The Annals of Statistics
2014, Vol. 42, No. 4, 1689–1691
DOI: 10.1214/14-AOS1217

ADDENDUM ON THE SCORING OF GAUSSIAN DIRECTED ACYCLIC GRAPHICAL MODELS

BY JACK KUIPERS, GIUSI MOFFA AND DAVID HECKERMAN

Regensburg University, Regensburg University and Microsoft Research

We provide a correction to the expression for scoring Gaussian directed acyclic graphical models derived in Geiger and Heckerman [*Ann. Statist.* **30** (2002) 1414–1440] and discuss how to evaluate the score efficiently.

Gaussian directed acyclic graph (DAG) models represent a particular type of Bayesian networks where the node variables are assumed to come from a multivariate Gaussian distribution. The Bayesian Gaussian equivalent (BGe) score was introduced in Geiger and Heckerman (1994, 2002), Heckerman and Geiger (1995) for learning such networks.

For brevity, we omit formal definitions and refer the reader to Geiger and Heckerman (2002), while following their notation in considering DAG models m with n nodes corresponding to the set of variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. Let m^h be the model hypothesis that the true distribution of $\mathbf{X}$ is faithful to the DAG model m, meaning that it satisfies only and all the conditional independencies encoded by the DAG. For a complete random data sample $d = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with N observations and a complete DAG model m_c, the marginal likelihood is [Geiger and Heckerman (2002), Theorem 2]

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

where $\mathbf{Pa}_i$ are the parent variables of the vertex i and $d^{\mathbf{Y}}$ is the data restricted to the coordinates in $\mathbf{Y} \subseteq \mathbf{X}$. The BGe score is the posterior probability of m^h which is proportional to the marginal likelihood in (1) and the graphical prior; see equation (2) of Geiger and Heckerman (2002).

Different DAGs which encode the same set of conditional independencies are said to belong to an equivalence class. Along with ensuring that all DAGs in the same equivalence class are scored equally, the modularity of the score allows the steps in structure MCMC [Madigan and York (1995)] to be evaluated much more efficiently. Order MCMC [Friedman and Koller (2003), on the related space of triangular matrices] as well as the edge reversal move of Grzegorczyk and Husmeier (2008) would not be possible without it.

Received February 2014.
MSC2010 subject classifications. 62-07, 62F15, 62H99.
Key words and phrases. Gaussian DAG models, Bayesian network learning, BGe score.

For Gaussian DAG models, the likelihood is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and precision matrix W. The need for global parameter independence, so that the expression of the score in (1) holds, implies that the prior distribution of $(\boldsymbol{\mu}, W)$ must be normal-Wishart [Geiger and Heckerman (2002)]. The parameter $\boldsymbol{\mu}$ is taken to be normally distributed with mean $\boldsymbol{\nu}$ and precision matrix $\alpha_\mu W$, for $\alpha_\mu > 0$. W is Wishart distributed with positive definite parametric matrix T (the inverse of the scale matrix) and degrees of freedom α_w, with $\alpha_w > n - 1$. As detailed in the supplementary material [Kuipers, Moffa and Heckerman (2014)], one finds

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

where l is the size of $\mathbf{Y}$, $A_{\mathbf{YY}}$ means selecting the rows and columns corresponding to $\mathbf{Y}$ of a matrix A,

$$
\Gamma_l\left(\frac{x}{2}\right) = \pi^{l(l-1)/4} \prod_{j=1}^{l} \Gamma\left(\frac{x+1-j}{2}\right) \tag{3}
$$

is the multivariate Gamma function and

$$
R = T + S_N + \frac{N\alpha_\mu}{(N+\alpha_\mu)}(\boldsymbol{\nu} - \bar{\mathbf{x}})(\boldsymbol{\nu} - \bar{\mathbf{x}})^{\mathrm{T}} \tag{4}
$$

is the posterior parametric matrix involving

$$
\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \qquad S_N = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \tag{5}
$$

the sample mean and sample variance multiplied by $(N-1)$.

The result in (2) is identical to equation (18) of Geiger and Heckerman (2002), once some factors are cancelled, apart from the manner in which the matrix elements are chosen. The result in Geiger and Heckerman (2002) replaces the $T_{\mathbf{YY}}$ and $R_{\mathbf{YY}}$ by $T_{\mathbf{Y}}$ and $R_{\mathbf{Y}}$, where $A_{\mathbf{Y}} = ((A^{-1})_{\mathbf{YY}})^{-1}$. Inverting the matrices before the elements are selected and then inverting again [as in Geiger and Heckerman (2002)] we found inconsistent behavior on simulated data.

We may further compare to equation (24) of Heckerman and Geiger (1995), which with the current notation becomes

$$
p(d^{\mathbf{Y}} \mid m_c^h) = \left(\frac{\alpha_\mu}{N+\alpha_\mu}\right)^{l/2} \frac{\Gamma_l((N+\alpha_w)/2)}{\pi^{lN/2}\Gamma_l(\alpha_w/2)} \frac{|T_{\mathbf{YY}}|^{\alpha_w/2}}{|R_{\mathbf{YY}}|^{(N+\alpha_w)/2}} \tag{6}
$$

while incorrectly defining the S_N in the R in (4) as the sample variance. However, the same terminology, with the correct formula for S_N, is used in Geiger and Heckerman (1994) whose equation (15) is otherwise identical to (6) aside from having π replaced by 2π.

The difference in the powers of the determinants between (2) and (6) could lead to a subtle, and hard to predict, change in the scores. There is also the same loss of l-dependence in the arguments of the multivariate gamma functions. The ratio of gamma functions for each node now actually decreases with l while the ratio from (2) increases instead. As discussed in the supplementary material [Kuipers, Moffa and Heckerman (2014)], using (6) instead of (2) effectively penalises each node with l parents by a factor $\sim N^l$, giving a substantial bias toward sparse DAGs. This bias is likely to be present in early works implementing the score of Heckerman and Geiger (1995) and possibly remains in legacy code.

SUPPLEMENTARY MATERIAL

Deriving and simplifying the BGe score (DOI: 10.1214/14-AOS1217SUPP; .pdf). We detail the steps used to derive (2) and simplify the ratios appearing in (1) to improve the numerical computation of the score.

REFERENCES

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 *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence* 235–243. Morgan Kaufmann, San Francisco, CA.

GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324

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 *Uncertainty in Artificial Intelligence* (*Montreal*, *PQ*, 1995) 274–284. Morgan Kaufmann, San Francisco, CA. MR1615024

KUIPERS, J., MOFFA, G. and HECKERMAN, D. (2014). Supplement to "Addendum on the scoring of Gaussian directed acyclic graphical models." DOI:10.1214/14-AOS1217SUPP.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63** 215–232.

J. KUIPERS
INSTITUT FÜR THEORETISCHE PHYSIK
UNIVERSITÄT REGENSBURG
D-93040 REGENSBURG
GERMANY
E-MAIL: jack.kuipers@ur.de

G. MOFFA
INSTITUT FÜR FUNKTIONELLE GENOMIK
UNIVERSITÄT REGENSBURG
JOSEF ENGERTSTRASSE 9
93053 REGENSBURG
GERMANY
E-MAIL: giusi.moffa@ukr.de

D. HECKERMAN
MICROSOFT RESEARCH
1100 GLENDON AVE SUITE PH1
LOS ANGELES, CALIFORNIA 90024
USA
E-MAIL: heckerma@microsoft.com