Local Robustness of Bayesian Parametric Inference and Observed Likelihoods

J.Q.Smith.
Department of Statistics,
The University of Warwick,
Coventry CV4 7AL

August 6, 2007

Abstract

Here a new class of local separation measures over prior densities is studied and their usefulness for examining prior to posterior robustness under a sequence of observed likelihoods, possibly erroneous, illustrated. It is shown that provided an approximation to a prior distribution satisfies certain mild smoothness and tail conditions then prior to posterior inference for large samples is robust, irrespective of whether the priors are grossly misspecified with respect to variation distance and irrespective of the form or the validity of the observed likelihood. Furthermore it is usually possible to specify error bounds explicitly in terms of statistics associated with the posterior associated with the approximating prior and asumed prior error bounds. These results apply in a general multivariate setting and are especially easy to interpret when prior densities are approximated using standard families or multivariate prior densities factorise.

1 Introduction

Let f_0 and g_0 denote two prior densities and f_n and g_n their corresponding posterior densities after observing a sample $\boldsymbol{x}_n$, $n \geq 1$. Suppose we implement a Bayesian analysis using a prior density f_0 - henceforth called our *functioning prior* - instead of g_0 -our *genuine prior*: i.e. the one we would specify if we had enough time and skill to elicit it perfectly. Our interest in this paper focuses on examining when Bayesian inference using a posterior density f_n after n observations based on the functioning prior f_0 provides a good approximation for inferences based on g_0 as a sample size grows.

Suppose the posterior density $f_n(\boldsymbol{\theta})$ can be used to calculate an optimal decision $d^*(f_n) \in \mathbb{D}$ associated with *any* utility functions U as a function of $(d, \boldsymbol{\theta})$ - for example a utility associated with the prediction of *any* possible type of future observable associated with the context of the model - by choosing

$d^*(f_n) = \arg\max_{d\in\mathbb{D}} \overline{U}(d, f_n)$, where $\overline{U}(d, f)$ denotes the expected utility under decision d utiltiy U and density f. Let $d_V(f, g) = \int_\Theta |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})| d\boldsymbol{\theta}$ denote the *variation distance* between densities f and g. Then if $d_V(f_n, g_n) < \varepsilon$ it is trivial to check that for any utility U in the class $\mathbb{U}$ of all measurable utility functions bounded below by 0 and above by 1, on a decision space $\mathbb{D}$

$$\left|\overline{U}(d^*(f_n), f_n) - \overline{U}(d^*(f_n), g_n)\right| \leq \sup_{d\in\mathbb{D}} \left|\overline{U}(d, f_n) - \overline{U}(d, g_n)\right| < \varepsilon$$

It follows that closeness in variation distance of the functioning posterior density $f_n$to the genuine posterior density g_n guarantees all decisions will be close to optimal in this sense. This is the most we could reasonably hope to achieve from approximate inference of this kind - see [15],[14] for a careful discussion of this issue. Henceforth in this paper we will measure adequacy of approximation in terms of this metric.

Bayesian robustness issues are commonly addressed through examining when both the sequences of posterior functioning densities$\{f_n\}_{n\geq 1}$ and genuine $\{g_n\}_{n\geq 1}$, are consistent [24]. Thus assume a random sample $\boldsymbol{x}_n$ each have a distribution with parameter $\boldsymbol{\theta}_0 \in \Theta$ in a finite dimensional family indexed by Θ. When $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$ are both consistent and continuous at $\boldsymbol{\theta}_0 \in \Theta^0$ where Θ^0 is the interior of Θ it is straightforward to prove [see e.g.[12] or [11], p18].that, $\lim_{n\to\infty} d_V(f_n, g_n) = 0$. almost surely $P_{\boldsymbol{\theta}_0}$. This in turn implies that f_n will provide a good working approximation for g_n for *all* reasonable estimation purposes in the sense above, given consistency..

For finite parametric inference this consistency condition is apparently not too restrictive. Thus for example it is shown in [24] p429 that consistency of a posterior sequence will automatically follow for both $\{f_n\}_{n\geq 1}$ and $\{g_n\boldsymbol{\theta})\}_{n\geq 1}$ provided that a consistent estimator of $\boldsymbol{\theta}$ can be constructed and the parametrisation of $\boldsymbol{\theta}$ respects Kullback- Leibler separations in the sense of Theorem 7.80 in [24]. However 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 provided our functioning posterior distribution concentrates near to a value $\boldsymbol{\theta}_0 \in \Theta$, then $\lim_{n\to\infty} d_V(f_n, g_n) = 0$. Ideally we would like this property for *any* given *observed* likelihood, even when the sample distribution family is not accurately specified and the data is not a random sample. In this paper we prove that, under mild smoothness and regularity conditions this is also true. In Section 2 we introduce a new family of separation measures that are closely related to density ratio separation measures [9], [30], [23], but redefined so that they apply locally. In Section 3 we examine various useful properties of these topologies and proceed in Section 4 to show 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 assert that her functioning and genuine prior would be close.

In Section 5 the "isoseparation property" is used to show that, provided we believe that the genuine prior density is in one of these coarse neighbourhoods of the functioning prior we obtain convergence in variation between the functioning and genuine posteriors. Furthermore we are also usually able to find *explicit*

bounds for the approximation as a function of parameters we already know: specifically the parameters associated with our beliefs about the accuracy of our functioning prior and certain statistics of our calculated posterior density. These bounds apply irrespective of whether the observed data is consistent with the family of sample distributions underpinning the observed likelihood and make the robustness operational. The bounds can even apply when the functioning posterior does not converge in distribution to a degenerate distribution In Sections 6 we examine the neighbourhoods of various well - known classes of models more closely and show that within these classes the neighbourhoods are extremely natural.

2 Density ratio balls and Isoseparation

Assume we receive a sequence of *observed* sample densities $\{p(\boldsymbol{x}_n|\boldsymbol{\theta})\}_{n\geq 1}$of an n -vector of data $\boldsymbol{x}_n = (x_1, x_2, \ldots x_n)$

For the purposes of this paper we will assume that the genuine prior $g_0(\boldsymbol{\theta})$ is strictly positive and continuous on the interior of its support - and so uniquely defined, and each observed likelihood, $p(\boldsymbol{x}_n|\boldsymbol{\theta})$, $n \geq 1$ is measurable with respect to $g_0(\boldsymbol{\theta})$. We will calculate our posterior density $g_n(\boldsymbol{\theta}) \triangleq g(\boldsymbol{\theta}|\boldsymbol{x}_n)$ after n observations using Bayes Rule using the usual formula. Thus let $\Theta(n) = \{\boldsymbol{\theta} \in \Theta : p(\boldsymbol{x}_n|\boldsymbol{\theta}) > 0\}$. Then for all $\boldsymbol{\theta} \in \Theta(n)$ our genuine posterior density is given by the formula

$$\log g_n(\boldsymbol{\theta}) = \log g_0(\boldsymbol{\theta}) + \log p_n(\boldsymbol{\theta}) - \log p_g(\boldsymbol{x}_n) \tag{1}$$

The predictive density

$$p_g(\boldsymbol{x}_n) = \int_{\boldsymbol{\theta}\in\Theta(n)} p(\boldsymbol{x}_n|\boldsymbol{\theta}) g_0(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

is usually calculated indirectly, either algebraically or numerically so as to ensure $g_n(\boldsymbol{\theta})$ integrates to 1. For all $\boldsymbol{\theta} \in \Theta\backslash\Theta(n)$ we simply set $g_n(\boldsymbol{\theta}) = 0$. In this paper we call this the *formal Bayesian updating formula* and henceforth assume that our inference is sufficiently regular that it is appropriate to use it. We shall say that $f(\boldsymbol{\theta})$ is a $1-\alpha$ *concentrate* on an open set $B(\boldsymbol{\theta}_0, \rho)$, centred at $\boldsymbol{\theta}_0$ and of radius ρ, if $F(\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \rho)) \leq \alpha$, writing $f(\boldsymbol{\theta}) \in C^{\alpha}(\boldsymbol{\theta}_0, \rho)$.

Let $f_n(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}|\boldsymbol{x}_n)$, $\boldsymbol{\theta} \in \Theta(n)$ represent our functioning posterior density, and suppose that the functioning posterior density $f_n(\boldsymbol{\theta}) \in C^{\alpha_n}(\boldsymbol{\theta}_0^n, \rho_n)$ converges in distribution to a point mass in the closure neighbourhood of $\boldsymbol{\theta}_0 \in \Theta(n)$. We would then like to be able to state that for $g_0 \in \mathcal{N}$ where $\mathcal{N}$ is a suitably chosen neighbourhood of f_0, $d_V(f_n, g_n) < \varepsilon_n(\alpha_n, \rho_n)$ where $\varepsilon_n(\alpha_n, \rho_n) > 0$ - *whatever* $\{p_n(\boldsymbol{\theta})\}_{n\geq 1}$ is - and where ε_n is an explicit function of prior parameters and functions of f_n.

It has long been established that this is not so in general. For example [6], [22] [1] prove that even when the functioning posterior convergences to a defective distribution the variation distance between $f_n(\boldsymbol{\theta})$ and $g_n(\boldsymbol{\theta})$ cannot be

guaranteed. Thus suppose $\boldsymbol{x}_n$ is a random sample of n normal $N(\theta, 1)$ random variables, the prior $f_0(\theta)$ is a normal $N(0, 1)$ density whilst the genuine prior $g_{0,\varepsilon}(\theta)$ takes the form

$$g_{0,\varepsilon}(\theta) = (1-\varepsilon)f_0(\theta) + \varepsilon h(\theta)$$

where $0 < \varepsilon < 1$ and $h(\theta)$ is the Cauchy density given by

$$h(\theta) = \pi^{-1}(1+\theta^2)^{-1}$$

Clearly $d_V(f_0, g_{0,\varepsilon}) < \varepsilon$. However the posterior density f_n under prior f_0 is $N(\mu_f, (n+1)^{-1})$ where $\mu_f = n(n+1)^{-1}\overline{x}_n$ whilst under prior $g_{0,\varepsilon}$ it can be shown (see for example [6]) that the posterior density $g_{n,\varepsilon} \simeq h_n$ when $\overline{x}_n^2$ is sufficiently large, which is approximately distributed $N(\overline{x}_n, n^{-1})$. It follows that for any fixed n and $\varepsilon > 0$, for a sequence $\{\overline{x}_n\}_{n\geq 1}$ diverging (to ∞ say)

$$\lim_{n\to\infty} d_V(f_n, g_{\varepsilon,n}) = 1$$

Note that this does not contradict the consistency result quoted above because, for any $\theta_0 \in \Theta$, $\lim_{n\to\infty} \overline{X}_n$ is finite with probability one whilst in our construction $\overline{x}_n$ diverges. But it demonstrates how fragile the demand for consistency is. For example it is elementary to construct an infinitely differentiable sample density $\widetilde{q}(x|\theta)$ - arbitrarily close to the normal density $q(x|\theta)$ given above in variation distance and to within drawing accuracy to $q(x|\theta)$ - where the distribution of $\overline{X}_n$ diverges with n with probability one. So in this sense under the slightest contamination of the sample distribution the two posterior densities may in fact diverge. We show below that this divergence phenomenon occurs because f_0 and $g_{0,\varepsilon}$ are not equicontinuous in an appropriate topology at the point $\boldsymbol{\theta}_0$ of convergence - here ∞.

We are also interested in determining explicit rates of convergence all of whose parameters are given from the problem description. It has been established that such issues cannot be directly addressed using variation distance alone. Thus Gustafson &Wasserman [9] proved that, for almost all parametric inference, the ratio

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

of the supremum over a neighbourhood $\mathcal{N}$ of f_0 the prior and posterior variation distance almost always *diverged* rather than converged with increasing sample size n, even when data are drawn from a "true" density indexed by the parameter $\boldsymbol{\theta}_0 \in \Theta^0$ .Furthermore this is 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 - 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. They further proved that when a random sample is drawn from observations whose sample density is

labelled by a $\boldsymbol{\theta}_0 \in \Theta^0$, under certain mild regularity conditions, the rate of divergence of this ratio is $n^{k/2}$, where k is $\dim \Theta$ with probability 1. So in this sense the precise specification of our prior appeared to become more and more critical the more information we gathered and is dramatic when the dimension of the parameter space is large!

However the problem here is that 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 $\mathcal{N}$ they specify [9]. So this Frechet derivative cannot be expected to do the job we require. Here we show $\sup_{g\in\mathcal{N}} \{d_V(f_n, g_n)\}$ can be bounded and typically decreases in powers of n for neighbourhoods $\mathcal{N}$ much less tight than the ones specified in [9].

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 Say that g lies in the $A-$*DeRobertis ball $B_A(f, \delta)$ *iff*

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

In fact if the functions f and g are continuous on their shared support - an assumption we will make in this paper unless specifically stating otherwise, this formula simplifies thus

Definition 2 *Let $A \subseteq \Theta$ be measurable. Then g is said to lie in the $A-$*DeRobertis ball $B_A(f, \delta)$ *of f iff*

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

*and write $d^R(f, g) \triangleq d_\Theta^R(f, g)$. Call $d_A^R(f, g)$ the $A-$*DS.

The separation measure defined by [7] and cited above is $d^R(f, g)$.

An equivalent separation measure to $d^R(f, g)$ is called the *density ratio separation* and will be denoted by $d^L(f, g)$. This has an obvious extension to an *$A-$density ratio separation* which we will denote by $d_A^L(f, g)$ - and for continuous densities is given by

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

where $d^L(f, g) = d_\Theta^L(f, g)$. Note that

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

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 these functions 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 $d(f,g) = d(g,f)$. On the other hand, unless $A = \Theta$, $\delta(f,g) = 0 \nRightarrow f = g$ nor that the induced neighbourhood system is equimeasurable: that is invariant to permutation - like transformations of the Θ space [see [19] for a precise definition of this term]. However on all measurable sets $A \subseteq \Theta$ and sets $\mathcal{F}$ with the property that for all $f, g \in \mathcal{F}$, $d_A^L(f,g)$, $d_A^L(f,g) < \infty$ $d_A^L(f,g)$ is a semi-metric, since it is easily checked that the triangle inequality is also satisfied.

Note that the smaller the set A the coarser the neighbourhoods defined by $d_A^L(f,g) < \eta$: if $A_1 \subset A_2$ then $d_{A_1}^L(f,g)$ is weaker than $d_{A_2}^L(f,g)$. Also note that when the lower bound and upper bound are attained

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

where

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

So in particular $d_A^L(f,g)$ is the difference of the two log densities at their maximum and minimum value within a set A. These separations therefore have a simple and transparent interpretation.

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

3.1 Isoseparation

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 is that for all $n \geq 1$,for $\boldsymbol{\theta} \in A \subseteq \Theta(n)$

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

So in particular, when the observed likelihood $p_n(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$, for all for any measurable subset A of Θ

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

Thus unlike the variation distance analogue 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 they do not get closer either: the assumptions you make about neighbourhoods endure for all time regardless of what is observed - provided $p(\boldsymbol{x}_n|\boldsymbol{\theta}) > 0$ for any $\boldsymbol{\theta} \in A$. In this case the data simply shifts us in an invariant way around the space of densities. Note that in general when $\Theta(n) \subset \Theta$

$$\sup_{g_o} \left\{ \frac{d_A^L(f_n, g_n)}{d_A^L(f_0, g_0)} \right\} \leq 1 \tag{9}$$

with equality if and only if there exists a pair $\left(\boldsymbol{\theta}_{u(A,f_0,g_0)}, \boldsymbol{\theta}_{l(A,f_0,g_0)}\right) \in \overline{\Theta}(n)$ where $\overline{\Theta}(n)$ is the closure of $\Theta(n)$ if $\left(\boldsymbol{\theta}_{u(A,f_0,g_0)}, \boldsymbol{\theta}_{l(A,f_0,g_0)}\right)$ are defined.

This property when $A = \Theta$ has in fact been known for a very long time. However $A = \Theta$ is perhaps 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 - see below. The most useful of these separations is when A is small because when studying posterior behaviour we are usually interested in small areas of the parameter space on to which the posterior is concentrating Because this paper largely focuses on convergence results here we let $A = B(\boldsymbol{\theta}_0, \rho)$ where $B(\boldsymbol{\theta}_0, \rho)$ denotes the open ball centred on $\boldsymbol{\theta}_0$ of radius ρ. We write $d^R_{\Theta_0,\rho}(f_n, g_n) \triangleq \sup\{d^R_{B(\boldsymbol{\theta}_0,\rho)}(f, g) : \boldsymbol{\theta}_0 \in \Theta_0\}$, $d^R_\rho(f, g) \triangleq \sup\{d^L_{B(\boldsymbol{\theta}_0,\rho)}(f, g) : \boldsymbol{\theta}_0 \in \Theta\}$, $(d^L_{\Theta_0,\rho}(f, g) \triangleq \sup\{d^L_{B(\boldsymbol{\theta}_0,\rho)}(f, g) : \boldsymbol{\theta}_0 \in \Theta_0\}$ and $d^L_\rho(f, g) \triangleq \sup\{d^L_{B(\boldsymbol{\theta}_0,\rho)}(f, g) : \boldsymbol{\theta}_0 \in \Theta\})$. Note that all these are separation measures in the sense above and that for all $\rho > 0$. when $\mathcal{F}$ is any subset of densities for which $d^L_\rho(f, g) < \infty$, $d^L_\rho(f, g)$ is in fact a metric.

Unlike $d^R_A(f, g)$, $d^R_{\Theta_0,\rho}(f, g)$ is a function of parametrisation we use. So in particular to obtain invariance of convergence to transformations $\mathbb{T} : \Theta \to \Theta$, of the parameter space we may require that the reparametrising map $\mathbb{T}$ is a diffeomorphism and open sets under the two parametrisations map invertibly to one another: see below. However this restriction is usually consistent with a parametric model where parametrisations are usually *specifically chosen* so the sample densities with similar parameter values are "close" and certainly where their interpretations are close. Demanding that neighbourhood system be invariant to arbitrary measurable reparametrisations, as does [30], appears inappropriate for the parametric purposes of this paper.

3.2 Isoseparation minus nuisance parameters

Note that if $\{p_n(\boldsymbol{\theta})\}_{n\geq 1}$ are not explicit functions of $\boldsymbol{\theta}_2$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $f_{0,1}$ 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 $\boldsymbol{\theta}_1$ then when the observed likelihood $p_n(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \Theta$, these marginal densities inherit the isoseparation property. Thus for all $n \geq 1$,for $\boldsymbol{\theta} \in A \subseteq \Theta(n)$

$$d^L_A(f_{n,1}, g_{n,1}) = d^L_A(f_{0,1}, g_{0,1})$$

This property becomes important in for example 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 3 *Suppose that the joint distribution of our observations X_n have a sample distribution depending on $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the vector of means of X_n and $\boldsymbol{\Sigma}$ a vector of other hyperparameter, for example variances and covariances, which together with $\boldsymbol{\mu}$ uniquely specify this distribution. To specify the*

prior on $\boldsymbol{\mu}$ it is common practice to extend this model so that

$$\boldsymbol{\mu} = \boldsymbol{\tau}(\boldsymbol{\phi}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\phi}$ is a low dimensional vector $\boldsymbol{\tau}$ is a known function (often linear) and $\boldsymbol{\varepsilon}$ is an error vector parametrised by a vector $\boldsymbol{\Lambda}$ of for example covariances Often our utility will be a function only of $\boldsymbol{\theta}$ through $\boldsymbol{\theta}_1$. The invariance property above then allows us to substitute $\boldsymbol{\theta}_1$ for $\boldsymbol{\theta}$ in all the examples of robust inference discussed below. Note that in simple linear models with unknown variances, $\boldsymbol{\mu}$ will often have a marginal density with inverse polynomial tails: for example multivariate Student t- distributions. We will see later that such priors give rise to d_A^L neighbourhoods exhibiting strong robustness properties.

3.3 Conditioning and Local Invariance

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

$$d_A^R(f_B, g_B) = d_A^R(f, g) \tag{10}$$

This property is important in its own right because it demonstrates an invariance under conditioning which the variation distances between conditioned densities do not possess. Note that in particular this implies that $d_A^R(f, g) = d^R(f_A, g_A)$

A third immediate property henceforth called *local invariance* we will use below is that if there exists a $\boldsymbol{\theta}_{u(A,f,g)}$and a $\boldsymbol{\theta}_{l(A,f,g)} \in A \subseteq B$ where these parameter values are defined in 6 then

$$d_A^R(f, g) = d_B^R(f, g)$$

3.4 Marginal Separation

Most current applications of Bayesian inference are in high dimensional, but also often highly structured spaces: breaking up the space into modules of locally specifies neighbourhoods. Common examples of these classes are state space time series, hierarchical models and Bayesian networks. It is therefore interesting to examine to what extent these separation measures respect the sorts of modularity these classes exhibit.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ be two candidate parameter values in $\Theta = \Theta_1 \times \Theta_2$ where $\boldsymbol{\theta}_1, \boldsymbol{\phi}_1 \in \Theta_1$ and $\boldsymbol{\theta}_2, \boldsymbol{\phi}_2 \in \Theta_2$, where the joint densities $f(\boldsymbol{\theta}) = f_1(\boldsymbol{\theta}_1) f_{2|1}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$, $g(\boldsymbol{\theta}) = g_1(\boldsymbol{\theta}_1) g_{2|1}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ and $f_1(\boldsymbol{\theta}_1), g_1(\boldsymbol{\theta}_1)$ are the marginal densities on Θ_1 of the two joint densities $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ respectively. Suppose, that $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ are continuous at $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\phi}}$ where $\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)$and $\widetilde{\boldsymbol{\phi}} = \left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\phi}}_2\right)$ where $\widetilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_{u(A,f_1,g_1)}$ and $\widetilde{\boldsymbol{\phi}}_1 = \boldsymbol{\theta}_{l(A,f_1,g_1)}$ and $\widetilde{\boldsymbol{\theta}}_2$ is any point satisfying $f_{2|1}(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1) \geq g_{2|1}(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1)$ and $\widetilde{\boldsymbol{\phi}}_2$ is any point satisfying $f_{2|1}(\widetilde{\boldsymbol{\phi}}_2|\widetilde{\boldsymbol{\phi}}_1) \leq$

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

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

and therefore

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

where $A_1 = \{\boldsymbol{\theta}_1 : \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in A$ for all $\boldsymbol{\theta}_2 \in B \subset \Theta_2$ for some open set B in $\Theta_2\}$, the property we require. Note that when $f_{2|1}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1) = g_{2|1}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1)$ this is an equality.

On the other hand if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_k)$ where all these subvectors $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_k\}$ 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 $\boldsymbol{\theta}_i$ margin of $f(g)$, $1 \leq i \leq n$. Note that if we a priori set $\boldsymbol{\mu} \coprod \boldsymbol{\Sigma}$ in the hierarchical model above then the equation above gives a simple decomposition of the prior separation between the functioning and genuine prior which will hold a posteriori. So in this sense the consequences of this commonly made prior independence assumption endures a posteriori.

3.5 Conditional Separation

There is a similar relationship that equates the distance between two joint densities and two of their conditionals. Thus without loss of generality assume that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k) \in \Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_k$ is such that Θ_k is the line segment joining $\boldsymbol{\theta}^u_{A,f,g}$ to $\boldsymbol{\theta}^u_{A,f,g}$. and $\widehat{\theta_1} = (, \theta_2, \ldots \theta_k)$ parametrises its orthogonal space. (Otherwise perform a rotation of the space so that this is so and note that all local distances we define above - under the appropriate transformations of A are invariant to such rotations and in particular the radius of all balls are preserved) Then we can see immediately from their definition, since proportionality constants cancel, that

$$d_A^L(f,g) = \sup_{\boldsymbol{\theta}\in\Theta} d_A^L(f(\theta_1|\widehat{\theta_1}), g(\theta_1|\widehat{\theta_1}) \geq \sup_{\boldsymbol{\theta}'\in\Theta} d_A^L(f(\theta_1'|\widehat{\theta_1'}), g(\theta_1'|\widehat{\theta_1'}) \tag{12}$$

for any invertible linear map $\boldsymbol{\theta} \to \boldsymbol{\theta}$. Further inequalities relating especially to Bayesian Networks are discussed below.

3.6 When Likelihoods are not held in Common

In Section 2 we discussed the lack of stability of consistency based limiting results to the misspecification of the likelihood. So far we have considered only the possibility that this likelihood is misspecified in the same way for both the functioning and genuine priors, but this need not be so. Write $p_n^f(\boldsymbol{\theta}) \triangleq \frac{p^f(\boldsymbol{x}_n|\boldsymbol{\theta})}{\sup_\theta p^f(\boldsymbol{x}_n|\boldsymbol{\theta})}$ and $p_n^g(\boldsymbol{\theta}) \triangleq \frac{p^g(\boldsymbol{x}_n|\boldsymbol{\theta})}{\sup_\theta p^g(\boldsymbol{x}_n|\boldsymbol{\theta})}$ where these likelihoods are no longer assumed the same. Thus suppose the priors f_0 and g_0 have different respective associated likelihoods

$$p_n^f(\boldsymbol{\theta}) = \prod_{i=1}^n q_i^f(\boldsymbol{\theta}),\ p_n^g(\boldsymbol{\theta}) = \prod_{i=1}^n q_i^g(\boldsymbol{\theta})$$

where now we do not assume that $q_i^f(\boldsymbol{\theta}) = q_i^g(\boldsymbol{\theta})$, $1 \leq i \leq n$.where $0 \leq q_1^f(\boldsymbol{\theta}) = p_1^f(\boldsymbol{\theta}) \leq 1$ and

$$0 \leq q_i^f(\boldsymbol{\theta}) = \frac{p^f(x_i|x_{1,\ldots}x_{i-1},\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} p^f(x_i|x_1,\ldots x_{i-1},\boldsymbol{\theta})} \leq 1 \tag{13}$$

with an analogous formula for $q_i^g(\boldsymbol{\theta})$. From Bayes Rule we now calculate that for all $f_0, g_0 \in \mathcal{F}$ where$\mathcal{F}$ is any subset of continuous densities given above is that for all $n \geq 1$,for $\boldsymbol{\theta} \in A \subseteq \Theta(n)$

$$\begin{aligned} d_A^L(f_n, g_n) &= \sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in A} \{\log f(\boldsymbol{\theta}) - \log g(\boldsymbol{\theta}) - \log f(\boldsymbol{\phi}) + \log g(\boldsymbol{\phi}) && (14)\\ &\quad + \log q_i^f(\boldsymbol{\theta}) - \log q_i^g(\boldsymbol{\theta}) - \log q_i^f(\boldsymbol{\phi}) + \log q_i^g(\boldsymbol{\phi})\} && (15)\\ &\leq d_A^L(f_0, g_0) + \delta_A^L(p^f(\boldsymbol{\theta}), p^g(\boldsymbol{\theta})) && (16)\\ &\leq d_A^L(f_0, g_0) + \sum_{i=1}^n \delta_A^L(q_i^f, q_i^g) && (17) \end{aligned}$$

where

$$\delta_A^L(q_i^f(\boldsymbol{\theta}), q_i^g(\boldsymbol{\theta})) = \sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in A} \Big(\log q_i^f(\boldsymbol{\theta}) - \log q_i^g(\boldsymbol{\theta})\Big) - \Big(\log q_i^f(\boldsymbol{\phi}) - \log q_i^g(\boldsymbol{\phi})\Big) \tag{18}$$

the natural analogue to d_A^L but applied to likelihoods. Henceforth let δ inherit the notation for d given above. Note that in the particular case when $q_i^f = q^f$ and $q_i^g = q^g$ as for example would be the case after observing a random identically distributed sample $\boldsymbol{x}_n$ with all observations equal then it is easy to check that

$$d_A^L(f_n, g_n) \geq n\delta_A^L(q^f, q^g) \tag{19}$$

So in this worst case scenario a data set can cause divergence between the functioning and genuine posterior with respect to these separation measures for any A for which $\sup_{g\in\mathcal{N}} \delta_A^L(q^f, q^g) \neq 0$. This is in fact reflected in the complementary result in [6] when we switch the roles of likelihood and prior. The message here is clear. Models (f, p^f) whose data generating process p^f

depend on the explanation whose related uncertainty is expressed through f are inherently less robust than models where this is not so: see [25] for a related point albeit made in a different context. The lack of large sample robustness of *any* likelihood based method is of course well established and this is simply a very special example of this phenomenon.

However note that if we can parametrise $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ so that $\boldsymbol{\theta}$ maps to $\boldsymbol{\theta}_1$ and

$$q_i^f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = q_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2[f]), q_i^g(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = q_i(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2[g])$$

for some functions q_i, $i = 1, 2, \ldots$ so that the discrepancy in the two likelihoods is explained by different values of hyperparameters, whose prior respective densities $f_{2,0}$ and $g_{2,0}$ are independent of $f_{1,0}$ and $g_{1,0}$ respectively then from (11) and the isoseparation property

$$d_A^L(f_n, g_n) = d_A^L(f_{1,0}, g_{1,0}) + d_A^L(f_{2,0}, g_{2,0}) \tag{20}$$

i.e. the awkward dependence on sample size is removed. So if it is possible to represent the associated uncertainty in this Bayesian way then we typically recover robustness: see below. The propriety of this type of construction is of course entirely dependent on the context of the inference.

3.7 Links to Simulated annealing and melding

To improve the behaviour characteristics of Bayes numerical algorithms it is common to "increase the temperature" of a given posterior or perform simulated melding. This can improve mixing whilst retaining the geometrical form of the posterior density and substitutes $f^* \propto f^\alpha$ for some value of α, $0 < \alpha < 1$ and has a simple interpretation within the distances discussed above. Thus we note that for all f, g for which $d_A^L(f, g) < \infty$, for this fixed value of α

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

This is therefore simply a linear contraction on the separation space defined by $d_A^L(f, g)$ and draws all densities closer to one another. On the other hand simulated annealing employ the same transformation but lets $\alpha \to \infty$ *.i.e.*increasingly separates the densities with respect to these separations until a degenerate distribution is attained.

Example 4 (The Power Steady Model) *A class of state space models based on increasing temperature was introduced in [26],[27]. These specified the predictive distributions of* $\{Y_t|y_1, y_2, \ldots y_{t-1}\}_{t\geq 1}$ *in terms of a recurrence of the form*

$$\begin{aligned} p(y_t|\boldsymbol{\theta}_t, y_1, y_2, \ldots y_{t-1}) &= p(y_t|\boldsymbol{\theta}_t),\ t \geq 1 \\ f_t(\boldsymbol{\theta}_t|y_1, y_2, \ldots y_{t-1}) &\propto f_{t-1}^\alpha(\boldsymbol{\theta}_{t-1}|y_1, y_2, \ldots y_{t-1}), t \geq 2 \\ f_1(\boldsymbol{\theta}_1) &\propto f_0^\alpha(\boldsymbol{\theta}_0) \end{aligned}$$

for some $0 < \alpha < 1$. One example of such processes is the Gaussian Steady DLM (Harrison and West) if priors hyperparameters of $f_0(\boldsymbol{\theta}_0)$ 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 $\{y_t\}_{t\geq 1}$is via the equations

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

where the proportionality constants can be calculated using the fact that densities integrate to unity. Suppose our interest is to the joint distribution of $\{Y_t|y_1, y_2, \dots y_{t-1}\}_{t\geq T}$ $T \geq 2$ and that we are confident that our sample distributions $\{p(y_t|\boldsymbol{\theta}_t)\}_{t\geq T}$ are correctly specified. Note that the margin $f_{T-1}(\boldsymbol{\theta}_{T-1}|y_1, y_2, \dots y_{T-1})$ of $\boldsymbol{\theta}_{T-1}|y_1, y_2, \dots y_{T-1}$ is sufficient for forecasting $\{Y_t|y_1, y_2, \dots y_{t-1}\}_{t\geq T}$. Let $g_{T-1}(\boldsymbol{\theta}_{T-1}|y_1, y_2, \dots y_{T-1})$ be the corresponding density using $f_0(\boldsymbol{\theta}_0)$ instead of $f_0(\boldsymbol{\theta}_0)$. Then from the isometry property and the melding property above

$$\begin{aligned} & d_A^L(f_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_T), g_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_T)) \\ = \; & d_A^L(f_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_{T-1}), g_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_{T-1})) \\ = \; & \alpha d_A^L(f_{T-1}(\boldsymbol{\theta}_{T-1}|y_1, y_2, \dots y_{T-1}), g_{T-1}(\boldsymbol{\theta}_{T-1}|y_1, y_2, \dots y_{T-1})) \end{aligned}$$

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.

$$d_A^L(f_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_T), g_T(\boldsymbol{\theta}_T|y_1, y_2, \dots y_T) = \alpha^T d_A^L(f_0(\boldsymbol{\theta}_0), g_0(\boldsymbol{\theta}_0))$$

Notice furthermore that the isoseparation property ensures that this result will still *hold whatever $\{p(y_t|\boldsymbol{\theta}_t)\}_{1\leq t<T}$ 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 misspecifation: see below.*

4 Smoothness and Local Density Ratio Separation

Here we introduce some smoothness conditions that would often be plausible for both our functioning and genuine prior to satisfy. Suppose that for $j = 1, 2$

$$\sup_{\boldsymbol{\theta}, \boldsymbol{\phi} \in B(\boldsymbol{\theta}_0; \rho))} |\log f_j(\boldsymbol{\theta}) - \log f_j(\boldsymbol{\phi})| \leq \varepsilon_j(\boldsymbol{\theta}_0, \rho)$$

Then in particular

$$d^L_{B(\boldsymbol{\theta}_0;\rho)}(f_1, f_2) \leq \varepsilon_1(\boldsymbol{\theta}_0, \rho) + \varepsilon_2(\boldsymbol{\theta}_0, \rho) \tag{21}$$

Let $\mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$ denote the set of densities g such that for all $\boldsymbol{\theta}_0 \in \Theta_0 \subseteq \Theta$

$$\varepsilon_g(\boldsymbol{\theta}_0, \rho) \leq M(\Theta_0)\rho^{0.5p(\Theta_0)} \tag{22}$$

$M(\Theta_0) < \infty$. Let $\mathcal{F}(\Theta_0, M(\Theta_0), 0)$ to be the set of functions for which

$$\sup_{\boldsymbol{\theta}_0 \in \Theta_0} \varepsilon_g(\boldsymbol{\theta}_0, \rho) \to 0 \text{ as } \rho \to 0$$

Then clearly for any set Θ_0 when $0 \leq p_1 < p_2 \leq 2$

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

The parameter p thus governs the roughness of the densities in these sets. For example if $p = 0$ then we require that $\log g$ is equicontinuous with $\log f$ and uniformly continuous over the set Θ_0, where M is the modulus of continuity whilst $\mathcal{F}(\Theta_0, M(\Theta_0), 2)$ is the set of log densities differentiable on Θ_0 with all derivatives bounded in modulus by M. Even with no strong contextual knowledge a Bayesian may well note that his functioning f_0 and believe that his genuine density g_0 both lie in this class for an arbitrary compact set Θ_0. Thus for example $f, g \in \mathcal{F}(M([-m, m]), 2)$ when f is Gaussian and g is believed to be differentiable with bounded derivative on the closed set $[-m, m]$.

It is now trivial to show the following property:

Lemma 5 *If there is a function $h(\boldsymbol{\theta})$ such that $f = f'h$ and $g = g'h$ where $f', g' \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $p(\Theta_0) > 0$ then for all $\theta_0 \in \Theta_0$*

$$\begin{aligned} d^L_{\Theta_0,\rho}(f, g) &\leq 2M(\Theta_0)\rho^{1/2p(\Theta_0)} \\ d^R_{\Theta_0,\rho}(f, g) &\leq \exp\left\{2M(\Theta_0)\rho^{1/2p(\Theta_0)}\right\} - 1 \end{aligned} \tag{23}$$

and if there is a function $h(\boldsymbol{\theta})$ such that $f = f'h$ and $g = g'h$ where $f', g' \in \mathcal{F}(\Theta_0, M(\Theta_0), 0)$ then

$$\begin{aligned} d^L_{\Theta_0,\rho}(f, g) &\to 0 \text{ as } \rho \to 0 \\ d^R_{\Theta_0,\rho}(f, g) &\to 0 \text{ as } \rho \to 0 \end{aligned}$$

Proof. *This follows from simply noting that by hypothesis, for all $\boldsymbol{\theta}_0 \in \Theta_0$, when $p(\Theta_0) > 0$*

$$\begin{aligned} d^L_{B(\boldsymbol{\theta}_0;\rho)}(f, g) &= \sup_{\boldsymbol{\theta},\boldsymbol{\phi} \in B(\boldsymbol{\theta}_0;\rho))} |(\log f(\boldsymbol{\theta}) - \log f(\boldsymbol{\phi})) - (\log g(\boldsymbol{\theta}) - \log g(\boldsymbol{\phi}))| \\ &= \sup_{\boldsymbol{\theta},\boldsymbol{\phi} \in B(\boldsymbol{\theta}_0;\rho))} |(\log f'(\boldsymbol{\theta}) - \log f'(\boldsymbol{\phi})) - (\log g'(\boldsymbol{\theta}) - \log g'(\boldsymbol{\phi}))| \\ &\leq \sup_{\boldsymbol{\theta},\boldsymbol{\phi} \in B(\theta_0;\rho))} |\log f'(\boldsymbol{\theta}) - \log f'(\boldsymbol{\phi})| + \sup_{\theta,\phi \in B(\theta_0;\rho)} |\log g'(\boldsymbol{\theta}) - \log g'(\boldsymbol{\phi})| \\ &\leq \varepsilon_{f'}(\boldsymbol{\theta}_0, \rho) + \varepsilon_{g'}(\boldsymbol{\theta}_0, \rho) \leq 2M(\Theta_0)\rho^{1/2p(\Theta_0)} \end{aligned}$$

The second inequality is a simple consequence of the identity 5..The final limits are obtained simply by substituting the definition of $\mathcal{F}(\Theta_0, M(\Theta_0), 0)$. ■

Thus in particular if we treat the term "sufficiently smooth" to mean that $f, g \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$ then $d^L_{\Theta_0,\rho}(f,g)$ can be made arbitrarily small by choosing the radius $\rho > 0$ sufficiently small. By imposing this weak equicontinuity conditions rather than the one used in [9] that $\{f, g : d^R(f,g) < \eta\}$ where η is chosen arbitrarily small, allows us to use these local separations to prove the expected large sample stability results. Furthermore the inequalities above also allow us the bound the rate at which this occurs: see below

Corollary 6 *If f, g are one dimensional densities such that $\log f$ and $\log g$ are both continuously differentiable and have derivatives bounded by M for all $\boldsymbol{\theta} \in \Theta$, then $d^L_\rho(f,g) \leq 2M\rho$.*

Proof. This simply uses the result above when $p = 2$. ■

Thus suppose Θ is one dimensional $\log f$ is continuously differentiable with derivative bounded by M for all $\boldsymbol{\theta} \in \Theta$, and f is misspecified only in terms of its location and scale parameter so that we know that the genuine prior $g(\theta) = f(\sigma^{-1}(\theta - \mu))$ where μ and $\sigma > 0$ are arbitrary .Then the lemma above tells us that $d^L_\rho(f,g)$ automatically tends to zero at a rate ρ: see the example below

Example 7 *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 that $\theta_0 \geq 0$, then

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

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

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

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. Then we know that the distance $d^L_\rho(f_1,f_2)$ of any genuine Student t prior f_2 with arbitrary prior mode μ_2 tends to zero at a rate of ρ provided the degrees of freedom of the

genuine prior and their spread parameter are bounded i.e. $0 < a^L \leq \alpha_2 \leq a^U < \infty$, $0 < s^L \leq \sigma_2 \leq s^U < \infty$. *Note that, by letting* $|\mu_2 - \mu_1| \to \infty$ *two prior densities, close with respect to these separations, can be arbitrarily far apart in variation distance.*

Thus the condition that $d_\rho^L(f_1, f_2)$ 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 we are centred near $\boldsymbol{\theta}_0$ where the derivative of $\log f - \log g$ might be unbounded - for example when $\boldsymbol{\theta}_0$ lies in the tail of a density f where either f or g has an exponential or faster tail - might f and g be a long distance apart for a small enough value of ρ. If we believe our genuine prior has inverse polynomial tails and choose a functioning prior that also shares this property then f and g will lie in the same neighbourhood for sufficiently small ρ. But even when this is not so, to obtain bounds on $d^L_{\Theta_0,\rho}(f_1, f_2)$ as $\rho \to 0$ we need only require that $\boldsymbol{\theta}_0 \in \Theta_o \subseteq \Theta$ in a suitable compact set Θ_0. to retrieve this property for continuous f_1,f_2: see below.

Example 8 *Consider a simple Bayesian model hierarchical where joint prior densities over parameters are specified through vector equations like*

$$\boldsymbol{\theta} = \boldsymbol{\phi} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\phi}$ *is some function of hyperparameters encoding the systematic mean variation in* $\boldsymbol{\theta}$ *and* $\boldsymbol{\varepsilon}$ *is a vector of error terms with zero mean and independent of* ϕ. *Commonly the prior density* f_ε *of the error term* ε *is chosen from some smooth family - for example a Student t (if an associated variance hyperparameter is given an inverted gamma distribution and integrated out). This will ensure that* $f_\varepsilon \in \mathcal{F}(\Theta, M, p)$ *for some suitable choices of these first two parameters. Here* $\boldsymbol{\varepsilon}$ *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* $\boldsymbol{\theta}$. *Let* $f_\phi(g_\phi)$ *and* $f_\theta(g_\theta)$ *denote the functioning (genuine) joint density of* $\boldsymbol{\phi}$ *and* $\boldsymbol{\theta}$ *respectively. Since*

$$\begin{aligned} f_\theta(\boldsymbol{\theta}) - f_\theta(\boldsymbol{\theta} - \boldsymbol{\delta}) &= \int \left(f_\varepsilon(\boldsymbol{\varepsilon} - \phi) - f_\varepsilon(\boldsymbol{\varepsilon} - \boldsymbol{\phi} - \boldsymbol{\delta})\right) f_\phi(\boldsymbol{\phi}) d\boldsymbol{\phi} \\ g_\theta(\boldsymbol{\theta}) - g_\theta(\boldsymbol{\theta} - \boldsymbol{\delta}) &= \int \left(f_\varepsilon(\boldsymbol{\varepsilon} - \phi) - f_\varepsilon(\boldsymbol{\varepsilon} - \boldsymbol{\phi} - \boldsymbol{\delta})\right) g_\phi(\boldsymbol{\phi}) d\boldsymbol{\phi} \end{aligned}$$

an automatic consequence of constructing the prior in this way is that $f_\theta, g_\theta \in \mathcal{F}(\Theta, M, p)$ 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 we have surreptitiously imposed the condition that our genuine and functioning prior are suitably smooth in the sense given above, which we will see in turn forms the basis for the sort of posterior closeness we require. Notice that this is also true even if we allow the error distribution to come from a different family, provided the genuine* $g_\varepsilon \in \mathcal{F}(\Theta, M, p)$. *Similar albeit*

slightly weaker robustness also applies if $f_\varepsilon \in \mathcal{F}(\Theta_0, M, p)$ for some subset $\Theta_0 \subset \Theta$. Interestingly the more levels of hierarchy we use to specify our model the more smoothing of this type we tend to introduce into the second level marginal densities.

5 Variation distance and local separations

5.1 Chebychev Type Bounds

We next show that the distance between the functioning and genuine prior $d^L_{B(\theta_0;\rho)}(f_0, g_0)$ being small for small ρ is almost a sufficient condition for posterior variation distance between these densities being close for sufficiently large sample size n when the observed likelihood is common to both models. Furthermore we can usually find explicit bounds for this variation distance between two posteriors in terms of features of the functioning posterior f_n - which we have calculated - together with the parameters of our prior bounds.

Notation 9 *Let $\mathcal{N}(f_0, \Delta, M(\Theta_0), p(\Theta_0))$ denote the set of g_0 such that $d^R(g_0, f_0) \leq \Delta$ where $\Delta < \infty$ and there exists a function k such that $f_0 = f_0' k$ and $g_0 = g_0' k$ where $f_0', g_0' \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $0 < p \leq 2$.*

When f_0 is bounded then heuristically the condition $g_0 \in \mathcal{N}(f_0, \Delta, M, p)$ stipulates that g_0 is "comparably smooth" to f_0 and has identical tail behaviour to f_0. Thus for example if f_0 had faster tail behaviour than g_0 it might smooth away significant masses under the likelihood that happens to centre in its tail (and vice versa). This condition provides us with a very coarse but nonetheless very useful upper bound for the variation distance between the corresponding two posterior densities.

There are various ways to bound the term $T_n(2, \rho)$. A coarse bound that does not require any condition on the observed likelihood other than boundedness and positivity is used below.

Theorem 10 *If $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ then for $0 < p \leq 2$*

$$d_V(f_n, g_n) \leq \inf_{\rho > 0} \{T_n(1, \rho) + 2T_n(2, \rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in C^{\alpha_n}(\boldsymbol{\theta}_0, \rho)\} \tag{24}$$

where

$$\begin{aligned} T_n(1, \rho) &= \exp\left\{2M\rho^{p/2}\right\} - 1 \\ T_n(2, \rho) &= (1 + \Delta)\alpha_n(\rho) \end{aligned}$$

Moreover if $f_n(\boldsymbol{\theta})$ converges in distribution to a point mass at distribution $\boldsymbol{\theta}_0$ then for $0 \leq p \leq 2$

$$\lim_{n \to \infty} \sup_{g_0 \in \mathcal{N}(f_0, \Delta, M(\boldsymbol{\theta}_0), p)} d_V(f_n, g_n) = 0$$

Proof. See the Appendix ■

When the smoothness parameter $p > 0$ we illustrate below that this often enables us to calculate explicit bounds for $d_V(f_n, g_n)$ and to calculate natural rates of convergence not depending on *any* assumptions about the form of the likelihood. Furthermore neither the condition $f_0', g_0' \in \mathcal{F}(\Theta_0, M, p)$, which just asks that f_0, g_0 are comparably smooth nor the condition $d^R(g_0, f_0) \leq \Delta$ which regulates the tails of the densities - require anything about *where* the two candidate prior densities concentrate their mass. In particular the prior variation distance which could be arbitrarily close to its maximum value of 1 and the results still hold. This is consistent with the generally held belief that, provided the functioning posterior concentrates, the choice of prior will not be critical. Here are some illustrations.

Example 11 *Suppose F_n is the one dimensional distribution function of a normal $N(\theta_0^n, \sigma_n^2)$. 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 $\rho_n \to 0$. Suppose we believe that $d^R(f_0, g_0) \leq \Delta$ and that $g_0 \in \mathcal{F}(\theta_0^n, M(\Theta_0), p)$ for some prespecified values of $(\Delta, M(\Theta_0), p)$. It follows that, provided $B(\theta_0^n, \rho) \subset \Theta_0$*

$$T_n(1, \rho) \leq \exp\left\{2M(\Theta_0)\sigma_n^{p(1-r)/2}\right\} - 1$$

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

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

we have that

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

Choosing $0 < r < 1$ appropriately we can now obtain an upper bound for the variation distance. Note that with the differentiability condition that $p = 2$ confirms that for any $0 < r < 1$.

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

Typically $\sigma_n \leq \sigma n^{-1/2}$ for some σ when this becomes

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

for any $r' = 1/2\{1 - r\} < 1/2$. Thus we retrieve the expected $n^{-1/2}$ speed of convergence in variation distance between the two posteriors, contrasting with $\sqrt{n}$ speed of divergence using the [9] analogue 2. Note here that it is the difference in mutual roughness of the approximation of f and g that governs the rate above.

Example 12 *Suppose F_n is any one dimensional functioning distribution function with mean θ_0^n and variance σ_n^2 and we believe that $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), 2)$, where $\Theta_0 \supset B(\theta_0; \rho_n)$. Then by Chebychev's inequality*

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

using the definition above and setting $r = 1/3$ when $p = 2$ give 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), $g_0 \in \mathcal{N}(f_0, \Delta, M, 2)$ lying in this neighbourhood typically is bounded by a rate $\sqrt[3]{n}$. Thus 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 $Var(\theta|x_n) = \sigma_n^2$. To obtain robustness intervals we can simply plug in these moments. Note that for a given data set this posterior variance σ_n^2 could increase, as in the example above. However this would be unexpected and furthermore, provided we can calculate (or reliably estimate) σ_n^2, as in the example above, we know when this is happening.

Example 13 *Even when the moments of f_n do not exist we can still obtain bounds, it is just that the rate of convergence associated with these bounds can be slower. Thus suppose $f_n(x) = f\left(\sigma_n^{-1}(\theta - \theta_0^n)\right)$ where f is a Cauchy density and note that $x > 0$ the Cauchy distribution function $F(x)$. has the property that $F(-x) < \frac{1}{2\pi}x^{-1}$ It then follows that if $g_0 \in \mathcal{N}(f_0, \Delta, M, 2)$ then letting $\tau_n = \sigma_n^{-r}$*

$$T_n(2, \rho) \leq \frac{\Delta}{\pi}\sigma_n^r$$

The best asymptotic bound for $d_V(f_n, g_n)$ of this type we can obtain using these inequalities is by setting $r = 0.5$ when

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

Example 14 *Now suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ and that the $\mu_{j,n}, \sigma_{jj,n}^2$ denote, respectively the mean and variance of θ_j, $1 \leq j \leq k$ under the functioning posterior density f_n. Then [29], p153) proves that, writing $\boldsymbol{\theta}_0^n = (\mu_{1,n}, \mu_{2,n}, \ldots \mu_{k,n})$*

$$F_n\left(\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0^n; \rho_n)\right) \geq F_n\left[\bigcap_{j=1}^{k}\left\{\left|\theta_j - \mu_{j,n}\right| \leq \sqrt{k}\rho_n\right\}\right] \geq 1 - k\rho_n^{-2}\sum_{j=1}^{k}\sigma_{jj,n}^2$$

so that

$$F_n\left(\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0^n; \rho_n)\right) \leq k\rho_n^{-2} \sum_{j=1}^{k} \sigma_{jj,n}^2$$

so that

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

where $\sigma_n^2 = k \max_{1 \leq j \leq k} \sigma_{j,n}^2$. *Under this notation we therefore have exactly analogous bounds to the univariate Chebychev bounds given above. In fact this bound is coarse and can be improved (see [20]). Note here that this bound increases linearly with the dimension k of the parameter space. Note also here that if we are interested in the margin of a subvector* $\boldsymbol{\theta}_1$ *of the parameter* $\boldsymbol{\theta}$ *this bound can be made tighter.*

Example 15 *In the particular case when the functioning posterior density* f_n *is Gaussian with mean vector* $\boldsymbol{\theta}_0^n$ *and covariance matrix* $\Sigma_n = \{\sigma_{ijn}^2\}$ *has an l structure - i.e. for values* $\lambda_{1,n}, \lambda_{2,n}, \ldots \lambda_{k,n}$ *where* $\lambda_{j,n} \in (-1, 1)$, $\sigma_{ijn}^2 = \lambda_{1,n}, \lambda_{2,n} \sigma_{iin} \sigma_{jjn}$ $1 \leq i \neq j \leq k$ *then by the inequality in the univariate example above and Theorem 2.2.4 in [29] (p20)*

$$T_n(2, \rho) \geq \sqrt{\frac{2k}{\pi}} \Delta \prod_{j=1}^{k} \sigma_{jjn}^r \exp -\sigma_{jjn}^{-2r}/2 \geq \sqrt{\frac{2k}{\pi}} \Delta \sigma_{0n}^{kr} \exp -\overline{\sigma}_n^{-2r}/2$$

where $\sigma_{0n} = \min_{1 \leq j \leq k} \sigma_{jjn}$ *and* $\overline{\sigma}_n^{-2r} = \sum_{j=1}^{k} \sigma_{jjn}^{-2r}$. *So in this case the* $\sqrt{n}$ *bounds of the univariate case still apply. However the l structure is difficult to verify.*

Thus, 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. So in this sense issues of the appropriateness of the model and its robustness to prior misspecifation are separated from one another.

In fact the smoothness conditions on $d_A^R(f_0, g_0)$ are almost necessary for convergence. Thus if local closeness of the type described above does not exist then it is possible to construct a sequence of likelihoods that converge in a very strong way to a true values $\boldsymbol{\theta}_0$ - so in particular we obtain uniformly consistent estimates of θ - but for which the genuine and functioning posterior densities remain at least an ε distance apart whatever the value of sample size n. The formal statement and proof of this property are rather technical and so is relegated to an appendix.

There is a second useful and often tighter bound based on another belief we might hold when we are concerned that f_0 and g_0 might have different tail characteristics Call a genuine prior *c-rejectable* if the ratio of marginal likelihood

$\frac{p_g(\boldsymbol{x})}{p_f(\boldsymbol{x})} < c$. If we believed that the genuine prior would explain the data better than the functioning prior this in turn would mean that we would expect this ratio to be small: and certainly not c- rejectable for a moderately large values of $c \geq 1$. Say density f $\Lambda-$*tail dominates* a density g if

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

When $g(\boldsymbol{\theta})$ is bounded then this condition requires that the tail convergence of g is no faster than f.

Notation 16 *Let $\mathcal{N}'(f_0, c, \Lambda, M(\Theta_0), p(\Theta_0))$ denote the set of g_0 is not c rejectable with respect to f_0 , f_0 $\Lambda-$tail dominates g_0 and there exists a function k such that $f_0 = f_0'k$ and $g_0 = g_0'k$ where $f_0', g_0' \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $0 < p \leq 2$.*

Then we have the following theorem

Theorem 17 *If $g_0 \in \mathcal{N}'(f_0, c, \Lambda, M(\Theta_0), p(\Theta_0))$ then for $0 < p \leq 2$*

$$d_V(f_n, g_n) \leq \inf_{\rho>0} \{T_n(1,\rho) + 2T_n'(2,\rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in C^{\alpha_n}(\boldsymbol{\theta}_0, \rho)\} \tag{25}$$

where

$$\begin{aligned} T_n(1,\rho) &= \exp\left\{2M\rho^{p/2}\right\} - 1 \\ T_n'(2,\rho) &= (1 + c\Lambda)\alpha_n(\rho) \end{aligned}$$

Moreover if $f_n(\boldsymbol{\theta})$ converges in distribution to a point mass at distribution $\boldsymbol{\theta}_0$ then for $0 \leq p \leq 2$

$$\lim_{n\to\infty} \sup_{g_0 \in \mathcal{N}(f_0, \Delta, M(\boldsymbol{\theta}_0), p)} d_V(f_n, g_n) = 0$$

Proof. See the Appendix ■

Note that if the genuine prior were c - rejectable for a large c we would probably want to abandon it. Here the prior tail dominance condition simply encourages us to use a flat tailed functioning prior so that if a tail observation is seen it allows the likelihood to dominate the posterior and this information is not killed off by the functioning prior tail. This formal result just technically confirms Bayesian modelling principles seen by many as good practical management anyway see for example [23]. Under these conditions, analogues of all the examples above can be calculated simply by substituting $c\Lambda$ for Δ throughout. Alternative bounds can also be calculated in terms of the Highest Posterior Density Bounds of f_n which are tighter but of less practical use - see [5] for related results.

Of course if we make assumptions about the form of the likelihood, then we can also improve these bounds: indeed that takes us closer to standard consistency results [11]. Note in particular the close link between the convergence rates above and those derived in for example [10] and [18] under consistency conditions. However in the spirit of using formal Bayes, we prefer to derive results that are not dependent on the form of the likelihood..

5.2 Robustness without Consistent estimators

It is not uncommon to know that a Bayesian analysis will not produce consistency. For example even after sampling a very large number of data points, aliasing problems can exist in a variety of hierarchical linear models, many latent class models can at best exhibit multiple local minima along the orbits of sometimes quite unexpected invariant group actions [28] and censoring in general often produces likelihoods with multiple local maxima. However in these cases it is often possible to exhibit stability of inference by slightly adapting the results above.

Theorem 18 *If* $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ *then for* $0 < p \leq 2$

$$d_V(f_n, g_n) \leq \inf_{\rho>0} \left\{ mT_n(1,\rho) + 2T_n(2,\rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in \cup_{j=1}^{m} C^{\alpha_n[j]}(\boldsymbol{\theta}_0[j], \rho) \right\} \tag{26}$$

where $T_n(1,\rho)$ *and* $T_n(2,\rho)$ *are defined in the last section and* $\alpha_n(\rho) = \sum_{j=1}^{m} \alpha[j]$ *and* $B(\boldsymbol{\theta}_0[j], \rho)$ *are all disjoint.*

Proof. See the Appendix ■

Thus for example when the posterior converges in distribution to a discrete a non- defective distribution on a support with m atoms, provided g_0 is in an appropriate neighbourhood of f_0 we retain robustness.

Example 19 *Suppose a random sample is drawn from a mixture of normal densities with respective means* μ_1, μ_2*, unit variances and associated mixing probability* $1/2$*. Suppose that the distribution on* (μ_1, μ_2) *is exchangeable. Then since for all* n

$$F_n(\mu_1 \in A_1, \mu_2 \in A_2) = F_n(\mu_1 \in A_2, \mu_2 \in A_1)$$

unless $\mu_1 = \mu_2$ $\{f_n\}_{n\geq 1}$ *will not converge to a point mass as* $n \to \infty$*. However it is easy to check that the posterior density will converge on the space spanned by* $\left(\mu_{(1)}, \mu_{(2)}\right)$ *where* $\mu_{(1)} = \min(\mu_1, \mu_2)$ *and* $\mu_{(2)} = \max(\mu_1, \mu_2)$*.It follows that* F_n *will concentrate its mass around two points so that the theorem above applies. Obviously in this example a simple reparametrisation will suffice but in more complex models like discrete latent class models identifying which points are equally likely is non-trivial. What the theorem above tells us is that though this is important for determining the actual inference it is less important an issue with regard to the robustness of that inference.*

6 Examples of De Robertis Separation

6.1 Exponential families and density balls

These local separations make sense in a parametric setting. Thus suppose $f_1(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\alpha}_1)$ and $f_2(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\alpha}_2)$ lie in the same regular exponential family.

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

for some measurable functions $\pi_1, \pi_2, .., \pi_k, t_1, t_{2,...,}t_k$ for some integer k where $\boldsymbol{\pi}(\boldsymbol{\alpha}) = (\pi_1, \pi_2, \ldots, \pi_k), \boldsymbol{t} = (t_1, t_{2,...,}t_k) \in \mathbb{T}$. For $1 \leq i \leq k$, and $j = 1, 2$ write

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

where, since the exponential family is regular, $\mathbb{T}$ does not depend on α. Note that if a set A is of the form $A = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{t}(\boldsymbol{\theta}) \in \mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2 \times \ldots \mathbb{A}_k\}$ and $\mu(\mathbb{A}_i)$ denotes the length of the interval $\mathbb{A}_i$ then,

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

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 $d^L(f_1, f_2) = \infty$ so the usual density ratio diverges. In particular two densities within the family with parameters arbitrarily close under Euclidean distance are usually infinitely separated under this separation. But under its local form we note that two models with parameters close in Euclidean distance have close local separation as well. For example suppose $\boldsymbol{t}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Then

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

In the special case when all points $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0; \rho)$ lie in the sample space (so that $\boldsymbol{\theta}_0$ is not near the boundary of Θ and the components of θ functionally independent within this ball).so does not depend on $\boldsymbol{\theta}_0$ the inequality above becomes an identity. So these separation give us what we might expect for a standard family, being a weighted distance between the components of the natural parameters of the two prior densities.

Example 20 (Exponential) *When f_j is an exponential $E(\lambda_j)$ density with rate is λ_j, $j = 1, 2$ then $\rho \leq \theta_0$*

$$d^L_{B(\theta_0,\rho)}(f_1, f_2) = 2\rho\,|\lambda_1 - \lambda_2|$$

whilst when $0 \leq \theta_0 < \rho$

$$\rho\,|\lambda_1 - \lambda_2| \leq d^L_{B(\theta_0,\rho)}(f_1, f_2) < 2\rho\,|\lambda_1 - \lambda_2|$$

Note that if prior $f_2 \in \mathcal{G}(f_1, \Delta, \Delta, 2)$ *for any fixed* λ_1 *provided that the parameter of our genuine prior* $|\lambda_1 - \lambda_2| \leq \Delta$.

Example 21 (Gaussian) *Suppose* f_j *is a normal* $N(\mu_j, \sigma_j^2)$ *density* $j = 1, 2$ *so that* $k = 2$. *Then for* $j = 1, 2$

$$\pi_{1,j} = \frac{\mu_j}{\sigma_j^2}, \pi_{2,j} = -\frac{1}{2\sigma_j^2}$$

We can conclude that under a conjugate analysis we can never learn anything about the difference between the reciprocal of the square of the coefficient of variation. and the difference in precision in these two distributions. This fact is of course easily checked from the usual recursions: see e.g.[3]. Note here that $\boldsymbol{t}(\boldsymbol{\theta}) = (t_1(\theta), t_2(\theta)) = (\theta, \theta^2)$. *If we define* $A = \{\theta : B(\theta_0, \rho)\}$ *then*

$$\begin{aligned} d^L_A(f_1, f_2) &= \sup_{\boldsymbol{t}(\theta),\boldsymbol{t}(\phi)\in\mathbb{A}} \left\{(\pi_{1,1} - \pi_{1,2})(\theta - \phi) + (\pi_{2,1} - \pi_{2,2})(\theta^2 - \phi^2))\right\} \\ &= \sup_{\boldsymbol{t}(\theta),\boldsymbol{t}(\phi)\in\mathbb{A}} \left\{(\theta - \phi)\left[(\pi_{1,1} - \pi_{1,2}) + (\pi_{2,1} - \pi_{2,2})(\theta + \phi))\right]\right\} \\ &\leq 2\rho \sup_{\boldsymbol{t}(\theta),\boldsymbol{t}(\phi)\in\mathbb{A}} |(\pi_{1,1} - \pi_{1,2}) + 2(\pi_{2,1} - \pi_{2,2})(\theta_0 + \rho))| \end{aligned}$$

which is bounded only if $\pi_{2,1} = \pi_{2,2}$ *(i.e. the variances of the two distributions are the same) or* θ_0 *lies in some bounded interval. Assuming the latter is so we can obtain the usual bounds provided that we are conservative in the sense that we ensure our choice of prior precision in our functioning prior is no larger than the prior precision of our genuine prior (i.e. use a prior as "vague" as possible). Note below that we obtain even better stability by choosing a functioning prior that has inverse polynomial tails. Finally note in this example we could have defined neighbourhoods using use a 2 dim. ball on* (θ, θ^2) *look at the image of this on* θ*- space. In this case we note that, for* θ_1 *and* θ_2 *to be close we not only require not only* $|\theta_1 - \theta_2| < 2\rho$ *but also*

$$|\theta_1 - \theta_2| < 2\rho \max\left\{1, \frac{1}{|\theta_1 + \theta_2|}\right\} < 2\rho \max\left\{1, \frac{1}{||\theta_0| - 2\rho|}\right\}$$

i.e. that the further the location of the densities from zero, the closer we require them to be.

6.2 Conjugate densities to Exponential families

Now suppose $f_1(\theta) = f(\theta|n_1, \boldsymbol{y}_1)$ and $f_2(\theta) = f(\theta|n_2, \boldsymbol{y}_2)$ lie in the same conjugate of the exponential family [3] p277.

$$f(\theta|\alpha) = c(n, \boldsymbol{y})h^n(\theta)\exp\left\{n\sum_{i=1}^{k} y_i t_i(\theta)\right\}$$

for some integer k where $\boldsymbol{y}_j = (y_{1,j}, y_{2,j}, \ldots, y_{k,j})$ for $j = 1, 2$ where, $\boldsymbol{t} = (t_1, t_{2,\ldots,} t_k) \in \mathbb{T}$ where $\mathbb{T}$ does not depend on $\mathbf{y}$.

Then provided all points $\theta \in B(\theta_0; \rho)$ lie in Θ_0 (so that θ_0 is not near the boundary)

$$\begin{aligned} d^L_{B(\theta_0;\rho)}(f_1, f_2) &= \sup \left\{ \sum_{i=1}^{k} (n_1 y_{i,1} - n_2 y_{i,2}) \left(t_i(\theta) - t_i(\phi)\right) + (n_1 - n_2)(\log h(\theta) - \log h(\phi)) : \theta, \phi \in B(\theta_0; \rho) \right\} \\ &\leq 2\rho \sum_{i=1}^{k} |n_1 y_{i,1} - n_2 y_{i,2}| + |n_1 - n_2| \left|\log h(\theta_0 + \rho) - \log h(\theta_0 - \rho)\right| \end{aligned}$$

so that

$$|C_1(n_1, n_2, \boldsymbol{y}_1 \boldsymbol{y}_2, \rho) - C_2(n_1, n_2, \rho)| \leq (2\rho)^{-1} d^L_{B(\theta_0;\rho)}(f_1, f_2) \leq |C_1(n_1, n_2, \boldsymbol{y}_1 \boldsymbol{y}_2, \rho) + C_2(n_1, n_2, \rho)|$$

where

$$\begin{aligned} C_1(n_1, n_2, \boldsymbol{y}_1 \boldsymbol{y}_2, \rho) &\triangleq \sum_{i=1}^{k} |n_1 y_{i,1} - n_2 y_{i,2}| \\ C_2(n_1, n_2, \rho) &\triangleq (2\rho)^{-1} |n_1 - n_2| \left|\log h(\boldsymbol{\theta}_0 + \rho) - \log h(\boldsymbol{\theta}_0 - \rho)\right| \\ &\simeq |n_1 - n_2| \left|\nabla h(\boldsymbol{\theta}_0)\right| = |n_1 - n_2| \left|\mathbb{E}_{\boldsymbol{\theta}_0} \mathbf{Y}^*\right| \end{aligned}$$

by properties of the exponential family (see e.g. [3] p. 203) where $\mathbf{Y}^*$ is the random vector of data whose exponential family density is the sample density to which this prior is conjugate. In general the non-linear function $h(\boldsymbol{\theta})$ can cause problems near the boundary of the parameter space.

Example 22 *Suppose $f_1(\theta|\alpha_1, \beta_1)$ and $f_2(\theta|\alpha_2, \beta_2)$ have respectively a Beta(α_1, β_1) Beta(α_2, β_2) density where*

$$f(\theta|\alpha, \beta) \propto \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}$$

Then

$$d^L_{B(\theta_0;\rho)}(f_1, f_2) = \sup \left\{ (\alpha_1 - \alpha_2) \log \left\{ \frac{\theta}{\phi} \right\} + (\beta_1 - \beta_2) \log \left\{ \frac{1 - \theta}{1 - \phi} \right\} : \theta, \phi \in B(\theta_0; \rho) \right\}$$

So if

$$\rho < \min \{\theta_0, 1 - \theta_0\}$$

then

$$d^L_{B(\theta_0;\rho)}(f_1, f_2) \leq |\alpha_1 - \alpha_2| \log \left\{ \frac{\theta_0 + \rho}{\theta_0 - \rho} \right\} + |\beta_1 - \beta_2| \log \left\{ \frac{1 - \theta_0 + \rho}{1 - \theta_0 - \rho} \right\}$$

with equality when $(\alpha_1 - \alpha_2)(\beta_1 - \beta_2) \leq 0$. Note that when

$$2\rho \ll \min \{\theta_0, 1 - \theta_0\}$$

then this upper bound is approximately

$$2\rho\left\{\frac{|\alpha_1-\alpha_2|+(|\beta_1-\beta_2|-|\alpha_1-\alpha_2|)\,\theta_0}{\theta_0\,(1-\theta_0)}\right\}$$

which is uniformly bounded for any given closed interval $\Theta_0 \subset [0,1]$. *On the other hand, near the boundary of the parameter space, distances become large, even for relatively similar parameter value. For example if* $\theta_0 = \rho$ *so that we examine probabilities* θ *close to zero, then, for all* $\rho > 0$

$$d^L_{B(\rho;\rho)}(f_1,f_2) = \infty$$

if $\alpha_1 \neq \alpha_2$. *This means - see a result in the appendix - that posteriors may well not converge in variation. To demonstrate this let* $r_n(\theta)$ *denote the digits* $\{0,1$ *or* $2\}$ *of the* r_n^{th} *term in the tertiary expansion of* θ. *Let* $\boldsymbol{r}_n = (r_1, r_2, \ldots r_n)$, $\boldsymbol{x}_n = (x_1, x_2, \ldots x_n)$, $n \geq 2$ *and let* $\boldsymbol{x}_n(\mathbf{r}_n)$ *denote the observation of the value of* $\boldsymbol{r}_{n-1}$ *together with a noisy observation of the value of* r_n *so that*

$$p(\boldsymbol{x}_n|\theta) = \begin{cases} (r_1, r_2, \ldots r_n) \text{ with probability } 1-\lambda_n \\ (r_1, r_2, \ldots [r_n+2] \bmod 3) \text{ with probability } \lambda_n \end{cases}$$

for some set of probabilities $\{\lambda_n : 0 < \lambda_n < 1, n \geq 1\}$. *Let* $A(\boldsymbol{x}_n) = \{\theta : p(\boldsymbol{x}_n|\theta) > 0\}$. *Clearly* θ *is uniformly estimable. For example, irrespective of the value of* $\{\lambda_n : n \geq 1\}$ *and possible values of* $\boldsymbol{x}_n$ *if*

$$\widetilde{\theta}(\boldsymbol{X}_n) = \sum_{i=1}^{n} \left(\frac{1}{3}\right)^{ix_i}$$

$n \geq 2$, *then*

$$\left|\widetilde{\theta}(\boldsymbol{X}_n) - \theta\right| \leq \left(\frac{1}{3}\right)^n$$

Now suppose that f_1, f_2 *are respectively prior beta densities* $Be(\alpha_1, \beta)$, $Be(\alpha_2, \beta)$ *where we choose* $\alpha_1(\varepsilon) > \alpha_2(\varepsilon)$ *so that* $d_V(f_1,f_2) < \varepsilon$ *for some suitable small value of* $\varepsilon > 0$. *By noting the continuity of the gamma function on the positive half plane, we can always do this using the inequality [8] that the Hellinger distance* $d^2_H(f_1,f_2) \leq d_V(f_1,f_2)$ *by choosing* $\alpha_1 - \alpha_2$ *sufficiently small since it is straightforward to verify that*

$$d^2_H(f_1,f_2) = 1 - \sqrt{\frac{\Gamma(\alpha_1+\beta)\Gamma(\alpha_2+\beta)\Gamma^2(\overline{\alpha})}{\Gamma^2(\overline{\alpha}+\beta)\Gamma(\alpha_1)\Gamma(\alpha_2)}}$$

where $\overline{\alpha} = 1/2(\alpha_1+\alpha_2)$. *Suppose for all* n *we now observe* $\boldsymbol{x}_n = (0,0,\ldots,0)$. *Then by construction, since*

$$d^R_{A(\boldsymbol{x}_n)}(f_1,f_2) = \sup_{\theta,\phi \in A(\boldsymbol{x}_n)} \left(\frac{\theta}{\phi}\right)^{\alpha_1-\alpha_2} - 1$$

this is unbounded, as $n \to \infty$. Using analogous arguments to those in the result in the appendix the distance $d_V(f_{1,n}, f_{2,n})$ the posterior densities is therefore bounded away from zero as $n \to \infty$. Thus, even two common densities on bounded support and close in prior variation distance but not exhibiting at the limit point the local smoothness we demand, may not converge as a sequence of increasingly informative experiments are performed.

Of course in this example despite diverging in total variation both $f_{1,n}$ and $f_{2,n}$ converge in distribution to one that assigns probability one to the event $\{\theta = 0\}$. So such failure of robustness might not be critical. However it might be. For example if the distribution to first failure time $\xi(\theta)$ was of primary interest and the probability of failure in any interval was θ then the posterior densities of $\xi(\theta)$ under priors f_1 and f_2 will look very different. Furthermore the prior probability under either model of observing data consistent with such a likelihood is zero. But in complicated models we do not expect a specified model to be absolutely accurate. For example BN's with hidden variables often have an observed likelihood that assigns a maximum likelihood of a conditional probability of the latent variable to lie on a boundary see e.g. [28] . We can therefore expect high dependence on the parameters of the product Beta priors when the analysis does not involve the use of a boundary searching diagnostic applied to the functioning prior f_1. These are rarely used in these contexts.

6.3 Factorisations of densities and graphical models

Unstructured high dimensional models are hard to analyse in general and it is now common practice to work within families of distributions which exhibit factorisations of their joint density. One popular example is the Bayesian Network (BN) of which the usual family of hierarchical models is a special case. A model is often elicited first through qualitative features that can be encoded in terms of such a factorization and specific functional forms of various conditionals added only subsequent to this framework being established.

Suppose our functioning prior $f(\boldsymbol{\theta})$ and genuine prior $g(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k) \in \Theta = \Theta_1 \times \Theta_2 \times \ldots \times \Theta_k$ are both constrained to respect the factorisation

$$\begin{aligned} f(\boldsymbol{\theta}) &= f(\theta_1) \prod_{i=2}^{k} f_{i|.}(\theta_i | \boldsymbol{\theta}_{pa_i}) \\ g(\boldsymbol{\theta}) &= g(\theta_1) \prod_{i=2}^{k} g_{i|.}(\theta_i | \boldsymbol{\theta}_{pa_i}) \end{aligned}$$

where for $2 \leq i \leq m$, $\boldsymbol{\theta}_{pa_i}$ is a subvector of $(\theta_1, \theta_2, \ldots \theta_{i-1})$ and write $\theta[1] = \theta_1 \in \Theta[1] = \Theta_1$and $\boldsymbol{\theta}[i] = (\theta_i, \boldsymbol{\theta}_{pa_i}) \in \Theta[i]$, $2 \leq i \leq k$.Then letting $A = A[1] \times A[2] \times \ldots \times A[k] \subseteq \Theta$ where $A[i] \subseteq \Theta[i]$, $1 \leq i \leq k$.and

$$\begin{aligned} t_1(\boldsymbol{\theta}[1], \boldsymbol{\phi}[1]) &= \log f_1(\theta_1) - \log g_1(\theta_1) + \log g_{1.}(\phi_1) - \log f_1(\phi_1) \\ t_i(\boldsymbol{\theta}[i], \boldsymbol{\phi}[i]) &= \log f_{i|.}(\theta_i | \boldsymbol{\theta}_{pa_i}) - \log g_{i|.}(\theta_i | \boldsymbol{\theta}_{pa_i}) + \log g_{i|.}(\phi_i | \boldsymbol{\phi}_{pa_i}) - \log f_{i|.}(\phi_i | \boldsymbol{\phi}_{pa_i}) \end{aligned}$$

we note directly from the definition of $d_A^L(f,g)$ that

$$\begin{aligned} d_A^L(f,g) &= \sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in A}\left\{\sum_{i=1}^{k} t_i(\boldsymbol{\theta}[i],\boldsymbol{\phi}[i])\right\} \\ &\leq \sum_{i=1}^{k}\sup_{\boldsymbol{\theta},\boldsymbol{\phi}\in A[i]}\{t_i(\boldsymbol{\theta}[i],\boldsymbol{\phi}[i])\} \end{aligned}$$

where $A[i] = A \cap \Theta[i]$. Letting

$$d_{A[i]}^L(f_{i|},g_{i|}) = \sup_{\boldsymbol{\theta}[i]\in A[i]}\left\{\log f_{i|}(\boldsymbol{\theta}[i]) - \log g_{i|}(\boldsymbol{\theta}[i]) - \log f_{i|}(\boldsymbol{\phi}[i]) + \log g_{i|}(\boldsymbol{\phi}[i])\right\}$$

we then have that

$$d_A^L(f,g) \leq d_{A[1]}^L(f_1,g_1) + \sum_{i=2}^{k} d_{A[i]}^L(f_{i|},g_{i|})$$

Applying the inequality 12 to each subset $\Theta[i]$ $i = 2,3,\ldots,k$ now gives that

$$d_A^L(f,g) \leq \sum_{i=2}^{k} d_{A[i]}^L(f_{[i]},g_{[i]})$$

where $f_{A[i]}, g_{A[i]}$ are respectively the margin of $f$and g on the space $\Theta[i]$. This gives a relatively simple upper bound of the full distance in terms of distances of clique marginal densities which in common models all often only consist of a low number of components. So for graphical models at least bounds on the convergence of the full posterior can be written in terms of characteristics of the low dimensional priors on the clique margins of the two densities. There is also a lower bound for $d_A^L(f,g)$ in terms of its components provided both densities are continuous on their support. Thus let

$$d_{A[i]}^{L*}(f_{i|},g_{i|}) \triangleq \sup_{\boldsymbol{\theta}[i]\in A[i]}\left\{\log f_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}^u\right) - \log g_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}^u\right) - \log f_{i|}\left(\phi_i,\boldsymbol{\phi}_{pa_i}^L\right) + \log g_{i|}(\left(\phi_i,\boldsymbol{\phi}_{pa_i}^L\right)\right\}$$

where

$$\theta_1^u, \phi_1^l \triangleq \arg\sup_{\theta_1,\phi_1\in A_1}\{\log f_1(\theta_1) - \log g_1(\theta_1) + \log g_{1.}(\phi_1) - \log f_1(\phi_1)\}$$

and for $2 \leq i \leq k$,

$$\theta_i^u, \phi_i^l \triangleq \arg\sup_{\boldsymbol{\theta}[i],\phi[i]\in A[i]}\left\{\log f_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}^u\right) - \log g_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}^u\right) - \log\left(\phi_i,\boldsymbol{\phi}_{pa_i}^L\right) + \log g_{i|}(\left(\phi_i,\boldsymbol{\phi}_{pa_i}^L\right)\right\}$$

Then we have that

$$d_{A_1}^L(f_1,g_1) + \sum_{i=2}^{k} d_{A[i]}^{L*}(f_{i|},g_{i|}) \leq d_A^L(f,g)$$

Finally suppose we want to set the prior bounds for the conditional densities in the factorisation above functionally independently of the particular parent configuration $\boldsymbol{\theta}_{pa_i}$.

Definition 23 *Say the neighbourhood* $\mathcal{N}(f)$ *of* $f(\boldsymbol{\theta}) = f(\theta_1)\prod_{i=2}^{k} f_{i|.}(\theta_i|\boldsymbol{\theta}_{pa_i})$ *is* uniformly A uncertain *if* $g \in \mathcal{N}(f)$ *respect the same factorisation as* f *and*

$$\sup_{g\in\mathcal{G}(f)}\sup_{\boldsymbol{\theta}_i,\phi_i\in A[i]}\left\{\log f_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}\right)-\log g_{i|}\left(\theta_i,\boldsymbol{\theta}_{pa_i}\right)-\log\left(\phi_i,\boldsymbol{\theta}_{pa_i}\right)+\log g_{i|}((\phi_i,\boldsymbol{\theta}_{pa_i})\right\}$$

is not a function of $\boldsymbol{\theta}_{pa_i}$ $2 \le i \le n$.

If we believe the genuine prior $g \in \mathcal{G}(f)$ is uniformly A uncertain then we can write

$$d_A^L(f,g) = \sum_{i=1}^{k} d_{A[i]}^{L*}(f_{i|}, g_{i|})$$

The separation between the joint densities f and g is then simply the sum of the separation between its component conditionals $f_{i|}$ and $g_{i|}$ $1 \le i \le k$. So in particular to calculate bounds for the joint density of the genuine posterior form prior smoothness conditions on each of the genuine and functioning conditionals and parameters of the posterior. Notice that these bounds will apply *even* when the likelihood destroys the factorisation of the prior. So the critical property we assume here is the fact that we believe that g respects the same factorisation as g. If we learn the value of $\boldsymbol{\theta}(I) = \{\theta_i : i \in I\}$ where I is some index set then the separation between the densities reduces to

$$d_A^L(f(.|\boldsymbol{\theta}(I)), g(.|\boldsymbol{\theta}(I))) = \sum_{i\notin I} d_{A[i]}^{L*}(f_{i|}, g_{i|})$$

If the factorisation of both $g_0(\boldsymbol{\theta})$ and $f_0(\boldsymbol{\theta})$ have a factorisation which respects a decomposable undirected graph [16] with cliques $C[1], C[2], \ldots C[m]$ and separators $S[2], S[3], \ldots S[m]$. Let $\boldsymbol{\theta}(C[i]) \in \Theta_i$ denote the subvector of components of $\boldsymbol{\theta}$ in the clique $C[i]$ and $\boldsymbol{\theta}(S[i]) \in \widetilde{\Theta}_i$ be a subvector of $\boldsymbol{\theta}(C[i])$ with shared components from earlier listed clique contained only in $C[j(i)]$where $1 \le j[i] < i \le k$. Then we note that since

$$\log g_0(\boldsymbol{\theta}) = \sum_{i=1}^{k} \log g_{0,i}(\boldsymbol{\theta}(C[i])) - \sum_{i=2}^{k} \log \widetilde{g}_{0,i}(\boldsymbol{\theta}(S[i]))$$

where $g_{0,i}(\boldsymbol{\theta}(C[i]))$ is the joint density of $\boldsymbol{\theta}(C[i])$ and $\widetilde{g}_{0,i}(\boldsymbol{\theta}(S[i]))$ is the joint density of $\boldsymbol{\theta}(S[i])$, with an identical equation for $f_0(\boldsymbol{\theta})$

$$\begin{aligned} d_A^L(f_0,g_0) &= \sum_{i=1}^{k} d_A^L(f_{0,i}, g_{0,i}) - \sum_{i=2}^{k} d_A^L(\widetilde{f}_{0,i}, \widetilde{g}_{0,i}) \\ &= \sum_{i=1}^{k} d_{A[i]}^L(f_{0,i}, g_{0,i}) - \sum_{i=2}^{k} d_{\widetilde{A}[i]}^L(\widetilde{f}_{0,i}, \widetilde{g}_{0,i}) \end{aligned}$$

where $A[i] = \Theta_i \cap A$ and $\widetilde{A}[i] = \widetilde{\Theta}_i \cap A$. Since $\boldsymbol{\theta}(S[i])$ is a margin of $\boldsymbol{\theta}(C[i])$

$$d^L_{\widetilde{A}[i]}(\widetilde{f}_{0,i}, \widetilde{g}_{0,i}) \leq d^L_{A[i]}(f_{0,i}, g_{0,i})$$

and $d^L_A(f_0, g_0)$ is simply a function of clique distances and these distances never change if we partially observe them through a strictly postive likelihood. This supports the practical observation that the differences between inferences in graphical models respecting the same underlying conditional independencies are usually not large if prior assumptions are perturbed. However models with different prior conditional independencies, here ones that a priori exhibit different clique structure can lead to very different inferences.

6.4 Discussion

It appears that to employ a proper prior whose mass is poorly positioned will give the same answers as getting the prior right provided we have enough data albeit under on three conditions. The first is that the same likelihood is shared by the two priors. The second is that functioning posterior converges: and not to a defective distribution unless the rate of convergence of the appropriate tail of the two distributions are comparable in some sense. Thirdly we require both priors to be comparably smooth: a property that is often induced systematically by the way priors are currently specified.

However there are three caveats to these comforting points when working in high dimensions. First, it is quite likely that the posterior density will not concentrate on to a ball so that unidentifiability issues prevent routine robustness: the bounds above being dominated by the most uncertain parameter and the multiple prior tolerance bounds being much more uncertain. Second the tail regions represent an increasingly larger proportion of the whole density, so issues associated with tails are more significant. Third if the functioning prior respects a factorisation, the assumption that the genuine prior also respects that factorisation is critical for robustness to be guaranteed.

An interesting question is whether the methods of the last section can be extended into non-parametric setting where robustness issues are known to be more fragile. Much current non-parametric Bayesian inference is performed with hierarchical models prior densities that at their lowest level are discrete with probability one. This makes it unlikely that they can be stable in the sense above for the full parameter space although the smoothing properties on the second level variables will often ensure smoothness and hence robustness on this margin can be recovered.

One non-parametric class that does not appear to suffer this problem is the Gaussian process prior. For example [21] assume a priori that

$$r(\theta) = 1 - \frac{g_0(\theta)}{f_0(\theta)}$$

is a sample path from a Gaussian process with zero mean. Notice that the condition $f_0', g_0' \in \mathcal{F}(\Theta_0, M(\Theta_0), 2)$, where these terms are given in 22 is equivalent

to requiring that $r(\theta)$ is contiuously differentiable with bounded derivative on Θ_0. The typical Gaussian covariance function these authors use appear to give rise to sample paths for which this condition typically holds true (i.e. with probability one). This is also true of the transformation used by [17]. Note that Gaussian processes with rougher paths, for example diffusions have roughness coefficient $p < 1$, and so prior to posterior inferences are less robust in this sense. Further discussion of these issues will be reported elsewhere.

Finally we note that the posterior robustness of variation distances to prior specification is the strongest we could reasonably demand. If we demand less strong forms of stability - see [14] for examples of these - then obviously more robustness can be achieved. However what we have been able to demonstrate above is that conditions that appear fairly mild, ensure this strong form of stability anyway.

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

7 Appendix

We begin proving three simple lemmas.

Lemma 24 *For any set measurable set A*

$$d_V(f,g) \leq 2\int_{\theta\notin A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\,\boldsymbol{\theta} + d_A^R(f,g)$$

Proof. First note that if

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

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

$$\xi_A(\boldsymbol{\theta}|f,g) \leq \sup_{\boldsymbol{\theta}\in A}\left|\frac{g_A(\boldsymbol{\theta})}{f_A(\boldsymbol{\theta})} - 1\right| \leq d_A^R(f_A,g_A) = d_A^R(f,g) \tag{27}$$

by invariance of the De Robertis local separation to conditioning. It follows that

$$\begin{aligned}
\int_{\theta\in A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\boldsymbol{\theta} &= \int_{\theta\notin A} |F(A)f_A(\boldsymbol{\theta}) - G(A)g_A(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \\
&\leq |F(A) - G(A)| \int_{\theta\notin A} g_A(\boldsymbol{\theta}) d\boldsymbol{\theta} + F(A)\int_{\theta\in A} |f_A(\boldsymbol{\theta}) - g_A(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \\
&= |F(A^c) - G(A^c)| + F(A)\int_{\theta\in A} |f_A(\boldsymbol{\theta}) - g_A(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \\
&\leq \int_{\theta\notin A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\,\boldsymbol{\theta} + F(A) d_A^R(f,g)
\end{aligned}$$

The result follows. ■

Let $\mathcal{G}$ denote an arbitrary set containing both f and g .

Lemma 25 *Suppose $f = f'k$ and $g = g'k$, $f\ g \in \mathcal{G}$ are such that $f', g' \subseteq \mathcal{F}(\Theta_0, M(\Theta_0), p)$, let $A(\rho) \subseteq B(\boldsymbol{\theta}_0; \rho) \subseteq \Theta_0$ and let $0 < p^* < p$. Then*

$$\sup_{f,g\in\mathcal{G}} \int_{\theta\in A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \leq \sup_{f,g\in\mathcal{G}} \int_{\boldsymbol{\theta}\notin A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\boldsymbol{\theta} + S(A)$$

where

$$\lim_{\rho\to 0} \rho^{-1/2p^*} S(A) = 0$$

Proof. *From the above*

$$\begin{aligned} \sup_{f,g\in\mathcal{G}} \int_{\theta\in A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\boldsymbol{\theta} - \sup_{f,g\in\mathcal{G}} \int_{\boldsymbol{\theta}\notin A} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|\, d\boldsymbol{\theta} &\leq \sup_{f',g'\in\mathcal{F}(\boldsymbol{\theta}_0, M(\boldsymbol{\theta}_0), p(\boldsymbol{\theta}_0))} d^R_A(f,g) \\ &\leq \left(\exp 2M(\boldsymbol{\theta}_0)\rho^{1/2p(\boldsymbol{\theta}_0)} - 1\right) \end{aligned} \tag{28}$$

since $A \subseteq B(\boldsymbol{\theta}_0; \rho)$ by 27 and 23 .Since for any $\beta < 1$

$$\lim_{y\to 0} \frac{\exp \alpha y - 1}{y^\beta} = 0$$

substituting $y = \rho^{1/2p}, \alpha = 2M(\Theta_0), \beta = p^*/p < 1$ and $0 \leq F(A) \leq 1$ now gives the result. ■

Lemma 26 *For $n \geq 1$,*

$$d_V(f_n, g_n) \leq \inf_{\boldsymbol{\theta}_0\in\Theta, \rho>0} \left\{T_n(1, \rho) + 2T^0_n(2, \rho)\right\}$$

where

$$\begin{aligned} T_n(1, \rho) &= d^R_{B(\boldsymbol{\theta}_0;\rho)}(f_0, g_0) \\ T^0_n(2, \rho) &= \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\boldsymbol{\theta}. \end{aligned}$$

Proof.

$$\begin{aligned} d_V(f_n, g_n) &= \int_\Theta |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\theta + \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \end{aligned}$$

where form the first lemma above

$$\int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\boldsymbol{\theta} \leq d^R_{B(\boldsymbol{\theta}_0;\rho)}(f_n, g_n) + \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})|\, d\boldsymbol{\theta}$$

by28 The result now follows. ■

There are various ways to bound the term $T^0_n(2, \rho)$. A coarse bound that does not require any condition on the observed likelihood is used below.

Theorem 27 *If* $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ *then for* $0 < p \leq 2$

$$d_V(f_n, g_n) \leq \inf_{\rho > 0} \{T_n(1,\rho) + 2T_n(2,\rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in C^{\alpha_n}(\boldsymbol{\theta}_0, \rho)\} \tag{29}$$

where

$$\begin{aligned} T_n(1,\rho) &= \exp\left\{2M\rho^{p/2}\right\} - 1 \\ T_n(2,\rho) &= (1+\Delta)\alpha_n(\rho) \end{aligned}$$

Moreover if $f_n(\boldsymbol{\theta})$ *converges in distribution to a point mass at distribution* $\boldsymbol{\theta}_0$ *then for* $0 \leq p \leq 2$

$$\lim_{n\to\infty} \sup_{g \in \mathcal{G}} d_V(f_n, g_n) = 0$$

Proof. The first part is immediate from the lemmas above and noticing that if $g_0 \in \mathcal{G}(f_0, \Delta, M, p)$ then

$$\begin{aligned} T_n(2,\rho) &\leq \int_{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| \, d\boldsymbol{\theta} \\ &\leq F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} \int_{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)} \left|1 - \frac{g_n(\boldsymbol{\theta})}{f_n(\boldsymbol{\theta})}\right| d\boldsymbol{\theta} \\ &= F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} \sup_{\boldsymbol{\theta} \in \Theta} \left|1 - \frac{g_n(\boldsymbol{\theta})}{f_n(\boldsymbol{\theta})}\right| \\ &\leq \alpha_n \Delta \end{aligned}$$

The result thus follows immediately from the lemmas above and the definition of a concentrate. The second part follows from the definitions above and the definition of convergence in distribution ensures that $\lim_{n\to\infty} \alpha_n = 0$.

A tighter bound can be found provided we assume that g_0 is not c rejectable.

■

Theorem 28 *If* $g_0 \in \mathcal{N}'(f_0, c, \Lambda, M(\Theta_0), p(\Theta_0))$ *then for* $0 < p \leq 2$

$$d_V(f_n, g_n) \leq \inf_{\rho > 0} \{T_n(1,\rho) + 2T'_n(2,\rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in C^{\alpha_n}(\boldsymbol{\theta}_0, \rho)\} \tag{30}$$

where

$$\begin{aligned} T_n(1,\rho) &= \exp\left\{2M\rho^{p/2}\right\} - 1 \\ T'_n(2,\rho) &= (1+c\Lambda)\alpha_n(\rho) \end{aligned}$$

Moreover if $f_n(\boldsymbol{\theta})$ *converges in distribution to a point mass at distribution* $\boldsymbol{\theta}_0$ *then for* $0 \leq p \leq 2$

$$\lim_{n\to\infty} \sup_{g_0 \in \mathcal{G}(f_0, \Delta, M(\boldsymbol{\theta}_0), p)} d_V(f_n, g_n) = 0$$

Proof. The first part is immediate from the lemmas above and noticing that if $g_0 \in \mathcal{G}'(f_0, c, \Lambda, M(\Theta_0), p(\Theta_0))$ then

$$\begin{aligned} T_n^0(2,\rho) &\leq F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} + G_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} \\ &= F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} + \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} \frac{g_n(\boldsymbol{\theta})}{f_n(\boldsymbol{\theta})} f_n(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \\ &= F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} + \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} \frac{p_{f_0}(\boldsymbol{x})}{p_{g_0}(\boldsymbol{x})} \frac{g_0(\boldsymbol{\theta})}{f_0(\boldsymbol{\theta})} f_n(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \\ &\leq F_n\{\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0;\rho)\} + c\Lambda \int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0;\rho)} f_n(\boldsymbol{\theta}) d(\boldsymbol{\theta}) \\ &\leq \alpha_n(1 + c\Lambda) \end{aligned}$$

The result thus follows immediately from the lemmas above and the definition of a concentrate. The second part follows from the definitions above and the definition of convergence in distribution ensures that $\lim_{n\to\infty} \alpha_n = 0$. ■

Theorem 29 *If $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ then for $0 < p \leq 2$*

$$d_V(f_n, g_n) \leq \inf_{\rho>0} \left\{ mT_n(1,\rho) + T_n(2,\rho) : B(\boldsymbol{\theta}_0, \rho) \subset \Theta_0, f_n(\boldsymbol{\theta}) \in \cup_{j=1}^m C^{\alpha_n[j]}(\boldsymbol{\theta}_0[j], \rho) \right\} \tag{31}$$

where $T_n(1,\rho)$ and $T_n(2,\rho)$ are defined above and $\alpha_n(\rho) = \sum_{j=1}^m \alpha[j]$ and $B(\boldsymbol{\theta}_0[j], \rho)$ are all disjoint.

Proof. Let $\overline{B} = \Theta \backslash \cup_{j=1}^m B(\boldsymbol{\theta}_0[j], \rho)$

$$\begin{aligned} d_V(f_n, g_n) &= \sum_{j=1}^m \int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0[j];\rho)} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| \, d\theta + \int_{\boldsymbol{\theta}\in\overline{B}} |f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta})| \, d\boldsymbol{\theta} \\ &\leq \sum_{j=1}^m d^R_{B(\boldsymbol{\theta}_0[j];\rho)}(f_n, g_n) + \alpha_n d^R(f_n, g_n) \end{aligned}$$

by 27. The result follows.

A useful result that demonstrates the necessity of these local smoothness conditions and convergence in total variation is given below. It shows that if a local DS does not converge as the radius ρ decreases then there is at least one sequence of likelihoods, concentrating round a single parameter value $\boldsymbol{\theta}_0$ such that the posterior distributions f_n and g_n do not converge. For simplicity we prove this result for certain types of regular f_0 and g_0 only. ■

Definition 30 *Say two densities $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ are $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$* -regular *if there are points $\boldsymbol{\theta}_u, \boldsymbol{\theta}_l \in \Theta_0$, where Θ_0 is a compact subset of Θ satisfying*

$$\begin{aligned} \log f(\boldsymbol{\theta}_u) - \log g(\boldsymbol{\theta}_u) &= \sup_{\boldsymbol{\theta}\in\Theta} \{\log f(\boldsymbol{\theta}) - \log g(\boldsymbol{\theta})\} \triangleq \lambda^+ \\ \log g(\boldsymbol{\theta}_l) - \log f(\boldsymbol{\theta}_l) &= \sup_{\boldsymbol{\theta}\in\Theta} \{\log g(\boldsymbol{\theta}) - \log f(\boldsymbol{\theta})\} \triangleq \lambda^- \end{aligned}$$

such that, for $n \geq 1$, *both* B_n^+ *and* B_n^- *contain open sets containing* $\boldsymbol{\theta}_u$ *and* $\boldsymbol{\theta}_l$ *respectively where*

$$\begin{aligned} B_n^+ &= \{\boldsymbol{\theta} : \sup_{\boldsymbol{\theta}\in\Theta} \{\log f(\boldsymbol{\theta}) - \log g(\boldsymbol{\theta})\} > \lambda^+(1-(2n)^{-1}) \\ B_n^- &= \{\boldsymbol{\theta} : \sup_{\boldsymbol{\theta}\in\Theta} \{\log g(\boldsymbol{\theta}) - \log f(\boldsymbol{\theta})\} > \lambda^-(1-(2n)^{-1}) \end{aligned} \tag{32}$$

Demanding the $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$ -regularity is a very weak condition. For example when Θ is one dimensional it is satisfied for any (f, g) such that $\log f(\theta) - \log g(\theta)$ is right (or left) continuous.

Definition 31 *Say a sequence* $\{p_n(\boldsymbol{\theta}) : n = 1, 2, 3 \ldots, \boldsymbol{\theta} \in \Theta\}$ *of sample densities is* strongly concentrating *on* $\boldsymbol{\theta}_0$ *if there is a sequence* $\{\rho_n : n \geq 1\}$ *such that* $0 < \rho_{n+1} \leq \rho_n, n \geq 1$ *where* $\rho_n \to 0$ *as* $n \to \infty$, *for which* $p_n(\boldsymbol{\theta}) = 0$, $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0; \rho_n) \cap \Theta$.

Note in particular that, directly from 1 $\{f_n(\boldsymbol{\theta})\}_{n\geq 1}$ *concentrates* on $\{\boldsymbol{\theta}_0^n\}_{n\geq 1}$ where we set $\{\rho_n\}_{n\geq 1}$ as defined above and $\alpha_n = 0$, $n \geq 1$ whenever$\{p_n\}_{n\geq 1}$ strongly concentrates on $\boldsymbol{\theta}_0$.

Definition 32 *Say that* (f, g) *are* DRS singular *at* $\boldsymbol{\theta}_0 \in \overline{\Theta}$ *where* $\overline{\Theta}$ *is the compactification of* Θ *if*

$$\lim_{\rho\to 0} d^L_{B(\boldsymbol{\theta}_0;\rho)}(f, g) > 0$$

Note that if $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, ..., \theta_{k,0})$ then we allow $\theta_{i,0} = \infty$ or $\theta_{i,0} = -\infty$. We then use the convention that

$$B(\boldsymbol{\theta}_0; \rho) = B_0(\boldsymbol{\theta}_0; \rho) \cap B_{-\infty}(\boldsymbol{\theta}_0; \rho) \cap B_\infty(\boldsymbol{\theta}_0; \rho)$$

where

$$\begin{aligned} B_0(\boldsymbol{\theta}_0; \rho) &= \bigcap \{\boldsymbol{\theta} : \theta_{i,0} - \rho < \theta_i < \theta_{i,0} + \rho, \text{ where } i \text{ such that } \theta_{i,0} \neq \infty, -\infty\} \\ B_{-\infty}(\boldsymbol{\theta}_0; \rho) &= \bigcap \{\boldsymbol{\theta} : \theta_i < -\rho^{-1}, \text{ where } i \text{ such that } \theta_{i,0} = -\infty\} \\ B_\infty(\boldsymbol{\theta}_0; \rho) &= \bigcap \{\boldsymbol{\theta} : \theta_i > \rho^{-1}, \text{ where } i \text{ such that } \theta_{i,0} = \infty\} \end{aligned}$$

Under this definition we have shown above that $f, g \in \mathcal{F}(\Theta_0, M(\Theta_0), 0)$ are not DRS singular at any point in $\boldsymbol{\theta}_0 \in \Theta_0$. However it is easily checked that the constructions used in [9] to demonstrate the divergence of the ratio 2 have a singularity of this type near the maximum of the likelihood. Furthermore when $\Theta = \mathbb{R}$ and f and g have different rates of convergence in their tails then both ∞ and $-\infty$ are singular points. For example in this case for every $\rho > 0$ we can find a pair $\theta(\rho), \phi(\rho) \in \Theta$ such that $\theta(\rho), \phi(\rho) \in B_\infty(\infty; \rho)$; i.e. as large as we like so that for any $M > 0$

$$|(\log f(\theta(\rho)) - \log g(\theta(\rho)) - (\log f(\phi(\rho)) - \log g(\phi(\rho))| > M$$

so that $\lim_{\rho\to 0} d^L_{B(\infty;\rho)}(f, g) = \infty$.

Definition 33 *Call sequence of pairs of sets* $\{(A_n^+, A_n^-) : n \geq 1\}$ (F, G) - test sequence *on* $\boldsymbol{\theta}_0 \in \Theta_0$ *and* $\{(\rho_n^+, \rho_n^-, \rho_n) : n \geq 1\}$ *when for* $n \geq 1$

$$\begin{aligned} A_n^+ &= B(\boldsymbol{\theta}_u, \rho_n^+) \cap B_n^+ \cap B(\boldsymbol{\theta}_0, \rho_n) \\ A_n^- &= B(\boldsymbol{\theta}_l, \rho_n^-) \cap B_n^- \cap B(\boldsymbol{\theta}_0, \rho_n) \end{aligned}$$

where $\rho_{n+1}^+ \leq \rho_n^+$, $\rho_{n+1}^- \leq \rho_n^-$ *and*

$$\max\{\rho_{n+1}^+, \rho_{n+1}^-\} \to 0 \text{ as } n \to 0 \tag{33}$$

and where

$$F(A_n^+ \cup A_n^-) = G(A_n^+ \cup A_n^-) \tag{34}$$

Demanding the $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$ -regularity is a very weak condition. For example when Θ is one dimensional it is satisfied for any (f, g) such that $\log f(\theta) - \log g(\theta)$ is right (or left) continuous. Since $f(\boldsymbol{\theta}_u), f(\boldsymbol{\theta}_l), g(\boldsymbol{\theta}_u), g(\boldsymbol{\theta}_l) > 0$, it enables us to assert that for $n \geq 1$

$$F(A_n^+), F(A_n^-), F(A_n^+), F(A_n^-) > 0 \tag{35}$$

Note also as a consequence of this definition, for $n \geq 1$, $A_{n+1}^+ \subseteq A_n^+, A_{n+1}^- \subseteq A_n^-$ and $A_n^+ \cap A_n^- = \emptyset$

For any given absolutely continuous F and G that are $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$ -regular with respect to their densities f, g and a sequence $\{\rho_n : n \geq 1\}$ decreasing to zero it is straightforward to construct a sequence $\{(\rho_n^+, \rho_n^-)\ n \geq 1\}$ so that $\{(A_n^+, A_n^-) : n \geq 1\}$ are a test sequence. Go by induction. First choose two arbitrary $(\rho_1^+(0), \rho_1^-(0))$. All we need do now is reduce one of $(\rho_1^+(0), \rho_1^-(0))$ so that (34) is satisfied, which, since $A_1^+ \cap A_1^- = \emptyset$, can be written as

$$\begin{aligned} & F(B(\boldsymbol{\theta}_u, \rho_1^+) \cap B(\boldsymbol{\theta}_0, \rho_1)) - G(B(\boldsymbol{\theta}_u, \rho_1^+) \cap B(\boldsymbol{\theta}_0, \rho_1)) \\ = \ & G(B(\boldsymbol{\theta}_l, \rho_1^-) \cap B(\boldsymbol{\theta}_0, \rho_1)) - F(B(\boldsymbol{\theta}_l, \rho_1^-) \cap B(\boldsymbol{\theta}_0, \rho_1)) \end{aligned} \tag{36}$$

Since by construction $F(B(\boldsymbol{\theta}_u, \rho)) - G(B(\boldsymbol{\theta}_u, \rho))$ is continuous in ρ and is also by construction clearly increasing in ρ with

$$\lim_{\rho \to 0} \{F(B(\boldsymbol{\theta}_u, \rho) \cap B(\boldsymbol{\theta}_0, \rho_1)) - G(B(\boldsymbol{\theta}_u, \rho) \cap B(\boldsymbol{\theta}_0, \rho_1))\} = 0$$

by the midpoint theorem if we have that

$$\begin{aligned} & F(B(\boldsymbol{\theta}_u, \rho_1^+(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) - G(B(\boldsymbol{\theta}_u, \rho_1^+(0) \cap B(\boldsymbol{\theta}_0, \rho_1))) \\ > \ & G(B(\boldsymbol{\theta}_l, \rho_1^-(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) - F(B(\boldsymbol{\theta}_l, \rho_1^-(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) \end{aligned}$$

we simply reduce $\rho_1^+(0)$ to a value ρ_1^+ where the equality is satisfied and set $\rho_1^- = \rho_1^-(0)$. Similarly if

$$\begin{aligned} & F(B(\boldsymbol{\theta}_u, \rho_1^+(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) - G(B(\boldsymbol{\theta}_u, \rho_1^+(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) \\ < \ & G(B(\boldsymbol{\theta}_l, \rho_1^-(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) - F(B(\boldsymbol{\theta}_l, \rho_1^-(0)) \cap B(\boldsymbol{\theta}_0, \rho_1)) \end{aligned}$$

noting that

$$\lim_{\rho \to 0} \{F(B(\boldsymbol{\theta}_l, \rho) \cap B(\boldsymbol{\theta}_0, \rho_1)) - G(B(\boldsymbol{\theta}_l, \rho)) \cap B(\boldsymbol{\theta}_0, \rho_1)\} = 0$$

setting $\rho_1^+ = \rho_1^+(0)$ we can find $\rho_1^- < \rho_1^-(0)$ so that the equality above is satisfied.

Now assume such that (ρ_m^+, ρ_m^-) satisfy the conditions above for $1 \leq m \leq n$ and choose $\rho_{n+1}^+(0) \leq \min\{\rho_n^+, n^{-1}\}$ and $\rho_{n+1}^-(0) \leq \min\{\rho_n^-, n^{-1}\}$ Now simply to reduce one of $(\rho_{n+1}^+, \rho_{n+1}^-)$ $n \geq 1$ so that 34 is satisfied using exactly the same argument as above but with the subscript n replacing the subscript 1. The inductive step is complete and the test set sequence now constructed.

Note that under this construction if for $n \geq 1$ $A_n \triangleq A_n^+ \cup A_n^-$ by the isoseparation property,

$$d_{A_n}^L(f_n, g_n) = d_{A_n}^L(f, g) = d^L(f_n, g_n) = d^L(f, g) \triangleq 2\lambda$$

where $2\lambda = \lambda^+ + \lambda^-$.

Now suppose we are lucky enough to learn from $\boldsymbol{x}_n$ that there exists a sequence $\{\rho_n : n \geq 1\}$ decreasing to zero as $n \to \infty$ that $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0; \rho_n)$ then we might hope that, at least the variation distance between the posterior densities of two close densities might converge. However this is not true in general for any pair of $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$ -regular priors (f, g) that are DRS singular at $\boldsymbol{\theta}_0 \in \Theta$.

To construct a case when variation convergence does not hold let

$$p_n(\boldsymbol{x}_n|\boldsymbol{\theta}) \propto \chi_{A_n}(\theta) \tag{37}$$

where

$$A_n \triangleq A_n^+ \cup A_n^-$$

and where $\{(A_n^+, A_n^-) : n \geq 1\}$ form a test sequence. Notice that since $A_n \subseteq A_{n+1}$, $n \geq 1$ an observed (uniform) independent sample for which

$$p_n(x_n|\boldsymbol{\theta}) = 0 \ \boldsymbol{\theta} \notin A_n$$

would give such a joint sampling distribution. Then, in this simple case the sequence of posterior densities can be written down explicitly as

$$\begin{aligned} f_n(\boldsymbol{\theta}) &= (F(A_n))^{-1} f(\boldsymbol{\theta}) \chi_{A_n}(\boldsymbol{\theta}) \\ g_n(\boldsymbol{\theta}) &= (G(A_n))^{-1} g(\boldsymbol{\theta}) \chi_{A_n}(\boldsymbol{\theta}) \end{aligned}$$

and it is straightforward to prove the following theorem in section 4

Theorem 34 *Suppose that (f, g) are $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_l)$ -regular and DRS singular at $\boldsymbol{\theta}_0 \in \Theta_0$.where Θ_0 is a compact subset of Θ. Then there exists an $\varepsilon > 0$ and a sequence of sample distributions concentrating on $\boldsymbol{\theta}_0$ such that for all $n > N(\varepsilon)$*

$$d_V(f_n, g_n) > \varepsilon$$

Proof. *Assume that* $\{(A_n^+, A_n^-) : n \geq 1\}$ *is a* (F, G) *- test sequence on* Θ *and* $p_n(\boldsymbol{x}_n|\boldsymbol{\theta})$ *is defined in 37. Notice that if* $d^L(f, g) > 0$ *then*

$$\min\{\lambda^+(f,g), \lambda^-(f,g)\} = \lambda^* > 0$$

Then

$$0.5 d_V(f_n, g_n) = \int_{A_n^+} \left(1 - y_n^+(\boldsymbol{\theta})\right) f_n(\boldsymbol{\theta}) d\,\boldsymbol{\theta} + \int_{A_n^-} \left(1 - y_n^-(\boldsymbol{\theta})\right) g_n(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

where, for $\boldsymbol{\theta} \in A_n^+$

$$\begin{aligned} y_n^+(\boldsymbol{\theta}) &= \exp\left(-\left\{\log f_n(\boldsymbol{\theta}) - \log g_n(\boldsymbol{\theta})\right\}\right) \\ &\leq \exp\left(-\lambda^+(1 - (2n)^{-1}\right) \triangleq y_n^+ \end{aligned}$$

and for $\boldsymbol{\theta} \in A_n^-$

$$\begin{aligned} y_n^-(\boldsymbol{\theta}) &= \exp\left(-\left\{\log f_n(\boldsymbol{\theta}) - \log g_n(\boldsymbol{\theta})\right\}\right) \\ &\leq \exp\left(-\lambda^-(1 - (2n)^{-1}\right) \triangleq y_n^- \end{aligned}$$

So in particular

$$\int_{A_n^+} y_n^+(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \left(1 - y_n^+\right) F_n(A_n^+)$$

and

$$\int_{A_n^+} y_n^-(\boldsymbol{\theta}) g_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq \left(1 - y_n^-\right) G_n(A_n^-)$$

Since by construction

$$F(A_n^+)\left(1 - y_n^+\right) \geq F(A_n^+) - G(A_n^+) = G(A_n^-) - F(A_n^-) \geq F(A_n^-)\left(\left(y_n^-\right)^{-1} - 1\right)$$

so since $F(A_n^-) = 1 - F(A_n^+)$

$$F(A_n^+) \geq \left(1 - y_n^-\right)\left\{1 - y_n^- y_n^+\right\}^{-1}$$

Similarly

$$G(A_n^-) \geq \left(1 - y_n^+\right)\left\{1 - y_n^- y_n^+\right\}^{-1}$$

So, substituting into the above gives

$$d_V(f_n, g_n) > \varepsilon(\lambda^+, \lambda^-)$$

where

$$\varepsilon(\lambda^+, \lambda^-) = \left(1 - y_n^+\right)\left(1 - y_n^-\right)\left\{1 - y_n^- y_n^+\right\}^{-1} > 0$$

■

Thus whenever priors f_0, g_0 exhibit a DRS singularity at some point $\boldsymbol{\theta}_0$ it is possible that $d_V(f_n, g_n)$ will not converge even if, in the very strong sense described above, information becomes more and more informative that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ as $n \to \infty$.

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] Blackwell, D. and Dubins, L.(1962) "Merging of opinions with increasing information" Annals of Mathematical Statistics, 33, 882 -886

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

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

[7] DeRobertis, L. (1978) "The use of partial prior knowledge in Bayesian inference" Ph.D. idssertation, Yale Univ.

[8] Devroye, L. and Gyorfi, L.(1985) "Non-parametric density estimation - the L_1 view" Wiley New York

[9] Gustafson, P. and Wasserman, L. (1995) "Local sensitivity diagnostics for Bayesiain inference" Annals Statist ,23 , 23, 2153 - 2167

[10] Ghosal,S, Lember,J. and van der Vaart, A.W. (2002) "On Bayesian badaption" Proceedings 8th Vilnius Conference Probability and Statistics, B Grigelionis et al eds.

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

[12] Ghosh, J.K. Ghosal, S. aad Samanta, T. (1994) "Stability and Convergence of posteriors in non-regular problems," In Statistical Decision theory and related topics 5 Springer 183 - 199.

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

[14] French, S. and Rios Insua, D.(2000) "Statistical Decision Theory" Kendall's Library of Statistics 9 Arnold

[15] Kadane, J.B. and Ghuang, D.T. (1978) "Stable decision problems" Ann. Statist.6, 1095 -111

[16] Lauritzen, S.L.(1996) "Graphicla Models" Oxford University Press.

[17] Lenk, P.J.(1988) "The logistic normal distribution for Bayesian nonparametric predctie densities" J.Amer. Statist. Assoc. 83(402) 509 - 516.

[18] Kruijer, W. and van der Vaart, A.W. (2005) "Posterior Convergence Rates for Dirichlet Mixtures of Beta Densities" Res.Rep. Dept. Mathematics, Univ. Amsterdam.

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

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

[21] Oakley, J. E. and O'Hagan, A. (2007). Uncertainty in prior elicitations: a nonparametric approach.Biometrika.(to appear).

[22] O'Hagan, A.(1979) On outlier rejection phenomena in Bayesian inference J.R. Statist. Soc. B 41, 358 - 367

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

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

[25] Shafer, G., Gillett,P. R. and Scherl,R. (2003) A new understanding of subjective probability and its generalization to lower and upper prevision International Journal of Approximate Reasoning. 33 1-49.

[26] Smith, J.Q.(1979) "A generalisation of the Bayesian steady forecasting model" J.R.Statist. Soc . B 41, 375 -87

[27] Smith, J.Q.(1981) "The multiparameter steady model" J.R.Statist. Soc . B 43,2, 255-260

[28] Smith,J.Q. and Croft, J. (2003) "Bayesian networks for discrete multivariae data: an algebraic approach to inference" J of Multivariate Analysis 84(2), 387 -402

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

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

[31] Wasserman, L.(1992b) "Recent methodological advances in robust Bayesian inference (with discussion)" In Bayesian Statistics 4 J.M. Bernado et al (eds) 483 - 502 Oxford University Press