Isoseparation and Robustness in Finite Parameter Bayesian Inference

Jim Q. Smith[*], Fabio Rigat[†]

February 1, 2008

Abstract

Under a new family of separations the distance between two posterior densities is the same as the distance between their prior densities whatever the observed likelihood when that likelihood is strictly positive. Local versions of such separations form the basis of a weak topology having close links to the Euclidean metric on the natural parameters of two exponential family densities. Using these local separation measures it is shown that when the tails of the approximating density have appropriate properties, the variation distance between an approximating posterior density to a genuine density can be bounded explicitly. These bounds apply irrespective of whether the prior densities are grossly misspecified with respect to variation distance and irrespective of the form or the validity of the observed likelihood.

Keywords: density ratio class, hierarchical Bayesian inference, high dimensional inference, local robustness, parametric Bayes, total variation.

1 Introduction

Let f_0 denote the *functioning prior* over a finite parameter vector $\theta \in \Theta$ - i.e. the density actually used in a Bayesian analysis - and g_0 the *genuine prior*:

[*]Department of Statistics, University of Warwick, Coventry CV4 7AL, UK; J.Q.Smith@warwick.ac.uk

[†]CRiSM, Department of Statistics, University of Warwick, Coventry CV4 7AL, UK; f.rigat@warwick.ac.uk

- i.e. the one that would be used if there was enough time and skill applied to elicit it perfectly. Denote the two corresponding posterior densities after observing a sample $\mathbf{x}_n$, $n \geq 1$ of n observations by f_n and g_n respectively. In this paper a new family of separation measures on prior densities is examined. Using the properties of these separations it is possible to derive conditions ensuring that the posterior density f_n is a good approximation for g_n when n is large, where following other authors closeness in posterior densities is measured by *variation distance* $d_V(f_n, g_n) = \int_\Theta |f_n(\theta) - g_n(\theta)| d\theta$.

The problem of posterior robustness is commonly addressed by first assuming that both the sequences of posterior functioning densities $\{f_n\}_{n\geq 1}$ and genuine $\{g_n\}_{n\geq 1}$, are consistent. For example, [16] proved (p.439) that such consistency will automatically follow provided there is a consistent estimator of θ in a finite dimensional parameter space Θ and the parametrisation of θ respects Kullback-Leibler separations in the sense of Theorem 7.80 in [16]. It follows that when each component of a random sample $\mathbf{x}_n$ is drawn from a distribution labelled by a parameter $\theta_0 \in \Theta$ and whenever $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$ are both continuous at $\theta_0 \in \Theta^0$ where Θ^0 is the interior of Θ, $\lim_{n\to\infty} d_V(f_n, g_n) = 0$ almost surely P_{θ_0}, e.g. [8], p.18. This in turn implies that f_n provides a good working approximation for g_n for *all* reasonable estimation purposes in the sense above.

These results rely heavily on the assumption that the sample family is precisely and correctly specified. Typically this is rarely credible or verifiable. A more useful result would be that $\lim_{n\to\infty} d_V(f_n, g_n) = 0$ whenever the functioning posterior distribution concentrates on a point θ_0 on the closure of the parameter space Θ.

In this paper we prove that under a new family of local separation measures, this property holds for *any* given *observed* likelihood, even when the sample distribution family is not accurately specified and the data is not a random sample. In Section 2 we introduce a of separation measures, closely related to density ratio separation measures [7], [20], [14], but redefined so that they apply locally. In Section 3 various useful properties of these separations are examined. In Section 4 it is shown that in the limit they can be used to compare the relative roughness of two prior densities and provide a very coarse topology where a Bayesian might plausibly believe that her functioning and genuine prior to be close.

In Section 5 the "isoseparation property" of the new separations is used to prove that, provided the genuine prior density lies in one of these coarse neighbourhoods of the functioning prior, closeness in variation between the

functioning and genuine posteriors is almost guaranteed. Furthermore *explicit* bounds for the approximation can be calculated and apply irrespective of whether the observed data is consistent with the family of sample distributions underpinning the observed likelihood. This makes the results here of considerable practical significance to the study of the robustness to prior specification in high dimensional parametric inference and in particular for hierarchical models.

2 Density ratio balls and Isoseparation

Henceforth for simplicity assume that all candidate genuine priors $g_0(\theta)$ and the functioning prior $f_0(\theta)$ are strictly positive and continuous on the interior of their shared support Θ so that they are uniquely defined. Assume a sequence of *observed* sample densities $\{p(\mathbf{x}_n|\theta)\}_{n\geq 1}$ of an n -vector of data $\mathbf{x}_n = (x_1, x_2, \ldots x_n)$ are measurable with respect to $g_0(\theta)$. Let $\Theta(n) = \{\theta \in \Theta : p(\mathbf{x}_n|\theta) > 0\}$ and assume that $p(\mathbf{x}_n|\theta)$ is continuous on $\Theta(n)$. The *formal Bayesian updating formula* calculates the posterior density $g_n(\theta) \triangleq g(\theta|\mathbf{x}_n)$ after n observations for all $\theta \in \Theta(n)$ using the equation

$$\log g_n(\theta) = \log g_0(\theta) + \log p(\mathbf{x}_n|\theta) - \log p_{g_0}(\mathbf{x}_n), \tag{1}$$

where the predictive density $p_{g_0}(\mathbf{x}_n) = \int_{\theta\in\Theta(n)} p(\mathbf{x}_n|\theta)g_0(\theta)d\theta$ is usually calculated indirectly, either algebraically or numerically, so as to ensure $g_n(\theta)$ integrates to 1. For all $\theta \in \Theta\backslash\Theta(n)$ we simply set $g_n(\theta) = 0$. In this paper we assume that our inference is sufficiently regular that it is appropriate to use this formula, both for g_0 and f_0.

Let $f_n(\theta) \triangleq f(\theta|\mathbf{x}_n)$, $\theta \in \Theta(n)$ represent our functioning posterior density after the first n observations and suppose that the functioning posterior density $f_n(\theta) \in C^{\alpha_n}(\theta_0^n, \rho_n)$ converges in distribution as $n \to \infty$ to a point mass in the closure neighbourhood of $\theta_0 \in \Theta(n)$. It has long been known - see e.g. [5], [13], [1] - that when the functioning posterior convergences to a defective distribution the variation distance between f_n and g_n cannot be guaranteed to converge to 0 if the tails of the densities f_n and g_n converge at different rates. However more recently [7] proved that, for most parametric models, the ratio of the supremum of prior and posterior variation distance

over a neighbourhood $\mathcal{N}$ of the density f_0

$$\sup_{g_0 \in \mathcal{N}} \left\{ \frac{d_V(f_n, g_n)}{d_V(f_0, g_0)} \right\}, \tag{2}$$

almost always *diverges* in n - usually at a of rate $n^{k/2}$, where k is $\dim \Theta$ with probability 1 even when data are drawn from a "true" density indexed by the parameter $\theta_0 \in \Theta^0$. Furthermore this remains true even when the class $\mathcal{N}$ is chosen so the tail characteristics of f_0 and g_0 are identical - thus precluding cases like the one cited above - and g_0 is constrained to be infinitely differentiable. Therefore this phenomenon cannot be explained by discrepancy in tails of f_0 and g_0 - contra [2] - nor does it occur because the deviation between g_0 and f_0 is discontinuous in a neighbourhood of a maximum likelihood estimate.

These results appear to demonstrate a disturbing lack of robustness in Bayesian inference. However it is shown below that this lack of stability occurs because the prior variation distance $d_V(f_0, g_0)$ has virtually no bearing on the posterior variation distance $d_V(f_n, g_n)$ even in the tight neighbourhood $\mathcal{N}$ given in [7]. By imposing weak equicontinuity conditions on g_0 it is shown that $\sup_{g \in \mathcal{N}} \{d_V(f_n, g_n)\}$ can be bounded and typically decreases in powers of n for neighbourhoods $\mathcal{N}$ much less tight than the ones specified in [7].

Definition 1 *Let $B[1], B[2] \subset A$ and $A \subseteq \Theta$ be measurable sets with respect to the common dominating measure of two distributions F and G with respective densities f and g. Define the* DR_A separation $d_A^R(f, g)$ *by*

$$d_A^R(f, g) \triangleq \sup_{B[1], B[2] \subseteq A} \left| \frac{F(B[1])G(B[2])}{F(B[2])G(B[1])} - 1 \right|. \tag{3}$$

In this paper because two compared densities f and g are assumed to be continuous on a shared support it is easily checked that $d_A^R(f, g)$ simplifies to

$$d_A^R(f, g) = \sup_{\theta, \phi \in A} \left| \frac{f(\theta)g(\phi)}{f(\phi)g(\theta)} - 1 \right| = \sup_{\theta, \phi \in A} \left(\frac{f(\theta)g(\phi)}{f(\phi)g(\theta)} \right) - 1 < \delta, \tag{4}$$

a formula used henceforth. Note that the separation measure defined by [6] and cited above is $d_\Theta^R(f, g)$.

An equivalent separation measure the $A-$*density ratio separation* $d_A^L(f, g)$ - is given by

$$d_A^R(f, g) = \exp d_A^L(f, g) - 1, \tag{5}$$

so that

$$d_A^L(f,g) = \sup_{\theta,\phi \in A} \{(\log f(\theta) - \log g(\theta)) - (\log f(\phi) - \log g(\phi))\}.$$

Because $d_A^L(f,g)$ and $d_A^R(f,g)$ are equivalent henceforth we freely move between them.

It is easy to check that all DR_A and $A-$density ratio separations are separation measures in the sense that for all continuous densities $f, g \in \mathcal{F}, d(f,g)$ takes values in $\mathbb{R} \cup \infty$ where for all $f, g \in \mathcal{F}$, $d(f,f) = 0, d(f,g) \geq 0$ and $d(f,g) = d(g,f)$.

The neighbourhoods of these separations have some nice convexity properties: for example they are closed under the standard multiplicative pool of f and g. Thus it is immediate from the definition above that for any set A and $0 \leq \alpha \leq 1$

$$d_A^L(f, g_{\alpha,f}) = \alpha d_A^L(f,g) \leq d_A^L(f,g),$$

where

$$g_{\alpha,f}(\theta) \triangleq \left(\int f^{(1-\alpha)}(\theta) g^{\alpha}(\theta) d\theta \right)^{-1} f^{(1-\alpha)}(\theta) g^{\alpha}(\theta).$$

Note that if $A_1 \subset A_2$ then $d_{A_1}^L(f,g)$ is weaker than $d_{A_2}^L(f,g)$. When the lower bound and upper bound are attained

$$d_A^L(f,g) = (\log f(\theta_{u(A,f,g)}) - \log g(\theta_{u(A,f,g)})) - (\log f(\theta_{l(A,f,g)}) - \log g(\theta_{l(A,f,g)})),$$

where

$$\begin{aligned} \theta_{u(A,f,g)} &= \arg\sup_A (\log f(\theta) - \log g(\theta)), \\ \theta_{l(A,f,g)} &= \arg\inf_A (\log f(\theta) - \log g(\theta)). \end{aligned} \tag{6}$$

So in particular $d_A^L(f,g)$ is easy to interpret, being the difference of the two log densities at their maximum and minimum values within a set A.

To prove the convergence results of this paper it is usually sufficient to consider only sets $A = B(\theta_0, \rho)$ where $B(\theta_0, \rho)$ is an open ball with centre θ_0 and of radius ρ. We write, $d_{\theta_0,\rho}^R(f,g) \triangleq d_{B(\theta_0,\rho)}^R(f,g)$, $d_{\theta_0,\rho}^L(f,g) \triangleq d_{B(\theta_0,\rho)}^L(f,g)$, $d_{\Theta_0,\rho}^R(f,g) \triangleq \sup\{d_{\theta_0,\rho}^R(f,g) : \theta_0 \in \Theta_0\}$ and $d_{\Theta_0,\rho}^L(f,g) \triangleq \sup\{d_{\theta_0,\rho}^L(f,g) : \theta_0 \in \Theta_0\}$.

When all densities f and g which lie in a subset of densities $\mathcal{F}$ have the property that $d_{\Theta,\rho}^L(f,g) < \infty$, then it easliy checked that $d_{\Theta,\rho}^L(f,g)$ is

in fact a metric. Note that unlike $d_A^R(f,g)$, $d_{\Theta_0,\rho}^R(f,g)$ is a function of the parametrisation we use. So in particular to obtain invariance of convergence to transformations $\mathbb{T} : \Theta \to \Theta$ of the parameter space, the reparametrising map $\mathbb{T}$ needs to be a diffeomorphism, a natural restriction when in a finitely parametrized family. Demanding that a neighbourhood system be invariant to arbitrary measurable reparametrizations, as does [20], appears inappropriate for the parametric purposes of this paper.

We next show that these separations are closely related to Euclidean distances on the natural parameter when the two densities compared come from an exponential family.

Example 2 *Let $f_1(\theta) = f(\theta|\alpha_1)$ and $f_2(\theta) = f(\theta|\alpha_2)$ lie in the same regular exponential family*

$$f(\theta|\alpha) = c(\pi(\alpha))h(\theta)\exp\left\{\sum_{i=1}^{k}\pi_i(\alpha)t_i(\theta)\right\},$$

for some measurable functions $(\pi_1, \pi_2, ..., \pi_k, t_1, t_2, ..., t_k)$ for some integer k where $\pi(\alpha) = (\pi_1, \pi_2, \ldots, \pi_k), \mathbf{t} = (t_1, t_{2,\ldots,}t_k) \in \mathbb{T}$ and $\mathbb{T}$ does not depend on α since the exponential family is regular. For $1 \le i \le k$, and $j = 1, 2$ write

$$\pi_i(\alpha_j) = \pi_{i,j}.$$

Note that if a set A is of the form $A = \{\theta \in \Theta : \mathbf{t}(\theta) \in \mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2 \times \ldots \mathbb{A}_k\}$ and $\mu(\mathbb{A}_i)$ denotes the length of the interval $\mathbb{A}_i$ then

$$\begin{aligned} d_A^L(f_1,f_2) &= \sup_{\theta,\phi\in A}\log\left\{\frac{f(\theta)g(\phi)}{f(\phi)g(\theta)}\right\}, \\ &= \sup_{\mathbf{t}(\theta),\mathbf{t}(\phi)\in\mathbb{A}}\left\{\sum_{i=1}^{k}(\pi_{i,1}-\pi_{i,2})(t_i(\theta)-t_i(\phi))\right\}, \\ &= \sum_{i=1}^{k}|\pi_{i,1}-\pi_{i,2}|\mu(\mathbb{A}_i). \end{aligned}$$

It follows that if $\mu(\mathbb{A}_i)$ is infinite for some (f_1, f_2) with $\pi_{i,1} \neq \pi_{i,2}$ then so is $d_\Theta^L(f_1, f_2)$ and the usual density ratio diverges. In particular, two densities within the regular exponential family with parameters arbitrarily close under Euclidean distance are usually infinitely far apart under $d_\Theta^L(.,.)$. But under

$d_A^L(f_1, f_2)$ *two such models with parameters close in Euclidean distance have close local separation as well. For example suppose* $\mathbf{t}(\theta) = \theta$. *Then*

$$d^L_{\theta_0,\rho}(f_1, f_2) \leq 2\rho\sqrt{\sum_{i=1}^{k}(\pi_{i,1} - \pi_{i,2})^2}.$$

In the special case when all points $\theta \in B(\theta_0; \rho)$ *lie in the sample space - so that* θ_0 *is not near the boundary of* Θ, *the components of* θ *are functionally independent within this ball and they do not depend on* θ_0 *- the inequality above becomes an identity. So for this family* $d_A^L(f_1, f_2)$ *simply corresponds to a weighted Euclidean distance between the components of the natural parameters of the two prior densities. The distances between prior densities conjugate to exponential families [3] also have an analogous simple closed form. However, in [18] it is shown that within this family* $d^L_{\theta_0,\rho}$ *distances have a dependence on* θ_0 *so that in a Euclidean neighbourhood at the boundary of the parameter space they can be unbounded. An example of this and a demonstration of a correponding lack of robustness for beta densities whose values of hyperparameters close to zero is given in [18].*

Example 3 *When* f_0 *and* g_0 *are respectively the prior densities of a collection of* n *independent normally distributed random vectors with respective mean vectors* μ_f, μ_g *and covariance matrices* Σ_f, Σ_g *it is easily checked that*

$$d^L_{\theta_0,\rho}(f, g) \leq d^1_{\theta_0,\rho}(f, g) + d^2_{\theta_0,\rho}(f, g),$$

where, if $\mathbf{e}$ *is a vector with all entries* 1,

$$\begin{aligned} d^1_{\theta_0,\rho}(f, g) &= \sup\left\{\left(\mu_f\Sigma_f^{-1} - \mu_g\Sigma_g^{-1}\right)(\theta - \phi)\,\mathbf{e}^T : \theta, \phi \in B(\theta_0, \rho)\right\}, \\ &\leq 2n\rho\left|\mu_f\Sigma_f^{-1} - \mu_g\Sigma_g^{-1}\right|, \end{aligned}$$

and

$$\begin{aligned} d^2_{\theta_0,\rho}(f, g) &= \sup\left\{trace\left(\Sigma_f^{-1} - \Sigma_g^{-1}\right)\left\{\theta\theta^T - \phi\phi^T\right\}/2 : \theta, \phi \in B(\theta_0, \rho)\right\}, \\ &\leq 2n\rho\left(n\left\|\theta_0\right\| + n\rho\right)\left|trace\left(\Sigma_f^{-1} - \Sigma_g^{-1}\right)\right|. \end{aligned}$$

So provided that

$$\left|\mu_f\Sigma_f^{-1} - \mu_g\Sigma_g^{-1}\right|, \left|trace\left(\Sigma_f^{-1} - \Sigma_g^{-1}\right)\right|, \left\|\theta_0\right\|,$$

are finite, the DR_A separation decreases to zero with ρ and mirrors Euclidean distance in the natural hyperparameters of this family. Therefore, the usual choice of low precision priors ensures that when ρ is small the local neighbourhoods of f are very coarse and contain most candidate genuine prior densities g_0 that might be entertained. In fact it will be demonstrated later that mixing on hyperparameters of a family often ensures that the neighbourhoods of the margins of θ become increasingly coarse even when $\|\theta_0\|$ is unbounded.

3 Some basic properties of $d_A^L(f,g)$ and $d_A^R(f,g)$

For any measurable subset A of Θ a striking property - here called the *isoseparation* property - of $d_A^L(f,g)$ (and hence $d_A^R(f,g)$) can be calculated directly from the formal Bayes Rule. Thus for all $f_0, g_0 \in \mathcal{F}$ where $\mathcal{F}$ is any subset of continuous densities given above, for all $n \geq 1$ and for $A \subseteq \Theta(n)$ we have

$$d_A^L(f_n, g_n) = d_A^L(f_0, g_0). \tag{7}$$

In particular, when the observed likelihood $p_n(\mathbf{x}_n|\theta) > 0$ for all $\theta \in \Theta$ and for any measurable subset A of Θ we have

$$\sup_{g_o} \left\{ \frac{d_A^L(f_n, g_n)}{d_A^L(f_0, g_0)} \right\} = 1. \tag{8}$$

Thus unlike the variation distance analogue (2) this ratio does not diverge for *any* neighbourhood $\mathcal{N}$ of f_0. Prior densities that are close under these topologies remain close a posteriori. However surprisingly - provided $p(\mathbf{x}_n|\theta) > 0$ and is continuous for any $\theta \in A$ - they do not get closer either: prior separations endure regardless of what is observed.

When $A = \Theta$ this property has in fact been known for a very long time ([6]). However, $A = \Theta$ is the least interesting of special cases because this separation is very fine, for example being a discrete topology on the class of densities in standard exponential families. The most useful of these separations is when the measure of A is small because, when studying posterior densities, interest often focuses on the small areas of the parameter space on to which the posterior functioning density concentrates its mass.

Let the observed likelihood $p(\mathbf{x}_n|\theta) > 0$ for all $\theta \in \Theta$. When $\{p(\mathbf{x}_n|\theta)\}_{n\geq 1}$ are not explicit functions of θ_2 where $\theta = (\theta_1, \theta_2)$, $f_{0,1}$ and $g_{0,1}$ are the functioning and genuine prior marginal and $f_{n,1}$ and $g_{n,1}$ are the functioning and

the genuine posterior marginal density of θ_1 then these marginal densities inherit the isoseparation property. Thus for all $n \geq 1$, for $\theta \in A \subseteq \Theta(n)$

$$d_A^L(f_{n,1}, g_{n,1}) = d_A^L(f_{0,1}, g_{0,1}).$$

This property becomes important in the study of hierarchical models, where the distribution of the first hidden level of variables together with the relevant sampling distribution is often sufficient for any prediction of observable quantity.

Example 4 *Suppose that the observations X_n have a joint sample distribution uniquely specified by $\theta_1 = (\mu, \mathbf{\Sigma})$ where μ is the vector of means of X_n and $\mathbf{\Sigma}$ a vector of other hyperparameters, for example variances and covariances. To specify the prior on μ it is common practice to extend this model so that*

$$\mu = \tau(\phi) + \varepsilon,$$

where ϕ is a low dimensional vector, τ is a known function - often linear - and ε is an error vector parametrised by a vector $\mathbf{\Lambda}$ of, for example, covariances. Often a utility will be a function only of θ through $\theta_1 = \mu$. The invariance property above then allows us to substitute θ_1 for θ in all the examples of robust inference discussed below.

Let $f_A(g_A)$ henceforth denote the densities of $f(g)$ conditioned on the event $\{\theta \in A \subset \Theta\}$. A second property called *conditioning invariance* is essentially a special case of the first. When we learn that $\{\theta \in B \subset \Theta\}$, for some measurable set B where $A \subseteq B$, then

$$d_A^R(f_B, g_B) = d_A^R(f, g). \tag{9}$$

In particular, this property implies the useful identity $d_A^R(f, g) = d^R(f_A, g_A)$.

In common with other separation measures such as Hellinger and Kullback-Leibler, DR_A has the property that the separation between two marginal densities is not larger than the separation between their corresponding joint densities.

Thus let $\theta = (\theta_1, \theta_2)$ and $\phi = (\phi_1, \phi_2)$ be two candidate parameter values in $\Theta = \Theta_1 \times \Theta_2$ where $\theta_1, \phi_1 \in \Theta_1$ and $\theta_2, \phi_2 \in \Theta_2$, where the joint densities $f(\theta) = f_1(\theta_1) f_{2|1}(\theta_2|\theta_1)$, $g(\theta) = g_1(\theta_1) g_{2|1}(\theta_2|\theta_1)$ and $f_1(\theta_1), g_1(\theta_1)$ are the

marginal densities on Θ_1 of the two joint densities $f(\theta)$ and $g(\theta)$ respectively. Then it is proved in the appendix that

$$d_A^L(f,g) \geq d_{A_1}^L(f_1,g_1), \tag{10}$$

where $A_1 = \{\theta_1 : \theta = (\theta_1,\theta_2) \in A \text{ for all } \theta_2 \in B \subset \Theta_2 \text{ for some open set } B \text{ in } \Theta_2\}$. Through this proof it is transparent that when $f_{2|1}(\theta_2|\theta_1) = g_{2|1}(\theta_2|\theta_1)$ this inequality becomes an equality. On the other hand if $\theta = (\theta_1,\theta_2,\ldots\theta_k)$ where the subvectors $\{\theta_1,\theta_2,\ldots\theta_k\}$ of parameters are mutually independent, then in common with Chernov or Kullback-Liebler separations it is easy to check that

$$d_A^L(f,g) = \sum_{i=1}^{k} d_{A_i}^L(f_i,g_i), \tag{11}$$

where f_i (g_i) denotes the θ_i margin of $f(g)$, $1 \leq i \leq n$. In particular, if in the hierarchical model of Example 4 a priori $\mu \amalg \mathbf{\Sigma}$, then by equation (11) and by the isoseparation property the posterior separation over these parameter vectors continues to be (11), regardless of what is observed.

To improve the mixing characteristics of Bayes numerical algorithms it is common to "increase the temperature" of a given posterior or to perform simulated melding ([21]). A density f is replaced by $f^* \propto f^\alpha$ for some value of α, $0 < \alpha < 1$. The interpretation of such a substitution is that for all f, g for which $d_A^L(f,g) < \infty$, then

$$d_A^L(f^*,g^*) = \alpha d_A^L(f,g).$$

Therefore, simulated melding corresponds to a simple linear contraction on the separation space defined by $d_A^L(f,g)$ and draws all densities closer to one another in this sense. On the other hand simulated annealing employs the same transformation but lets $\alpha \to \infty$, increasingly separating the densities.

Example 5 (The Power Steady Model) *A class of state space models based on increasing temperatures was introduced in [17] and [15]. The predictive distributions of* $\{Y_t|y_1,y_2,\ldots y_{t-1}\}_{t\geq 1}$ *are specified in terms of a recurrence of the form*

$$\begin{aligned} p(y_t|\theta_t,y_1,y_2,\ldots y_{t-1}) &= p(y_t \mid \theta_t), t \geq 1, \\ f_t(\theta_t|y_1,y_2,\ldots y_{t-1}) &\propto f_{t-1}^\alpha(\theta_{t-1}|y_1,y_2,\ldots y_{t-1}), t \geq 2, \\ f_1(\theta_1) &\propto f_0^\alpha(\theta_0). \end{aligned}$$

for some $0 < \alpha < 1$. *One example of such processes is the Gaussian Steady DLM ([9]) if the prior hyperparameters of* $f_0(\theta_0)$ *are set to appropriate limiting values. Note that, although the state space distribution is not fully specified, the one step ahead predictives - and hence the whole predictive distribution of* $\{y_t\}_{t\geq 1}$- *are available as*

$$\begin{aligned} p(y_t|, y_1, y_2, \dots y_{t-1}) &\propto \int_{\theta_t \in \Theta_t} p(y_t|\theta_t) f_t^{\alpha}(\theta_{t-1}|y_1, y_2, \dots y_{t-1}) d\theta, \\ f_{t-1}(\theta_{t-1}|y_1, y_2, \dots y_{t-1}) &\propto p(y_{t-1}|\theta_{t-1}) f_{t-1}(\theta_{,t-1}\,|y_1, y_2, \dots y_{t-2}) \end{aligned}$$

where the proportionality constants can be calculated using the fact that densities integrate to unity. Suppose interest is in the joint distribution of $\{Y_t|y_1, y_2, \dots y_{t-1}\}_{t\geq T}$ $T \geq 2$. *Note that the margin* $f_{T-1}(\theta_{T-1}|y_1, y_2, \dots y_{T-1})$ *of* $\theta_{T-1}|y_1, y_2, \dots y_{T-1}$ *is sufficient for forecasting* $\{Y_t|y_1, y_2, \dots y_{t-1}\}_{t\geq T}$. *Let* $g_{T-1}(\theta_{T-1}|y_1, y_2, \dots y_{T-1})$ *be the corresponding density using* $g_0(\theta_0)$ *instead of* $f_0(\theta_0)$. *Then from the isoseparation property and the melding property above we have*

$$\begin{aligned} &d_A^L(f_T(\theta_T|y_1, y_2, \dots y_T), g_T(\theta_T|y_1, y_2, \dots y_T)) =, \\ &d_A^L(f_T(\theta_T|y_1, y_2, \dots y_{T-1}), g_T(\theta_T|y_1, y_2, \dots y_{T-1})) =, \\ &\alpha d_A^L(f_{T-1}(\theta_{T-1}|y_1, y_2, \dots y_{T-1}), g_{T-1}(\theta_{T-1}|y_1, y_2, \dots y_{T-1})). \end{aligned}$$

It follows that the quality of the approximation using the functioning prior instead of the genuine prior improves exponentially fast in T *with respect to all these separation measures, i.e.*

$$d_A^L(f_T(\theta_T|y_1, y_2, \dots y_T), g_T(\theta_T|y_1, y_2, \dots y_T) = \alpha^T d_A^L(f_0(\theta_0), g_0(\theta_0)).$$

Notice furthermore that the isoseparation property ensures that this result will still *hold whatever* $\{p(y_t|\theta_t)\}_{1\leq t<T}$ *is and whether or not this sequence were supplemented or corrupted, for example by censoring. It follows that in the long run this class of models is very robust with respect to prior misspecification.*

4 Roughness and Local Density Ratio Separation

In this section it is shown that prior closeness $d^L_{\theta_0,\rho}(f_0, g_0)$ for small radii ρ is essentially a condition that f_0 and g_0 are "similarly rough" and has

virtually no relationship with prior variation distance between f_0 and g_0. Later we show that it is these local separations $d^L_{\theta_0,\rho}(f_0, g_0)$ which control the posterior variation distances of the two densities. It is first necessary to define what is meant in this paper by the term "similarly rough". Say a function $f \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$ if

$$\sup_{\theta,\phi \in B(\theta_0;\rho)} |\log f(\theta) - \log f(\phi)| \leq M(\Theta_0)\rho^{0.5p(\Theta_0)},$$

where Θ_0 is a subset of Θ, the domain of f, and where $0 < p \leq 2$. Then the smallness of the parameter p governs the roughness of these function densities in these families. In particular, for any set Θ_0 when $0 < p_1 < p_2 \leq 2$ then

$$\mathcal{F}(\Theta_0, M(\Theta_0), p_2) \subset \mathcal{F}(\Theta_0, M(\Theta_0), p_1).$$

For example $\mathcal{F}(\Theta_0, M(\Theta_0), 2)$ denotes the set of functions whose logarithm is differentiable on Θ_0 with all derivatives bounded in modulus by M. Say that $g \in \mathcal{N}(f, \Theta_0, M(\Theta_0), p(\Theta_0))$ if there is a continuous function $h(\theta)$ such that $f = f'h$ and $g = g'h$ where $f', g' \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $p(\Theta_0) > 0$ for all $\theta_0 \in \Theta_0$. Note that if $g \in \mathcal{N}(f, \Theta_0, M(\Theta_0), p(\Theta_0))$ then for all $\theta_0 \in \Theta_0$

$$\begin{aligned} d^L_{\theta_0,\rho}(f, g) &= \sup_{\theta,\phi \in B(\theta_0;\rho)} |(\log f'(\theta) - \log f'(\phi)) - (\log g'(\theta) - \log g'(\phi))|, \\ &\leq \sup_{\theta,\phi \in B(\theta_0;\rho)} |\log f'(\theta) - \log f'(\phi)| + \sup_{\theta,\phi \in B(\theta_0;\rho)} |\log g'(\theta) - \log g'(\phi)|, \\ &\leq 2M(\Theta_0)\rho^{1/2p(\Theta_0)}, \end{aligned}$$

so that

$$d^R_{\Theta_0,\rho}(f, g) \leq \exp\left\{2M(\Theta_0)\rho^{1/2p(\Theta_0)}\right\} - 1. \tag{12}$$

Even with no strong contextual knowledge a Bayesian may well plausibly believe that her genuine prior density g_0 and her functioning prior density f_0 are similarly rough in the sense that $g_0 \in \mathcal{N}(f_0, \Theta_0, M(\Theta_0), p(\Theta_0))$ for all compact sets Θ_0 of sufficiently small measure and for an appropriate choice of M and p. When this is so for all small sets Θ_0, $d^L_{\Theta_0,\rho}(f_0, g_0)$ can be made arbitrarily small by choosing the radius $\rho > 0$ sufficiently small. By demanding this type of weak equicontinuity condition - that $\{f, g : d^R_\Theta(f, g) < \eta\}$ where η is chosen arbitrarily small- rather than the one used in [7], it is shown below that the expected large sample stability results. Furthermore the inequalities introduced above allow us to bound the rate at which this occurs. Note that

if, $g_0 \in \mathcal{N}(f_0, \Theta_0, M(\Theta_0), 2))$ and $\Theta = \mathbb{R}$ then $d^L_{\Theta,\rho}(f_0, g_0) \leq 2M\rho$. In this case if f_0 is misspecified only in terms of its location and scale parameters, so that the genuine prior $g_0(\theta) = f_0(\sigma^{-1}(\theta - \mu))$ for some μ and $\sigma > 0$, then $d^L_\rho(f_0, g_0)$ must tend to zero at a rate ρ, as shown in the following example.

Example 6 *Two one dimensional Student t densities $f_j(\theta) = f(\theta|\alpha_j)$, $j = 1, 2$ where*

$$f(\theta|\alpha) = \frac{\Gamma(0.5[\alpha+1])}{\sqrt{\alpha\pi}\Gamma(0.5\alpha)}(1 + \alpha^{-1}\sigma^{-2}(\theta-\mu)^2)^{-0.5(\alpha+1)},$$

have

$$d^L_A(f_1, f_2) = 1/2 \sup_{\theta,\phi\in A} \left| \sum_{j=1}^{2} (\alpha_j + 1) \log\left\{1 + \xi(\theta, \phi, \alpha_j, \mu_j, \sigma_j^2)\right\} \right|,$$

where $\xi(\theta, \phi, \alpha, \mu, \sigma^2) = (\theta - \phi)(\theta + \phi + 2\mu)(\alpha\sigma^2 + (\phi - \mu)^2)^{-1}$. Assuming without loss of generality that $\theta_0 \geq 0$, it follows that

$$\sup_{\theta,\phi\in B(\theta_0;\rho)} \left|\xi(\theta, \phi, \alpha, \sigma^2)\right| \leq \frac{4\rho(\theta_0 - \mu + \rho)}{(\alpha\sigma^2 + (\theta_0 - \mu + \rho)^2)} \leq \frac{2\rho}{\sigma\sqrt{\alpha}},$$

where the last inequality is obtained by identifying a maximum by differentiating. Thus

$$d^L_{\Theta\rho}(f_1, f_2) \leq \sum_{j=1,2} (\alpha_j + 1) \log\left(1 + \frac{2\rho}{\sigma_j\sqrt{\alpha_j}}\right) \leq \rho M,$$

where

$$M = \sum_{j=1,2} 2\sigma_j^{-1}(\alpha_j^{1/2} + \alpha_j^{-1/2}).$$

Suppose our functioning prior f_1 has the Student t density given above. It follows that the distance $d^L_{\Theta,\rho}(f_1, f_2)$ of any genuine Student t prior f_2 with arbitrary prior mode μ_2 tends to zero at a rate ρ provided the degrees of freedom of the genuine prior and their spread parameter are bounded i.e. $0 < a^L \leq \alpha_2 \leq a^U < \infty$, $0 < s^L \leq \sigma_2 \leq s^U < \infty$. In particular, by letting $|\mu_2 - \mu_1| \to \infty$ for a sufficiently small choice of radius ρ, two prior student t densities will be close even when their variation distance is arbitrarily large.

Thus the condition that $d^L_{\Theta,\rho}(f_0, g_0)$ is small for small ρ is a mild one to impose for flat tailed bounded densities: whole families of densities with different locations and scales can lie in the same neighbourhood. Only when the masses of the two densities concentrate near θ_0, where the derivative of $\log f_0 - \log g_0$ might be unbounded might f_0 and g_0 be a long distance apart for a small enough value of ρ. This happens for instance when θ_0 lies in the tail of a density f_0 where either f_0 or g_0 has an exponential or faster tail. Even in this case, provided the mass of the functioning posterior concentrates on to a compact subset $\Theta_0 \subset \Theta$ with high probability, sufficient smoothness will usually still exist to ensure convergence in total variation, as illustrated in the example below.

Example 7 *Consider a Bayesian hierarchical model where joint prior densities over parameters are specified through vector equations like*

$$\theta = \varphi + \varepsilon,$$

where φ is some function of hyperparameters encoding the systematic mean variation in θ and ε is a vector of error terms with zero mean and independent of ϕ. Commonly the functioning prior density f_ε of the error term ε is chosen from some smooth family - for example a Student t arising from a Gaussian whose associated variance hyperparameter is given an inverted Gamma distribution and integrated out. Assume this choice is such that $f_\varepsilon \in \mathcal{F}(\Theta, M, p)$ for some suitable choices of the two parameters (M, p). Here ε can be considered as a nuisance parameter vector in the sense given in Section 3 because the likelihood would not be a function of it given θ. Let $f_\varphi(g_\varphi)$ and $f(g)$ denote the functioning (genuine) joint prior densities of φ and θ respectively. Then since

$$\begin{aligned} f(\theta) - f(\theta - \delta) &= \int \left(f_\varepsilon(\varepsilon - \varphi) - f_\varepsilon(\varepsilon - \varphi - \delta)\right) f_\varphi(\varphi) d\varphi, \\ g(\theta) - g(\theta - \delta) &= \int \left(f_\varepsilon(\varepsilon - \varphi) - f_\varepsilon(\varepsilon - \varphi - \delta)\right) g_\varphi(\varphi) d\varphi, \end{aligned}$$

an automatic consequence of constructing the prior in this way is that

$$g \in \mathcal{N}(f, \Theta_0, M(\Theta_0), p(\Theta_0)),$$

irrespective *of the density of the mean signals f_φ and g_φ, even if this is governed by a discrete process - for example the realisation of a Dirichlet process.*

Thus the condition that the genuine prior density g lies in a neighbourhood of the functioning prior density f is often simply a consequence of the way hierarchical priors are conventionally chosen. It is shown below that this largely determines the robustness of posterior inferences with respect to the variation metric. Whether the use of such conventions which implicitly impose robustness are justified is of course entirely dependent on the modelling context.

5 Variation distance and local separations

If A is a measurable subset of Θ and we write

$$\xi_A(\theta|f,g) \triangleq \left|\frac{g_A(\theta)}{f_A(\theta)} - 1\right|,$$

where $f_A(\theta) = \frac{f(\theta)}{F(A)}$ and $g_A(\theta) = \frac{g(\theta)}{G(A)}$ are respectively the conditional densities of θ under $f(\theta)$ and $g(\theta)$ given $\theta \in A$ then

$$\xi_A(\theta|f,g) \leq \sup_{\theta\in A}\left|\frac{g_A(\theta)}{f_A(\theta)} - 1\right| \leq d_A^R(f_A,g_A) = d_A^R(f,g). \tag{13}$$

This enables us to relate DR_A to total variation. Note that for any $A \subset \Theta_0$, $d_V(f_n,g_n) = T_n[1] + T_n[2]$ where

$$\begin{aligned}
T_n[1] &= \int_{\theta\notin A} |f_n(\theta) - g_n(\theta)|\, d\theta, \\
&= \int_{\theta\notin A} |F_n(A)f_{n,A}(\theta) - G_n(A)g_{n,A}(\theta)|\, d\theta, \\
&\leq |F_n(A) - G_n(A)| \int_{\theta\notin A} g_{n,A}(\theta)d\theta + F_n(A)\int_{\theta\in A} |f_{n,A}(\theta) - g_{n,A}(\theta)|\, d\theta, \\
&\leq |F(A^c) - G(A^c)| + \int_{\theta\notin A} |\xi_A(\theta|f_n,g_n)|\, f_{n,A}(\theta)d\theta, \\
&\leq T_n[2] + \sup_{\theta\in A}\xi_A(\theta|f_n,g_n), \\
&\leq T_n[2] + d_A^R(f_n,g_n) = T_n[2] + d_A^R(f_0,g_0),
\end{aligned}$$

by the isoseparation property. Similarly

$$T_n[2] = \int_{\theta\notin A} |f_n(\theta) - g_n(\theta)|\, d\theta,$$

$$\begin{aligned} &\leq \int_{\theta\in\Theta^c} |\xi_\Theta(\theta|f_n,g_n)|\, f_n(\theta)d\theta, \\ &\leq \sup_{\theta\in A^c} \xi_{A^c}(\theta|f_n,g_n)\,\{1-F_n(A_n)\} \leq \alpha_n\Delta, \end{aligned} \tag{14}$$

again by isoseparation, where $\Delta = d^R_\Theta(f_0,g_0)$ and $\alpha_n = \{1-F_n(A_n)\}$. In particular

$$d_V(f_n,g_n) \leq d^R_{A_n}(f_0,g_0) + 2T_n[2]. \tag{15}$$

By choosing $\{A_n\}_{n\geq 1}$ as a function of the statistics of the functioning posterior in such a way that $\alpha_n = \{1-F_n(A_n)\} \to 0$ as $n\to\infty$ but for which $d^R_{A_n}(f_n,g_n)\to 0$ as $n\to\infty$ ensures convergence. Furthermore explicit bounds for the total variation between f_n and g_n to be calculated directly from the inequality (15). Constructing appropriate sequences $\{A_n\}_{n\geq 1}$ for a given statistical model is usually straightforward when $\Delta<\infty$: $A_n = B(\theta_{n,0},\rho_n)$ is simply set to be a sequence of open balls centred at the functioning posterior mean and whose radius $\rho_n\to 0$ as $n\to\infty$. For example it is usually sufficient to use well known Chebychev inequalities with the result that, from equation (12), $g_0\in\mathcal{N}(f_0,A_n,M(A_n),p(A_n))$ implies that $d^R_{A_n}(f_0,g_0)\leq \exp\left\{2M\rho_n^{p/2}\right\}-1$.

Note that neither of the two conditions $g_0\in\mathcal{N}(f_0,A_n,M(A_n),p(A_n))$ and $\Delta<\infty$ imply anything about *where* the two candidate prior densities concentrate their mass. Illustrations of how such bounds are given below. In all these examples it is assumed that the prior mutual roughness condition $g_0\in\mathcal{N}(f_0,A_n,M(A_n),p(A_n))$ holds for given values of $M(\Theta_0),p(\Theta_0)$ for the sequence $\{A_n = B(\theta_{n,0},\rho_n)\}$ where $A_n\subseteq\Theta_0$ and that $d^R_\Theta(f_0,g_0)\leq\Delta$.

Example 8 *For $n\geq 1$ let F_n denote the one dimensional Gaussian $N(\theta_0^n,\sigma_n^2)$ distribution function. Let $\tau_n=\sigma_n^{-r}$ for some $0<r<1$, and let $\rho_n=\sigma_n\tau_n$. Note that if $\sigma_n^2\to 0$ then $\tau_n\to\infty$ and $\rho_n\to 0$. It follows that*

$$d^R_{A_n}(f_0,g_0)\leq \exp\left\{2M(\Theta_0)\sigma_n^{p(1-r)/2}\right\}-1.$$

Also since (see e.g. [12], p.279) the standard normal distribution function Φ satisfies for all $x>0$

$$\Phi(-x) < (2\pi)^{-1/2}\, x^{-1}\exp -x^2/2.$$

It follows that

$$\begin{aligned} T_n[2] &= d^R_\Theta(f_0,g_0)F_n(\theta\notin B(\theta_0;\rho)). \leq 2\Delta\Phi(-\tau_n), \\ &< \sqrt{\frac{2}{\pi}}\Delta\sigma_n^r\exp -\sigma_n^{-2r}/2. \end{aligned}$$

Choosing $0 < r < 1$ *appropriately gives an upper bound for the variation distance. Note that with the differentiability condition* $p = 2$ *this confirms that for any* $0 < r < 1$

$$\lim_{n\to\infty} \sup_{g\in\mathcal{N}} \left\{ \frac{d_V(f_n, g_n)}{\sigma_n^{(1-r)}} \right\} = 0.$$

Typically $\sigma_n \leq \sigma n^{-1/2}$ *for some* σ*, so that*

$$\lim_{n\to\infty} \sup_{g\in\mathcal{N}} \left\{ n^{r'} d_V(f_n, g_n) \right\},$$

for any $r' = 1/2\{1 - r\} < 1/2$*. Thus the expected* $n^{-1/2}$ *speed of convergence in variation distance between the two posteriors is retrieved. This contrasts with the* $\sqrt{n}$ *speed of divergence obtained by [7] in their analogue (2). Note here that it is the difference in mutual roughness between the prior densities* f_0 *and* g_0 *that governs the latter rate of divergence.*

Example 9 *Suppose* F_n *is any one dimensional functioning posterior distribution function with mean* $\theta_{n,0}$ *and variance* $\sigma_n^2 < \infty$ *. Then by Chebychev's inequality*

$$T_n[2] \leq \Delta F_n(\theta \notin B(\theta_0; \rho_n)) \leq \Delta \frac{\sigma_n^2}{\rho_n^2}.$$

using the definition above and setting $r = 1/3$ *when* $p = 2$ *we have that*

$$\begin{aligned} d_V(f_n, g_n) &\leq \exp\left\{2M\sigma_n^{2/3}\right\} - 1 + 2\Delta\frac{\sigma_n^2}{\rho_n^2}, \\ &= \exp\left\{2M\sigma_n^{2/3}\right\} - 1 + 2\Delta\sigma_n^{2/3}. \end{aligned}$$

It follows that for any *one dimensional functioning posterior density with a finite mean and variance the variation distance between posteriors* (f_n, g_n)*, lying in this neighbourhood typically is bounded by a rate* $\sqrt[3]{n}$*. For example, it is common for the posterior density* $f_n(\theta)$ *of a mean parameter* θ *to be Student t so that*

$$f_n(\theta) \propto \left[1 + \frac{(x - \theta_0^n)^2}{(\alpha_n - 2)\sigma_n^2}\right]^{-(\alpha_n+1)/2},$$

where $\alpha_n = \alpha_0 + n/2$*,* $n > 4$*,* $\mathbb{E}(\theta|x_n) = \theta_0^n$ *and* $Var(\theta|x_n) = \sigma_n^2$*. Plugging in these moments gives the required robustness intervals. For a given data set this posterior variance* σ_n^2 *could increase but this increase will be apparent.*

Example 10 *Even when the moments of f_n do not exist bounds can still be calculated, although the rate of convergence associated with these bounds is sometimes slower. Suppose $f_n(x) = f\left(\sigma_n^{-1}(\theta - \theta_0^n)\right)$ where f is a Cauchy density and note that for $x > 0$ the Cauchy distribution function $F(x)$ has the property that $F(-x) < \frac{1}{2\pi}x^{-1}$. It then follows that by letting $\tau_n = \sigma_n^{-r}$*

$$T_n[2] \leq \frac{\Delta}{\pi}\sigma_n^r.$$

To obtain the best asymptotic bound for $d_V(f_n, g_n)$ of this type using these inequalities set $r = 0.5$ when

$$\lim_{n\to\infty}\sup_{g\in\mathcal{N}}\left\{\frac{d_V(f_n, g_n)}{\sqrt{\sigma_n}}\right\} \leq M + 2\Delta.$$

Example 11 *Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ and let $\mu_{j,n}, \sigma_{j,n}^2$ denote, respectively, the mean and variance of θ_j, $1 \leq j \leq k$ under the functioning posterior density f_n. Then [19] p.153 proves that writing $\theta_0^n = (\mu_{1,n}, \mu_{2,n}, \ldots \mu_{k,n})$,*

$$F_n\left(\theta \in B(\theta_0^n; \rho_n)\right) \geq F_n\left[\bigcap_{j=1}^{k}\left\{|\theta_j - \mu_{j,n}| \leq \sqrt{k}\rho_n\right\}\right] \geq 1 - k\rho_n^{-2}\sum_{j=1}^{k}\sigma_{j,n}^2,$$

so that

$$F_n\left\{\theta \notin B(\theta_0^n; \rho_n)\right\} \leq k\rho_n^{-2}\sum_{j=1}^{k}\sigma_{j,n}^2,$$

implies

$$T_n(2, \rho) \leq \Delta\frac{\sigma_n^2}{\rho_n^2}.$$

where $\sigma_n^2 = k\max_{1\leq j\leq k}\sigma_{j,n}^2$. Thus exactly analogous bounds to the univariate Chebychev bounds can also be calculated for multivariate problems. In fact this bound is coarse and can be improved (see [11]). Note that the given bound increases only linearly with the dimension k of the parameter space. If the utility is a function only of the margin of a subvector θ_1 of the parameter θ then because of the property (10) such bounds can be tightened. In the case above such bounds can be calculated directly from prior bounds and the means and variances only of those components θ_1 in the margin of interest.

A second useful bound can be calculated when a Bayesian wants to entertain the possibility that the prior densities f_0 and g_0 might have different tail characteristics violating the condition $\Delta < \infty$ used in constructing the previous bound. Call g_0 *c-rejectable* if the ratio of marginal likelihoods $\frac{p_{f_0}(\mathbf{x})}{p_{g_0}(\mathbf{x})} \geq c$. If the genuine prior is believed to explain the data better than f_0 then this ratio would be predicted a priori to be small and certainly not c-rejectable for a moderately large values of $c \geq 1$ (for example $c = 2$). Say density f $\Lambda-$*tail dominates* a density g if

$$\sup_{\theta \in \Theta} \frac{g(\theta)}{f(\theta)} = \Lambda < \infty. \tag{16}$$

When $g(\theta)$ is bounded then this condition requires that the tail convergence of g is no faster than f. Then if it is believed that g_0 is not c-rejectable and equation (16) holds then

$$\begin{aligned} T_n[2] &\leq F_n\{A_n\} + G_n\{A_n\}, \\ &= F_n\{A_n\} + \int_{\theta \notin A_n} \frac{g_n(\theta)}{f_n(\theta)} f_n(\theta) d(\theta), \\ &= F_n\{A_n\} + \int_{\theta \notin A_n} \frac{p_f(\mathbf{x})}{p_g(\mathbf{x})} \frac{g_0(\theta)}{f_0(\theta)} f_n(\theta) d(\theta), \\ &\leq F_n\{A_n\} + c\Lambda \int_{\theta \notin A_n} f_n(\theta) d(\theta) \leq \alpha_n(1 + c\Lambda). \end{aligned}$$

Here the prior tail dominance condition simply encourages the use of a flat tailed functioning prior so that if data is observed in its tail the likelihood will tend to dominate the posterior and this information is not killed off by the functioning prior tail. This formal result just technically confirms practical Bayesian modelling principles of using as flat a tailed functioning prior as possible: see for example [14]. Under these conditions, analogues of all the examples above can be calculated simply by substituting $1 + c\Lambda$ for Δ throughout. For other related bounds see [4].

Provided that f_0 and g_0 are close with respect to these new separation measures whilst the family of models may be inconsistent with the data, the functioning posterior distribution nevertheless will tend to provide a good approximation of the genuine posterior. All similar priors will give similar, if possibly erroneous, posterior densities.

It is interesting that useful bounds can sometimes be obtained even when the functioning posterior does not converge in distribution ([18]). On the

other hand, whilst the continuity of f_0 and g_0 can be relaxed, their mutual roughness conditions given above are almost necessary for posterior robustness. If f_0 and g_0 do not lie in a local separation neighbourhood about a particular location θ_0 then no matter how small the radius of that neighbourhood it is possible to construct a sequence of likelihoods that converge in a very strong way to a true value θ_0. In particular, uniformly consistent estimates of θ can be obtained but nevertheless the genuine and functioning posterior densities remain at least a bounded variation distance apart whatever the value of sample size n. A formal statement and proof of this property is given in [18] based on the counterexample used in [7].

6 Discussion

The local separation measures introduced in this paper appear natural ones to employ when examining how Bayesian models learn. It appears that to employ a proper prior whose mass is poorly positioned will give approximately the same posterior density as getting the prior right provided the sample size is large enough - in a way that can be measured - under three caveats. The first is that the same likelihood is shared by the two priors. This commonly assumed but very strong condition is absolutely critical ([18]) : if this is not the case, then posterior inferences will typically diverge with increasing sample size. The second condition is that the functioning posterior usually needs to converge to a set of small measure. If the convergence is to a defective distribution then the local DR_A distances in the tails of the genuine and functioning priors need to be comparable: a condition well - known in the literature. Thirdly both priors need to be comparably rough: a property that is often *implicitly* but perhaps not always thoughtfully induced by the way priors are currently specified.

7 Appendix

To prove the inequality (10) assume that $f(\theta)$ and $g(\theta)$ are continuous at $\widetilde{\theta}$ and $\widetilde{\phi}$ where $\widetilde{\theta} = \left(\widetilde{\theta}_1, \widetilde{\theta}_2\right)$ and $\widetilde{\phi} = \left(\widetilde{\theta}_1, \widetilde{\phi}_2\right)$ where $\widetilde{\theta}_1 = \theta_{u(A,f_1,g_1)}$ and $\widetilde{\phi}_1 = \theta_{l(A,f_1,g_1)}$ and $\widetilde{\theta}_2$ is any point satisfying $f_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1) \geq g_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)$ and $\widetilde{\phi}_2$ is any point satisfying $f_{2|1}(\widetilde{\phi}_2|\widetilde{\phi}_1) \leq g_{2|1}(\widetilde{\phi}_2|\widetilde{\phi}_1)$. Note that such points $\widetilde{\theta}_2$ and

$\widetilde{\phi}_2$ of the conditional densities must exist because $f_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)$ and $g_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)$ are densities. Then for all continuous joint densities f, g and sets $A \subseteq \Theta$

$$\begin{aligned} d_A^R(f,g) &= \sup_{\theta,\phi \in A} \left(\frac{f_1(\theta_1) f_{2|1}(\theta_2|\theta_1) g_1(\phi_1) g_{2|1}(\phi_2|\phi_1)}{f_1(\phi_1) f_{2|1}(\phi_2|\phi_1) g_1(\theta_1) g_{2|1}(\theta_2|\theta_1)} \right) - 1, \\ &\geq \frac{f_1(\widetilde{\theta}_1) f_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1) g_1(\widetilde{\phi}_1) g_{2|1}(\widetilde{\phi}_2|\widetilde{\phi}_1)}{f_1(\widetilde{\phi}_1) f_{2|1}(\widetilde{\phi}_2|\widetilde{\phi}_1) g_1(\widetilde{\theta}_1) g_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)} - 1, \\ &= \sup_{\theta_1,\phi_1 \in A_1} \left(\frac{f_1(\theta_1) g_1(\phi_1)}{f_1(\phi_1) g_1(\theta_1)} \right) - 1 = d_{A_1}^R(f_1, g_1), \end{aligned}$$

and therefore

$$d_A^L(f,g) \geq d_{A_1}^L(f_1, g_1). \tag{17}$$

where $A_1 = \{\theta_1 : \theta = (\theta_1, \theta_2) \in A$ for all $\theta_2 \in B \subset \Theta_2$ for some open set B in $\Theta_2\}$, the property we require.

References

[1] Andrade, J. A. A. and O'Hagan, A. (2006). "Bayesian robustness modelling using regularly varying distributions". Bayesian Analysis 1, 169-188.

[2] Berger, J. (1992) in discussion of Wasserman, L.(1992b) "Recent methodological advances in robust Bayesian inference (with discussion)", in Bayesian Statistics 4 J.M. Bernado et al (eds) 495 - 496 Oxford University Press.

[3] Bernardo, J.M. and Smith, A.F.M.(1996) "Bayesian Theory", Wiley Chichester

[4] Daneseshkhah,A (2004) "Estimation in Causal Graphical Models", PhD Thesis University of Warwick.

[5] Dawid, A.P. (1973) "Posterior expectations for large observations", Biometrika, 60, 664-667.

[6] DeRobertis, L. (1978) "The use of partial prior knowledge in Bayesian inference", Ph.D. dissertation, Yale University.

[7] Gustafson, P. and Wasserman, L. (1995) "Local sensitivity diagnostics for Bayesian inference", Annals of Statistics , 23, 2153-2167.

[8] Ghosh, J.K. and Ramamoorthi, R.V.(2003) "Bayesian Nonparametrics", Springer.

[9] West, M. and Harrison, P.J.(1997) "Bayesian Forecasting and Dynamic Models", Springer.

[10] Marshall, A.W. and Olkin, (1979) "Inequalities: Theory of Majorisation and its Applications", Academic Press.

[11] Monhor, D. (2007) "A Chebyshev Inequality for Multivariate Normal Distribution", Probability in the Engineering And Informational Sciences, 21-2, 289-300.

[12] Moran, P.A.P.(1968) "An Introduction to Probability Theory", Oxford University Press.

[13] O'Hagan, A.(1979) "On outlier rejection phenomena in Bayesian inference", Journal of the Royal Statistical Society B 41, 358-367.

[14] O'Hagan, A and Forster, J. (2004) "Bayesian Inference", Kendall's Advanced Theory of Statistics, Arnold.

[15] Peterka, V. (1981) "Bayesian system identification". In: Trends and Progress in System Identification, P. Eykhoff, Ed., p. 239-304. Pergamon Press, Oxford.

[16] Schervish, M.J. (1995) "The Theory of Statistics", Springer Verlag New York.

[17] Smith, J.Q.(1979) "A generalisation of the Bayesian steady forecasting model", Journal of the Royal Statistical Society B 41, 375-87.

[18] Smith, J.Q. "Local Robustness of Bayesian Parametric Inference and Observed Likelihoods", CRiSM Research Report 07-09.

[19] Tong, Y.L.(1980) "Probability Inequalities in Multivariate Distributions" Academic Press New York.

[20] Wasserman, L.(1992a) "Invariance properties of density ratio priors" Annals of Statistics, 20, 2177-2182.

[21] Poole, A. and Raftery, A.(2000) "Inference for Deterministic Simulation Models: The Bayesian Melding Approach" Journal of the American Statistical Association, 95, 1244-1255.

Acknowledgements This paper has greatly benefitted from discussions with Ali Daneshkah, Jim Griffin, Jon Warren, Wilfred Kendall, Sigurd Assing and Stephen Walker.