Large Incomplete Sample Robustness in Bayesian Networks

Jim Q. Smith
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Alireza Daneshkhah
Department of Management Science
University of Strathclyde
Glasgow, G1 1QE, UK

Abstract

Under local DeRobertis (LDR) separation measures, the posterior distances between two densities is the same as between the prior densities. Like Kullback - Leibler separation they also are additive under factorization. These two properties allow us to prove that the precise specification of the prior will not be critical with respect to the variation distance on the posteriors under the following conditions. The genuine and approximating prior need to be similarly rough, the approximating prior has concentrated on a small ball on the margin of interest, not on the boundary of the probability space, and the approximating prior has similar or fatter tails to the genuine prior. Robustness then follows for all likelihoods, even ones that are misspecified. Furthermore, the variation distances can be bounded explicitly by an easy to calculate function of the prior LDR separation measures and simple summary statistics of the functioning posterior. In this paper we apply these results to study the robustness of prior specification to learning Bayesian Networks.

1 Introduction

Discrete Bayesian networks (BNs) are now widely used as a framework for inference. The usual Bayesian methodology requires the selection of prior distributions on the space of conditional probabilities and various authors have suggested ways to do this (see (Cowell et al, 2000) and references therein). When data sets are complete, the usual analysis is conjugate and it is straightforward to appreciate the effect of prior specification on the subsequent inferences. However it is now more common to be working on problems where data entries are randomly or systematically missing. In this case conjugacy is then lost, models can become unidentifiable and sensitive to outliers. In such circumstances it is much less clear what features of the prior drive the inferential conclusions. Of course good modelers use various forms of sensitivity analyses to examine potential prior influence. However it is hard to do this systematically and to be sure that the posterior densities used really are robust to prior specifications, even when the sample size n is large. Indeed results on local sensitivity in Gustafson and Wasserman (1995) appeared to suggest that the hoped for robustness is a vain one.

A new family of separation measures has now been discovered which encode neighbourhoods of a prior that are on the one hand plausibly large and on the other are sufficient to enable the modeler to determine posterior variation neighbourhoods within which all posterior densities arising from the prior neighbourhood must lie. These posterior total variation neighbourhoods can be bounded explicitly in terms of the parameters of the prior separations and the

sort of summary statistics we would calculate anyway from the joint posterior distribution of the model actually implemented: such as posterior means and covariances. In many situations it is possible to demonstrate that these bounds between the functioning posterior and genuine posterior decrease quickly with sample size, irrespective of the likelihood - even when that likelihood is misspecified.

Under local DeRobertis (LDR) separation measures, the posterior distances between two densities is the same as the prior densities. Analogously to KL separation they also are additive under factorization so are easy to calculate or bound for most high dimensional models.

After reviewing some of the important properties of LDR in the next section we illustrate how these techniques can be used to examine analytically the robustness of inference to various forms of prior misspecification in graphical models (GMs) in Section 3.

2 Local De Robertis Separation

Let g_0 denote our *genuine prior* density and f_0 denote the *functioning prior* we actually use: usually chosen from some standard family- often products of Dirichlets - and let f_n and g_n denote their corresponding posterior densities after observing a sample $\mathbf{x}_n = (x_1, x_2, \dots x_n)$, $n \geq 1$, with *observed* sample densities $\{p_n(\mathbf{x}_n|\boldsymbol{\theta})\}_{n\geq 1}$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. The genuine prior is unknown but we hope that it lies in some appropriate neighbourhood of f_0 so that inferences based on f_0 will be approximately right.

In many situations, because of missingness, these sample densities are typically sums of products of the conditional probabilities defining the GM so both posterior densities f_n and g_n usually have a very complicated analytic form. The functioning posterior density is therefore approximated either by drawing samples or making some algebraic computations.

Let $\Theta(n) = \{\boldsymbol{\theta} \in \Theta : p(\mathbf{x}_n|\boldsymbol{\theta}) > 0\}$, assume that $g_0(\boldsymbol{\theta})$, $f_0(\boldsymbol{\theta})$ are strictly positive and continuous on the interior of their shared support - and so uniquely defined - and assume each observed likelihood, $p_n(\mathbf{x}_n|\boldsymbol{\theta})$, $n \geq 1$ is measurable with respect to $g_0(\boldsymbol{\theta})$ and $f_0(\boldsymbol{\theta})$. From Bayes rule, for all $\boldsymbol{\theta} \in \Theta(n)$ our posterior densities $g_n(\boldsymbol{\theta}) \triangleq g(\boldsymbol{\theta}|\mathbf{x}_n)$, $f_n(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}|\mathbf{x}_n)$ are given by

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

$$\log f_n(\boldsymbol{\theta}) = \log f_0(\boldsymbol{\theta}) + \log p_n(\mathbf{x}_n|\boldsymbol{\theta}) - \log p_f(\mathbf{x}_n)$$

where $p_g(\mathbf{x}_n) = \int_{\Theta(n)} p(\mathbf{x}_n|\boldsymbol{\theta}) g_0(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and $p_f(\mathbf{x}_n) = \int_{\Theta(n)} p(\mathbf{x}_n|\boldsymbol{\theta}) f_0(\boldsymbol{\theta}) d\boldsymbol{\theta}$, whilst whenever $\boldsymbol{\theta} \in \Theta \backslash \Theta(n)$ we simply set $g_n(\boldsymbol{\theta}) = f_n(\boldsymbol{\theta}) = 0$.

For any subset $A \subseteq \Theta(n)$ let

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

Note that this is a transparent way of measuring the discrepancy between two densities on a set A. It is non-negative, symmetric, and clearly only zero when f and g are proportional to each other - i.e. when $f(\boldsymbol{\theta}) \propto g(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in A$ and $f(\boldsymbol{\phi}) \propto g(\boldsymbol{\phi})$, $\boldsymbol{\phi} \in A$. The separations have been studied when $A = \Theta(n)$ (see e.g., DeRobertis (1978); O'Hagan and Forster (2004)) but then the neighbourhoods are far too small for practical purposes. Here we focus on cases where A is chosen to be small. This allows not only the associated neighbourhoods to be realistically large but also leads to the types of strong convergence results we need.

The reason these separation measures are so important is that for *any* sequence $\{p(\mathbf{x}_n|\boldsymbol{\theta})\}_{n\geq 1}$ - however complicated -

$$d_A^L(f_n, g_n) = d_A^L(f_0, g_0) \tag{1}$$

It follows that for all sets $A \subseteq \Theta(n)$ the quality of the approximation of f_n to g_n - as measured by such a separation - is identical to the quality of the approximation of f_0 to g_0. In particular distances between two posterior densities can be calculated effortlessly from two different candidate prior densities. Unlike the functioning posterior density with missingness, the functioning prior and sometimes the genuine prior

lying in standard families and then the LDR separations can then often be expressed explicitly and always explicitly bounded. It can be shown that these separation measures are essentially the only ones with the *isoseparation property* (1) (Smith, 2007).

The fact that there are features in any prior which always endure into the posterior suggests that the priors we choose will "always" have a critical impact on inference and this will indeed be so for small sample size n. However for moderately large n the posterior f_n we calculate often places most of its mass within a set $A_n = B(\mu_n, \rho_n)$ where $B(\mu_n, \rho_n)$ denotes the open ball centred on μ_n of radius ρ_n. Write $d^L_{\Theta_0,\rho}(f,g) \triangleq \sup\{d^L_{B(\mu_n,\rho)}(f,g) : \mu_n \in \Theta_0\}$ and $d^L_{\rho}(f,g) \triangleq \sup\{d^L_{B(\mu_n,\rho)}(f,g) : \mu_n \in \Theta\}$. It has long been known that a necessary condition for robustness is that in some sense the functioning prior is "similarly smooth" to the genuine one. We therefore demand the following mild condition regulating the mutual roughness of the functioning and genuine prior. Assume that $f_0, g_0 \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, where $\mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $M(\Theta_0) < \infty, 0 < p(\Theta_0) \leq 2$ denotes the set of densities f such that for all $\boldsymbol{\theta}_0 \in \Theta_0 \subseteq \Theta$

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

$$M(\Theta_0)\rho^{0.5p(\Theta_0)} \tag{2}$$

Thus for example when $p(\Theta_0) = 2$ we demand that $\log f_0$ and $\log g_0$ both have bounded derivatives within the set Θ_0 of interest. Under these conditions Smith and Rigat (2008) show that

$$d^L_{\Theta_0,\rho}(f,g) \leq 2M(\Theta_0)\rho^{1/2p(\Theta_0)}. \tag{3}$$

It follows that as the mass of the functioning prior converges on a ball of decreasing radius within Θ_0, $d^L_{\Theta_0,\rho}(f,g)$ converges to zero at a rate governed by the roughness parameter $p(\Theta_0)$. In particular 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 $\theta_0 \in \Theta_0$, then $d^L_{\rho}(f,g) \leq 2M\rho$.

Suppose the analysis of a Bayesian network is used to support decisions but the user's utility function is unknown to the modeler. If we can ensure that the *variation distance* $d_V(f_n, g_n) = \int_\Theta \mid f_n(\boldsymbol{\theta}) - g_n(\boldsymbol{\theta}) \mid d\boldsymbol{\theta}$, between f_n and g_n is small then this is sufficient to deduce that the impact of using f_n instead of g_n will not be large. For example if $d_V(f_n, g_n) < \epsilon$ then it is trivial to check that for any utility U in the class $\mathcal{U}$ of all measurable utility functions bounded below by 0 and above by 1, on a decision space $\mathcal{D}$ (Kadane and Chuang, 1978)

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

for $d^*(h) = \arg\max_{d\in\mathcal{D}} \overline{U}(d,h)$ and $d \in \mathcal{D}$ where $\overline{U}(d^*(h), h) = \int_\Theta U(d, \boldsymbol{\theta}) h(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

So provided that $d_V(f_n, g_n) < \varepsilon$ where $\varepsilon > 0$ is small, the consequence - measured by utility - of erroneously using f_n instead of g_n is similarly small. Conversely - unlike for the KL separation - if $d_V(f_n, g_n)$ does not tend to zero as $n \to \infty$, there is at least some utility function for which the decisions based on f_n will remain much worse than those of g_n. This has made posterior discrepancy measured through variation distance a popular choice and so is the one we focus on. In this paper we therefore investigate the conditions under which BN models are robust in this sense.

In fact the condition that the distance between the functioning and genuine prior $d^L_{B(\boldsymbol{\theta}_0;\rho)}(f_0, g_0)$ 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 regardless of the value of the observed likelihood, provided that the functioning posterior concentrates its mass on a small set for large n. Below is one useful result of this type. A useful result of this type is given below.

Definition 1. Call a genuine prior g *c-rejectable* with respect to a functioning f if the ratio of marginal likelihood $\frac{p_f(\mathbf{x})}{p_g(\mathbf{x})} \geq c$.

We should believe the genuine prior will explain the data better than the functioning prior.

This in turn means that we should expect this ratio to be small and certainly not c- rejectable for a moderately large values of $c \geq 1$. Note that if the genuine prior were c - rejectable for a large c we would probably want to abandon it. For example using standard Bayesian selection techniques it would be rejected in favour of f. We need to preclude such densities from our neighbourhood.

Say density f $\Lambda-$ *tail dominates* a density g if

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

When $g(\theta)$ is bounded then this condition requires that the tail convergence of g is no faster than f. Here the prior tail dominance condition simply encourages us not to use a prior density with an overly sharp tail: a recommendation made on other grounds by for example O'Hagan and Forster (2004). The following result now holds.

Theorem 1. *If the genuine prior g_0 is not c rejectable with respect to f_0, f_0 $\Lambda-$tail dominates g_0 and $f_0, g_0 \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, then for $0 < p \leq 2$*

$$d_V(f_n, g_n) \leq T_n(1, \rho_n) + 2T_n(2, \rho_n) \qquad (4)$$

where

$$T_n(1, \rho_n) = \exp d^L_{\mu, \rho_n}(f, g) - 1 \leq \exp\left\{2M\rho_n^{p/2}\right\} - 1$$

and $T_n(2, \rho_n) = (1 + c\Lambda)\alpha_n(\rho_n)$, where $\alpha_n(\rho_n) = F_n(\theta \notin B(\theta_0, \rho_n))$ and $F_n(.)$ stands for the cumulative distribution function of θ .

Proof. See Appendix in Smith (2007). □

It is usually easy to bound $T_n(2, \rho_n)$ explicitly using Chebychev type inequalities (see Smith, 2007 for more details). One useful bound, sufficient for our present context, is given below. It assumes that we can calculate or approximate well the posterior means and variances of the vector of parameters under the functioning prior. These posterior summaries are routinely calculated in most Bayesian analyses.

Example 1. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ and $\mu_{j,n}, \sigma^2_{jj,n}$ denote, respectively, the mean and variance of θ_j, $1 \leq j \leq k$ under the functioning posterior density f_n. Then Tong (1980, p153) proves that, writing $\boldsymbol{\mu}_n = (\mu_{1,n}, \mu_{2,n}, \ldots \mu_{k,n})$

$$\begin{aligned} & F_n\left(\boldsymbol{\theta} \in B(\boldsymbol{\mu}_n; \rho_n)\right) \\ \geq \; & F_n\left[\bigcap_{j=1}^{k} \left\{|\theta_j - \mu_{j,n}| \leq \sqrt{k}\rho_n\right\}\right] \\ \geq \; & 1 - k\rho_n^{-2} \sum_{j=1}^{k} \sigma^2_{jj,n} \end{aligned}$$

so that $F_n\left(\boldsymbol{\theta} \notin B(\boldsymbol{\mu}_n; \rho_n)\right) \leq k\rho_n^{-2} \sum_{j=1}^{k} \sigma^2_{jj,n}$ implying

$$T_n(2, \rho_n) \leq c\Lambda\sigma_n^2\rho_n^{-2},$$

where $\sigma_n^2 = k \max_{1 \leq j \leq k} \sigma^2_{j,n}$. In many cases we can show that $\sigma_n^2 \leq n^{-1}\sigma^2$ for some value σ^2. Note that this gives an explicit upper bound on $T_n(2, \rho_n)$ which tends to zero provided ρ_n is chosen so that $\rho_n^2 \leq n^r\rho$ where $0 < r < 1$.

For a fixed (small) ρ, provided σ_n^2 is sufficiently small $d_V(f_n, g_n)$ will also be small. Indeed when $p = 2$ it will tend to zero at any rate slower than the rate σ_n^2 converges to zero. The other component of our bound $T_n(1, \rho_n)$ can also be calculated or bounded for most standard multivariate distributions. A simple illustration of this bound, where both the functioning prior and genuine prior are drawn from the same family, is given below.

Example 2. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, $\theta_i, \alpha_i > 0, \sum_{i=1}^{k} \theta_i = 1$ - so that Θ is the k simplex. Let the two prior densities $f_0(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_f)$ and $g_0(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_g)$ be *Dirichlet* so that

$$f_0(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_f) \propto \prod_{i=1}^{k} \theta_i^{\alpha_{i,f}-1}, \;\; g_0(\boldsymbol{\theta} \mid \boldsymbol{\alpha}_g) \propto \prod_{i=1}^{k} \theta_i^{\alpha_{i,g}-1}$$

Let $\boldsymbol{\mu}_n = (\mu_{1,n}, \mu_{2,n}, \ldots, \mu_{k,n})$ denote the mean of the functioning posterior density f_n. Then it can be easily checked that if $\rho_n < \mu_n^0 = \min\{\mu_{i,n} : 1 \leq i \leq k\}$, then $d^L_{\boldsymbol{\mu}_n; \rho_n}(f_0, g_0)$ is bounded above by

$$\sum_{i=1}^{k} |\alpha_{i,f} - \alpha_{i,g}| \left\{\log(\mu_{i,n} + \rho_n) - \log(\mu_{i,n} - \rho_n)\right\}$$

$$\leq 2k\rho_n \left(\mu_n^0 - \rho_n\right)^{-1} \overline{\alpha}(f_0, g_0)$$

where $\overline{\alpha}(f_0, g_0) = k^{-1} \sum_{i=1}^{k} |\alpha_{i,f} - \alpha_{i,g}|$ is the average distance between the hyperparameters of the functioning and genuine priors. So $T_n(1, \rho_n)$ is uniformly bounded whenever $\boldsymbol{\mu}_n$ remains in a given fixed closed interval Θ_0 for all n and converges approximately linearly in n. Note that in the cases above, provided we ensure $\rho_n^2 \leq n^r \rho$, $0 < r < 1$ then both $T_n(1, \rho_n)$ and $T_n(2, \rho_n)$ - and hence $d_V(f_n, g_n)$ - tends to zero. However if f_n tends to concentrate its mass on the boundary of Θ near one of the cell probabilities being zero, then even when the average distance $\overline{\alpha}(f, g)$ between the hyperparameters of the priors are small, it can be shown that at least some likelihoods will force the variation distance between the posterior densities to stay large for increasing ρ_n. See Smith (2007) for a proof and an explicit example of this phenomenon. Typically the smaller the probability the slower any convergence in variation distance will be.

Example 3. Sometimes it is convenient, particularly with covariate information, to smoothly transform a vector of probabilities. One commonly used transformation in BNs is the logistic transformation (Spiegelhalter and Laritzen, 1990). Like the variation distance the LDR is invariant to diffeomorphic transformations like this one. When the learning has proceeded on this transformed scale it is often expedient to use this scale directly in the use of Theorem 1. Note that under the logistic transformation we can identify the problem area of inference in the example above - i.e. where the posterior concentrates near a zero in one of the component probabilities, corresponds exactly to the well known sensitivity to tail behaviour when outliers are observed (O'Hagan (1979); Andrade and O'Hagan (2006)). Any family of distributions on the transformed scale having sub-exponential tails - for example multivariate t-distribution has better robustness properties both in term of the LDR and the tail domination condition above than super-exponential tails families - like the Gaussian, and should be preferred in this context (O'Hagan and Forster, 2004).

Of course the usual priors in discrete GMs are typically *products* of many such Dirichlet densities. However our local separation for these products is similarly easily explicitly bounded: see below.

It is interesting to note that lower bounds on variation distances can calculated given that $d^L_{\boldsymbol{\mu}_n;\rho_n}(f_0, g_0)$ stay unbounded above as $n \to \infty$. Thus Smith (2007) show that whenever $d^L_{\boldsymbol{\mu}_n;\rho_n}(f_0, g_0)$ does not converge to zero as $\rho_n \to 0$, in general. Of course our genuine prior g_0 need not be Dirichlet even if the functioning prior is. However, the general conditions above ensure that except when posterior distribution of a single vector of probabilities under the functioning prior tend to zero in some component or unless the prior we should use is much rougher (or smoother) than f_0 with large n we will obtain approximately the right answer in the sense described above.

Note that if two priors are close with respect to LDRs, even when the likelihood is inconsistent with the data, the functioning posterior distribution nevertheless will tend to provide a good approximation of the genuine posterior as the functioning posterior concentrates. All similar priors will give similar (if possibly erroneous) posterior densities.

We now proceed to investigate the properties of $d^L_{\boldsymbol{\mu}_n;\rho_n}(f_0, g_0)$ for graphical models.

3 Isoseparation and BN's

3.1 Some General Results for Multivariate BN's

We begin with some general comments about multivariate robustness.

In Smith and Rigat (2008) it is proved that if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ are 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 $f(\boldsymbol{\theta})$, $g(\boldsymbol{\theta})$ are continuous in Θ and $f_1(\boldsymbol{\theta}_1), g_1(\boldsymbol{\theta}_1)$ represent the marginal densities on Θ_1 of the two joint densities $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ respectively, then

$$d^L_{A_1}(f_1, g_1) \leq d^L_A(f, g) \tag{5}$$

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 $\Theta_2\}$. So in particular marginal densities are never more separated than their joint densities. Thus if we are interested only in particular margins of the probabilities in a BN and we can show that the functioning prior converges on that margin, then even if the model is unidentified provided $f_0, g_0 \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, we will still be able to assert - using an argument exactly analogous to that in the proof of Theorem 1 that with large n the functioning prior will be a good surrogate for the genuine one. This is important since we know that BNs with interior systematically hidden variables are unidentified. However if our utility function is a function only of the manifest variables we can ensure that the variation distance between two posterior marginal densities $f_{1,n}$ $g_{1,n}$ become increasing close - usually at a rate of at least $\sqrt[3]{n}$ - in variation. So in such a case lack of robustness only exists on prior specifications of functions of probabilities of the conditional distributions of the hidden variables conditional on the manifest variables.

Next we note that the usual convention is to use BNs whose probabilities all exhibit prior local and global independence (LGI). Immediately from the definition of $d_A^L(f, g)$ if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_k)$ with functioning prior $f(\boldsymbol{\theta})$ and genuine prior $g(\boldsymbol{\theta})$ both with the property that subvectors $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_k\}$ of parameters are mutually independent so that

$$f(\boldsymbol{\theta}) = \prod_{i=1}^{k} f_i(\boldsymbol{\theta}_i), \quad g(\boldsymbol{\theta}) = \prod_{i=1}^{k} g_i(\boldsymbol{\theta}_i)$$

where $f_i(\boldsymbol{\theta}_i)$ $(g_i(\boldsymbol{\theta}_i))$ are the functioning (genuine) margin on $\boldsymbol{\theta}_i$, $1 \leq i \leq k$, then

$$d_A^L(f, g) = \sum_{i=1}^{k} d_{A_i}^L(f_i, g_i) \tag{6}$$

It follows that - all other things being equal - our local prior distances grow linearly with the number of parameters needed to specify a BN. In particular models encoding more conditional independences are intrinsically more stable and the effects of possibly erroneous prior information will endure longer than more complex models encoding less conditional independences. It has long been known that Bayesian selection methods, for example based on Bayes Factors automatically select simpler models when they provide similar explanation of the observed data than more complex models. But here we have a complementary point. The choice of the complex model will tend to give less reliable posteriors if we are not absolutely sure of our priors.

Example 4. Suppose a discrete BN G on $\{X_1, X_2, \ldots, X_m\}$ where X_i has t levels and parents Pa_i, taking on s_i different parent configurations, $1 \leq i \leq m$. Make the common assumption that our genuine and functioning prior both exhibits LGI: i.e. all $s = \prod_{i=1}^{m} s_i$ parameter vectors $\boldsymbol{\theta}_i | pa_i$ are mutually independent under both f and g. If we believe the LDR separation between the s component densities of the functioning and genuine prior is δ_A then $d_A^L(f, g) = s\delta_A$. Note that the quality of the approximation will depend on the number of parent configurations in the model. Thus if G^1 has all components independent, G^2 is a tree, G^3 is complete and f^j, g^j are the prior densities under G^j, $j = 1, 2, 3$ then

$$d_A^L(f^1, g^1) = m\delta_A, \;\; d_A^L(f^2, g^2) = \{mt - t + 1\}\,\delta_A$$

$$d_A^L(f^3, g^3) = \{t^m - 1\}\,\{t - 1\}^{-1}\,\delta_A$$

The last most general separation bound increases exponentially with m. By (5) this in turn implies that BN's containing a large clique are most unreliable in the sense that data size has to be enormous before we can be confident our inferences are approximately reliable in the sense measured by LDR. Note that in this setting the bound given by our first example on the second component $T_n(2, \rho_n)$ in our theorem is a function of the mean and variances of the component vectors of probabilities (or in some analyses their logistic transform). These are routinely sampled anyway so good estimates can just be plugged in our formula and together with the bounds above this provides explicit operational uncertainty bounds on our variation distances.

Example 5. If the BN is decomposable with cliques $C[j]$, $j = 1, 2, \ldots, m$ then if we require LGI to hold in all Markov equivalent graphs then it is proved that the joint distribution of the clique probabilities on the vector of probability tables over each clique must have a Dirichlet distribution (with consistent distributions over separators). This in turn implies all conditional probabilities used in a BN will also be Dirichlet for both the genuine and functioning prior allowing us to calculate explicit expressions for distances between components. Here we note again that prior distances are expressed through a Euclidean distance on the hyperparameters of the genuine and functioning prior then posterior variation instabilities can occur in the limit only if our posterior density concentrates near zero on some component. Although this phenomenon is unusual for many likelihoods where components are missing at random this is not the case when some components are systematically missing (Smith and Croft, 2003). Indeed when estimating probabilities on phylogenetic trees where only the root and leaf nodes are observed and all probabilities are free it is the norm in practice to find the distribution of at least some of the internal hidden nodes concentrating near zero on some of the probabilities. In these cases, whilst it can be shown that the estimates of the marginal manifest probabilities are usually stable under large samples and the prior may well have a large effect on the inferences about the internal explanatory probabilities, even when the probabilities are identifiable and samples are very large. Unfortunately these probabilities are often the ones of scientific interest!

3.2 Sensitivity to Departures in Parameter Independence

Although LGI is a useful expedient, if a prior is elicited using contextual information - as it should be- systematic biases in the elicitation processes due to poor calibration or selection bias will break these assumptions dramatically. The issue then is to what extent using the assumption of LGI matters. One possible extension away from LGI that naturally occurs under selection biases is for the vector of probabilities in the problem to mirror the dependence structure of the BN G. A special case of this is when we drop the local independence assumption. So suppose a functioning prior $f(\boldsymbol{\theta})$ and a 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 same factorisation

$$f(\boldsymbol{\theta}) = f(\theta_1) \prod_{i=2}^{k} f_{i|.}(\theta_i \mid \boldsymbol{\theta}_{pa_i})$$

$$g(\boldsymbol{\theta}) = g(\theta_1) \prod_{i=2}^{k} g_{i|.}(\theta_i \mid \boldsymbol{\theta}_{pa_i}),$$

where for $2 \leq i \leq k$, the parents $\boldsymbol{\theta}_{pa_i}$ of θ_i is a subvector of $(\theta_1, \theta_2, \ldots \theta_{i-1})$. Write $\boldsymbol{\theta}[1] = \theta_1 \in \Theta[1] = \Theta_1$ and $\boldsymbol{\theta}[i] = (\theta_i, \boldsymbol{\theta}_{pa_i}) \in \Theta[i]$, $2 \leq i \leq k$. Let $A = A[1] \times A[2] \times \ldots \times A[k] \subseteq \Theta$ where $A[i] \subseteq \Theta[i]$, $1 \leq i \leq k$. Then it is straightforward to show 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 $\Theta[i]$ of the i^{th} variable and its parents (Smith, 2007). Note therefore that our local separations increase no faster than linearly in the number of probabilities. It is natural to set these bounds so that they are functionally independently of the particular parent configuration $\boldsymbol{\theta}_{pa_i}$.

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

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

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

If we believe the genuine prior $g \in \mathcal{G}(f)$ is uniformly A uncertain then we can write $d_A^L(f, g) = \sum_{i=1}^{k} d_{A[i]}^{L*}(f_{i|.}, g_{i|.})$ (see Smith, 2007).

The separation between the joint densities f and g is then simply the sum of the separation between its component conditionals $f_{i|.}$ and $g_{i|.}$, $1 \leq i \leq k$. So in particular we can calculate bounds for the joint density of the genuine posterior from 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 a priori that f 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

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

There is therefore a degree of stability to deviations in parameter independence assumptions.

Finally consider the general case where the hyperprior is totally general but the modeler believes that the dependence between parameters has been caused by the expert first assuming all component probabilities as mutually independent and then observing a particular data set $\mathbf{y}$ with sample mass function $q(\mathbf{y}|\boldsymbol{\theta}) > 0$ and forming her new dependent posterior. If we assume that deviation in this process is only caused by the misspecification of the initial independence prior then by the isoseparation property, the LDR discrepancy between genuine and functioning prior should be set at the same deviation parameters as the independence priors. So on this strong assumption we regain the stability existing under LGI.

4 Discussion

For any BNs whose densities factorise, the LDR separations are a valuable way of understanding exactly what forces the final posterior inferences. Robustness under large n will typically exist for sparse graphs with no component probabilities close to zero. On the other hand graphical models with many boundary probabilities and/or a large number of edges will exhibit enduring large approximation errors measured in total variation distance. This gives yet another reason why restricting inference with BN's to graphs with only a small number of edges is a good idea.

We note that the same techniques can be used to study inference in continuous and mixed BN's and also for all other GMs encoding a single factorization. We are currently implementing these techniques and the bounds appear to provide genuinely helpful supplementary diagnostic information to what is often a complex estimation exercise.

References

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

R. G. Cowell, A. P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter. 2000. *Probabilistic Networks and Expert Systems*. Springer Verlag.

L. DeRobertis. 1978. *The use of partial prior knowledge in Bayesian inference*. Ph.D. dissertation, Yale Univ.

P. Gustafson and L. Wasserman. 1995. Local sensitivity diagnostics for Bayesiain inference. *Annals Statist*, 23:2153–2167.

J. B. Kadane and D. T. Chuang. 1978. Stable decision problems. *Ann. Statist*, 6:1095–1111.

A. O'Hagan. 1979. On outlier rejection phenomena in Bayesian inference. *J. R. Statist. Soc. B*, 41:358–367.

A. O'Hagan and J. Forster. 2004. *Bayesian Inference*. Kendall's Advanced Theory of Statistics, Arnold.

D. J. Spiegelhalter and S. L. Lauritzen .1990. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605.

J. Q. Smith. 2007. Local Robustness of Bayesian Parametric Inference and Observed Likelihoods. *CRiSM Res Rep*, 07-08.

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

J. Q. Smith and F. Rigat. 2008. Isoseparation and Robustness in Finitre Parameter Bayesian Inference. *CRiSM Res Rep*, 07-22.

Y. L. Tong. 1980. *Probability Inequalities in Multivariate Distributions*. Academic Press New York.