On the Robustness of Bayesian Networks to Learning from Non-conjugate Sampling

J. Q. Smith[*,a,1], A. Daneshkhah[**,a,b,1,1]

[a] *Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom*
[b] *Department of Management Science, University of Strathclyde, Glasgow UK, G1 1QE*

Abstract

Under local DeRobertis separation measures, the posterior distances between two densities is the same as between the prior densities. Like Kullback - Leibler separation they are also additive under factorisation. 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 priors need to be similarly rough, the approximating prior must have concentrated on a small ball on the margin of interest, not on the boundary of the probability space, and the approximating prior must have similar or fatter tails to the genuine prior. Robustness then follows for all likelihoods, even ones that are misspecified. Furthermore, the total variation distances can be bounded explicitly by an easy to calculate function of the prior local DeRobertis 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.

Key words: Bayesian networks, Bayesian robustness, isoseparation property, local DeRobertis separation measures, total variation distance

[*]Principal corresponding author
[**]Corresponding author
Email addresses: j.q.smith@warwick.ac.uk (J. Q. Smith), alireza.daneshkhah@strath.ac.uk (A. Daneshkhah)

Preprint submitted to International Journal of Approximate Reasoning December 9, 2008

1. Introduction

Discrete Bayesian networks 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 [2] 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 whether certain 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 [6] appeared to suggest that the hoped for robustness is a vain one.

Recently (see [11]) it has been demonstrated that MAP model selection can be highly sensitive to the setting of priors over the probabilities in a discrete Bayesian network *even* if we restrict ourselves only to using conjugate product Dirichlets. However in this paper we are concerned not with model selection but the potential misleading effects of choosing a misspecified prior. That we can have robustness to prior misspecification and instability to model selection can be illustrated though the following example.

Suppose that it is know that either all components of a vector $\mathbf{x}$ of observations will all be strictly positive with a known density $p^+(\mathbf{x})$ or all components are negative with a known density $p^-(\mathbf{x})$. Let model $M(\alpha)$ assigns a probability α where $0 < \alpha \neq \frac{1}{2} < 1$ to all observations being positive whilst model $M(\frac{1}{2})$ assigns a probability $\frac{1}{2}$ to this event. MAP selection after observing the first component x_1 of the vector $\mathbf{x}$ is clearly sensitive to the choice of α. For example if $x_1 > 0$ then we will choose $M(\alpha)$ over M_2 if and only if $\alpha > \frac{1}{2}$. However after observing x_1 - whether x_1 is positive or negative - all predictions about future observations made by $M(\alpha)$ $0 < \alpha \neq \frac{1}{2} < 1$ will be identical to those made using M_1. In the sense used in this paper the inference posterior to x_1 is totally insensitive to the assignment of the value of the hyperparameter α.

A new family of separation measures has now been discovered which en-

code 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 neighbourhoods 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 separation measures, the posterior distances between two densities is the same as the prior densities. Analogously to Kullback - Leibler separation they also are additive under factorisation so are easy to calculate or bound for most high dimensional models.

After reviewing some of the important properties of local DeRobertis separation 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 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 distribution family - for example the products of Dirichlets - and let f_n and g_n denote their corresponding posterior densities after observing a sample $\mathbf{x}_n = (x_1, x_2, \ldots x_n)$, $n \geq 1$, with *observed* sample densities $\{p_n(\mathbf{x}_n|\boldsymbol{\theta})\}_{n\geq 1}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. 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 graphical model 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 $g_n(\boldsymbol{\theta}) \triangleq g(\boldsymbol{\theta}|\mathbf{x}_n)$ and $f_n(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}|\mathbf{x}_n)$ 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

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

$$p_f(\mathbf{x}_n) = \int_{\Theta(n)} p(\mathbf{x}_n|\boldsymbol{\theta}) f_0(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

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

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

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., [3]; [10]) 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 local DeRobertis 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) [12].

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 $\mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, $M(\Theta_0) < \infty, 0 < p(\Theta_0) \leq 2$ denote 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 [14] show that

$$d^L_{\Theta_0,\rho}(f,g) \leq 2M(\Theta_0)\rho^{0.5p(\Theta_0)}. \tag{3}$$

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 and 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 [8] 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 $\mathcal{D}$

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

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 Kullback - Leibler 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 Bayesian network 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. A useful result of this type is given below.

Definition 1. Call a genuine prior g c - *rejectable* with respect to a functioning prior 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.

Definition 2. 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. 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 [10]. 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 $0 < p \leq 2$*

$$d_V(f_n, g_n) \leq T_n(1, \rho) + 2T_n(2, \rho) \tag{4}$$

where

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

and

$$T_n(2, \rho) = (1 + c\Lambda)\alpha_n(\rho_n),$$

where $\alpha_n(\rho_n) = F_n(\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \rho_n))$ and $F_n(.)$ stands for the cumulative distribution function of $\boldsymbol{\theta}$.

Proof. See Appendix in [12]. □

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

$$\lim_{n\to\infty} \sup_{g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)} d_V(f_n, g_n) = 0$$

where $\mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ denote the set of g_0 such that

$$\exp\{d^L_{\Theta_0,\rho}(f_0, g_0)\} \leq 1 + \Delta$$

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 the condition $g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p)$ heuristically 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). The condition provides us with a very coarse but nonetheless very useful upper bound for the variation distance between the corresponding two posterior densities which is presented in Theorem 1 (see also [12]).

It is usually easy to bound $T_n(2,\rho)$ explicitly using Chebychev type inequalities (see [12] 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 $\boldsymbol{\mu}_n = (\mu_{1,n}, \mu_{2,n}, \ldots \mu_{k,n})$

$$F_n\left(\boldsymbol{\theta} \in B(\boldsymbol{\mu}_n; \rho_n)\right) \geq F_n\left[\bigcap_{j=1}^{k}\left\{|\theta_j - \mu_{j,n}| \leq \sqrt{k}\rho_n\right\}\right] \geq 1 - k\rho_n^{-2}\sum_{j=1}^{k}\sigma^2_{jj,n} \tag{5}$$

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

$$\rho_n < \mu_n^0 = \min\left\{\mu_{i,n} : 1 \leq i \leq k\right\},$$

then $d^L_{\boldsymbol{\mu}_n,\rho_n}(f_0, g_0)$ is bounded above by

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

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

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 ρ. 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 $d_V(f_n, g_n)$ - tend to zero. Note that this contrast strongly with the demonstrated instability of MAP model selection (see [11]) to how the scale parameter of the Dirichlet. We can conclude here that provided the scale parameter is set to be in the right ball park - so that $\overline{\alpha}$ is bounded - the number n of observations is large and sample proportions in each cell are not close to zero, then the misspecification of the scale parameter will have very little effect on the inferences about *future* observations. After a large number of observations the sensitivity of the relative efficacy of models with different scale parameters, early in the prediction process as measured by the Bayes factors scores no longer matters if our interest is prediction.

On the other hand 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 [12] for a proof and an explicit example of this phenomenon. Typically the smaller the probability the slower any convergence in variation distance will be.

We now show the impact of the ρ_n and misspecification of the hyperprameters of the functioning prior on the variation bound through an example.

Let us consider the functioning prior follows a *Beta* distribution with the following density function

$$f_0(\theta \mid \boldsymbol{\alpha}_f) \propto \theta^{\alpha_{1f}-1}(1-\theta)^{\alpha_{2f}-1}$$

and the corresponding posterior distribution for a sample drawn from a Binomial distribution with size n is given by

$$f_n(\theta \mid \boldsymbol{\alpha}_f, x) \propto \theta^{(\alpha_{1f}+x)-1}(1-\theta)^{(n+\alpha_{2f}-x)-1}$$

where x is the number of successes observed in the sample.

It is known that the posterior mean and and variance of the Beta distribution, given below, respectively,

$$\theta_0^n = \frac{(\alpha_{1f}+x)}{(\alpha_{1f}+\alpha_{2f})+n}, \quad \sigma_n^2 = \frac{\alpha_{1f}\alpha_{2f}}{(\alpha_{1f}+\alpha_{2f}+n)^2(\alpha_{1f}+\alpha_{2f}+n+1)}$$

exist and are finite, and we believe that the genuine prior

$$g_0 \in \mathcal{N}(f_0, \Delta, M(\Theta_0), p=2).$$

By Chebychev's inequality [12] $T_n(2,\rho)$ satisfies

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

The variation distance bounds associated with $\rho_{n1} = (\sigma_n^2)^{1/3}$ and $\rho_{n2} = (\sigma_n^2)^{1/4}$ are given respectively by

$$d_{V1}(f_n, g_n) \leq \exp\{2M(\Theta_0)(\sigma_n^2)^{1/3}\} - 1 + \Delta(\sigma_n^2)^{1/3} \tag{6}$$

and

$$d_{V2}(f_n, g_n) \leq \exp\{2M(\Theta_0)(\sigma_n^2)^{1/4}\} - 1 + \Delta(\sigma_n^2)^{1/2} \tag{7}$$

Thus, as $n \to \infty$, $\sigma_n^2 \to 0$ and as a result both of the bounds calculated above tend to zero, i.e. $d_{V1}(f_n, g_n) \to 0$ and $d_{V2}(f_n, g_n) \to 0$.

Figure 1 illustrates the variation distance bounds given in (6) and (7) associated with $f_0(\theta \mid \boldsymbol{\alpha}_f) = Beta(4,6)$ with two different choices of $\rho_n = (\sigma_n^2)^{1/4}$ (presented by "o" in the figure) and $\rho_n = (\sigma_n^2)^{1/3}$ (presented by "*" in the figure). Figure 2 illustrates the variation distance bound associated with $f_0(\theta \mid \boldsymbol{\alpha}_f) = Beta(16,24)$ with the similar choices of ρ_n's as above.

It can be concluded that the convergence rates of the bounds associated with both functioning priors considered above are faster for $\rho_n^2 = (\sigma_n^2)^{1/2}$, and the convergence rates of the bounds are faster for $f_0(\theta \mid \boldsymbol{\alpha}_f) = Beta(4,6)$ despite their larger values for the small sample size.

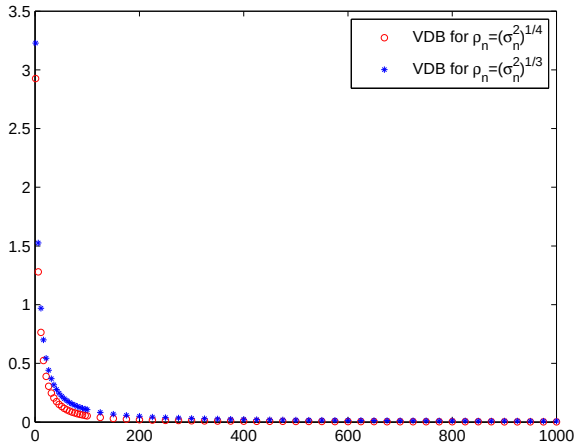

Figure 1: The variation distance bound associated with $f_0(\theta) = Beta(4, 6)$, $M(\theta_0) = 5, p = 2, \Delta = 1.2, \rho_n^3 = \sigma_n^2$.

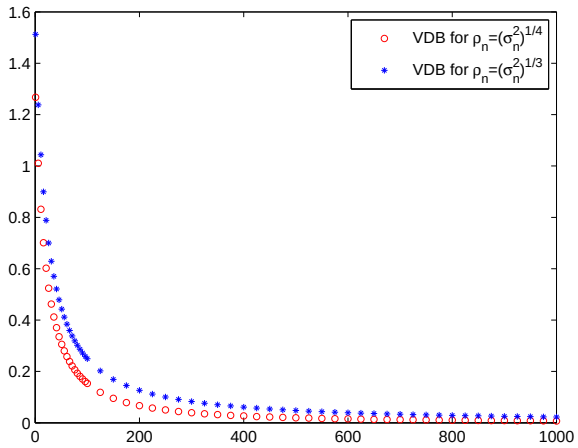

Figure 2: The variation distance bound associated with $f_0(\theta) = Beta(16, 24)$, $M(\theta_0) = 5, p = 2, \Delta = 1.2, \rho_n^3 = \sigma_n^2$.

Example 3. Sometimes it is convenient, particularly with covariate information, to smoothly transform a vector of probabilities. One commonly used transformation in Bayesian networks is the logistic transformation [15]. Like the variation distance the local DeRobertis 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 ([9]; [1]). 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 local DeRobertis and the tail domination condition above than super-exponential tails families - like the Gaussian, and should be preferred in this context [10].

Of course the usual priors in discrete graphical models 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 be calculated given that $d^L_{\boldsymbol{\mu}_n,\rho_n}(f_0, g_0)$ stay unbounded above as $n \to \infty$. Thus [12] 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 to 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 local DeRobertis separation measures, 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 Bayesian Network's

3.1. Some General Results for Multivariate Bayesian Network's

We begin with some general comments about multivariate robustness.

In [14] 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

$$d^L_{A_1}(f_1, g_1) \leq d^L_A(f, g) \tag{8}$$

where $A_1 = \{\boldsymbol{\theta}_1 : \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in A \text{ for all } \boldsymbol{\theta}_2 \in B \subset \Theta_2 \text{ for some open set } B \text{ in } \Theta_2\}$. 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 Bayesian network and we can show that the functioning prior converges on that margin, then even if the model is unidentified provided that $f_0, g_0 \in \mathcal{F}(\Theta_0, M(\Theta_0), p(\Theta_0))$, 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 Bayesian networks 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 Bayesian networks whose probabilities all exhibit prior local and global independence. Immediately from the definition of $d^L_A(f, g)$ 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) marginal densities on $\boldsymbol{\theta}_i$,

$1 \leq i \leq k$, then

$$d_A^L(f,g) = \sum_{i=1}^{k} d_{A_i}^L(f_i, g_i) \tag{9}$$

It follows that - all other things being equal - our local prior distances grow linearly with the number of parameters needed to specify a Bayesian network. In particular models encoding more conditional independences are intrinsically more stable and the effects of possibly erroneous prior information will endure less long 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 Bayesian network 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 priors both exhibit local and global independence: 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 local DeRobertis separation between the s component densities of the functioning and genuine priors 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

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

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

The last most general separation bound increases exponentially with m. By (8) this in turn implies that Bayesian network'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 local DeRobertis. 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 Bayesian network is decomposable with cliques $C[j]$, $j = 1, 2, \ldots, m$ then if we require local and global independence 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 Bayesian network will also be Dirichlet for both the genuine and functioning priors 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 priors 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 [13]. 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 local and global independence 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 local and global independence matters. One possible extension away from local and global independence that naturally occurs under selection biases is for the vector of probabilities in the problem to mirror the dependence structure of the Bayesian network 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 $f_{[i]}, g_{[i]}$ are respectively the marginal densities of f and g on the space $\Theta[i]$ of the i^{th} variable and its parents [12]. 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 3. 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 $g \in \mathcal{N}(f)$ respects the same factorisation as f and

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

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

If we believe the genuine prior $g \in \mathcal{N}(f)$ 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 [12]).

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 local DeRobertis discrepancy between genuine and functioning priors should be set at the same deviation parameters as the independence priors. So on this strong assumption we regain the stability existing under local and global independence.

Example 6 (Gaussian Bayesian Network). A Gaussian Bayesian network is a directed acyclic graph (DAG) model as defined by

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{k} p(x_i \mid \mathbf{pa}_i, \theta_i)$$

where each variable X_i is continuous, and each local likelihood is the linear regression model presented by

$$p(x_i \mid \mathbf{pa}_i, \theta_i) = N(m_i + \sum_{x_j \in \mathbf{pa}_i} b_{ji} x_j, 1/v_i) \tag{10}$$

Given this form, a missing arc from X_j to X_i is equivalent to $b_{ji} = 0$ in the DAG model. The local parameters are given by $\theta_i = (m_i, b_i, v_i)$, where $b_i = (b_{1i}, \ldots, b_{i-1,i})$ of regression coefficients. Furthermore, m_i is the conditional mean of X_i and v_i is the conditional variance of X_i.

The joint likelihood function $p(\mathbf{x} \mid \boldsymbol{\theta})$ is a k-dimensional multivariate normal distribution with a mean vector $\boldsymbol{\mu}$ and a symmetric positive definite precision matric W,

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{k} p(x_i \mid \mathbf{pa}_i, \theta_i) = N(\boldsymbol{\mu}, W)$$

It is shown that there is a one-to-one mapping between $\boldsymbol{\theta} = \bigcup_{i=1}^{k} \theta_i$ and $\{\boldsymbol{\mu}, W\}$ (see [4]). In [4] Geiger and Heckerman give certain sets of local and global independence assumptions that if believed demand the use of a normal - Wishart prior for $\{\boldsymbol{\mu}, W\}$, and this has encouraged the use of a functioning prior to be drawn from this family. However suppose there was concern that the conditions required for the characterisation were not compelling but only held approximately in terms of the local DeRobertis separation. Proceeding with our functioning prior by change of variables, we can get the prior distribution for $\{m_n, b_n, v_n\}$ from the prior distribution presented above for $\{\boldsymbol{\mu}, W\}$ as

$$f_i(m_i, b_i, v_i) = Wishart(1/v_i|\alpha_W + k - i, T_{22} - T_{12}' T_{11}^{-1} T_{12})$$

$$\times N(b_i|T_{11}^{-1}T_{12}, T_{22}/v_k) N(m_i|\mu_{0i}, \alpha_{\boldsymbol{\mu}}/v_i), \quad i = 1, \ldots, k$$

where the first block, T_{11} corresponds to $X_1, \ldots, X_i$ and the second block , T_{22} corresponds to $X_{i+1}, \ldots, X_k$. It should be noticed that the only independence assumption expressed by this product is that m_i and b_i are independent given v_i (see [7]).

In this example, we assume the regression coefficients, b_i and the conditional variances of X_i's, v_i are known, and we focus on the Bayesian sensitivity about the conditional means of X_i's when the global parameter independence satisfies. So assume our functioning prior is

$$f_0(\mathbf{m} \mid \mathbf{v}) = \prod_{i=1}^{k} p(m_i \mid v_i) = \prod_{i=1}^{k} N(m_{0i}, \alpha_{\boldsymbol{\mu}}/v_i)$$

where $m_{0i} = \mu_{0i} - \sum_{j=1}^{i-1} b_{ji}\mu_{0j}$.

By combining the likelihood of the parameters given in (10) with the prior distribution of m_i, $p(m_i) = N(m_{0i}, \alpha_{\boldsymbol{\mu}}/v_i)$, the posterior distribution associated with each m_i is:

$$p(m_i \mid \mathcal{D}_i = (x_{i1}, x_{i2}, \ldots, x_{in})) = N(\mu_{i,n}, \sigma_{ii,n}^2) \tag{11}$$

where $\mu_{i,n} = \frac{\alpha_{\boldsymbol{\mu}} m_{0i} + n\bar{x}_i}{\alpha_{\boldsymbol{\mu}} + n}$, $\sigma_{ii,n}^2 = \{(\alpha_{\boldsymbol{\mu}} + n)/v_i\}^{-1}$, $\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$ (see [4], [7]).

In other words, the functioning posterior density f_n is given by

$$\mathbf{m} \mid \mathcal{D} \sim N_k(\boldsymbol{\mu}_n, \Sigma_n)$$

where $\mathbf{m} = (m_1, m_2, \ldots, m_k)$, $\boldsymbol{\mu}_n = (\mu_{1,n}, \ldots, \mu_{k,n})$, $\Sigma_n = diag\{(\sigma^2_{ii,n})^k_{i=1}\}$, $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_k)$.

So letting $\boldsymbol{\mu}^n_0 = \boldsymbol{\mu}_n$, and using Theorem 7.2.2 and Corollary 1 (in [16], p153), we can conclude that

$$F_n(\mathbf{m} \in B(\boldsymbol{\mu}^n_0; \rho_n)) \geq$$

$$F_n\left[\bigcap_{i=1}^{k}\{|m_i - \mu_{i,n}| \leq (\sqrt{k}\rho_n\sigma^{-1}_{ii,n})\sigma_{ii,n}\}\right] \geq 1 - k^{-1}\rho_n^{-2}\sum_{i=1}^{k}\sigma^2_{ii,n}$$

so that

$$F_n(\mathbf{m} \notin B(\boldsymbol{\mu}^n_0; \rho_n)) \leq k^{-1}\rho_n^{-2}\sum_{i=1}^{k}\sigma^2_{ii,n}$$

Suppose we believe that $d^R(f_0(\mathbf{m} \mid \mathbf{v}), g_0(\mathbf{m} \mid \mathbf{v})) \leq \Delta$, and that $g_0 \in \mathcal{F}(\boldsymbol{\mu}^n_0, M(\mathcal{M}_0), p)$ for some prespecified values of $(\Delta, M(\mathcal{M}_0), p)$. Therefore,

$$T_n(2, \rho) \leq \Delta\frac{\sigma^2_n}{\rho^2_n}$$

where $\sigma^2_n = \max_{1\leq i\leq k}\sigma^2_{ii,n} = \frac{v_n}{k(n+\alpha\boldsymbol{\mu})}$, $v_n = \max_{1\leq i\leq k} v_i$. Let $\tau_n = \sigma^{-r}_n$ for some $0 < r < 1$, and let $\rho_n = \sigma_n\tau_n = \left(\frac{v_n}{k(n+\alpha\boldsymbol{\mu})}\right)^{\frac{1-r}{2}}$. Note that if $\sigma^2_n(v_n) \to 0$ then $\tau_n \to \infty$ and $\rho_n \to 0$. Provided that, $B(\boldsymbol{\mu}^n_0; \rho_n) \subset M(\mathcal{M}_0)$

$$T_n(1, \rho) \leq \exp\{2M(\mathcal{M}_0)(\frac{v_n}{k(n+\alpha\boldsymbol{\mu})})^{p(1-r)/2}\} - 1$$

Therefore, it can be seen that

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

We use the example reported in [5] in which the duration of time that a machine works is their interest. The machine consists of 7 components and the time to failure of each of these components is considered as a random variable, X_i, $i = 1, \ldots, 7$, and connected together in a DAG shown in Figure 3.

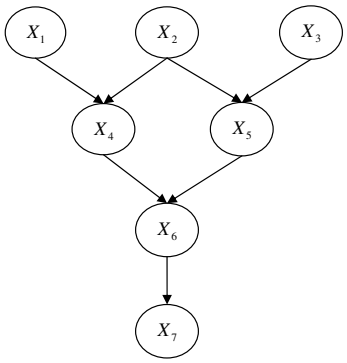

Figure 3: The DAG representation of the Gaussian Bayesian network discussed above.

They assumed that each component of the machine is functioning follows a normal distribution, and the joint probability distribution of these components, $\mathbf{X} = \{X_1, X_2, \ldots, X_7\}$ is a multivariate normal distribution $N(\mu, W)$ with the following parameters.

$$\mu = \begin{pmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 4 \\ 5 \\ 8 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 2 & 8 & 8 \\ 0 & 0 & 2 & 0 & 2 & 4 & 4 \\ 1 & 2 & 0 & 6 & 4 & 20 & 20 \\ 0 & 2 & 2 & 4 & 10 & 28 & 28 \\ 2 & 8 & 4 & 20 & 28 & 97 & 97 \\ 2 & 8 & 4 & 20 & 28 & 97 & 99 \end{pmatrix}$$

The functioning prior distribution for $\mathbf{m}$ is

$$f_0(\mathbf{m} \mid \mathbf{v}) = \prod_{i=1}^{7} N(m_{0i}, \alpha/v_i)$$

where $\mathbf{v} = (1, 1, 2, 6, 10, 97, 99)$ (the diagonal elements of Σ), $\alpha = 65$, $\mathbf{m}_0 = (1.5, 2.75, 2.5, 1.25, 4.5, 5.25, 9)$ (the prior mean, $\mathbf{m}_0$ and α can be elicited from the experts or presented based on the ground information).

The variation distance bound based on the information mentioned above for different values of $r = 0.25, 0.5, 0.75$ are given in Figure 4. It can be

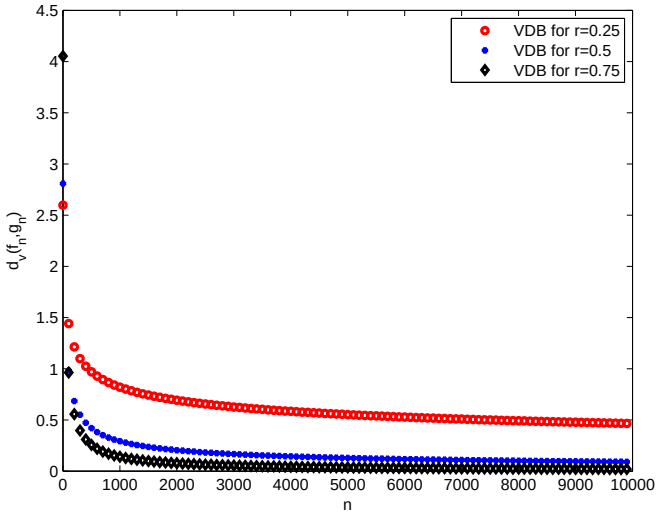

Figure 4: The variation distance bounds associated with $\alpha_{\boldsymbol{\mu}} = 65, M(\mathcal{M}_0) = 5, p = 2, \Delta = 1.2$ and $r = 0.25, 0.5, 0.75$, respectively.

concluded that the convergence rates of the bounds will be increased as the r increases.

4. Discussion

For any Bayesian network whose density factorises in terms of the conditional probabilities, the local DeRobertis separations are a valuable way of understanding exactly what forces the final posterior inference. 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 Bayesian network'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 Bayesian network's and also for all other graphical models encoding a single factorisation. 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

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

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

[3] L. DeRobertis, The use of partial prior knowledge in Bayesian inference, PhD dissertation, Yale Univesity, 1978.

[4] D. Geiger, D. Heckerman, Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions, Annals Statist, 30 (2002) 1412-1440.

[5] M.A. Gómez-Villegas, P. Main, R. Susi, Sensitivity of Gaussian Bayesian networks to inaccuracies in their parameters, in: M. Jaeger, T.D. Nielsen (Eds.), Proceedings of the 4^{th} European Workshop on Probabilistic Graphical Models, Hirtshals, Denmark, 2008, pp. 265-272.

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

[7] D. Heckerman, D. Geiger, Likelihoods and Parameter Priors for Bayesian Networks, Technical Report, MSR-TR-95-54, Microsoft Research, Advanced Technology Division, 1995.

[8] J.B. Kadane, D.T. Chuang, Stable decision problems, Ann. Statist, 6 (1978) 1095-1111.

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

[10] A. O'Hagan, J. Forster, Bayesian Inference, Kendall's Advanced Theory of Statistics, Arnold, 2004.

[11] T. Silander, P. Kontkanen, P. Myllymäki, On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter, In: R. Parr, L. van der Gaag (Eds.), Proceedings of the 23^{rd} Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 360-367.

[12] J.Q. Smith, Local Robustness of Bayesian Parametric Inference and Observed Likelihoods, CRiSM Research Report number 07-08, The University of Warwick, 2007.

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

[14] J.Q. Smith, F. Rigat, Isoseparation and Robustness in Finitre Parameter Bayesian Inference, CRiSM Research Report number 07-22, The University of Warwick, 2008.

[15] D.J. Spiegelhalter, S.L. Lauritzen, Sequential updating of conditional probabilities on directed graphical structures, Networks, 20 (1990) 579-605.

[16] Y.L. Tong, Probability Inequalities in Multivariate Distributions, Academic Press New York, 1980.