International Journal of Approximate Reasoning 51 (2010) 558–572

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

On the robustness of Bayesian networks to learning from non-conjugate sampling

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

[a] Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom
[b] Department of Statistics, Faculty of Mathematical Sciences and Computer, Shahid Chamran University, Ahvaz 6135714463, Iran

ARTICLE INFO

Article history:
Received 9 December 2008
Received in revised form 1 June 2009
Accepted 25 August 2009
Available online 28 January 2010

Keywords:
Bayesian networks
Bayesian robustness
Isoseparation property
Local DeRobertis distance
Total variation distance

ABSTRACT

Recent results concerning the instability of Bayes Factor search over Bayesian Networks (BN's) lead us to ask whether learning the parameters of a selected BN might also depend heavily on the often rather arbitrary choice of prior density. Robustness of inferences to misspecification of the prior density would at least ensure that a selected candidate model would give similar predictions of future data points given somewhat different priors and a given large training data set. In this paper we derive new explicit total variation bounds on the calculated posterior density as the function of the closeness of the genuine prior to the approximating one used and certain summary statistics of the calculated posterior density. We show that the approximating posterior density often converges to the genuine one as the number of sample point increases and our bounds allow us to identify when the posterior approximation might not. To prove our general results we needed to develop a new family of distance measures called local DeRobertis distances. These provide coarse non-parametric neighbourhoods and allowed us to derive elegant explicit posterior bounds in total variation. The bounds can be routinely calculated for BNs even when the sample has systematically missing observations and no conjugate analyses are possible.

1. Introduction

Bayesian networks are now widely used as a framework for inference. Once the structure of a BN has been selected it is necessary to choose a distribution over its parameters, a set of conditional probabilities in the case of a discrete BN or mean vector and covariance matrix in the case of a continuous BN. Various authors have suggested ways to do this (see [6,10] and references therein). For example suppose a Bayesian graphical expert system needs to be developed – perhaps for medical diagnosis. The necessary selection of a BN which explains well the variation in the training data has been widely discussed in the machine learning literature. But it is also important to ensure that the selected BN model together with its chosen prior hyperparameters will predict well the values of new units: in the example above future patients diagnosed by the medical expert system. This second issue is the focus of this paper.

It is challenging to set the prior distributions appropriately over the parameters of a chosen BN to a given context. Furthermore, especially when working in problems where data on some variables is systematically missing – when inferences are made based on numerical methods or approximations – it is very difficult to appreciate the effect on inference this choice of prior over the parameters of a chosen BN might have. Without some appropriate diagnostic, the modeler would be justifiably uneasy about the veracity of her inferences. Of course it is possible to perform sensitivity analyses to check the effect

* Corresponding author.
E-mail addresses: j.q.smith@warwick.ac.uk (J.Q. Smith), adaneshkhah@scu.ac.ir, alireza.daneshkhah@strath.ac.uk (A. Daneshkhah).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.01.013

of a few different combinations of hyperparameters within a chosen parametric family of prior densities. But the results of such a study can be no more than indicative of possible sensitivities. After all why should the modeler believe a prior should lie in a given parametric family?

There used to be a common misconception that inferences would be robust to the way a prior over parameters was chosen provided the priors densities were reasonably diffuse and the sample size moderately large, irrespective of whether or not that prior was chosen from conventional parametric families of distribution. However two strands of research over recent decades have undermined this belief: new results about Bayesian model selection and new results concerning local sensitivity. We outline some of the more pertinent results from these two different fields below.

There are some other works which addressed the sensitivity analysis of the selected BN from different perspectives. In [17], the author examines the use of global sensitivity analysis by calculating the bounds of some posterior quantities of interests when the prior distribution varies in some class of distributions. He also develops some numerical method based on the importance sampling to calculate the requested bounds. In another study reported in [4], the Bayesian robustness is examined when the prior belongs to a class defined in terms of the so called generalized moment conditions which the problem can be then reduced to one of linear semi-infinite programming (LSIP). A numerical method based on the accelerated central cutting plane algorithm to solve LSIP problems is introduced and illustrated by an example.

A popular choice of BN model selection is to use the maximum – a – posteriori (MAP) score to discriminate between competing models. In order to evaluating the score of different BNs, the marginal likelihood of each model needs to be calculated which in turn requires a prior density over the parameters for each BN in the candidate set. For discrete Bayesian networks, the model parameters are its defining vectors of conditional probabilities and it is usual to use the conjugate product Dirichlet priors on these parameters. This conjugate family has some justification because this family exhibits certain invariance properties over the class of BNs (see [10] for details). However even if all cells in the joint probability tables are assigned uniformly as is required for the BDeu score [5] there still remains an additional parameter to fix the equivalent sample size parameter α. Recently both from the theoretical [24,23] and from the practical point of view [18], model selection has been found to be very sensitive to how this hyperparameter is set. So in the problem of model selection, even when a Bayesian analysis is restricted to one using standard families of prior densities on the parameters of a BN and when sampling is complete, the choice of this prior over the BN's conditional probabilities can have a critical impact on the ensuing inference. Various solutions to this problem have been proposed, most recently one by [23] who develops a fast approximate method of simultaneously maximising over α and the space of BN's to select a model.

Now it is true that the robustness to the misspecification of the prior over the parameters of a BN for model selection can be quite different to robustness of inferences within a selected BN. To illustrate this distinction it is sufficient to consider the following very simple example.

Suppose that it is known that either all components of a vector $\mathbf{x}$ of observations will be strictly positive with a known density $p^+(\mathbf{x})$ or all components are negative with a known density $p^-(\mathbf{x})$. Let model $M(\alpha)$ assign a probability α where $0 < \alpha < 1$ to all observations being positive. Here we can think of α of the hyperparameter of a prior density and the corresponding distribution of $\mathbf{x}$ the marginal likelihood of the observations after integrating out the parameters θ of the model. After observing the first observation x_1 all these models will give the same predictions, forecasting all future observations using p^+ if $x_1 > 0$ and p^- if $x_1 < 0$. So for prediction problems – the issue we address in this paper – the problem is completely robust to the possible misspecification of the hyperparameter α once x_1 – our training data – has been observed. On the other hand MAP model selection will score $M(\alpha)$ increasingly highly in α for $x_1 > 0$ and decreasingly in α for $x_1 < 0$. So a model that is clearly suboptimal from a selection point of view can be entirely adequate for forecasting. This is an extreme example of a phenomenon where Bayes Factors can score models lowly simply because of initial poor calibration of hyperparameters which after a few data points recalibrates to make future forecasts almost as reliable as they could be. Nevertheless the sensitivity of model selection to the choice of prior is disturbing. After an appropriate structure of a BN has been selected it leads us to question whether the exact form of its prior over its parameters will have an enduring effect on later inferences.

In a second strand of research mathematical statisticians have addressed the issue of the robustness of Bayesian inference to misspeciation of a prior in a given model: here our selected BN. Thus suppose the structure of this BN has been selected, by whatever means, and we are interested in using it for forecasting future units: in our running example for diagnosing future patients. The formal Bayesian approach would require us to carefully elicit all expert judgements and express these judgements faithfully as a joint probability distribution over the parameters of the BN. Were this ever to be done it would be somewhat unrealistic to believe that the most appropriate prior over the parameters of our chosen BN – here called our *genuine prior* and denoted by g_0 – would exactly lie in a convenient conjugate family. However this process is usually extremely costly and for pragmatic reasons a prior density – here called our *functioning prior* and denoted by f_0 – will often be chosen to approximate g_0 – usually from a convenient parametric family. So the best that can reasonably be assumed is that the genuine prior g_0 will lie within a non-parametric neighbourhood of the prior f_0 used in the analysis. After observing n vectors of data points, provided issues of unidentifiability are avoided, even when certain sets of values of these variables in the BN are missing and f_n cannot be calculated in closed form, most practitioners would hope that if the posterior – here called our *functioning posterior* f_n – associated with the prior f_0 concentrated on to a small ball of values in the parameter space then f_n would also be a good approximation of the *genuine posterior* g_n – i.e. the one we would have obtained if we had thought as hard as we could about the complex scenario in front of us – provided n was very large.

However a startling result in [12] proved that no currently used non-parametric neighborhoods of prior distributions, based on prior total variation distances or ϕ-divergence, including Kullback–Leibler, Hellinger, directed divergence and