if this is not the case, the analyst usually has the tools (such as Markov Chain Monte Carlo) to generate a sample from $P_0$. It is very simple to use this approximation with such a sample.

The following example describes a situation where two statisticians needed an assessment of robustness and turned to contaminations classes. We will use their approach and demonstrate how $DR$ can be used within the context of contamination classes to assess robustness.
Example 3 Greenhouse and Wasserman (1995, GW hereafter) evaluated the decision of Korn et al. (1993) to stop a clinical trial early for unacceptable toxicity from a robust Bayesian standpoint. The trial was investigating whether G-CSF, a biologic agent, would decrease the incidence of mucositis, a common side-effect of 5-FU/LV chemotherapy. In a previous study of 5-FU/LV without G-CSF, 12 of the 176 patients experienced grade 4 leukopenia. The design did not accommodate an early stopping rule. When three of the first four patients (out of a planned total of 35) presented with grade 4 leukopenia the investigators asked for the opinion of statisticians whether it is advisable to stop the trial early for safety.

We will follow GW and model the incidence of leukopenia in a given patient using a Bernoulli distribution with parameter $\theta$. The conjugate prior for $\theta$ is the beta family with shape parameter $\alpha$ and scale parameter $\beta$, denoted by $B(\alpha, \beta)$. They advocate the following contamination class

$$\Gamma_a = \{ \pi = (1 - \epsilon)\pi_0 + \epsilon q, \quad q \in Q_a \},$$

where $\pi_0$ is $B(1.56, 8.44)$ and $Q_a$ is the class of all probability distributions. Note that $\pi_0$ represents an effective sample size of 10 and has mode 0.07 (around the observed rate of the previous trial) with mean 0.016 and variance 0.012. Using $\epsilon = 0.1$ they report that the $E(\theta|x)$ varies between 0.78 to 0.96 and $P(\theta > 0.2|x)$ varies between 0.31 and 0.62. Based on their results they claim that their analysis is posterior robust.

Here we will work with

$$\Gamma_b = \{ \pi = (1 - \epsilon)\pi_0 + \epsilon q, \quad q \in Q_b \},$$

where $Q_b$ is the family of all beta distributions. We argue that $\Gamma_a$ contains
some unreasonable distributions such as the ones contaminated by degenerate densities. On the other hand beta family contains several distributions on the unit interval which can approximate a wide variety of beliefs. It has the additional advantage of lending itself to simple computational schemes.

Using the lemma, we calculate \( \sup_{\pi \in \Gamma_b} DR = 0.56 \). It is certainly the case that \( \sup_{\pi \in \Gamma_a} DR \geq 0.56 \). While this is not an immediate indication of lack of robustness, it certainly raises some concerns. At first this conclusion might seem to disagree with that of GW. However they limit their investigation to two key features of the posterior: \( E(\theta|x) \) and \( P(\theta > 0.2|x) \). In fact, without reporting any quantitative results, they mention that \( P(\theta > 0.5|x) \) is sensitive to the choice of prior. Since \( DR \) is an overall measure of robustness, it reflects the lack of robustness in some features here.

4 Conclusion

In this article we provided a measure of the robustness of Bayesian inferences with respect to the changes in the prior. Our measure uses the divergence or Kullback-Leibler distance between two probability distributions and lends itself to easily-computed analytic approximations in the case of \( \epsilon \)-contamination classes, which constitute a popular way of carrying out robust Bayesian analysis.

An important feature of our method is that it does not focus on a single feature of the posterior and is an overall measure. Current practice in Bayesian sensitivity analysis usually involves a sizable effort without a clear “stopping rule.” We believe that cases like Example 3 are the ones that will
benefit most from the use of divergence ratio. We propose that divergence ratio can be used as a first-step filter to the sensitivity analysis. The fact that it is computationally easy, even in the case of contamination classes as shown in Lemma 1, makes this proposal more feasible since this first-step check for sensitivity can be carried out with little effort. An assessment of robustness can start with an evaluation of $DR$ over the chosen class of priors. If $DR$ is small, as in Examples 1 and 2, there is little concern for lack of robustness; it is likely that all the posteriors in the family are close to each other as a whole and we have posterior robustness. If $DR$ raises a red flag (or at least some concern, such as in Example 3), then a detailed sensitivity analysis is justified to identify the particulars of lack of robustness. Further, $DR$, together with $D_\pi$, can be used to assess if the chosen class of priors is rich enough to convince ourselves that our sensitivity analysis is adequate. A relatively small $D_\pi$, along with a moderately large $DR$ could be indicative of this problem, whereas sensitivity analysis using the information from posterior only could be misleading in such a situation.

Summarizing posterior robustness with a single number simplifies the presentation and provides a quick understanding of the problem, but this comes at the expense of possibly overlooking important details. We do not intend $DR$ to be used as the only measure of robustness in assessing sensitivity to the prior. Ideally, nothing should replace a detailed, attentive assessment of robustness. In reality, however, we are forced to deal with multiple projects which compete for our time and effort and can benefit from using an overall summary measure like $DR$. 

12
References


meta-analysis: A Bayesian data-augmentation to account for issues exemplified in the passive smoking debate.” *Statistical Science*, 12, 221–250.


Appendix

Here we provide the proof of Lemma 1 in Section 3. Let $\Gamma_\epsilon$ be as defined in (3) and let $\pi_1$ and $\pi_2$ be in $\Gamma_\epsilon$. We first work with $D(\pi_1||\pi_2)$. First note that

\[
\frac{\pi_1}{\pi_2} = \frac{\pi_1/\pi_0}{\pi_2/\pi_0} = \frac{(1-\epsilon)\pi_0 + q_1}{(1-\epsilon)\pi_0 + q_2} = \frac{1 + \epsilon(u_1 - 1)}{1 + \epsilon(u_2 - 1)}
\]

where

\[
u_i = q_i/\pi_0.
\]

Hence

\[
\log \frac{\pi_1}{\pi_2} = \log \{1 + \epsilon(u_1 - 1)\} - \log \{1 + \epsilon(u_2 - 1)\}
\]

and

\[
D(\pi_1||\pi_2) = E_{\pi_1} \log \{1 + \epsilon(u_1 - 1)\} - E_{\pi_1} \log \{1 + \epsilon(u_2 - 1)\}
\]

For small $\epsilon$ provided that $q_i(\theta)/\pi_0(\theta)$ is finite for all $\theta$, we have $\epsilon(u_i - 1)$ small. We can then use a second-order Taylor approximation to get

\[
\log \{1 + \epsilon(u_i - 1)\} \approx \epsilon(u_i - 1) - \frac{\epsilon^2}{2}(u_i - 1)^2, \quad i = 1, 2.
\]

Using this approximation, we can evaluate (9). Consider $i = 1$ first:

\[
E_{\pi_1} \log \{1 + \epsilon(u_1 - 1)\} \approx \epsilon \int \pi_1 u_1 - \frac{\epsilon^2}{2} \left[ \int \pi_1 u_1^2 - 2 \int \pi_1 u_1 \right] - \left( \epsilon + \frac{\epsilon^2}{2} \right)
\]

(10)
It follows from (8) that the two integrals in (10) are given by

\[
\int \pi_1 u_1 = (1 - \epsilon) + \epsilon E_{\pi_0} u_1^2 \\
\int \pi_1 u_2^2 = (1 - \epsilon) E_{\pi_0} u_1^2 + \epsilon E_{\pi_0} u_1^3
\]

which, when substituted in (10), gives

\[
E_{\pi_1} \log \{1 + \epsilon(u_1 - 1)\} \approx \left( \epsilon^3 - \frac{\epsilon^2}{2} \right) + \left( \frac{\epsilon^2}{2} + \frac{3\epsilon^3}{2} \right) E_{\pi_0} u_1^2 - \frac{\epsilon^3}{2} E_{\pi_0} u_1^3 (11)
\]

Similar steps can be traced for the second term:

\[
E_{\pi_1} \log \{1 + \epsilon(u_2 - 1)\} \approx E_{\pi_1} \left[ \epsilon(u_2 - 1) - \frac{\epsilon^2}{2} (u_2 - 1)^2 \right] (12)
\]

\[
\approx (\epsilon + \epsilon^3) \int \pi_1 u_2 - \frac{\epsilon^2}{2} \int \pi_1 u_2^2 - \left( \epsilon + \frac{\epsilon^2}{2} \right) (13)
\]

and the two integrations can be carried out as follows:

\[
\int \pi_1 u_2 = (1 - \epsilon) + \epsilon E_{\pi_0} u_1 u_2 \\
\int \pi_1 u_2^2 = (1 - \epsilon) E_{\pi_0} u_1^2 + \epsilon E_{\pi_0} u_1 u_2
\]

which when substituted back in (13) leads to

\[
E_{\pi_1} \log \{1 + \epsilon(u_2 - 1)\} \approx \\
\left( \epsilon^3 - \frac{\epsilon^2}{2} \right) + (\epsilon^2 + \epsilon^3) E_{\pi_0} u_1 u_2 - \left( \frac{\epsilon^2}{2} - \frac{\epsilon^3}{2} \right) E_{\pi_0} u_2^2 - \frac{\epsilon^3}{2} E_{\pi_0} u_1 u_2 (14)
\]

Now using (11) and (14) in (9), we get

\[
D(\pi_1 || \pi_2) = \frac{\epsilon^2}{2} E_{\pi_0} (u_1 - u_2)^2 + \epsilon^3 \text{ terms} (15)
\]

Further manipulation shows that $E_{\pi_0} (u_1 - u_2) = 0$, hence

\[
D(\pi_1 || \pi_2) = \frac{\epsilon^2}{2} \text{Var}_{\pi_0} (u_1 - u_2) + \epsilon^3 \text{ terms} (16)
\]
Calculation of $D_P$ follows along similar lines. The key identities are

$$p_i = \frac{L(\pi_0 + \epsilon(q_i - \pi_0))}{m_0 + \epsilon(m_i - m_0)}$$

(17)

$$\approx p_0 + \epsilon p_0 \left( u_i - \frac{m_i}{m_0} \right) + \epsilon^2 p_0 \left( \frac{m_i}{m_0} - 1 \right)^2$$

$$-\epsilon^2 \frac{p_0}{m_0} (u_i - 1)(m_i - m_0)$$

(18)

for $i = 1, 2$. Here $L$ is the likelihood and $m_0$ is the marginal corresponding to $\pi_0$ as defined in Section 2. Also $m_i(x) = \int L(x|\theta)q_i(\theta)$ refers to the marginal density of $x$ resulting from using $q_i$ as the prior. For the approximation, we use a Taylor series expansion on $1/(m_0 + \epsilon(m_i - m_0))$ and only keep terms that contain up to $\epsilon^2$.

By noting that $E_{P_0} u_i = m_i/m_0$ (recall that $E_{\pi_0} u_i = 0$) we find

$$D_P = \frac{\epsilon^2}{2} \text{Var}_{P_0}(u_1 - u_2) + \epsilon^3 \text{terms}$$

(19)

where $P_0$ is the posterior corresponding to $\pi_0$. Therefore

$$DR = \frac{D(P_1||P_2)}{D(\pi_1||\pi_2)} = \frac{\text{Var}_{P_0}(u_1 - u_2)}{\text{Var}_{\pi_0}(u_1 - u_2)}$$

(20)

and

$$\sup_{\pi \in \Gamma_\epsilon} DR = \sup_{q_1, q_2 \in Q} \frac{\text{Var}_{P_0}(\frac{q_1 - q_2}{\pi_0})}{\text{Var}_{\pi_0}(\frac{q_1 - q_2}{\pi_0})}.$$  

(21)

This completes the proof.
Table 1 Values of $D_P$ and $DR$ for different observed values of $X = x$

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4.5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_P$</td>
<td>0.0239</td>
<td>0.0295</td>
<td>0.0311</td>
<td>0.4279</td>
<td>4.2190</td>
</tr>
<tr>
<td>$DR$</td>
<td>0.1281</td>
<td>0.1587</td>
<td>0.1672</td>
<td>2.3006</td>
<td>22.6828</td>
</tr>
</tbody>
</table>

Table 2 Values of $D_\pi = D(\pi_0||\pi_i)$, $D_P = D(P_0||P_i)$ and $DR = D_P/D_\pi$, where $i = 1, 2, 3$ correspond to more informative, empirical and more uninformative priors respectively

<table>
<thead>
<tr>
<th></th>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
<th>$\pi_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_\pi$</td>
<td>0.215</td>
<td>0.647</td>
<td>0.552</td>
</tr>
<tr>
<td>$D_P$</td>
<td>0.042</td>
<td>0.052</td>
<td>0.010</td>
</tr>
<tr>
<td>$DR$</td>
<td>0.195</td>
<td>0.080</td>
<td>0.002</td>
</tr>
</tbody>
</table>