Ann Inst Stat Math (2012) 64:495–519
DOI 10.1007/s10463-011-0334-9

Isoseparation and robustness in parametric Bayesian inference

Jim Q. Smith · Fabio Rigat

Received: 28 May 2009 / Revised: 15 September 2010 / Published online: 25 September 2011

Abstract This paper introduces a new family of local density separations for assessing robustness of finite-dimensional Bayesian posterior inferences with respect to their priors. Unlike for their global equivalents, under these novel separations posterior robustness is recovered even when the functioning posterior converges to a defective distribution, irrespectively of whether the prior densities are grossly misspecified and of the form and the validity of the assumed data sampling distribution. For exponential family models, the local density separations are shown to form the basis of a weak topology closely linked to the Euclidean metric on the natural parameters. In general, the local separations are shown to measure relative roughness of the prior distribution with respect to its corresponding posterior and provide explicit bounds for the total variation distance between an approximating posterior density to a genuine posterior. We illustrate the application of these bounds for assessing robustness of the posterior inferences for a dynamic time series model of blood glucose concentration in diabetes mellitus patients with respect to alternative prior specifications.

Keywords Density ratio class · Hierarchical Bayesian inference · Local robustness · Total variation · Power steady model · Diabetes mellitus

J. Q. Smith · F. Rigat (✉)
Department of Statistics, University of Warwick,
Coventry CV4 7AL, UK
e-mail: F.Rigat@warwick.ac.uk

J. Q. Smith
e-mail: J.Q.smith@warwick.ac.uk

F. Rigat
Novartis Vaccines and Diagnostics,
Siena, Italy

1 Introduction

Assessing robustness of inferential procedures and decision rules with respect to the specification of likelihoods, prior distributions and loss functions is an essential yet challenging aspect of any statistical data analysis. Huber (1997) reviews several main achievements in the definition and computation of robustness criteria up to the mid 1990s. For the same period Gustafson (1996) illustrates the state of the art in evaluating robustness from a Bayesian perspective, whereas Kadane et al. (1996) and Martin et al. (1996) address the issue of robustness with respect to the specification of loss functions for Bayesian decision problems. Fernandez et al. (1996) and recently Copas et al. (2010) examine notions of robustness with respect to changes in the likelihood function. Gustafson and Bose (1996) consider the sensitivity of posterior inferences with respect to perturbations of hierarchically defined prior distributions. Abraham and Cadre (2004) study the rate of convergence of various measures of posterior global robustness with respect to sample size, recovering their respective convergence rates.

Let $\theta \in \Theta$ be a finite-dimensional parameter vector indexing the data sampling probability distribution. In this paper we let $f_0(\theta)$ denote the *functioning prior*, that is the probability density actually used in a Bayesian analysis, and $g_0(\theta)$ the genuine prior, that is the density that would be used if there was enough time and skills applied to elicit it at best (O'Hagan (2006)). Denote the two corresponding posterior densities after observing a sample $\mathbf{x}_n$ of $n \geq 1$ observations by $f_n(\theta)$ and $g_n(\theta)$, respectively. In this context the problem of posterior robustness is commonly addressed by first assuming that both sequences of posterior functioning and genuine densities, $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$, are consistent (Schervish 1995). It then follows that when each component of a random sample $\mathbf{x}_n$ is drawn from a distribution with parameter value $\theta \in \Theta$ 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, the posterior total variation distance (TVD)

$$d_V(f_n, g_n) = \int_\Theta |f_n(\theta) - g_n(\theta)| \, \mathrm{d}\theta,$$

converges to zero almost surely P_{θ_0} as $n \to \infty$ (Ghosh and Ramamoorthi 2003). This means that f_n provides a good asymptotic approximation for g_n for estimation purposes when the sampling family is precisely and correctly specified. On the other hand, it is well known that when the functioning posterior convergences to a defective distribution, $d_V(f_n, g_n)$ cannot be guaranteed to vanish if the tails of the densities f_n and g_n converge at sufficiently different rates (Dawid 1973; O'Hagan 1979; Andrade and O'Hagan 2006).

More recently Gustafson and Wasserman (1995) proved that, for most parametric models, when the genuine prior g_0 belongs to a tight neighborhood $\mathcal{N}$ of the density f_0, the supremum

$$\sup_{g_0 \in \mathcal{N}} \left\{ \frac{d_V(f_n, g_n)}{d_V(f_0, g_0)} \right\}, \tag{1}$$

almost always diverges in n with probability 1, usually at rate $n^{\dim(\Theta)/2}$. Divergence of (1) occurs even when the data are drawn from a "true" density indexed by the parameter $\theta_0 \in \Theta^0$ or when $\mathcal{N}$ is chosen so that the tail characteristics of f_0 and g_0 are identical and g_0 is constrained to be infinitely differentiable, as in Berger (1992).

In this paper we prove that under a new family of local separations the TVD between genuine and functioning posteriors vanishes even when the sampling distribution family is not accurately specified and the data is not a random sample. In particular, we show that divergence of (1) arises 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 neighborhoods $\mathcal{N}$ defined in Gustafson and Wasserman (1995). Furthermore, we show that these local separations can be used to provide the basis of an on-line diagnostic measure of robustness which is available in closed form for most statistical models and does not involve any additional hyper-parameters. This makes our results of considerable practical significance to the study of robustness to prior specifications in high dimensional parametric inference and in particular for hierarchical models.

In Sect. 2 we introduce the new local separations and we contrast them with the global density ratio separation measures of Wasserman (1992), Gustafson and Wasserman (1995) and of O'Hagan and Forster (2004). Examples 1 and 2 illustrate the close links between the new separations and the Euclidean distance on the natural parameter space for exponential family densities. In Sect. 3 three properties of these local separations are examined and Examples 3 and 4 demonstrate their application to derive closed-form discrepancy measures for selected models. In Sect. 4 it is shown that in the limit these local separations provide a coarse topology which can be used in practice to compare the relative roughness of alternative prior densities. Examples 5 and 6 illustrate that comparatively rough commonly used densities are in fact close under the new local separations. In Sect. 5 it is proved that closeness in TVD between the functioning and genuine posterior densities is guaranteed when the genuine prior lies in one such coarse neighborhood of the functioning prior. Closed-form upper bounds for the TVD between posterior densities are derived using the new local separations and Examples 7–10 demonstrate the derivation of these bounds for selected models. In Sect. 6 the TVD upper bounds are applied to calibrate a power steady model (Smith (1979)) under two alternative prior specifications for a time series of blood glucose concentration measurements taken from a patient affected by diabetes mellitus. Here is it shown that, consistently with the theory outlined in the paper, the TVD decays exponentially with the sample size. This example also illustrates that posterior convergence can be achieved after as few as ten data points. Section 7 concludes the paper with a critical discussion of its main results.

2 Density ratio balls and isoseparation

Henceforth for simplicity we 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. We also assume that the sequence of sampling densities $\{p(\mathbf{x}_n|\theta)\}_{n\geq 1}$ of an n-vector of data $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ is measurable with respect to $g_0(\theta)$. Let $\Theta(n) = \{\theta \in \Theta : p(\mathbf{x}_n|\theta) > 0\}$ and for

simplicity let $p(\mathbf{x}_n|\theta)$, seen as a function of θ, be 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 using the equation

$$g_n(\theta) = \begin{cases} \frac{g_0(\theta)p(\mathbf{x}_n|\theta)}{p_{g_0}(\mathbf{x}_n)} & \text{if } \theta \in \Theta(n), \\ 0 & \text{if } \theta \in \Theta \backslash \Theta(n), \end{cases} \tag{2}$$

where the predictive density

$$p_{g_0}(\mathbf{x}_n) = \int_{\theta \in \Theta(n)} p(\mathbf{x}_n|\theta) g_0(\theta)\, \mathrm{d}\theta$$

is calculated, either algebraically or numerically, so as to ensure that $g_n(\theta)$ integrates to 1.

Let $f_n(\theta) \triangleq f(\theta|\mathbf{x}_n)$, defined as in (2) but with $g_0(\theta)$ substituted with $f_0(\theta)$, be the functioning posterior density after the first n observations. This paper focuses on the case when the posterior density which is actually calculated, $f_n(\theta)$, converges in distribution as $n \to \infty$ to a point mass in the closure neighborhood of $\theta_0 \in \Theta(n)$. We start by defining the two equivalent local divergence measures $d_A^R(f, g)$ and $d_A^L(f, g)$.

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

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

Since in this paper the densities f and g are assumed to be continuous on a shared support, Eq. (3) simplifies to

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

Note that the global separation measure defined by DeRobertis (1978) is simply $d_{\Theta}^R(f, g)$. Under (4) an equivalent measure, which we call the A-density ratio separation $d_A^L(f, g)$, is given by

$$d_A^L(f, g) = \sup_{\theta, \phi \in A} \{(\log f(\theta) - \log g(\theta)) - (\log f(\phi) - \log g(\phi))\}. \tag{5}$$

When the lower and upper bounds of the difference $\log f - \log g$ are attained $d_A^L(f, g)$ is easy to interpret, being the difference of the two log densities at their maximum and minimum values within a set A. Equation (5) defines a separation measure in the sense that for all continuous densities $f, g \in \mathcal{F}$, $d_A^L(f, g)$ takes values in

$\mathbb{R} \cup \infty$, $d_A^L(f, f) = 0$, $d_A^L(f, g) \geq 0$ and $d_A^L(f, g) = d_A^L(g, f)$. Also, when $f, g \in \mathcal{F}$ have finite separation $d_A^L(f, g)$, then the latter is a metric within A.

To prove the convergence results of this paper it is sufficient to consider sets of the form $A = B(\theta_0, \rho)$ where $B(\theta_0, \rho)$ is an open ball with center θ_0 and radius ρ. In this case we write

$$
\begin{aligned}
d^R_{\theta_0,\rho}(f, g) &\triangleq d^R_{B(\theta_0,\rho)}(f, g),\\
d^L_{\theta_0,\rho}(f, g) &\triangleq d^L_{B(\theta_0,\rho)}(f, g),\\
d^R_{\Theta_0,\rho}(f, g) &\triangleq \sup\{d^R_{\theta_0,\rho}(f, g) : \theta_0 \in \Theta_0\},\\
d^L_{\Theta_0,\rho}(f, g) &\triangleq \sup\{d^L_{\theta_0,\rho}(f, g) : \theta_0 \in \Theta_0\}.
\end{aligned}
$$

Note that $d^R_{\theta_0,\rho}(f, g)$ is a function of the parameterization we use so that invariance of convergence to transformations of the parameter space $\mathbb{T} : \Theta \to \Theta$ obtains only if the map $\mathbb{T}$ is a diffeomorphism. This is a natural restriction within a finitely parameterized family, whereas demanding that a neighborhood system be invariant to arbitrary measurable reparameterizations, as in Wasserman (1992), appears inappropriate in the context of this paper. The next two examples show that (5) is closely related to the Euclidean distance on the natural parameter space when the two prior densities compared belong to univariate or multivariate exponential families.

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

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

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

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

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

$d^L_{\Theta}(.\,,\,.)$ but under $d^L_A(f_1, f_2)$ they have proportionally close local separations. For instance, if $\mathbf{t}(\theta) = \theta$ 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 θ_0 is not near the boundary of Θ, the components of θ are functionally independent within the ball $B(\theta_0, \rho)$ and they do not depend on θ_0. Under these conditions, the above inequality becomes an identity so that $d^L_{\theta_0,\rho}(f_1, f_2)$ is simply 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 of Bernardo (1996) also have an analogous simple closed form. However, this family of distances has a dependence on θ_0 so that in a Euclidean neighborhood at the boundary of the parameter space they can be unbounded. An example of this and a demonstration of a corresponding lack of robustness for beta densities whose hyper-parameters values are close to zero is given in Smith (2007).

Example 2 When the functioning and genuine priors f and g are n-dimensional Gaussian densities 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,

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

and

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

So provided that the terms $|\mu_f \Sigma_f^{-1} - \mu_g \Sigma_g^{-1}|$, $|\mathrm{trace}(\Sigma_f^{-1} - \Sigma_g^{-1})|$, $\|\theta_0\|$ are finite, $d^L_{\theta_0,\rho}(f, g)$ decreases to zero with the radius ρ and the density ratio separation mirrors Euclidean distance in the natural parameters of the multivariate Gaussian family.

Example 2 indicates that the usual choice of low precision Gaussian priors ensures that when the radius ρ is small the local neighborhoods of f are very coarse and contain most candidate genuine prior densities that might be entertained. In fact we demonstrate later that mixing on hyper-parameters of a family often ensures that the neighborhoods of the margins of θ become increasingly coarse even when $\|\theta_0\|$ is unbounded.

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

In this section, we examine three basic properties of the DR_A and of the A-separation measures. The first and second properties will be used to prove the main theorems of this paper, whereas the third property provides an interpretation of a popular class of posterior sampling algorithms.

3.1 Isoseparation

For any measurable subset $A \subseteq \Theta(n)$ a striking property, here called the isoseparation property of $d_A^L(f_n, g_n)$ and of $d_A^R(f_n, g_n)$, can be calculated directly from the formal Bayes rule (2). For all $f_0, g_0 \in \mathcal{F}$, for all $n \geq 1$ we have

$$d_A^L(f_n, g_n) = d_A^L(f_0, g_0), \tag{6}$$

so that we have

$$\sup_{g_0 \in \mathcal{F}} \left\{ \frac{d_A^L(f_n, g_n)}{d_A^L(f_0, g_0)} \right\} = 1. \tag{7}$$

Unlike its global version (1), the ratio of local separations (7) does not diverge for any neighborhood $\mathcal{N}$ of f_0. Prior densities that are close under these topologies remain close a posteriori. On the other hand, prior separations endure regardless of the observed data. When $A = \Theta$ this property has in fact been known for a very long time (DeRobertis 1978). However this global separation is very fine, for example providing a discrete topology on the class of all densities within standard exponential families. From a practical Bayesian perspective the most useful of these separations corresponds to the small subsets A of the parameter space on to which the posterior functioning density concentrates its mass as data is gathered.

We next consider the case 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 marginal priors and $f_{n,1}$ and $g_{n,1}$ are the functioning and the genuine marginal posterior densities of θ_1. It is then easy to check that these marginal densities inherit the isoseparation property. Thus for any $n \geq 1$, 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 isoseparation property of marginal posteriors is 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 predicting any observable quantity, as shown in the next example.

Example 3 Suppose that the observations X_n have a joint sample distribution uniquely specified by $\theta_1 = (\mu, \mathbf{\Sigma})$ where μ is a vector of means of X_n and $\mathbf{\Sigma}$ is a vector of

other hyper-parameters, e.g. 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 parameterized by a matrix Λ of, e.g. covariances. When a utility function only depends on θ through $\theta_1 = \mu$, the marginal isoseparation property allows us to substitute θ_1 for θ for evaluating the robustness of Bayesian inferences under this model.

3.2 Separation measures under conditioning and marginalization

Let f_A and g_A henceforth denote the densities f and g conditioned on the event $\{\theta \in A \subset \Theta\}$. A second useful property of DR_A is that, when we learn that $\{\theta \in A \subseteq B \subset \Theta\}$ for some measurable set B, then $d_A^R(f_B, g_B) = d_A^R(f, g)$. In particular, this second property implies that $d_A^R(f_A, g_A) = d_A^R(f, g)$ and since the densities f_A, g_A have support A, we have $d_A^R(f_A, g_A) = d_\Theta^R(f_A, g_A)$. Combining the latter two equations gives

$$d_A^R(f, g) = d_\Theta^R(f_A, g_A), \tag{8}$$

so that, in common with other separation measures such as Hellinger and Kullback-Leibler, the DR_A separation between two marginal densities is no larger than the corresponding separation between their joint densities. In general, (8) provides a simple direct relationship between the DR_A local separations and the corresponding global distances between conditional densities. Now let $\theta = (\theta_1, \theta_2)$ and $\phi = (\phi_1, \phi_2)$ be two candidate parameter values in $\Theta = \Theta_1 \times \Theta_2$ with $\theta_1, \phi_1 \in \Theta_1$ and $\theta_2, \phi_2 \in \Theta_2$. When the joint densities $f(\theta)$ and $g(\theta)$ can be factored as

$$f(\theta) = f_1(\theta_1) f_{2|1}(\theta_2|\theta_1),$$
$$g(\theta) = g_1(\theta_1) g_{2|1}(\theta_2|\theta_1),$$

where $f_1(\theta_1)$, $g_1(\theta_1)$ are the marginals on Θ_1 of $f(\theta)$ and $g(\theta)$ and $f_{2|1}(\theta_2|\theta_1)$, $g_{2|1}(\theta_2|\theta_1)$ are the respective conditional densities, then it is proved in the appendix that

$$d_A^L(f, g) \geq d_{A_1}^L(f_1, g_1), \tag{9}$$

where $A_1 = \{\theta_1 : \theta = (\theta_1, \theta_2) \in A \text{ for all } \theta_2 \in B \subset \Theta_2 \text{ for some open set } B\}$. In this sense marginal densities are never further apart from each other than their corresponding joint densities. Therefore, when $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ where the subvectors $\{\theta_1, \theta_2, \ldots, \theta_k\}$ are mutually independent, then in common with Chernov and

Kullback-Leibler separations, we have

$$d_A^L(f, g) = \sum_{i=1}^{k} d_{A_i}^L(f_i, g_i), \tag{10}$$

where $f_i(g_i)$ denotes the θ_i margin of $f(g)$. For instance, if in Example 3 we let $\mu \amalg \mathbf{\Sigma}$ a priori, then by equation (6) the posterior separation over these vectors continues to be (10), regardless of the observed data.

3.3 Tempering property

A third property of the A-density ratio separation $d_A^L(f, g)$, which we call the tempering property, allows a simple interpretation of the class of sampling algorithms raising a density to a non-negative power, which typically depends on a temperature coefficient (Geyer and Thompson (1995); Poole and Raftery (2000)). In this case a density f is replaced by $f^* \propto f^\alpha$ for some value $0 < \alpha < 1$. This typically improves mixing of posterior simulation in cases of multimodality as the heated posterior density f^* has broader modes and fatter tails than f. A useful interpretation of such a substitution is that for all f, g for which $d_A^L(f, g) < \infty$, then

$$d_A^L(f^*, g^*) = \alpha d_A^L(f, g). \tag{11}$$

Therefore, letting $\alpha \to 0$ corresponds to linear contractions 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 (Kirkpatrick et al. 1983) employs the same transformation but lets $\alpha \to \infty$, increasingly separating the densities. The following example illustrates an application of this property in time series analysis.

Example 4 (The Power Steady Model) A class of state space models using temperature-based transitions was introduced in Smith (1979) and Peterka (1981). The model for the random variables $\{y_t\}_{t\geq 1}$ is specified as

$$\begin{aligned} &\theta_0 \sim f_0(\theta_0), \\ &p(y_t|\theta_t, y_{1:(t-1)}) = p(y_t \mid \theta_t), \quad t \geq 1 \\ &f_t(\theta_t|y_{1:(t-1)}) \propto f_{t-1}^\alpha(\theta_{t-1}|y_{1:(t-1)}), \quad t \geq 2. \end{aligned} \tag{12}$$

conditionally on some $0 < \alpha < 1$. High values of α, corresponding to temperatures closer to one, imply that the effect of the initial conditions θ_0 endures whereas low values of α allow a higher influence of the data on the time dependent posterior inferences and predictions. One example of such processes is the Gaussian steady dynamic linear model (West and Harrison 1997).

When a dynamic time series model is specified as in (12), then from the isoseparation property (8) and the tempering property (11) we have that

$$\begin{aligned} & d_A^L(f_T(\theta_T|y_{1:T}), g_T(\theta_T|y_{1:T})) \\ & = d_A^L(f_T(\theta_T|y_{1:(T-1)}), g_T(\theta_T|y_{1:(T-1)})) \\ & = \alpha d_A^L(f_{T-1}(\theta_{T-1}|y_{1:(T-1)}), g_{T-1}(\theta_{T-1}|y_{1:(T-1)})) \\ & = \alpha^T d_A^L(f_0(\theta_0), g_0(\theta_0)). \end{aligned} \tag{13}$$

It follows that the quality of the approximation using the functioning prior instead of the genuine prior improves exponentially in T with respect to the A-density ratio local separation measure. Furthermore, the isoseparation property ensures that this result still holds whatever $\{p(y_t|\theta_t)\}_{1\le t<T}$ 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. In Sect. 6 we use a multivariate power-steady model of the form (12) to illustrate the application of the A-density local separations in a clinical setting. Prior to applying these separations, in Sects. 5 and 6 we establish their relations to the roughness of prior densities and to the TVD metric.

4 Roughness and local density ratio separation

Section 2 provides a first intuitive interpretation of $d_A^L(f, g)$ with respect to the maxima and minima attained by the densities f and g within the reference set A. For sets of the form $A = B(\theta_0, \rho)$, here it is shown that prior closeness with respect to $d_{\theta_0,\rho}^L(f, g)$ for small radii ρ implies that f and g are "similarly rough" and has virtually no relationship with prior variation distance between the densities f and g. We also show that these local separations control the posterior variation distance between the two densities. We set out to address the issue of comparative roughness by invoking an idea analogous to path roughness and applying it to the logarithm of the functioning prior density.

Definition 2 A continuous density f is said to have $(\Theta_0, M(\Theta_0), p(\Theta_0))$ *roughness*—written $f \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$—if

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

for all open balls $B(\theta_0, \rho) \subseteq \Theta$ with center $\theta_0 \in \Theta_0 \subseteq \Theta$ and radius $\rho > 0$, where $0 < p(\Theta_0) \le 2$.

In particular, when $0 < \rho < 1$, for any set Θ_0 such that $0 < p_1(\Theta_0) < p_2(\Theta_0) \le 2$ we have

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

The larger the parameter $p(\Theta_0)$ and the smaller $M(\Theta_0)$, the smoother the densities are on the set Θ_0. Although the parameter $p(\Theta_0)$ can in principle depend on Θ_0, so that the smoothness of a density f is different in different regions of the parameter space, this dependence will rarely be needed in practice. On the other hand, if for instance f is a log-convex density, so that the derivative of the log f is unbounded, then log f will be bounded by $M(\Theta_0)$ for some suitably chosen closed and bounded set $\Theta_0 \subset \Theta$. Note that $\mathcal{F}(\Theta_0, M(\Theta_0), 2)$ in fact denotes the set of functions whose logarithm is differentiable and with all derivatives bounded in modulus by $M(\Theta_0)$ on Θ_0. The next definition and Theorem 1 illustrate the relationship between the above notion of roughness and the A-separation measure $d^L_{\theta_0,\rho}(f, g)$.

Definition 3 A probability density g belongs to the set $\mathcal{N}(f, \Theta_0, M(\Theta_0), p(\Theta_0))$ if there is a continuous function $h(\theta)$ such that $f = f^*h$ and $g = g^*h$ where $f^*, g^* \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$.

Theorem 1 *If* $g \in \mathcal{N}(f, \Theta_0, M(\Theta_0), p(\Theta_0))$ *then we have*

$$d^R_{\Theta_0,\rho}(f, g) \leq \exp\left\{2M(\Theta_0)\rho^{0.5p(\Theta_0)}\right\} - 1. \tag{14}$$

Proof Inequality (14) is verified by noting that

$$\begin{aligned} d^L_{\theta_0,\rho}(f, g) &= \sup_{\theta,\phi\in B(\theta_0;\rho)} |(\log f^*(\theta) - \log f^*(\phi)) - (\log g^*(\theta) - \log g^*(\phi))|, \\ &\leq \sup_{\theta,\phi\in B(\theta_0;\rho)} |\log f^*(\theta) - \log f^*(\phi)| + \sup_{\theta,\phi\in B(\theta_0;\rho)} |\log g^*(\theta) - \log g^*(\phi)|, \\ &\leq 2M(\Theta_0)\rho^{0.5p(\Theta_0)}. \end{aligned}$$

The latter inequality together with (3) prove (14). □

Even without strong contextual knowledge, a Bayesian modeler may plausibly believe that her genuine and functioning prior densities, g_0 and 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 the functions $M(\Theta_0)$ and $p(\Theta_0)$. When this is the case, the prior local separation $d^L_{\Theta_0,\rho}(f_0, g_0)$ can be made arbitrarily small by choosing a sufficiently small radius $\rho > 0$. In the next section it is also shown that, under this weak equicontinuity condition, large sample convergence of posterior separations results. Furthermore when $p = 2$ the inequality (14) allows us to bound the rate at which this convergence occurs.

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

Example 5 Let $f_j(\theta) = f(\theta|\alpha_j, \mu_j, \sigma_j)$, $j = 1, 2$ be two univariate Student t densities

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

In this case we have

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

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

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

Thus

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

where

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

where the set M here does not depend on Θ_0. It follows that the global distance $d_{\Theta,\rho}^L(f_1, f_2)$ of any genuine Student t prior f_2 with arbitrary prior mode μ_2 from a functioning prior f_1 of the same form tends to zero at a rate ρ provided the degrees of freedom of the genuine prior and its spread parameter are bounded, i.e.

$$0 < a^L \leq \alpha_2 \leq a^U < \infty,$$
$$0 < s^L \leq \sigma_2 \leq s^U < \infty.$$

Therefore, by letting $|\mu_2 - \mu_1| \to \infty$ for a sufficiently small choice of the radius ρ, any two Student t prior densities will be locally close even when their variation distance is arbitrarily large.

The next section shows that requiring a small $d_{\Theta,\rho}^L(f_0, g_0)$ when the radius ρ is small is a mild condition to impose for flat-tailed bounded distributions since whole families of densities with different locations and scales can lie in the same neighborhood. Only when the masses of the two densities concentrate near a point θ_0 where the derivative of $\log f_0 - \log g_0$ might be unbounded, can f_0 and g_0 be a long distance apart for small enough values of ρ. For instance, this happens 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 exist to ensure convergence in TVD.

Example 6 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 hyper-parameters 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, e.g. a Student t arising from a Gaussian whose associated variance hyper-parameter is given an inverted Gamma distribution and integrated out. Assume this choice is such that $f_\varepsilon \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$ for some suitable choices of the two parameters $(M(\Theta_0), p(\Theta_0))$. Here ε can be considered as a nuisance parameter vector in the sense given in Sect. 3 because the likelihood would not be a function of it given θ. Let $f_\varphi(g_\varphi)$ and $f_\theta(g_\theta)$ denote the functioning (genuine) prior densities of φ and θ respectively. If these priors are constructed be marginalising over φ, then

$$f_\theta(\theta) - f_\theta(\theta - \delta) = \int (f_\varepsilon(\varepsilon - \varphi) - f_\varepsilon(\varepsilon - \varphi - \delta)) f_\varphi(\varphi)\, \mathrm{d}\varphi,$$

$$g_\theta(\theta) - g_\theta(\theta - \delta) = \int (f_\varepsilon(\varepsilon - \varphi) - f_\varepsilon(\varepsilon - \varphi - \delta)) g_\varphi(\varphi)\, \mathrm{d}\varphi.$$

An automatic consequence is therefore that

$$g_\theta \in \mathcal{N}(f_\theta, \Theta_0, M(\Theta_0), p(\Theta_0)),$$

irrespective of the density of the mean signals f_φ and g_φ, even if this is governed by a discrete process, e.g. the realization of a Dirichlet process.

This example illustrates that the genuine prior density g_θ lies in a neighborhood of the functioning prior density f_θ simply as a consequence of the way hierarchical priors are conventionally constructed. In the next section we show that this construction in turn largely determines the robustness of posterior inferences arising from hierarchical models with respect to the TVD metric. Whether it is justified to construct prior densities which implicitly impose robustness, is of course entirely dependent on the modeling context.

5 Variation distance and local separations

Convergence in variation distance of the functioning and genuine posterior densities is dependent on two conditions, namely differences in roughness between the genuine and functioning priors in the neighborhood of functioning posterior mass concentration and discrepant prior tail behavior. Here we exploit a surprising connection between closeness with respect to TVD and to the local DR_A separation to derive upper bounds for the former. Although these bounds are not tight, they nevertheless have four important advantages. First, they exhibit the same power dependence on sample size that has been met in other studies, as illustrated in Examples 7–9. Second,

they can be expressed explicitly in terms of parameters of the prior neighborhoods and of simple summary statistics of the functioning prior. Third, they apply even when the family of sample distributions is misspecified. By this we mean that, although in this case the functioning posterior may locate its mass wrongly, the posterior associated with the genuine prior using the same misspecified sample distribution will nevertheless be located within the bounds of the functioning posterior. This type of robustness, which is fundamentally a subjective Bayesian one, is not commonly a property of other robustness bounds. Fourth, our bounds decompose as a sum of two components. The first measures to what extent the posterior is centered at a point in the parameter space being unusually rough, often being near the boundary of the parameter space. The second measures the extent to which the tails of the functioning and genuine priors differ.

If A is a measurable subset of Θ and $n \geq 0$ we write

$$\xi_A(\theta|f_n, g_n) \triangleq \left| \frac{g_{n,A}(\theta)}{f_{n,A}(\theta)} - 1 \right|, \tag{15}$$

where $f_{n,A}(\theta) = \frac{f_n(\theta)}{F_n(A)}$ and $g_{n,A}(\theta) = \frac{g_n(\theta)}{G_n(A)}$ are respectively the conditional densities given $\theta \in A$. Then

$$\xi_A(\theta|f_n, g_n) \leq \sup_{\theta \in A} \xi_A(\theta|f_n, g_n) \leq d_A^R(f_{n,A}, g_{n,A}) = d_A^R(f_n, g_n). \tag{16}$$

This enables us to relate DR_A to TVD using the following theorem.

Theorem 2 *For any sequence $\{A_n\}_{n\geq 1}$ of measurable subsets of Θ*

$$d_V(f_n, g_n) \leq F_n(A_n) d_{A_n}^R(f_0, g_0) + 2\{1 - F_n(A_n)\} d_\Theta^R(f_0, g_0), \tag{17}$$

Proof For any measurable $A_n \subset \Theta_0 \subseteq \Theta$, the TVD between the functioning and genuine posterior densities after observing n data points can be written as

$$d_V(f_n, g_n) = T_n[1] + T_n[2],$$

where

$$\begin{aligned} T_n[1] &= \int_{\theta \in A_n} |f_n(\theta) - g_n(\theta)| \, \mathrm{d}\theta, \\ &= \int_{\theta \in A_n} |F_n(A_n) f_{n,A_n}(\theta) - G_n(A_n) g_{n,A_n}(\theta)| \, \mathrm{d}\theta, \\ &\leq |F_n(A_n) - G_n(A_n)| \int_{\theta \in A_n} g_{n,A_n}(\theta) \, \mathrm{d}\theta + F_n(A_n) \\ &\quad \times \int_{\theta \in A_n} |f_{n,A_n}(\theta) - g_{n,A_n}(\theta)| \, \mathrm{d}\theta, \end{aligned}$$

$$\begin{aligned}
&\leq |F_n(A_n^c) - G_n(A_n^c)| + F_n(A_n) \int_{\theta \in A_n} \xi_{A_n}(\theta | f_n, g_n) f_{n,A_n}(\theta)\, \mathrm{d}\theta, \\
&\leq T_n[2] + F_n(A_n) \sup_{\theta \in A_n} \xi_{A_n}(\theta | f_n, g_n), \\
&\leq T_n[2] + F_n(A_n) d^R_{A_n}(f_n, g_n) = T_n[2] + F_n(A_n) d^R_{A_n}(f_0, g_0), \qquad (18)
\end{aligned}$$

by the isoseparation property (6). Similarly

$$\begin{aligned}
T_n[2] &= \int_{\theta \in A_n^c} |f_n(\theta) - g_n(\theta)|\, \mathrm{d}\theta, \\
&= \int_{\theta \in A_n^c} \xi_\Theta(\theta | f_n, g_n) f_n(\theta)\, \mathrm{d}\theta, \quad \text{by (15) with } A = \Theta, \\
&\leq \sup_{\theta \in A_n^c} \xi_\Theta(\theta | f_n, g_n)\{1 - F_n(A_n)\}, \\
&\leq \sup_{\theta \in \Theta} \xi_\Theta(\theta | f_n, g_n)\{1 - F_n(A_n)\}, \\
&\leq \{1 - F_n(A_n)\} d^R_\Theta(f_0, g_0), \qquad (19)
\end{aligned}$$

again by isoseparation. The inequalities (18) and (19) imply (17). □

It follows from Theorem 2 that by choosing $\{A_n\}_{n \geq 1}$ as a function of the statistics of the functioning posterior in such a way that $F_n(A_n) \to 1$ and when $d^R_{A_n}(f_n, g_n) \to 0$ as $n \to \infty$, convergence in total variation is ensured. Constructing appropriate sequences $\{A_n\}_{n \geq 1}$ for a given statistical model is usually straightforward when $d^R_\Theta(f_0, g_0) < \infty$. For instance, $A_n = B(\theta_n, \rho_n)$ can be set to be a sequence of open balls centered at the functioning posterior mean θ_n and whose radius $\rho_n \to 0$ as $n \to \infty$. In this case it is usually sufficient to use well known Chebyshev inequalities, ensuring that equation (14) holds, as shown in the following examples. Here it is assumed that the prior mutual roughness condition holds for given values of $M(\Theta_0)$ and $p(\Theta_0)$ for the sequence $\{A_n = B(\theta_n, \rho_n) \subseteq \Theta_0\}$ and that $d^R_\Theta(f_0, g_0) < \infty$.

Example 7 For $n \geq 1$ let F_n denote the one dimensional Gaussian distribution with mean θ_n and variance σ_n^2 and let $\rho_n = \sigma_n^{1-r}$ for some $0 < r < 1$. Note that if $\sigma_n^2 \to 0$ then $\rho_n \to 0$. 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. Moran 1968, p. 279) the standard Normal cumulative distribution function Φ satisfies for all $x > 0$

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

It follows that

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

so that choosing $0 < r < 1$ appropriately gives an upper bound for the variation distance. Note that under the differentiability condition $p = 2$ for any $0 < r < 1$ we have

$$\lim_{n\to\infty} \sup_{g \in \mathcal{N}_0(r)} \left\{ \frac{d_V(f_n, g_n)}{\sigma_n^{(1-r)}} \right\} = 0,$$

where, for brevity of notation, we write $\mathcal{N}_0(r) = \mathcal{N}(f_0, \Theta_0, M(\Theta_0), r)$. For instance, when $\sigma_n \leq \sigma n^{-1/2}$ for some $\sigma > 0$, we have that

$$\lim_{n\to\infty} \sup_{g \in \mathcal{N}_0} (r)\{n^{r^*} d_V(f_n, g_n)\} = 0,$$

for any $r^* = 1/2\{1-r\} < 1/2$. Thus the expected $1/\sqrt{n}$ speed of convergence in TVD between the two Gaussian posteriors is retrieved. This result contrasts with the $\sqrt{n}$ speed of divergence obtained by Gustafson and Wasserman (1995) using the ratio of global variation distances (1). Note that in fact it is the difference in mutual roughness between the prior densities f_0 and g_0 that governs the latter rate of divergence. The next example generalizes our posterior convergence result to univariate location-scale densities.

Example 8 Suppose F_n is any one dimensional functioning posterior distribution function with mean θ_n and variance $\sigma_n^2 < \infty$. Then by Chebyshev's inequality

$$T_n[2] \leq d^R_{\Theta}(f_0, g_0) F_n(\theta \notin B(\theta_n; \rho_n)) \leq d^R_{\Theta}(f_0, g_0) \frac{\sigma_n^2}{\rho_n^2}.$$

Using (17), with $\rho_n = \sigma_n^{2/3}$ when $p = 2$ and since $F_n(A_n) \leq 1$, we have that

$$\begin{aligned} d_V(f_n, g_n) &\leq \exp\left\{2M(\Theta_0)\sigma_n^{2/3}\right\} - 1 + 2 d^R_{\Theta}(f_0, g_0) \frac{\sigma_n^2}{\rho_n^2}, \\ &= \exp\left\{2M(\Theta_0)\sigma_n^{2/3}\right\} - 1 + 2 d^R_{\Theta}(f_0, g_0) \sigma_n^{2/3}. \end{aligned} \tag{20}$$

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) is typically bounded by a rate $\sqrt[3]{n}$. For instance, it is common for the marginal posterior density $f_n(\theta)$ of

a mean parameter θ to be Student t so that

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

where $\alpha_n = \alpha_0 + n/2, n > 4, \mathbb{E}(\theta|x_n) = \theta_n$ and $Var(\theta|x_n) = \sigma_n^2$. For a given choice of $M(\Theta_0)$ and when the prior separation $d^R_{\Theta}(f_0, g_0)$ is finite, plugging the latter variance in (20) gives the required robustness bounds.

The next example shows that bounds can still be calculated even when the moments of f_n do not exist, although their rate of convergence is sometimes slower.

Example 9 Suppose $f_n(x) = f(\sigma_n^{-1}(\theta - \theta_n))$ 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 follows that

$$T_n[2] \leq \frac{d^R_{\Theta}(f_0, g_0)}{\pi}\sigma_n^r.$$

To obtain the best asymptotic bound for $d_V(f_n, g_n)$ using (17), set $\rho_n = \sigma_n^{1/2}$ so that $r = 0.5$ and

$$\lim_{n\to\infty} \sup_{g\in\mathcal{N}_0(0.5)} \left\{\frac{d_V(f_n, g_n)}{\sqrt{\sigma_n}}\right\} \leq M(\Theta_0) + 2d^R_{\Theta}(f_0, g_0).$$

The next example demonstrates that analogous Chebyshev bounds to the univariate case can also be calculated for multivariate problems.

Example 10 Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ be a k-dimensional parameter vector indexing the distribution of a random variable of interest Y and let $\mu_{j,n}, \sigma^2_{j,n}$ denote, respectively, the posterior mean and variance of the margin $\theta_j, 1 \leq j \leq k$ under the functioning prior f_0 upon observing $(y_1, \ldots, y_n)$. Then Tong (1980, p.153) proves that

$$1 - F_n(\theta \in B(\theta_n; \rho_n)) \leq 1 - F_n\left(\bigcap_{j=1}^{k}\{|\theta_j - \mu_{j,n}| \leq \sqrt{k}\rho_n\}\right) \leq k\rho_n^{-2}\sum_{j=1}^{k}\sigma^2_{j,n},$$

which implies

$$T_n[2] \leq d^R_{\Theta}(f_0, g_0)\frac{\sigma_n^2}{\rho_n^2}, \tag{21}$$

where $\sigma_n^2 = k \max_{1\leq j\leq k} \sigma^2_{j,n}$.

Note that the latter bound increases linearly with the dimension k of the parameter space. These Chebyshev bounds are in fact coarse and they can be somewhat improved

using rather more complicated expressions (see Monhor 2007). Also, if example (10) is enriched by a utility function indexed only by the margin θ_1 of the parameter θ, then by (9) the posterior bounds (21) can be calculated directly from prior bounds and the means and variances of the margin of interest.

When f_0 and g_0 have sufficiently different tail characteristics, $d^R_\Theta(f_0, g_0)$ is unbounded so that the result of Theorem 2 is no longer useful. However, by introducing the following two definitions we can derive and alternative bound for the term $T_n[2]$.

Definition 4 Call g_0 *k-rejectable* if $\frac{p_{f_0}(\mathbf{x})}{p_{g_0}(\mathbf{x})} \geq k$.

By Definition 3, 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 g_0 would certainly not be k-rejectable for moderately large values of k.

Definition 5 The density f is said to $\Lambda-$ *tail dominate* a density g if

$$\sup_{\theta\in\Theta} \frac{g(\theta)}{f(\theta)} = \Lambda < \infty. \tag{22}$$

When $g(\theta)$ is bounded, (22) requires that the tails of g decay no faster than those of f.

Corollary 1 *If g_0 is not k-rejectable and* (22) *holds, then* (19) *becomes*

$$\begin{aligned} T_n[2] &\leq F_n(A_n^c) + G_n(A_n^c), \\ &= F_n(A_n^c) + \int_{\theta\in A_n^c} \frac{g_n(\theta)}{f_n(\theta)} f_n(\theta)\, d\theta, \\ &= F_n(A_n^c) + \int_{\theta\in A_n^c} \frac{p_f(\mathbf{x})}{p_g(\mathbf{x})} \frac{g_0(\theta)}{f_0(\theta)} f_n(\theta)\, d\theta, \\ &\leq F_n(A_n^c) + k\Lambda \int_{\theta\in A_n^c} f_n(\theta)\, d\theta, \\ &\leq F_n(A_n^c)(1 + k\Lambda). \end{aligned} \tag{23}$$

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. This formal result technically confirms practical Bayesian modeling principles suggesting the use of flat tailed functioning priors (O'Hagan and Forster (2004)). Under these conditions, analogues of Examples 7–10 above can be calculated simply by substituting $d^R_\Theta(f_0, g_0) = 1 + k\Lambda$ throughout. Further related bounds are derived in Daneseshkhah (2004).

6 Application to the analysis of glucose concentration data

In this section we demonstrate the use of the TVD upper bounds (17) for assessing robustness of posterior inferences under a power steady model analogous to Example 5. In this case, having chosen the form of a time series likelihood function, the posterior inferences are dependent on the prior hyper-parameters and in particular on the value of the power-steady coefficient α. Describing posterior dependence with respect to the latter is critical because α is not identifiable from the data so that its value must be fixed in advance (Smith 1979; West and Harrison 1997; O'Hagan 2006). In particular, for practical experimental design it may be important to set the parameter α so as to allow for posterior convergence under alternative priors using realistic data sample sizes.

At any time $t \geq 1$ we assume that a real-valued random variable Y_t has a Gaussian distribution with mean μ_t and variance σ_t^2. At time $t = 1$ these moments are given the standard conjugate priors

$$\begin{aligned} \mu_1 \mid m, \sigma_1^2 &\sim N(m, \sigma_1^2), \\ \sigma_1^2 \mid a, b &\sim IG(a, b). \end{aligned}$$

Upon observing the realization $Y_1 = y_1$, the conditional posterior densities are

$$\begin{aligned} \mu_1 \mid m, y_1, \sigma_1^2 &\sim N\left(\frac{m + y_1}{2}, \frac{\sigma_1^2}{2}\right), \\ \sigma_1^2 \mid a, b, m, \mu_1, y_1 &\sim IG\left(a + 1, b + \frac{(\mu_1 - m)^2 + (y_1 - \mu_1^2)}{2}\right). \end{aligned}$$

As in Example 5, the conditional prior for the mean at time $t = 2$ is then derived as

$$f(\mu_2 \mid m, y_1, \sigma_1^2) \propto f^{\alpha}(\mu_1 \mid m, y_1, \sigma_1^2),$$

where $\alpha \in (0, 1)$ is fixed, yielding the prior

$$\mu_2 \mid m, y_1, \sigma_1^2 \sim N\left(\frac{m + y_1}{2}, \frac{\sigma_1^2}{2\alpha}\right).$$

The conditional prior for the variance at time $t = 2$ is derived by using the transformation $\sigma_2^2 = \frac{\sigma_1^2}{2\alpha}$, giving

$$\sigma_2^2 \mid a, b, m, \mu_1, y_1 \sim IG\left(a + 1, \frac{1}{2\alpha}\left(b + \frac{(\mu_1 - m)^2 + (y_1 - \mu_1)^2}{2}\right)\right).$$

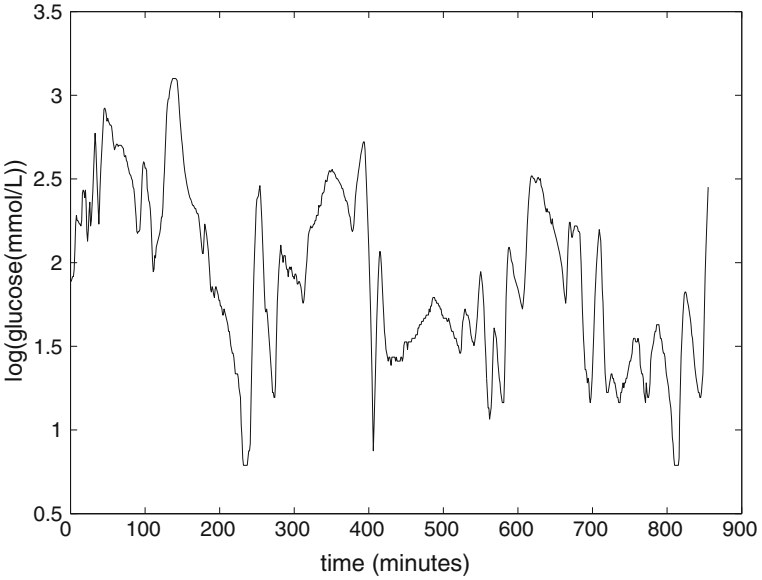

Fig. 1 Natural logarithm of the blood glucose concentration measurements for a hyperglycemic patient. Measures were taken automatically during the day on a minute-by-minute basis. Normal concentrations lie roughly in the interval $(1, 2)$, corresponding to $(3, 7)$ mmol/L, whereas this data presents the large values typical of this disease

By iterating this construction, the conditional posterior densities at time t are

$$\mu_t \mid m, y_{1:t}, \sigma_t^2 \sim N\left(m_t, \frac{\sigma_t^2}{2}\right),$$

$$\sigma_t^2 \mid a, b, m, \mu_{1:t}, y_{1:t} \sim IG\left(a + t, \frac{b}{(2\alpha)^{t-1}} + \sum_{j=1}^{t} \frac{(\mu_j - m_j)^2 + (y_j - \mu_j)^2}{2(2\alpha)^{t-j}}\right),$$

where $m_t = \frac{m}{2^t} + \sum_{j=1}^{t} \frac{y_j}{2^{t-j+1}}$.

Under this model, robustness of the posterior inferences can be effectively measured calculating the upper bounds for the variation distance (17) for alternative specifications of the prior hyper-parameters (m, a, b) and for different values of the coefficient α. Note that these bounds do in fact apply irrespective of whether the observations actually arise from the postulated likelihood function and of the form of the genuine and functioning priors, so that they could be applied to assess robustness in a much broader context.

Here we illustrate posterior robustness for a sample of minute-by-minute glucose concentration measurements (in mmol/L) taken from the blood of a patient affected by diabetes mellitus. On the natural logarithmic scale, glucose concentrations for a healthy subject lie roughly within the interval $(1, 2)$, whereas this data presents the large values typical of a hyperglycemic patient, as shown in Fig. 1.

The top-left panel in Fig. 2 shows the estimated time-varying mean along with the end-points of its 95% equal-tailed posterior probability intervals with $\alpha = 0.5$, inducing mild memory of the prior conditions over time, and using the functioning prior ($m = 1.6, a = 1, b = 0.001$). The latter hyper-parameter values are calibrated to the glucose concentrations which can be reasonably expected for a healthy subject and correspond to a 99% joint prior probability set $A_0 = \{(0.48, 3.32), (0.01, 0.98)\}$ respectively for the mean and variance of the log-glucose concentration. These prior settings correspond to what might be reasonably assumed by a modeler who does not expect the data to arise from a hyperglycemic patient. The top-right panel in the same figure shows the posterior estimates for the time-varying variance under the same prior. The posterior functionals represented in these two panels were approximated using the standard Gibbs sampler (Gelfand and Smith 1990). Note that under this model the posterior inferences for the time-varying means and variances become progressively more precise as data is accrued. In this application interest lies in assessing posterior robustness with respect to the alternative prior ($m_a = 2, a_a = 1, b_a = 0.01$), inducing a 99% joint prior probability set $\{(-0.16, 4.30), (0.02, 9.89)\}$. These settings are motivated by observing that higher mean glucose concentrations and higher variability are expected for data arising from a diabetic patient, so that we take these hyper-parameter values as identifying the genuine prior.

For the power steady model examined in this section, the TVD upper bound (17) simplifies to

$$d_V(f_t, g_t) \leq F_t(A_0) d^R_{A_0}(f_0, g_0)^{\alpha^t} + (1 - F_t(A_0)) d^R_{\Theta}(f_0, g_0), \tag{24}$$

where we have used Eq. (13). Unlike for the inequality (17), in (24) we evaluate both terms of the TVD upper bound with respect to 99% functioning prior probability set A_0. Since our functioning prior is relatively diffuse, figure 2 shows that $F_t(A_0) = 1$ for all $t > 0$ and for $\alpha = 0.5$. Should this not be the case for different values of α, the multiplier of $d^R_{\Theta}(f_0, g_0)$ in (24) can be made to converge to zero by defining an appropriate sequence of intervals $\{A_t\}$.

Under (24), the total variation upper bound is finite for any value of the parameter α when $d^R_{A_0}(f_0, g_0)$ and $d^R_{\Theta}(f_0, g_0)$ are finite. By evaluating the DR_{A_0} and the DR_{Θ} divergences over finite grids of values respectively within A_0 and over the largest possible numerical approximation of Θ, we find that $d^R_{A_0}(f_0, g_0) \approx 1{,}459$ and $d^R_{\Theta}(f_0, g_0) \approx 9{,}400$, so that the right-hand side of (24) has a finite value depending on the power-steady coefficient α. The bottom plot in Fig. 2 shows the number of observations required for the TVD upper bound (24) to become numerically zero for $\alpha = 0.01, 0.02, \ldots, 0.99$. As expected, lower values of α correspond to a lower weight of the initial conditions (m, a, b), so that the posterior distributions are practically indistinguishable after roughly 10 updates only. As α is increased, the initial conditions are given progressively more weight and the increase in the number of samples required to vanish the TVD upper bound is faster than linear. Values of α below 0.7 ensure that the functioning and genuine posterior densities converge using roughly two hours of measurements. As α approaches the value 0.95, equivalence in

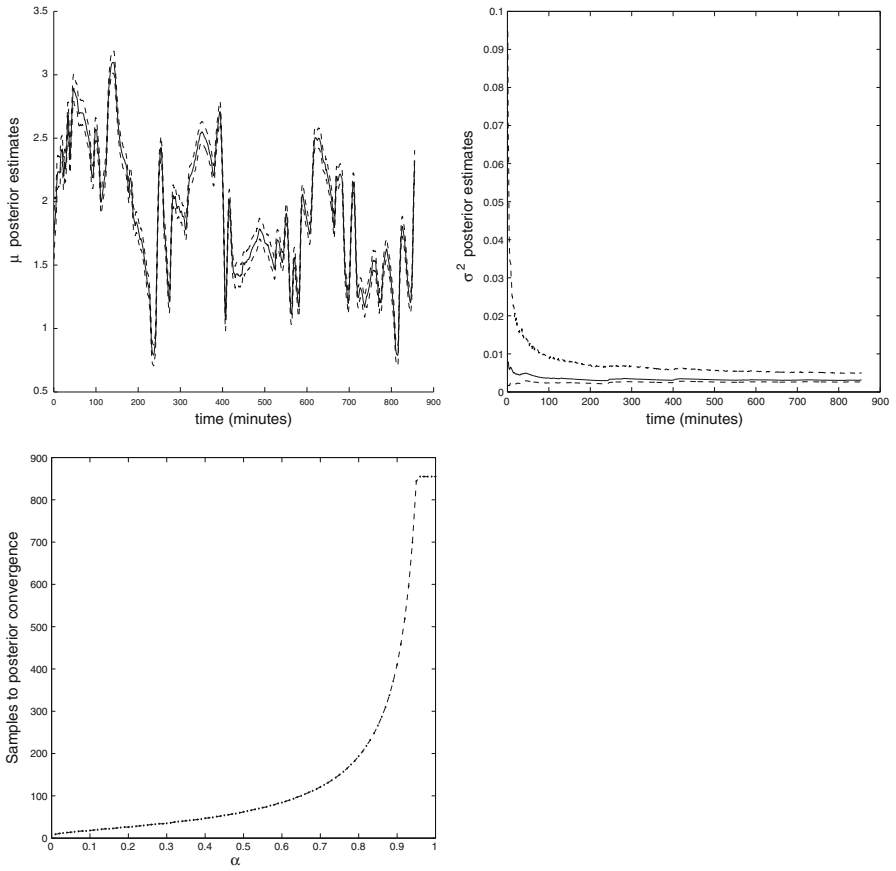

Fig. 2 Posterior estimates of the time-varying means (*top left*) and variances (*top right*) for the glucose concentrations using the functioning prior hyper-parameters ($m = 5, a = 1, b = 1.125$) and $\alpha = 0.50$. *Whole curves* represent the estimated posterior means of μ_t and σ_t^2 whereas *dashed curves* delimit their approximate 95% marginal posterior intervals. The *bottom plot* represents the number of observations required to ensure that the posterior distributions of the means and variances are numerically indistinguishable in variation distance. Increasing the coefficient α gives progressively more weight to the initial conditions. Values of α below 0.7 ensure that the functioning and genuine posterior densities converge using roughly two hours of measurements whereas for $\alpha > 0.95$ their TVD at time $t = 855$ does not vanish

total variation cannot be ensured for any of the posterior distributions of the means and variances at the 855 time points.

7 Discussion

The local separations introduced in this paper respond to a basic intuition: when a finite-dimensional posterior distribution converges on to a small set in the interior of the domain of its parameters, posterior robustness critically depends on the prior smoothness on this particular set. In this case, varying the location of the prior mass

has little impact on the posterior estimates. Provided that the priors f_0 and g_0 are close with respect to these new local separations, the functioning posterior distribution provides a good approximation to the genuine posterior even when the family of sampling models is inconsistent with the data. All similar priors will give similar, if possibly erroneous, posterior densities. In particular, using 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, as shown in Sect. 6, under three conditions. The first is that the same likelihood is shared by the two priors. This commonly assumed but very strong condition is absolutely critical in that if this is not the case posterior inferences will typically diverge with increasing sample size (Smith 2007). From this perspective, it is important to note that the local separations introduced in this paper measure posterior robustness with respect to the likelihood of the observed data and not by averaging with respect to the sample space (Berger and Wolpert 1984). This makes the TVD upper bounds (17) a general yet computable criterion for measuring posterior robustness in practical applications. The second condition is that the functioning posterior 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. Thirdly both priors need to be comparably rough, as emphasized in Sect. 4. This key property is often implicitly but not always thoughtfully induced by the way hierarchical priors are currently specified.

Under these three conditions, Theorem 2 gives a TVD upper bound in terms of differences in prior roughness within neighborhoods of high functioning posterior probability and in terms of the prior tail behavior. We note that the counterexamples constructed in Gustafson and Wasserman (1995), which demonstrate that conventional smoothness bounds and convergence in prior variations are not sufficient to ensure posterior convergence in most common models, all exhibit divergent local De Robertis separations because they are too rough in the sense of Definition 2. Theorem 2 also shows that the second necessary condition for posterior convergence in total variation is that the tails of the functioning prior are not too tight, so that the information in the likelihood is not dominated by the chosen prior. Otherwise, the mass of the functioning posterior may well remain far away from high likelihood values where posteriors from a genuine prior—close in variation distance but with a thicker tail—might concentrate its mass, as noticed by Dawid (1973).

Useful variation bounds can sometimes be obtained even when the functioning posterior does not converge in distribution (Smith 2007). On the other hand, whilst the continuity of f_0 and g_0 can be relaxed, their mutual roughness conditions given above are in practice necessary for posterior robustness. If f_0 and g_0 do not lie within a local separation neighborhood about a particular location θ_0 then no matter how small is the radius of that neighborhood it is possible to construct a sequence of likelihoods that converge 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 Smith (2007) based on a counterexample used in Gustafson and Wasserman (1995).

8 Appendix

To prove (9) assume that $f(\theta)$ and $g(\theta)$ are continuous at $\widetilde{\theta}$ and $\widetilde{\phi}$ where $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ and $\widetilde{\phi} = (\widetilde{\theta}_1, \widetilde{\phi}_2)$ where

$$\widetilde{\theta}_1 = \arg\sup_A(\log f_1(\theta) - \log g_1(\theta)),$$
$$\widetilde{\phi}_1 = \arg\inf_A(\log f_1(\theta) - \log g_1(\theta)),$$

the value $\widetilde{\theta}_2$ 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 $(\widetilde{\theta}_2, \widetilde{\phi}_2)$ must exist because $f_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)$ and $g_{2|1}(\widetilde{\theta}_2|\widetilde{\theta}_1)$ are probability 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

$$d_A^L(f,g) \geq d_{A_1}^L(f_1,g_1). \tag{25}$$

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

Acknowledgments This paper has greatly benefited from discussions with Ali Daneshkah, Jim Griffin, Jon Warren, Wilfred Kendall, Sigurd Assing and Stephen Walker. The glucose concentration data analysed in Sect. 6 has been kindly provided by Tim Holt, Warwick Medical School.

References

Abraham, C., Cadre, B. (2004). Asymptotic Bayesian robustness in Bayesian decision theory. *The Annals of Statistics, 32*, 1341–1366.

Andrade, J. A. A., O'Hagan, A. (2006). Bayesian robustness modelling using regularly varying distributions. *Bayesian Analysis, 1*, 169–188.

Berger, J. O. (1992). Recent methodological advances in robust Bayesian inference (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith (Eds.), *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting* (pp. 495–496). Oxford: Clarendon Press [In discussion of Wasserman, L.(1992b)].

Berger, J. O., Wolpert, R. L. (1984). The likelihood principle. In S. S. Gupta (Ed.), *IMS Lecture Notes* (Vol. 6.) Hayward, CA: IMS.

Bernardo, J. M., Smith, A. F. M. (1996). *Bayesian theory*. Chichester: Wiley.

Copas, J., Eguchi, S. (2010). Likelihood for statistically equivalent models. *Journal of the Royal Statistical Society B, 72*, 193–217.

Daneseshkhah, A. (2004). Estimation in causal graphical models. PhD Thesis, University of Warwick, UK.
Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika, 60*, 664–667.
De Robertis, L. (1978). The use of partial prior knowledge in Bayesian inference. PhD Thesis, Yale University, CT, USA.
Fernandez, C., Osiewalski, J., Steel, M. (1996) Classical and Bayesian inference robustness in multivariate regression models. *Journal of the American Statistical Association, 92*, 1434–1444.
Gelfand, A. E., Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.
Geyer, C. J., Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association, 90*, 909–920.
Ghosh, J. K., Ramamoorthi, R. V. (2003). Bayesian nonparametrics. *Springer Series in Statistics*. New York: Springer.
Gustafson, P. (1996). Aspects of Bayesian robustness in hierarchical models. In J. O. Berger, B. Betro, E. Moreno, L. R. Pericchi, F. Ruggeri, G. Salinetti, L. Wasserman (Eds.), *IMS Lecture Notes* (Vol. 29, pp. 81–100). Hayward, CA: IMS.
Gustafson, P., Bose, S. (1996). Aspects of Bayesian robustness in hierarchical models. In J. O. Berger, B. Betro, E. Moreno, L. R. Pericchi, F. Ruggeri, G. Salinetti, L. Wasserman (Eds.), *IMS Lecture Notes* (Vol. 29, pp. 63–80). Hayward, CA: IMS.
Gustafson, P., Wasserman, L. (1995). Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics, 23*, 2153–2167.
Huber, P. J. (1997). Robustness: where are we now? In Y. Dodge (Ed.), *IMS Lecture Notes* (Vol. 31, pp. 487–498). Hayward, CA: IMS.
Kadane, J., Srinivasan C., Salinetti, G. (1996). Bayesian robustness and stability. In J. O. Berger, B. Betro, E. Moreno, L. R. Pericchi, F. Ruggeri, G. Salinetti, L. Wasserman (Eds.), *IMS Lecture Notes* (Vol. 29, pp. 81–100). Hayward, CA: IMS.
Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Science, 220*, 671–680.
Martin, J., Rios Insua, D., Ruggeri, F. (1998). Issues in Bayesian loss robustness. *The Indian Journal of Statistics (A), 60*, 405–416.
Monhor, D. (2007). A Chebyshev inequality for multivariate normal distributions. *Probability in the Engineering and Informational Sciences, 21*, 289–300.
Moran, P. A. P. (1968). *An introduction to probability theory*. Oxford: Oxford University Press.
O'Hagan, A. (1979). On outlier rejection phenomena in Bayesian inference. *Journal of the Royal Statistical Society B, 41*, 358–367.
O'Hagan, A. (2006). Eliciting expert beliefs in substantial practical applications. *The Statistician, 47*, 21–35.
O'Hagan, A., Forster, J. (2004). Bayesian inference. In *Kendall's Advanced Theory of Statistics* (Vol. 2B). London: Arnold.
Peterka, V. (1981). Bayesian system identification. In P. Eykhoff (Ed.), *Trends and Progress in System Identification* (pp. 239–304). Oxford: Pergamon Press.
Poole, A., Raftery, A. (2000). Inference for deterministic simulation models: The Bayesian melding approach *Journal of the American Statistical Association, 95*, 1244–1255.
Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.
Smith, J. Q. (1979). A generalisation of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society B, 41*, 375–387.
Smith, J. Q. (2007). Local robustness of Bayesian parametric inference and observed likelihoods. *CRiSM Research Report 07-09*. University of Warwick, UK.
Tong, Y. L. (1980). *Probability inequalities in multivariate distributions*. New York: Academic Press.
Wasserman, L. (1992). Invariance properties of density ratio priors. *The Annals of Statistics, 20*, 2177–2182.
West, M., Harrison, P. J. (1997). Bayesian forecasting and dynamic models. In *Springer Series in Statistics*. New York: Springer.