# SUPPLEMENT TO

# "ADDENDUM ON THE SCORING OF GAUSSIAN DIRECTED ACYCLIC GRAPHICAL MODELS"

# DERIVING AND SIMPLIFYING THE BGE SCORE

By Jack Kuipers*, Giusi Moffa* and David Heckerman†

*Regensburg University** *and Microsoft Research*†

We derive the expression for scoring Gaussian directed acyclic graphical models and discuss how to evaluate the score efficiently.

**A. Deriving the score.** We first proceed through the steps needed to derive the Bayesian Gaussian equivalent (BGe) score as introduced in Geiger and Heckerman (1994); Heckerman and Geiger (1995); Geiger and Heckerman (2002). We use the same notation as Geiger and Heckerman (2002) and let $n$ be the dimension of the variables $\mathbf{X}$, with $N$ the number of observations of $\mathbf{x}$ in a dataset $d = \{\mathbf{x}_1, \dots \mathbf{x}_N\}$. For a Gaussian directed acyclic graph (DAG) model, we wish to evaluate the terms in the product

$$p(d \mid m^h) = \prod_{i=1}^{n} \frac{p(d^{\mathbf{Pa}_i \cup \{X_i\}} \mid m_c^h)}{p(d^{\mathbf{Pa}_i} \mid m_c^h)} \tag{A.1}$$

where $\mathbf{Pa}_i$ are the parent variables of the vertex $i$ and $d^{\mathbf{Y}}$ is the data restricted to the coordinates in $\mathbf{Y} \subseteq \mathbf{X}$. In the following we leave out the explicit dependence on the model hypothesis $m^h$ and the complete data model $m_c^h$ in the formulae.

A.1. *Definitions.* The data is assumed to be normally distributed with mean $\boldsymbol{\mu}$ and precision matrix $W$

$$\begin{aligned} p(d \mid \boldsymbol{\mu}, W) &= \frac{|W|^{\frac{N}{2}}}{(2\pi)^{\frac{nN}{2}}} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\mu}-\mathbf{x}_i)^{\mathrm{T}} W (\boldsymbol{\mu}-\mathbf{x}_i)} \\ &= \frac{|W|^{\frac{N}{2}}}{(2\pi)^{\frac{nN}{2}}} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left[\left(\sum_{i=1}^{N}(\boldsymbol{\mu}-\mathbf{x}_i)(\boldsymbol{\mu}-\mathbf{x}_i)^{\mathrm{T}}\right) W\right]} \end{aligned} \tag{A.2}$$

*MSC 2010 subject classifications:* 62-07, 62F15, 62H99

*Keywords and phrases:* Gaussian DAG models, Bayesian network learning, BGe score

where we reordered the exponent using the cyclicity of the trace. The prior on $W$ is taken to be a Wishart distribution, $W \sim \mathcal{W}_n(T^{-1}, \alpha_w)$, where $\alpha_w > n-1$ is the degrees of freedom and $T$ is the positive definite parametric matrix. We follow the standard practice, as in Press (1982), of using the scale matrix (the inverse of the parametric matrix) in the argument of the distribution, which has the following form

$$p(W) = \frac{|W|^{\frac{\alpha_w - n - 1}{2}}}{Z_{\mathcal{W}}(n, T, \alpha_w)} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[TW]} \tag{A.3}$$

with normalising constant

$$Z_{\mathcal{W}}(n, T, \alpha_w) = \frac{2^{\frac{\alpha_w n}{2}} \Gamma_n\left(\frac{\alpha_w}{2}\right)}{|T|^{\frac{\alpha_w}{2}}} \tag{A.4}$$

involving the multivariate gamma function

$$\Gamma_n\left(\frac{\alpha_w}{2}\right) = \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^{n} \Gamma\left(\frac{\alpha_w + 1 - j}{2}\right) \tag{A.5}$$

Next assume a normal prior on $\boldsymbol{\mu}$ with mean $\boldsymbol{\nu}$ and precision matrix $\alpha_\mu W$, with $\alpha_\mu > 0$,

$$\begin{aligned} p(\boldsymbol{\mu}|W) &= \frac{1}{Z_{\mathcal{N}}(n, W, \alpha_\mu)} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\nu})^{\mathrm{T}} \alpha_\mu W (\boldsymbol{\mu}-\boldsymbol{\nu})} \\ &= \frac{1}{Z_{\mathcal{N}}(n, W, \alpha_\mu)} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left[(\boldsymbol{\mu}-\boldsymbol{\nu})(\boldsymbol{\mu}-\boldsymbol{\nu})^{\mathrm{T}} \alpha_\mu W\right]} \end{aligned} \tag{A.6}$$

again using the cyclicity of the trace and where the normalising constant is

$$Z_{\mathcal{N}}(n, W, \alpha_\mu) = \frac{(2\pi)^{\frac{n}{2}}}{(\alpha_\mu)^{\frac{n}{2}} |W|^{\frac{1}{2}}} \tag{A.7}$$

Jointly $\boldsymbol{\mu}, W$ therefore follow a normal-Wishart prior distribution

$$p(\boldsymbol{\mu}, W) = \frac{|W|^{\frac{\alpha_w - n}{2}}}{Z(n, \alpha_\mu, T, \alpha_w)} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\nu})^{\mathrm{T}} \alpha_\mu W (\boldsymbol{\mu}-\boldsymbol{\nu})} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[TW]} \tag{A.8}$$

with

$$Z(n, \alpha_\mu, T, \alpha_w) = \frac{2^{\frac{(\alpha_w+1)n}{2}} \pi^{\frac{n}{2}} \Gamma_n\left(\frac{\alpha_w}{2}\right)}{(\alpha_\mu)^{\frac{n}{2}} |T|^{\frac{\alpha_w}{2}}} \tag{A.9}$$

A.2. *The posterior distribution.* The first step is to calculate the posterior probability $p(\boldsymbol{\mu}, W|d) = p(d, \boldsymbol{\mu}, W)/P(d)$ of the parameters $\boldsymbol{\mu}, W$ given the data $d$. Start by writing the joint probability in the form $p(d, \boldsymbol{\mu}, W) = p(d|\boldsymbol{\mu}, W)p(\boldsymbol{\mu}, W)$ and taking the log

$$\begin{aligned}&\ln p(d, \boldsymbol{\mu}, W) = \cdots \\ &-\frac{1}{2}\mathrm{Tr}\left[\left(\sum_{i=1}^{N}(\boldsymbol{\mu}-\mathbf{x}_i)(\boldsymbol{\mu}-\mathbf{x}_i)^{\mathrm{T}} + \alpha_\mu(\boldsymbol{\mu}-\boldsymbol{\nu})(\boldsymbol{\mu}-\boldsymbol{\nu})^{\mathrm{T}} + T\right)W\right]\end{aligned} \tag{A.10}$$

Define

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\,, \qquad S_N = \sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathrm{T}} \tag{A.11}$$

so that $S_N$ is the sample variance multiplied by $(N-1)$. By expanding and comparing terms, one has

$$\sum_{i=1}^{N}(\boldsymbol{\mu}-\mathbf{x}_i)(\boldsymbol{\mu}-\mathbf{x}_i)^{\mathrm{T}} = S_N + N(\boldsymbol{\mu}-\bar{\mathbf{x}})(\boldsymbol{\mu}-\bar{\mathbf{x}})^{\mathrm{T}} \tag{A.12}$$

and also

$$\begin{aligned}&N(\boldsymbol{\mu}-\bar{\mathbf{x}})(\boldsymbol{\mu}-\bar{\mathbf{x}})^{\mathrm{T}} + \alpha_\mu(\boldsymbol{\mu}-\boldsymbol{\nu})(\boldsymbol{\mu}-\boldsymbol{\nu})^{\mathrm{T}} \\ =\ &(N+\alpha_\mu)(\boldsymbol{\mu}-\boldsymbol{\nu}')(\boldsymbol{\mu}-\boldsymbol{\nu}')^{\mathrm{T}} + \frac{N\alpha_\mu}{(N+\alpha_\mu)}(\bar{\mathbf{x}}-\boldsymbol{\nu})(\bar{\mathbf{x}}-\boldsymbol{\nu})^{\mathrm{T}}\end{aligned} \tag{A.13}$$

with

$$\boldsymbol{\nu}' = \frac{N\bar{\mathbf{x}} + \alpha_\mu\boldsymbol{\nu}}{(N+\alpha_\mu)} \tag{A.14}$$

Further define

$$R = T + S_N + \frac{N\alpha_\mu}{(N+\alpha_\mu)}(\bar{\mathbf{x}}-\boldsymbol{\nu})(\bar{\mathbf{x}}-\boldsymbol{\nu})^{\mathrm{T}} \tag{A.15}$$

then we may rewrite the joint probability from (A.10) as

$$\ln p(d, \boldsymbol{\mu}, W) = \cdots - \frac{1}{2}\mathrm{Tr}\left[\left(R + (N+\alpha_\mu)(\boldsymbol{\mu}-\boldsymbol{\nu}')(\boldsymbol{\mu}-\boldsymbol{\nu}')^{\mathrm{T}}\right)W\right] \tag{A.16}$$

or explicitly

$$p(d, \boldsymbol{\mu}, W) = \frac{|W|^{\frac{N+\alpha_w-n}{2}}}{(2\pi)^{\frac{nN}{2}}Z(n, \alpha_\mu, T, \alpha_w)}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\nu}')^{\mathrm{T}}(N+\alpha_\mu)W(\boldsymbol{\mu}-\boldsymbol{\nu}')}\mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[RW]} \tag{A.17}$$

In order to obtain the marginal $P(d)$ the above expression needs to be integrated with respect to $\boldsymbol{\mu}, W$. Comparing to (A.8) we can see that the functional form inside the integral corresponds to that of a normal-Wishart distribution with the following transformations

$$\begin{aligned} \alpha_\mu &\rightarrow N+\alpha_\mu \\ \alpha_w &\rightarrow N+\alpha_w \\ \boldsymbol{\nu} &\rightarrow \boldsymbol{\nu}' \\ T &\rightarrow R \end{aligned} \tag{A.18}$$

Therefore its integral is given by the normalization constant, leading to

$$p(d) = \frac{Z(n, N+\alpha_\mu, R, N+\alpha_w)}{(2\pi)^{\frac{nN}{2}} Z(n, \alpha_\mu, T, \alpha_w)} = \left(\frac{\alpha_\mu}{N+\alpha_\mu}\right)^{\frac{n}{2}} \frac{\Gamma_n\left(\frac{N+\alpha_w}{2}\right)}{\pi^{\frac{nN}{2}}\Gamma_n\left(\frac{\alpha_w}{2}\right)} \frac{|T|^{\frac{\alpha_w}{2}}}{|R|^{\frac{N+\alpha_w}{2}}} \tag{A.19}$$

This formula agrees with both equation (24) of Heckerman and Geiger (1995) and equation (18) of Geiger and Heckerman (2002) in the case when the full data is considered rather than a node specific subset.

From the above it also follows that the posterior distribution of the parameters $\boldsymbol{\mu}, W$ remains normal-Wishart

$$p(\boldsymbol{\mu}, W|d) = \frac{|W|^{\frac{N+\alpha_w-n}{2}}}{Z(n, N+\alpha_\mu, R, N+\alpha_w)} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\nu}')^{\mathrm{T}}(N+\alpha_\mu)W(\boldsymbol{\mu}-\boldsymbol{\nu}')} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[RW]} \tag{A.20}$$

with parameters as in (A.18).

A.3. *Subsets.* Finally we need to perform the same steps but when we restrict to a subset $\mathbf{Y}$ of size $l$ of the $n$ coordinates of the data, $\mathbf{Y} \subseteq \mathbf{X}$. These subsets are the data corresponding to either the node being scored and its parents, or its parents alone, and the resulting terms are needed for the complete evaluation of the BGe score involving the marginal likelihood in (A.1). We form a mean vector $\boldsymbol{\mu}_{\mathbf{Y}}$ from the components of $\boldsymbol{\mu}$ that are in $\mathbf{Y}$ and similarly partition the matrix $W$. Denoting the complement of $\mathbf{Y}$ by $\tilde{\mathbf{Y}}$, and with the elements appropriately reordered

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{Y}} \\ \boldsymbol{\mu}_{\tilde{\mathbf{Y}}} \end{pmatrix}, \qquad W = \begin{pmatrix} W_{\mathbf{YY}} & W_{\mathbf{Y}\tilde{\mathbf{Y}}} \\ W_{\tilde{\mathbf{Y}}\mathbf{Y}} & W_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}} \end{pmatrix}, \qquad W_{\mathbf{Y}\tilde{\mathbf{Y}}} = W_{\tilde{\mathbf{Y}}\mathbf{Y}}^{\mathrm{T}} \tag{A.21}$$

Since the data $d$ is assumed to be normally distributed with mean $\boldsymbol{\mu}$ and precision matrix $W$, the subset of the data on the $\mathbf{Y}$ coordinates is then

normally distributed with mean $\boldsymbol{\mu}_{\mathbf{Y}}$ and precision matrix $W_{\mathbf{Y}}$ where

$$W_{\mathbf{Y}} = \left(\left(W^{-1}\right)_{\mathbf{YY}}\right)^{-1} = W_{\mathbf{YY}} - W_{\mathbf{Y}\tilde{\mathbf{Y}}}\left(W_{\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}}\right)^{-1} W_{\tilde{\mathbf{Y}}\mathbf{Y}} \tag{A.22}$$

Since for a multivariate normal we just need to keep the corresponding elements of the covariance matrix, we simply invert $W$, select the elements and then invert again as in (A.22).

Next we need the distribution of $\boldsymbol{\mu}_{\mathbf{Y}}$ given that $\boldsymbol{\mu}$ is normally distributed with mean $\boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_{\mathbf{Y}} \\ \boldsymbol{\nu}_{\tilde{\mathbf{Y}}} \end{pmatrix}$ and precision matrix $\alpha_\mu W$. Again transforming to the covariance matrix, selecting elements and transforming back leads to a normal distribution with mean $\boldsymbol{\nu}_{\mathbf{Y}}$ and precision matrix $\alpha_\mu W_{\mathbf{Y}}$. All we really need then is the distribution of $W_{\mathbf{Y}}$ given that $W$ follows a Wishart distribution with parametric matrix $T$ and degrees of freedom $\alpha_w$. For this we use Theorem 5.1.4 of Press (1982)

- If $W \sim \mathcal{W}_n(T^{-1}, \alpha_w)$ is Wishart distributed, then

$$W_{\mathbf{Y}} \sim \mathcal{W}_l\left(\left(T_{\mathbf{YY}}\right)^{-1}, \alpha_w - n + l\right) \tag{A.23}$$

  where the degrees of freedom has been reduced and we simply select the relevant entries from the parametric matrix.

This result is also included in Theorem 5 of Geiger and Heckerman (2002).

The prior distribution for the subset of the parameters is then

$$\begin{aligned} p(\boldsymbol{\mu}_{\mathbf{Y}}, W_{\mathbf{Y}}) &= \frac{|W_{\mathbf{Y}}|^{\frac{\alpha_w - n}{2}}}{Z(l, \alpha_\mu, T_{\mathbf{YY}}, \alpha_w - n + l)} \\ &\quad \cdot \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\nu}_{\mathbf{Y}})^{\mathrm{T}} \alpha_\mu W_{\mathbf{Y}} (\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\nu}_{\mathbf{Y}})} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[T_{\mathbf{YY}} W_{\mathbf{Y}}]} \end{aligned} \tag{A.24}$$

while the likelihood for the corresponding subset of data $d^{\mathbf{Y}}$ from (A.2) is

$$\begin{aligned} p(d^{\mathbf{Y}} | \boldsymbol{\mu}_{\mathbf{Y}}, W_{\mathbf{Y}}) &= \frac{|W_{\mathbf{Y}}|^{\frac{N}{2}}}{(2\pi)^{\frac{lN}{2}}} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left[\left(\sum_{i=1}^{N} (\boldsymbol{\mu}_{\mathbf{Y}} - \mathbf{x}_{\mathbf{Y}i})(\boldsymbol{\mu}_{\mathbf{Y}} - \mathbf{x}_{\mathbf{Y}i})^{\mathrm{T}}\right) W_{\mathbf{Y}}\right]} \\ &= \frac{|W_{\mathbf{Y}}|^{\frac{N}{2}}}{(2\pi)^{\frac{lN}{2}}} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}\left[\left((S_N)_{\mathbf{YY}} + N(\boldsymbol{\mu}_{\mathbf{Y}} - \bar{\mathbf{x}}_{\mathbf{Y}})(\boldsymbol{\mu}_{\mathbf{Y}} - \bar{\mathbf{x}}_{\mathbf{Y}})^{\mathrm{T}}\right) W_{\mathbf{Y}}\right]} \end{aligned} \tag{A.25}$$

reorganised using (A.12). We may follow the same steps as in section A.2 to obtain the joint probability

$$\begin{aligned} p(d^{\mathbf{Y}}, \boldsymbol{\mu}_{\mathbf{Y}}, W_{\mathbf{Y}}) &= \frac{|W_{\mathbf{Y}}|^{\frac{N + \alpha_w - n}{2}}}{(2\pi)^{\frac{lN}{2}} Z(l, \alpha_\mu, T_{\mathbf{YY}}, \alpha_w - n + l)} \\ &\quad \cdot \mathrm{e}^{-\frac{1}{2}\left(\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\nu}'_{\mathbf{Y}}\right)^{\mathrm{T}} (N + \alpha_\mu) W_{\mathbf{Y}} \left(\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\nu}'_{\mathbf{Y}}\right)} \mathrm{e}^{-\frac{1}{2}\mathrm{Tr}[R_{\mathbf{YY}} W_{\mathbf{Y}}]} \end{aligned} \tag{A.26}$$

where handily we may simply select the corresponding elements of the posterior parametric matrix $R$ in (A.15).

Finally we marginalise by integrating out $(\boldsymbol{\mu}_{\mathbf{Y}}, W_{\mathbf{Y}})$

$$\begin{aligned} p(d^{\mathbf{Y}}) &= \frac{Z(l, N+\alpha_\mu, R_{\mathbf{YY}}, N+\alpha_w-n+l)}{(2\pi)^{\frac{lN}{2}} Z(l, \alpha_\mu, T_{\mathbf{YY}}, \alpha_w-n+l)} \\ &= \left(\frac{\alpha_\mu}{N+\alpha_\mu}\right)^{\frac{l}{2}} \frac{\Gamma_l\left(\frac{N+\alpha_w-n+l}{2}\right)}{\pi^{\frac{lN}{2}}\Gamma_l\left(\frac{\alpha_w-n+l}{2}\right)} \frac{|T_{\mathbf{YY}}|^{\frac{\alpha_w-n+l}{2}}}{|R_{\mathbf{YY}}|^{\frac{N+\alpha_w-n+l}{2}}} \end{aligned} \tag{A.27}$$

This result differs from equation (18) of Geiger and Heckerman (2002) in how the matrix elements are selected and differs from equation (24) of Heckerman and Geiger (1995) in how the $\alpha_w$ parameter changes when looking at subsets. Both differences probably derive from how the arguments of the Wishart distribution in (A.23) are updated for subsets following from Theorem 5.1.4 of Press (1982) itemized above.

**B. Simplifying the score.** As can be seen in (A.1), the score for each node $i$ involves finding the probability of the data restricted to the node's parents compared to the data restricted to the parents and the node itself. Let $\mathbf{P}$ denote the parent set of size $p$ and $\mathbf{Q}$ the parent set plus the node, of size $p+1$. Since the difference in size is exactly 1, the ratio of scores involves similar matrix elements and prefactors which can be simplified. Starting with the prefactors:

$$\frac{p(d^{\mathbf{Q}})}{p(d^{\mathbf{P}})} = \left(\frac{\alpha_\mu}{N+\alpha_\mu}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{N+\alpha_w-n+p+1}{2}\right)}{\pi^{\frac{N}{2}}\Gamma\left(\frac{\alpha_w-n+p+1}{2}\right)} \frac{|T_{\mathbf{QQ}}|^{\frac{\alpha_w-n+p+1}{2}} |R_{\mathbf{PP}}|^{\frac{N+\alpha_w-n+p}{2}}}{|T_{\mathbf{PP}}|^{\frac{\alpha_w-n+p}{2}} |R_{\mathbf{QQ}}|^{\frac{N+\alpha_w-n+p+1}{2}}} \tag{B.1}$$

since the ratios of multivariate gamma functions each leave a single term. The term in front of the ratios of determinants need only be calculated once for each value of $p = 0, \ldots, n-1$ and stored for computational efficiency.

B.1. *Simplifying ratios of determinants.* The more expensive step is calculating the determinants. For the ratio however one can first partition the larger matrices

$$R_{\mathbf{QQ}} = \begin{pmatrix} a & \mathbf{b}^{\mathrm{T}} \\ \mathbf{b} & D \end{pmatrix}, \qquad D = R_{\mathbf{PP}} \tag{B.2}$$

so that

$$|R_{\mathbf{QQ}}| = a\left|D - \mathbf{b}a^{-1}\mathbf{b}^{\mathrm{T}}\right| = |D|\left(a - \mathbf{b}^{\mathrm{T}}D^{-1}\mathbf{b}\right) \tag{B.3}$$

and hence

$$\frac{|R_{\mathbf{QQ}}|^{\frac{N+\alpha_w-n+p+1}{2}}}{|R_{\mathbf{PP}}|^{\frac{N+\alpha_w-n+p}{2}}} = |D|^{\frac{1}{2}} \left(a - \mathbf{b}^{\mathrm{T}} D^{-1} \mathbf{b}\right)^{\frac{N+\alpha_w-n+p+1}{2}} \tag{B.4}$$

Due to the difference in powers, $|D|$ still needs to be calculated. Typically, determinants are determined via an LU decomposition, which for the symmetric matrices considered here reduces to the Cholesky decomposition. For a $p \times p$ matrix, this takes $\frac{p^3}{3}$ operations. Although the inverse can be found from the Cholesky decomposition (Krishnamoorthy and Menon, 2011) this is more expensive than taking a second $p \times p$ determinant in the first expression in (B.3). Only for small $p$ would this in turn be notably more efficient than simply calculating $|R_{\mathbf{QQ}}|$. In principle one could resort to Coppersmith and Winograd (1990) or faster (Williams, 2012) algorithms for matrix multiplication and inversion to obtain $D^{-1}$ with no overhead on calculating its determinant asymptotically. Here, luckily we can avoid such complications since from the Cholesky decomposition

$$D = LL^{\mathrm{T}}\,, \qquad \mathbf{b}^{\mathrm{T}} D^{-1} \mathbf{b} = \mathbf{c}^{\mathrm{T}} \mathbf{c}\,, \qquad \mathbf{c} = L^{-1} \mathbf{b} \tag{B.5}$$

we know $L$ and can solve $L\mathbf{c} = \mathbf{b}$ for $\mathbf{c}$ using back-substitution with $p^2$ operations. Using the block partitioning therefore speeds up the calculation of (B.4) by approximately a factor of two.

Using the block partitioning also means we completely avoid matrices when $p = 1$. When $p = 2$ it is much faster to use the explicit form for the determinant and inverse of a $2 \times 2$ matrix rather than decompose and back-solve. For $p = 3$ the Laplace expansion may provide the determinant more than twice as quickly as decomposition. In this case there should be a slight additional speed up from using the first expression in (B.3).

These simplifications provide further advantages above the general factor of two from the block partitioning, particularly for the sparse DAGs considered in Friedman and Koller (2003) and Grzegorczyk and Husmeier (2008). Since the scoring is likely to be the most expensive part of general MCMC schemes on DAGs (like the structure MCMC of Madigan and York, 1995) the efficiency gained here allows longer chains to be run.

B.2. *Diagonal prior parametric matrix.* A suggested choice for the prior parametric matrix $T$, if $\alpha_w > n + 1$, is to approximate the multivariate $t$-distribution of $\mathbf{x}$ by a normal (Geiger and Heckerman, 2002). Assuming independent normals, leads to

$$T = tI_n\,, \qquad t = \frac{\alpha_\mu(\alpha_w - n - 1)}{(\alpha_\mu + 1)} \tag{B.6}$$

being diagonal (the scaling is to match the covariance of both distributions). Of course, if $T$ is diagonal, then calculating determinants of its subblocks is straightforward. When the elements are equal as here then

$$\frac{|T_{\mathbf{QQ}}|^{\frac{\alpha_w-n+p+1}{2}}}{|T_{\mathbf{PP}}|^{\frac{\alpha_w-n+p}{2}}} = t^{\frac{\alpha_w-n+2p+1}{2}} \tag{B.7}$$

so that this term takes little computational time. Such a choice for $T$ therefore reduces the computation of the score by an additional factor of two.

B.3. *Bias in the score of Heckerman and Geiger (1995).* Finally we compare the terms in (B.1) here to those obtained from the marginal likelihoods in Heckerman and Geiger (1995) since that version is likely to have been used in earlier papers employing the BGe score. After simplifying the corresponding ratio of multivariate gamma functions, one obtains

$$\frac{p(d^{\mathbf{Q}})}{p(d^{\mathbf{P}})} = \left(\frac{\alpha_\mu}{N+\alpha_\mu}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{N+\alpha_w-p}{2}\right)}{\pi^{\frac{N}{2}}\Gamma\left(\frac{\alpha_w-p}{2}\right)} \frac{|T_{\mathbf{QQ}}|^{\frac{\alpha_w}{2}}|R_{\mathbf{PP}}|^{\frac{N+\alpha_w}{2}}}{|T_{\mathbf{PP}}|^{\frac{\alpha_w}{2}}|R_{\mathbf{QQ}}|^{\frac{N+\alpha_w}{2}}} \tag{B.8}$$

Aside from the difference in the powers of the determinants, the remaining ratio of gamma functions decreases with $p$ unlike the ratio in (B.1). To compare the difference in the prefactor better, we consider their ratio

$$g(p) = \frac{\Gamma\left(\frac{N+\alpha_w-n+p+1}{2}\right)}{\Gamma\left(\frac{\alpha_w-n+p+1}{2}\right)} \cdot \frac{\Gamma\left(\frac{\alpha_w-p}{2}\right)}{\Gamma\left(\frac{N+\alpha_w-p}{2}\right)} \tag{B.9}$$

For simplicity, assume that $n$ is odd so that $g=1$ for $p=\frac{n-1}{2}$. Around this central value

$$g(\tilde{p}) = \frac{\Gamma\left(\frac{2N+2\alpha_w-n+1}{4}+\frac{\tilde{p}}{2}\right)}{\Gamma\left(\frac{2N+2\alpha_w-n+1}{4}-\frac{\tilde{p}}{2}\right)} \cdot \frac{\Gamma\left(\frac{2\alpha_w-n+1}{4}-\frac{\tilde{p}}{2}\right)}{\Gamma\left(\frac{2\alpha_w-n+1}{4}+\frac{\tilde{p}}{2}\right)}, \qquad \tilde{p} = p - \frac{n-1}{2} \tag{B.10}$$

Using the recurrence relation for gamma functions

$$g(\tilde{p}) = \frac{\left(\frac{2N+2\alpha_w-n+1+2\tilde{p}}{4}-1\right)\cdots\left(\frac{2N+2\alpha_w-n+1+2\tilde{p}}{4}-\tilde{p}\right)}{\left(\frac{2\alpha_w-n+1+2\tilde{p}}{4}-1\right)\cdots\left(\frac{2\alpha_w-n+1+2\tilde{p}}{4}-\tilde{p}\right)} \tag{B.11}$$

we have a simple product of $\tilde{p}$ terms for $\tilde{p}=1,\ldots,\frac{n+1}{2}$, and the inverse for negative $\tilde{p}$. The exact behaviour of $g$ obviously depends on the dimension $n$,

the parameter $\alpha_w$ and the number of observations $N$, but if we assume the latter to be significantly larger than the others ($N \gg n, \alpha_w$) then

$$g(\tilde{p}) \sim N^{\tilde{p}}, \qquad g(p) \sim N^{p-\frac{n-1}{2}} \tag{B.12}$$

with the same behaviour for even $n$.

Practically this means that the score of Heckerman and Geiger (1995) penalises each node in a DAG with $p$ parents by roughly $N^p$ compared to the score in (B.1). Sparse DAGs with few parents are then artificially favoured. Even without assuming $N$ to be much larger than $n$ and $\alpha_w$, the ratio $g(\tilde{p})$ still increases with $\tilde{p}$ and a bias remains towards sparse DAGs.

## References.


Coppersmith, D. and Winograd, S. (1990). Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation* **9** 251–280.

Friedman, N. and Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50** 95–125.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Tenth Conference on Uncertainty in Artificial Intelligence* 235–243.

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

Grzegorczyk, M. and Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* **71** 265–305.

Heckerman, D. and Geiger, D. (1995). Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Eleventh Conference on Uncertainty in Artificial Intelligence* 274–284.

Krishnamoorthy, A. and Menon, D. (2011). Matrix inversion using Cholesky decomposition. Preprint, arXiv:1111.4144.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63** 215-232.

Press, S. J. (1982). *Applied multivariate analysis*. Krieger publishing company.

Williams, V. V. (2012). Multiplying matrices faster than Coppersmith-Winograd. In *Forty-fourth annual ACM symposium on Theory of Computing* 887–898.



Jack Kuipers
Institut für Theoretische Physik
Universität Regensburg
D-93040 Regensburg, Germany
E-mail: jack.kuipers@ur.de

Giusi Moffa
Institut für funktionelle Genomik
Universität Regensburg
Josef Engertstrasse 9
93053 Regensburg, Germany
E-mail: giusi.moffa@ukr.de

David Heckerman
Microsoft Research
1100 Glendon Ave Suite PH1
Los Angeles CA 90024
E-mail: heckerma@microsoft.com