NON LOCAL ALTERNATIVE PRIORS FOR GAUSSIAN DIRECTED ACYCLIC GRAPHICAL MODELS
Revised Version

Luca La Rocca[1] and Guido Consonni[2]

[1] Dipartimento di Scienze Sociali, Cognitive e Quantitative
Università di Modena e Reggio Emilia
Viale Allegri 9, 42121 Reggio Emilia, Italy
(e-mail: luca.larocca@unimore.it)
[2] Dipartimento di Economia Politica e Metodi Quantitativi
Università di Pavia
Via S. Felice 7, 27100 Pavia, Italy
(e-mail: guido.consonni@unipv.it)

ABSTRACT. Prior choice for Bayesian model comparison can be problematic for several reasons. In particular, for the comparison of two nested models, it was recently pointed out that typical prior choices may produce an unsatisfactory learning behavior of the Bayes factor. More in detail, if the sub-model is not true the accumulation of evidence is exponentially fast in favor of the encompassing model, whereas it is only sub-linear in favor of the sub-model under the assumption that the latter is true. To alleviate this imbalance, it was suggested that the prior under the encompassing model be modified so that it vanishes over the sub-space corresponding to the sub-model, thus obtaining a Non Local Alternative Prior (NLAP). In this work, we develop NLAPs for the comparison of Gaussian directed acyclic graphical models, and contrast their performance with that of traditional priors.

1 INTRODUCTION

Graphical models are a powerful tool for studying the structure of dependencies between three or more variables. For instance, in a simple application discussed by Wermuth (1993), the Directed Acyclic Graph (DAG) on the left of Figure 1 can be used to express, for patients with hypertension, the research hypothesis that systolic blood pressure depends on age only through the effect of age on weight, whereas the DAG on the right of Figure 1 can be used to express the alternative hypothesis that age also has a direct effect on systolic blood pressure. Wermuth (1993) makes a distinction between substantive research hypotheses and statistical association models, but for the goals of this work we can safely drop it, and identify models with hypotheses. We shall compare the two models in Figure 1 within a Bayesian framework, using in particular the Bayes Factor (BF). With even prior odds (in lack of specific prior knowledge) for the two models, the BF can be turned into the posterior probability of the simpler model, which represents an easily interpretable measure of evidence.

Bayesian model comparison is still an active area of research, especially with regard to prior choice; see Pericchi (2005) for a review. In this work, focussing on the above described application, we deal with prior choice for Gaussian DAG models, which we briefly introduce in the next section. For a general presentation of graphical model theory, terminology and notation the Reader is referred to Cowell *et al*. (1999).

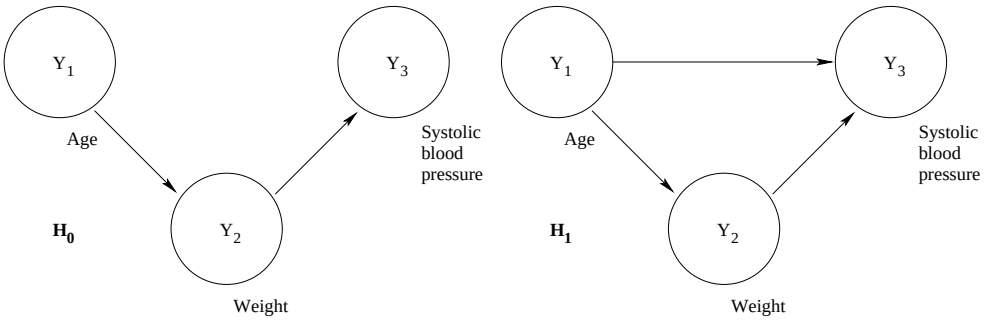

Figure 1. DAGs for the null (research) and alternative hypotheses in the simple application on patients with hypertension discussed by Wermuth (1993).

Variable	Y_1 (Age)	Y_2 (Weight)	Y_3 (Systolic blood pressure)
Y_1 (Age)	1.000	0.369	-0.007
Y_2 (Weight)	0.390	1.000	0.348
Y_3 (Systolic blood pressure)	0.139	0.371	1.000
Mean	32.74	0.42	128.31
Standard deviation	11.67	0.04	13.47

Table 1. Summary statistics for $n = 98$ patients with hypertension, as reported by Wermuth (1993): marginal correlations (lower half), partial correlations (upper half), means and standard deviations (age in years, weight relative to height, systolic blood pressure in millimeters of mercury).

2 GAUSSIAN DAG MODELS

A *directed graph* $D = (V,E)$ consists of a finite set of *vertices* $V = \{1,\ldots,q\}$ together with a set of *edges* $E \subseteq V \times V$ such that $(j,k) \in E$ implies $(k,j) \notin E$. A *cycle* in D is a sequence of vertices $j_0, j_1, \ldots, j_m$ such that $(j_{\ell-1}, j_\ell) \in E$ for all $\ell = 1,\ldots,m$ and $j_0 = j_m$. A directed graph containing no cycles is called a DAG. Two DAGs are depicted in Figure 1, with circles representing vertices and arrows representing edges.

Given a DAG D, the Gaussian DAG model defined by D is the family of q-variate normal distributions M_D such that their density function factorizes as

$$f(y_1,\ldots,y_q|\alpha,\beta,\gamma) = \prod_{j=1}^{q} f(y_j|y_{\mathrm{pa}(j)};\alpha_j,\beta_j,\gamma_j), \tag{1}$$

where $\alpha = (\alpha_j)_{j=1}^q$, $\beta = (\beta_j)_{j=1}^q$, $\gamma = (\gamma_j)_{j=1}^q$, $\mathrm{pa}(j) = \{k \in V : (k,j) \in E\}$ are the *parents* of j (in D), $y_{\mathrm{pa}(j)} = [y_k]'_{k\in\mathrm{pa}(j)}$ is a (column) vector of length $\#\mathrm{pa}(j)$, and $f(y_j|y_{\mathrm{pa}(j)};\alpha_j,\beta_j,\gamma_j)$ is a univariate normal density with mean $\alpha_j + \beta'_j y_{\mathrm{pa}(j)}$ and variance γ_j^{-1}.

If an i.i.d. sample $Y = [[Y_{i,j}]_{j=1}^q]_{i=1}^n$ of size n is available (an $n \times q$ matrix) the following expression for the likelihood under M_D is obtained from (1) by algebraic manipulation:

$$L(\alpha,\beta,\gamma|Y) = \prod_{j=1}^{q} \left(\frac{\gamma_j}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{n\gamma_j}{2}\left[C_{j,j}+\beta'_j C_{\mathrm{pa}(j),\mathrm{pa}(j)}\beta_j - 2C'_{\mathrm{pa}(j),j}\beta_j + \left(\bar{Y}_j - \alpha_j - \bar{Y}'_{\mathrm{pa}(j)}\beta_j\right)^2\right]}, \tag{2}$$

where $\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{i,j}$ is the sample mean of the j-th variable, C is the sample covariance matrix, that is, $C_{j,k} = \frac{1}{n}\sum_{i=1}^{n}(Y_{i,j} - \bar{Y}_j)(Y_{i,k} - \bar{Y}_k)$, and $C_{J,K}$ denotes the submatrix of C indexed by $j \in J$ and $k \in K$, for any two subsets J and K of V.

In the following two sections we discuss prior choices for the likelihood (2), in view of model comparison.

3 Conjugate Analysis

Since $(\alpha_j, \beta_j, \gamma_j)$, $j = 1, \ldots q$, in (2) are variation independent, a convenient (and widespread) prior choice is to let them be stochastically independent; this assumption is called *global parameter independence* by Cowell *et al.* (1999) and Geiger and Heckerman (2002). With this choice, the marginal likelihood of M_D given Y can be written as

$$L(M_D|Y) = \prod_{j=1}^{q} f(Y_j|Y_{\mathrm{pa}(j)}), \tag{3}$$

where $f(Y_j|Y_{\mathrm{pa}(j)}) = \int\int\int f(Y_j|Y_{\mathrm{pa}(j)}; \alpha_j, \beta_j, \gamma_j) p(\alpha_j, \beta_j, \gamma_j) d\alpha_j d\beta_j d\gamma_j$, if $p(\alpha_j, \beta_j, \gamma_j)$ is the prior density of $(\alpha_j, \beta_j, \gamma_j)$, Y_j denotes the data for variable j, and $Y_{\mathrm{pa}(j)}$ the data for variables in $\mathrm{pa}(j)$. It is easy to see that $(\alpha_j, \beta_j, \gamma_j)$, $j = 1, \ldots q$, are also independent *a posteriori*.

Given two DAG models M_{D_0} and M_{D_1}, the Bayes Factor (BF) of M_{D_1} against M_{D_0} is

$$BF_{10}(Y) = \frac{L(M_{D_1}|Y)}{L(M_{D_0}|Y)} \tag{4}$$

and from (4), assuming even prior odds for the two models, the posterior probability of M_{D_0} is obtained as $P(M_{D_0}|Y) = 1/(1 + BF_{10}(Y))$.

With reference to Gaussian DAG models, a conjugate prior is available (and often used) for $(\alpha_j, \beta_j, \gamma_j)$. Let $\eta_j = [\alpha_j \ \beta_j']'$ and $p_j = \#\mathrm{pa}(j) + 1$, then take

$$\eta_j|\gamma_j \sim N_{p_j}\left(e_j, \gamma_j^{-1} E_j^{-1}\right), \qquad \gamma_j \sim G\left(\frac{n_0}{2}, \frac{s_j}{2}\right), \tag{5}$$

where $N_p(\mu, \Sigma)$ denotes a p-variate normal distribution with mean vector μ and covariance matrix Σ, $G(\alpha, \beta)$ a gamma distribution with mean $\alpha\beta^{-1}$, e_j is a prior mean vector, E_j a positive definite prior precision matrix, n_0 has the meaning of a prior sample size, and s_j the meaning of a prior sum of squares. Because of conjugacy, letting $\tilde{Y}_{\mathrm{pa}(j)} = [1_n \ Y_{\mathrm{pa}(j)}]$, where 1_n is the vector with n ones, the posterior of (η_j, γ_j) is as in (5) with the following updated hyperparameters: $E_j^\star = \tilde{Y}'_{\mathrm{pa}(j)}\tilde{Y}_{\mathrm{pa}(j)} + E_j$, $e_j^\star = E_j^{\star -1}(\tilde{Y}'_{\mathrm{pa}(j)} Y_j + E_j e_j)$, $n_0^\star = n_0 + n$, and $s_j^\star = s_j + (Y_j - \tilde{Y}_{\mathrm{pa}(j)} e_j)'(I_n - \tilde{Y}_{\mathrm{pa}(j)} E_j^{\star -1} \tilde{Y}'_{\mathrm{pa}(j)})(Y_j - \tilde{Y}_{\mathrm{pa}(j)} e_j)$, where I_n denotes the $n \times n$ identity matrix. Moreover, the *local marginal likelihood* for vertex j can be written as

$$f(Y_j|Y_{\mathrm{pa}(j)}) = \frac{\Gamma(n_0^\star/2)|E_j|^{\frac{1}{2}} s_j^{\frac{n_0}{2}}}{\pi^{\frac{n}{2}}\Gamma(n_0/2)|E_j^\star|^{\frac{1}{2}} s_j^{\star \frac{n_0^\star}{2}}}, \tag{6}$$

where Γ denotes the gamma function and $|A|$ the determinant of a matrix A. Equation (6) is equation (9.44) of O'Hagan (1994) rewritten in our notation, and the expression we use for $s_j^\star$ can be otained from (9.47) of O'Hagan (1994) using (17) of Henderson and Searle (1981). This expression for $s_j^\star$ shows that the distribution of Y_j given $Y_{\mathrm{pa}(j)}$ is an n-variate Student t distribution with n_0 degrees of freedom, location vector $\tilde{Y}_{\mathrm{pa}(j)} e_j$, and scale matrix $n_0^{-1} s_j (I_n - \tilde{Y}_{\mathrm{pa}(j)} E_j^{\star -1} \tilde{Y}'_{\mathrm{pa}(j)})^{-1}$; cf. Fernandez *et al*. (2001). It also shows that (6) depends on data only through $\tilde{Y}'_{\mathrm{pa}(j)} \tilde{Y}_{\mathrm{pa}(j)}$, $\tilde{Y}'_{\mathrm{pa}(j)} Y_j$ and $Y_j' Y_j$, that is, only through n, $\bar{Y}_{\mathrm{fa}(j)}$ and $C_{\mathrm{fa}(j),\mathrm{fa}(j)}$, where $\mathrm{fa}(j) = \{j\} \cup \mathrm{pa}(j)$ is the *family* of j (in D).

4 MOMENT PRIORS

Suppose we want to compare the two Gaussian DAG models defined in Figure 1. Let D_0 be the DAG on the left, and D_1 the DAG on the right. Typically, from a Bayesian perspective, we shall select a conjugate prior for each of the two models, using (5) and global parameter independence, then we shall find $BF_{10}(Y)$ by means of equations (3), (4) and (6).

A general discussion of consistent prior choice for DAG models with the same vertex set is given by Geiger and Heckerman (2001). Here it will be enough to remark that M_{D_0} can be obtained from M_{D_1} by imposing $\beta_{31} = 0$, so that the two models are nested: the prior for M_{D_0} can thus be derived from the prior for M_{D_1} by conditioning or marginalization (given γ to preserve tractability). Whatever the specific choice, the prior p_1 under M_{D_1} will be a Local Alternative Prior (LAP):

$$\exists \varepsilon > 0 : \forall \zeta > 0 : \exists (\alpha, \beta, \gamma) \in I_0(\zeta) : p_1(\alpha, \beta, \gamma) \geq \varepsilon, \tag{7}$$

where $I_0(\zeta) = \{(\alpha, \beta, \gamma) : |\beta_{31}| < \zeta\}$. In words, the prior density will be bounded away from zero in any neighborhood of the nested model, no matter how small it may be.

Johnson and Rossell (2008) recently criticized LAPs on two grounds. On a conceptual ground, prior distributions should convey some notion of *separation* between models, if they are to be used for model comparison. On a pragmatic ground, LAPs lead to an imbalance in the learning behavior of the BF: if the sampling distribution belongs to the encompassing model only, $BF_{10}(Y)$ will be an infinite in probability of order e^{Kn}, as $n \to \infty$, for some $K > 0$, so that M_{D_1} will be chosen exponentially fast; on the other hand, if the sampling distribution also belongs to the nested model, $BF_{10}(Y)$ will be an infinitesimal in probability of order $n^{-L/2}$, as $n \to \infty$, where $L > 0$ is the difference in dimension between the two models, so that M_{D_0} will be chosen polynomially fast. In our example, as well as in all other cases where $L = 1$, the learning behavior will be sub-linear, when the nested model should be chosen, and this will make hard for the research hypothesis to be confirmed.

To alleviate the above described imbalance, Johnson and Rossell (2008) suggested that the prior under the encompassing model be a Non Local Alternative Prior (NLAP):

$$\forall \varepsilon > 0 : \exists \zeta > 0 : \forall (\alpha, \beta, \gamma) \in I_0(\zeta) : p_1(\alpha, \beta, \gamma) < \varepsilon. \tag{8}$$

In concrete, suppose a conventional LAP $p_1^L(\alpha, \beta, \gamma)$ is given. Define a positive continuous function $g(\beta, h)$ that is zero whenever $\beta_{31} = 0$, where h is an additional hyperparameter. Then

$K(h)^{-1}g(\beta,h)p_1^L(\alpha,\beta,\gamma)$ defines a NLAP, where $K(h)$ is the normalizing constant of the new density, provided it exists. The choice $g(\beta,h)=\beta_{31}^{2h}$, with h a strictly positive integer, gives rise to a family of *moment priors*, whose rate of learning becomes $n^{-h-L/2}$ if M_{D_0} is true. Johnson and Rossell (2008) also introduced *inverse moment priors*, achieving an exponential rate of learning when the nested model is true, but we do not consider them here, because they do not provide us with a general-purpose method for prior modification. Notice that moment priors obtained from conjugate priors are still conjugate priors, and that a convenient choice of g preserves parameter independence.

If we start from the conjugate prior defined by (5) and global parameter independence, provided $2h<n_0$, we obtain a valid moment prior with normalizing costant

$$K(h)=\sum_{\ell=0}^{h}\frac{(2h)!\cdot b_{31}^{(2h-2\ell)}\cdot v_{31}^{\ell}\cdot s_j^{\ell}}{2^{\ell}\cdot(2h-2\ell)!\cdot\ell!\cdot\prod_{m=1}^{\ell}(n_0-2m)}, \tag{9}$$

where $b_{31}=e_{32}$ is the prior mean of β_{31}, and $v_{31}=(E_3^{-1})_{2,2}$ the prior variance of $\gamma_3^{1/2}\beta_{31}$, given γ_3; equation (9) can be obtained by computing normal and inverse gamma moments. Then, since $p_1^L(\alpha_j,\beta_j,\gamma_j)f_1^L(Y_j|Y_{\mathrm{pa}(j)};\alpha_j,\beta_j,\gamma_j)=p_1^L(\alpha_j,\beta_j,\gamma_j|Y_j,Y_{\mathrm{pa}(j)})f_1^L(Y_j|Y_{\mathrm{pa}(j)})$ due to global parameter independence (cf. Cowell *et al*. (1999, p. 195)), we obtain the following expression for the moment prior local marginal likelihood:

$$f_1^M(Y_j|Y_{\mathrm{pa}(j)})=\frac{K^{\star}(h)}{K(h)}f_1^L(Y_j|Y_{\mathrm{pa}(j)}), \tag{10}$$

where $K^{\star}(h)$ is as in (9) with updated hyperparameters (and $f_1^L(Y_j|Y_{\mathrm{pa}(j)})$ is as in (6)).

In the next section, using simulated data, we contrast the conjugate and moment priors in terms of evidence for the true model, as sample size grows.

5 Simulation study

In a hypothesis-driven data analysis spirit, suppose we observe the data in Table 1 and these prompt us to compare the two models in Figure 1. We will use the data in Table 1 for prior elicitation, and collect additional data until we come to a conclusion, which we do when the posterior probability of M_{D_0} raises above, or drops below, a decision threshold. Here, data collection will be simulated by generating 100 samples of increasing size from three independent standard normal variables ε_1, ε_2, and ε_3, and letting

$$y_1=30+10\varepsilon_1,\quad y_2=0.4+0.001y_1+0.04\varepsilon_2,\quad y_3=80+\beta_{13}^{\bullet}y_1+100y_2+10\varepsilon_3, \tag{11}$$

where $\beta_{13}^{\bullet}=0$, in case M_{D_0} is true, and $\beta_{13}^{\bullet}=-0.1$, in case M_{D_0} is false (each sample is used twice). We analyze the simulated data with conjugate priors, letting $e_j=\hat{\eta}_j$, $E_j=\tilde{X}'_{\mathrm{pa}(j)}\tilde{X}_{\mathrm{pa}(j)}$, $s_j=X_j'X_j-X_j'\tilde{X}_{\mathrm{pa}(j)}\hat{\eta}_j$, and $n_0=98$, where $\hat{\eta}_j=(\tilde{X}'_{\mathrm{pa}(j)}\tilde{X}_{\mathrm{pa}(j)})^{-1}\tilde{X}'_{\mathrm{pa}(j)}X_j$ and X denotes the prior data in Table 1. Then, keeping the prior under M_{D_0} fixed, we turn the prior under M_{D_1} into the corresponding moment prior, with $h=1$, and repeat the analyses.

Figure 2 shows our findings. The moment prior learns that M_{D_0} is true much faster than the conjugate prior, whose evidence is not yet overwhelming even with huge sample sizes. However, the performance of the moment prior is inferior to that of the conjugate prior when M_{D_0} is false, but this only happens for moderate sample sizes.

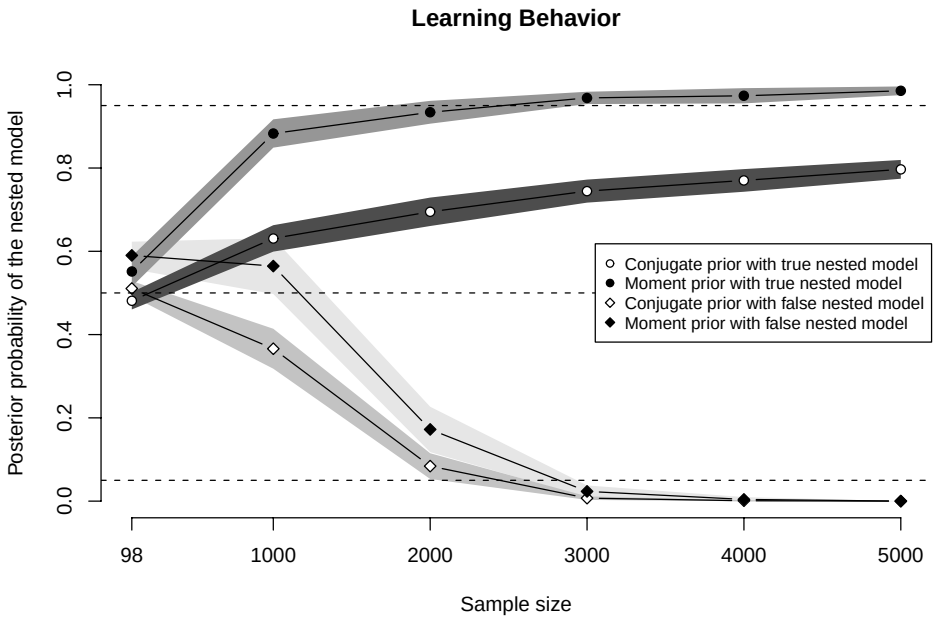

Figure 2. Learning behavior on simulated data of the conjugate and moment priors: $P(M_{D_0}|Y)$ is plotted for both priors, as a function of n, under two sampling conditions. Lines interpolate simulation averages, and shaded regions represent 95% ($\pm$ two standard errors) confidence bands. The dashed horizontal lines mark possible decision thresholds at 0.05, 0.50 and 0.95.

REFERENCES

COWELL, R.G., DAWID, A.P., LAURITZEN, S.L., SPIEGELHALTER, D.J. (1999): *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Sciences. Springer-Verlag, New York.

FERNANDEZ, C., LEY, E., STEEL, M.F.J. (2001): Benchmark priors for Bayesian model averaging. *Journal of Econometrics, 100*, 381–427.

GEIGER, D., HECKERMAN, D. (2002): Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics, 30*, 1412–1440.

HENDERSON, H.V., SEARLE, S.R. (1981): On deriving the inverse of a sum of matrices. *SIAM Review, 23*, 53–60.

JOHNSON, V.E., ROSSELL, D. (2008): *Non-local prior densities for default Bayesian hypothesis tests*. Working Paper 42, Department of Biostatistics, MD Anderson Cancer Center, University of Texas.

O'HAGAN, A. (1994): *Bayesian Inference*. Volume 2B of Kendall's Advanced Theory of Statistics. Arnold, London.

PERICCHI, L.R. (2005): Model selection and hypothesis testing based on objective probabilities and Bayes factors. In: Dey, D.K. and Rao C.R. (Eds.), *Handbook of Statistics 25: Bayesian Thinking Modeling and Computation*. North Holland, Amsterdam, 115–149.

WERMUTH, N. (1993): Association structures with few variables: characteristics and examples. In: K. Dean (Ed.), *Population Health Research: Linking Theory and Methods*. Sage, London, 181–203.