

ACADEMIC  
PRESSAvailable at  
**WWW.MATHEMATICSWEB.ORG**  
POWERED BY SCIENCE @ DIRECT®

Journal of Multivariate Analysis 85 (2003) 253–266

Journal of  
Multivariate  
Analysis<http://www.elsevier.com/locate/jmva>

# Bayesian graphical model determination using decision theory

Jukka Corander<sup>1</sup>*Rolf Nevanlinna Institute, P.O. Box 4, University of Helsinki, FIN-00014, Finland*

Received 29 December 1999

---

## Abstract

Bayesian model determination in the complete class of graphical models is considered using a decision theoretic framework within the regular exponential family. The complete class contains both decomposable and non-decomposable graphical models. A utility measure based on a logarithmic score function is introduced under reference priors for the model parameters. The logarithmic utility of a model is decomposed into predictive performance and relative complexity. Axioms of decision theory lead to the judgement of the plausibility of a model in terms of the posterior expected utility. This quantity has an analytic expression for decomposable models when certain reference priors are used and the exponential family is closed under marginalization. For non-decomposable models, a simulation consistent estimate of the expectation can be obtained. Both real and simulated data sets are used to illustrate the introduced methodology.

© 2003 Elsevier Science (USA). All rights reserved.

*AMS 1991 subject classifications:* 62H99; 62H17

*Keywords:* Bayesian model determination; Entropy; Exponential family; Graphical models; Multinomial distribution; Multinomial distribution; Reference analysis; Utility

---

## 1. Introduction

During the past two decades, numerous researchers have shown the value and versatility of graphical models in multivariate analysis; for a comprehensive list of

---

*E-mail address:* [jukka.corander@rni.helsinki.fi](mailto:jukka.corander@rni.helsinki.fi).

<sup>1</sup>The author thanks Mattias Villani and the two anonymous referees whose comments and suggestions led to a substantial improvement of the original manuscript. This research has been supported in part by Academy of Finland, Grant 50203.

references, see [17] or [29]. However, only recently, sound methods for graphical model determination, dealing simultaneously with the uncertainty involved in the dependence structure and the parameters, have been discussed in the statistical literature. After a landmark paper by Dawid and Lauritzen [8], where the general principles for Bayesian analysis of decomposable graphical models were developed, Bayesian strategies for model determination have been considered by Madigan and Raftery [19], Madigan and York [20], Dellaportas and Forster [9], Giudici and Green [11], and Giudici et al. [12].

Model determination strategies have mostly been considered within the class of decomposable graphical models, since the simple structure of such models reduces considerably the complexity of the model determination task. The exception is the paper by Dellaportas and Forster [9] where also non-decomposable models were considered for multinomial data. Here, we develop a method for model determination in the complete class of graphical models, using a decision theoretic framework within the regular exponential family. In practical application of the proposed method we focus on multinomial and multinormal distributions.

Numerous authors have advocated the view of statistical inference as a special case of decision theory, one of the most impressive works on this area being the book by Bernardo and Smith [4]. Advantages of such an approach to solving the problem of learning about graphical model structure given the observed data are: (i) it provides a formal framework which necessitates only a manageable amount of effort in specifying a priori information, (ii) it facilitates a convenient communication of the results of learning, and (iii) it guarantees a coherent solution regardless of the amount of information available in the data.

In a decision theoretic formulation the plausibility of a graphical model in the light of data is measured in terms of its posterior expected utility. We propose the use of a utility function which is a proper score function of a logarithmic form, for a detailed derivation of these concepts, see [4]. The expected utility “scores” a graphical model for its ability in predicting data, while weighting the ability against the complexity of the model, under the current uncertainty about the parameters.

Inspired in particular by the work of Bernardo [3], we enter into a criterion for graphical model determination, which can be used both in the absence and presence of a priori information concerning model parameters. We concentrate on the former case and derive the criterion under *reference* priors. For a general discussion on such priors, see [15].

There has been a considerable interest in the statistical community to define Bayesian methods for model determination which necessitate as little effort as possible by the potential user. See, for instance, [2,16,23,26]. Apart from the methods based on asymptotic approximations, such that the one by Schwarz [26], the common feature of the “automatic Bayesian” model determination methods is that their application to complex problems, such that the one considered in the current paper, is seldom described or even discussed.

Methods relying on asymptotic approximations are usually of the “penalized maximum likelihood” type, with various forms of penalty functions, and it is widely recognized that they have poor performance for small data sets. Here it will be

shown that the penalized maximum likelihood methods can be viewed as asymptotic approximations to the currently considered posterior expected utility, similarly as Smith and Spiegelhalter [27] showed that they can be viewed in the framework of Bayes factors (for the definition of Bayes factor, see [14]).

The present paper is organized as follows. The general concepts of graphical models are presented in Section 2. Model determination within the decision theoretic framework will be discussed in Section 3. Application of the introduced methodology to multinomial and multinormal data is considered in Sections 4 and 5, respectively. Some concluding remarks are given in Section 6.

## 2. Graphical models

We consider graphical modeling of distributions within the regular exponential family, for which a detailed theory is developed in [1,5].

Consider a finite set  $\Delta$  of  $k$  stochastic variables with a joint sample space  $\mathcal{X}$ , which is either a discrete set  $\mathcal{S}_\Delta$  or  $\mathbb{R}^k$ . The joint distribution of  $\Delta$  is characterized by a parameter  $\theta \in \Theta$  and a statistic  $\mathbf{t}(x), x \in \mathcal{X}$ , such that the inner product  $\langle \theta, \mathbf{t}(x) \rangle \in \mathbb{R}$ . The density of  $x$  with respect to a  $\sigma$ -finite measure  $\mu$  on  $\mathcal{X}$ , is assumed to be

$$f(x|\theta) = \exp\{\langle \theta, \mathbf{t}(x) \rangle - \kappa(\theta)\} \tag{2.1}$$

so that the probabilistic behavior of the variables  $\Delta$  is described by an exponential model. It is assumed that (2.1) is strictly positive for all  $x \in \mathcal{X}$ . For a subset  $a \subset \Delta$ , the marginal density of  $x_a \in \mathcal{X}_a$  is denoted by  $f_a(x_a|\theta_a)$ , where  $\theta_a$  is a subvector of  $\theta$ .

A generalization of the considered situation would allow the set  $\Delta$  to be partitioned into variables of discrete and continuous type. Within the regular exponential representation (2.1) there is then the possibility of using CG distributions introduced by Lauritzen and Wermuth [18]. However, the CG family does not obey closure under marginalization, which leads to technical difficulties and necessitates substantial further work in constructing a practically applicable method for model determination.

Let  $\mathcal{G}$  be the class of undirected graphical models, where each  $G \in \mathcal{G}$  restricts  $\theta$  to take values in an affine subspace  $\Theta_G$  of  $\Theta$ . The terms graphical model and graph are used interchangeably in the sequel. For further details and explanations concerning graphical models, see [30] or [17].

For any particular value  $\theta \in \Theta$ , the projection  $\theta(G) = \text{proj}(\theta|G)$  of  $\theta$  onto  $\Theta_G$  is defined as the unique parameter value, which minimizes the Kullback–Leibler divergence

$$D(\theta, \theta(G)) = \int f(x|\theta) \log \frac{f(x|\theta)}{f(x|\theta(G))} \mu(dx) \tag{2.2}$$

from  $f(x|\theta)$  to  $f(x|\theta(G))$ . The *entropy* of  $f(\cdot|\theta)$  is

$$h(f(\cdot|\theta)) = - \int f(x|\theta) \log f(x|\theta) \mu(dx). \tag{2.3}$$

Let  $\mathcal{C}(G)$  denote the set of *cliques* of a graph  $G$ . When  $G$  is *decomposable* the *separators*  $\mathcal{S}$  of the cliques can be obtained by a sequence of decompositions of  $G$  into  $\mathcal{C}(G)$ , see [17]. For such graphs the density  $f(x|\boldsymbol{\theta}(G))$  can be written as

$$f(x|\boldsymbol{\theta}(G)) = \frac{\prod_{c \in \mathcal{C}(G)} f_c(x_c|\boldsymbol{\theta}_c)}{\prod_{s \in \mathcal{S}(G)} f_s(x_s|\boldsymbol{\theta}_s)}. \quad (2.4)$$

For a decomposable model, divergence (2.2) can be written in terms of entropies of the marginal distributions, as

$$\sum_{c \in \mathcal{C}(G)} h(f_c(\cdot|\boldsymbol{\theta}_c)) - \sum_{s \in \mathcal{S}(G)} h(f_s(\cdot|\boldsymbol{\theta}_s)) - h(f(\cdot|\boldsymbol{\theta})). \quad (2.5)$$

The standard condition for the decomposability of a graph  $G$  is that the graph should not contain any chordless cycles of length four or larger. When  $G$  is non-decomposable, the density  $f(x|\boldsymbol{\theta}(G))$  cannot be directly presented in terms of marginal densities as in (2.4), and the projection of  $\boldsymbol{\theta}$  cannot be expressed in a closed form, but it has to be found by some iterative method. However, there is a possibility to utilize the concept of decomposability to represent the affine restrictions imposed by a non-decomposable model, which is done in the following definition (originally introduced in [6]). Essentially the same idea was discovered by Rudas [25], and used for the maximum likelihood estimation of graphical log-linear models.

**Definition 2.1.** For any two graphical models  $G, G'$  in  $\mathcal{G}$ , we let  $G \subset G'$ , if the edges present in  $G$  are also present in  $G'$ . We say that  $G'$  is a supermodel of  $G$ . For any non-decomposable model  $G$ , let  $\mathcal{A}_G$  denote the class of minimal decomposable supermodels, consisting of all  $G'$  for which the following three conditions hold:

- (1)  $G'$  is decomposable,
- (2)  $G \subset G'$ ,
- (3) there does not exist any decomposable  $G''$ , such that  $G \subset G'' \subset G'$ .

The affine restrictions to  $\boldsymbol{\theta}$  imposed by a non-decomposable  $G$  are such that the density  $f(x|\boldsymbol{\theta}(G))$  satisfies factorization (2.4) *simultaneously* for all  $G' \in \mathcal{A}_G$ . Assume  $\mathcal{A}_G$  contains the  $m$  graphs  $G_1, \dots, G_m$ . It follows from the general results of Csiszar [7], that the projection  $\boldsymbol{\theta}(G)$  can be obtained as a limit of a cyclical projection  $\boldsymbol{\theta}(G_1), \boldsymbol{\theta}(G_1)(G_2), \dots, \boldsymbol{\theta}(G_{m-1})(G_m), \boldsymbol{\theta}(G_m)(G_1), \dots$  to the subspaces  $\Theta_{G_1}, \dots, \Theta_{G_m}$ . As discussed in [25], this representation in terms of decomposable models has advantages over alternative, rather standard projection methods, such that those described in [28,30]. The above cyclical projection method facilitates both a simpler coding of the problem and a faster convergence to the parameter value minimizing (2.2).

### 3. Graphical model determination using decision theory

Let  $\mathbf{x}$  denote a set of  $n$  exchangeable observations  $x_1, \dots, x_n$  whose probabilistic behavior is governed by (2.1). It is desired to determine the degrees of plausibility of the elements of  $\mathcal{G}$  under the current uncertainty about  $\theta$ . The stated problem can formally be described as a decision problem, where the action to be taken is the choice of a model from  $\mathcal{G}$ . As discussed in [3,4], under quantitative coherence, we then need to specify: (i) the utility function  $\bar{u}(G)$ , measuring the desirability of choosing  $G$ , as a function of the plausibility of the parameter values  $\theta \in \Theta_G$ ; (ii) the prior distribution  $\pi(\theta)$  of  $\theta$ ; (iii) the model  $G$  which maximizes the expected posterior utility  $\bar{u}(G|\mathbf{x})$ .

To specify the utility structure for the stated decision problem, we use the general utility theory considered in [4]. A logarithmic utility function measures the appropriateness of an approximation to the density  $f(x|\theta)$  in terms of

$$\alpha \log f(x) + \beta(x), \quad x \in \mathcal{X}, \tag{3.1}$$

where  $\alpha > 0$  and  $\beta(\cdot)$  is an arbitrary real-valued function. Let  $c(G)$  denote a cost function representing the relative complexity of  $G$ . Since the data is assumed to be generated from  $f(x|\theta)$ , the expected logarithmic utility of  $G$  before the data are actually observed, is defined as

$$\bar{u}(G) = \alpha n \int f(x|\theta) \log f(x|\theta(G)) \mu(dx) + n \int \beta(x) f(x|\theta) \mu(dx) - c(G), \tag{3.2}$$

where the first part is proportional to the negative entropy of  $f(\cdot|\theta)$  minus the negative Kullback–Leibler divergence from  $f(x|\theta)$  to  $f(x|\theta(G))$ .

The maximization of the logarithmic utility does not depend on the actual form of the function  $\beta(\cdot)$ , since the posterior expectation of the term  $n \int \beta(x) f(x|\theta) \mu(dx)$  does not depend on  $G$ . Hence, the optimal decision is to choose the  $G$  which maximizes the posterior expected logarithmic utility

$$\bar{u}(G|\mathbf{x}) = \alpha n \int \left[ \int f(x|\theta) \log f(x|\theta(G)) \mu(dx) \right] \pi(\theta|\mathbf{x}) d\theta - c(G). \tag{3.3}$$

The decision criterion thus involves the negative *posterior expected entropy* of the distribution with density  $f(\cdot|\theta(G))$ .

When  $\mathcal{G}$  is large, it may be impractical or not feasible to calculate the expected utilities for all models. In such situations, the model determination may be carried out by an efficient heuristic search algorithm, such as the one considered in [19]. In the search among models, and in communicating the model determination results, it is useful to consider *relative* expected logarithmic utilities on an exponential scale

$$\frac{\exp\{\bar{u}(G|\mathbf{x})\}}{\sum_{G \in \mathcal{G}} \exp\{\bar{u}(G|\mathbf{x})\}}, \tag{3.4}$$

where  $\mathcal{G}$  can be replaced by a subclass of models under consideration. The relative utilities also facilitate a weighting of the models in the light of data, which is useful for prediction purposes. The positive effect of taking into account several models by their posterior weights in prediction has been illustrated, for instance, in [19].

In contrast to the Markov chain Monte Carlo (MCMC) model determination methods considered in [9,11,12,20], models are visited only once in the search algorithm of Madigan and Raftery [19]. In an MCMC approach the plausibility of a model is measured by the number times that particular model is visited in a Markov chain over the model space. This is a potential problem, since when  $\mathcal{G}$  is large, the Markov chain may remain for long periods in non-representative parts of  $\mathcal{G}$ , making convergence to the target distribution slow. On the contrary, the algorithm of Madigan and Raftery [19] should be capable of finding efficiently the relevant part of the model space, even for large  $\mathcal{G}$ .

Let  $G^*$  denote the model for which  $\Theta_{G^*} = \Theta$ . The difference between the expected utilities  $\bar{u}(G^*|\mathbf{x})$  and  $\bar{u}(G|\mathbf{x})$  is

$$\alpha n \int \left[ \int f(x|\theta) \log \frac{f(x|\theta)}{f(x|\theta(G))} \mu(dx) \right] \pi(\theta|\mathbf{x}) d\theta - (c(G^*) - c(G)), \quad (3.5)$$

where the first part is the expected log-likelihood ratio under  $f(\cdot|\theta)$ . As noted by Bernardo [3], this quantity measures the amount of information about future observations which would be needed to recover  $G^*$  from  $G$ . Further, the utility constant  $\alpha$  may be interpreted as the value of one unit of information about data generated from  $f(x|\theta)$ .

Recalling the definition of the density of  $x$  according to a decomposable graph, we see that for such a model, the expected logarithmic utility  $\bar{u}(G|\mathbf{x})$  involves only expectations of entropies of marginal distributions.

$$\bar{u}(G|\mathbf{x}) = \alpha n \left[ \sum_{c \in \mathcal{C}(G)} h(f_c(\cdot|\theta_c)) - \sum_{s \in \mathcal{S}(G)} h(f_s(\cdot|\theta_s)) \right] \pi(\theta|\mathbf{x}) d\theta - c(G). \quad (3.6)$$

It will be shown in the subsequent sections that these expected entropies have an analytic expression under a reference prior. For non-decomposable models, the expected value of  $\int f(x|\theta) \log f(x|\theta(G)) \mu(dx)$  has to be calculated by simulating values from  $\pi(\theta|\mathbf{x})$  and projecting each of these into  $\Theta_G$  to obtain a Monte Carlo approximation. The computational burden due to this task might at first sight appear prohibitive. Note, however, that when the expectation needs to be calculated for several models, the same set of simulated values from  $\pi(\theta|\mathbf{x})$  can be used for all models. Furthermore, it is necessary to calculate the Monte Carlo approximation only for the part of  $G$  involving chordless cycles of length four or larger. In large graphs this is a real advantage, since a major part of the graph may be decomposable.

The computational burden may also be reduced by performing the search in  $\mathcal{G}$  at two stages; at first among the decomposable models, and then, given the graphs with the highest utilities, among the non-decomposable ones. Such a procedure will reduce considerably the number of non-decomposable models that need to be investigated.

Comparison of the expected utilities of different models is particularly simple for certain decomposable models. Consider the difference between  $\bar{u}(G|\mathbf{x})$  and  $\bar{u}(G'|\mathbf{x})$  where  $G' \subset G$ . If both these models are decomposable, and  $G'$  is obtained from  $G$  by

removing a single edge  $\{\delta, \delta'\}$ , the utility difference will simplify considerably. Note that the edge  $\{\delta, \delta'\}$  is contained in a single clique  $c$  of  $G$ , and let  $c_\delta = c \setminus \{\delta'\}$ ,  $c_{\delta'} = c \setminus \{\delta\}$  and  $c_0 = c \setminus \{\delta, \delta'\}$ . Then, the utility difference equals

$$\alpha n \int [h(f_{c_\delta}(\cdot | \theta_{c_\delta})) + h(f_{c_{\delta'}}(\cdot | \theta_{c_{\delta'}})) - h(f_{c_0}(\cdot | \theta_{c_0})) - h(f_c(\cdot | \theta_c))] \pi(\theta | \mathbf{x}) d\theta - (c(G) - c(G')). \tag{3.7}$$

We now consider the choice of  $\alpha$  and  $c(G)$ , which, together with the choice of prior  $\pi(\theta)$  for  $\theta$ , are crucial for the practical implementation of the model determination procedure. An obvious requirement to be met by a model determination procedure in the current context is that of asymptotic consistency, which means that as the sample size approaches infinity, “the true model” among those in  $\mathcal{G}$  will be chosen with probability one. Although the truth of a model is merely a mathematical artifact, such a formulation provides valuable guidance in specifying a procedure which is applicable in practice. Notice that all models in  $\mathcal{G}$  with at least one missing edge are nested in the model corresponding to the complete graph, which by assumption gives the data generating mechanism. Therefore, the true model may be taken to correspond to the simplest possible graph  $G$  for which  $\theta \in \Theta_G$ .

Let  $q(G)$  denote the number of non-restricted elements in  $\theta \in \Theta_G$ , and let  $n \rightarrow \infty$ , so that the posterior distribution approaches a Dirac spike at the value  $\hat{\theta}$  maximizing the log-likelihood  $l(\mathbf{x} | \theta)$  over  $\Theta$ . At this limit, the expected logarithmic utility reduces to

$$\alpha l(\mathbf{x} | \hat{\theta}(G)) - c(G). \tag{3.8}$$

It is clear from (3.3) that the choice of  $\alpha$  does not affect the order of the models with respect to  $\bar{u}(G | \mathbf{x})$ , since  $\alpha$  is constant over  $\mathcal{G}$ . On the other hand, for the simplicity of interpretation of the relative utilities (3.4), one may set  $\alpha = 1$ . Now, any choice of  $c(G)$  which involves  $q(G)$ , and not  $n$ , will lead (asymptotically) to a non-consistent Akaike-type penalized maximum likelihood criterion. Whereas, if  $c(G)$  is set equal to  $q(G)(\log n)/2$ , the logarithmic utility will tend to the Schwarz criterion [26], which is consistent in the present context. However, since the expected utility will be lower than  $l(\mathbf{x} | \hat{\theta}(G))$  for finite samples, the cost  $q(G)(\log n)/2$  will produce even more conservative results than the Schwarz criterion.

One possible solution to the problem of specifying the cost  $c(G)$  is to seek for the least conservative asymptotically consistent criterion. Such a criterion, leading to  $c(G) = q(G) \log \log n$ , was introduced in [13], for the determination of the order of autoregression. Heuristics suggest that their result holds more generally in the exponential family, as does that of Schwarz [26], but this is sincerely difficult to demonstrate rigorously.

Although the asymptotic behavior of  $\bar{u}(G | \mathbf{x})$  is used to provide insight into the problem of choosing  $c(G)$ , it should be noted that  $\bar{u}(G | \mathbf{x})$  is not derived as an approximation to any other quantity. The choice of  $c(G)$  corresponds to the choice of the degree of uncertainty in a prior distribution to be used for the calculation of the marginal likelihood of a model. The marginal likelihood is defined as the

expectation of the likelihood with respect to prior  $\int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ , and it is the standard Bayesian model determination criterion used in all papers concerning graphical model determination listed in the first section.

A disadvantage of approximations of type (3.8) is that they are linear with respect to the model complexity for a fixed  $n$ . When  $n$  is small, these approximations tend to breakdown since they do not take into account the curvature in the likelihood around  $\hat{\boldsymbol{\theta}}(G)$ . In this respect  $\bar{u}(G|\mathbf{x})$  behaves as a marginal likelihood, since the expectation of the logarithmic utility with respect to posterior penalizes a model for increased complexity in a non-linear manner.

As noted in [3], in scientific applications of statistical inference there is usually a pragmatic need for a model-based prior, which has a minimal effect to the posterior inference relative to the data. Priors with such a desired property are often called *reference* or non-informative, reflecting the fact they are intended to be utilized without further subjective assessment of prior hyper parameters. This is a particularly important point, since it can be extremely difficult to specify reasonable subjective priors when the number of variables is large, see the discussion in [11].

When the expected utility  $\bar{u}(G|\mathbf{x})$  is based on a reference prior, and on the choices  $c(G) = q(G) \log \log n$ ,  $\alpha = 1$ , we define it as the *reference criterion* for graphical model determination. To summarize the advantages of this criterion: (i) it provides a unified approach to graphical model determination in the complete class of graphical models, (ii) it is consistent while having desirable small sample properties, (iii) it does not require further setting of prior hyperparameters, and (iv) it facilitates a simple communication of the model determination results.

#### 4. Multinomial case

We now consider graphical model determination for multinomial data. Let the finite set  $\mathcal{I}_\delta$  index the possible outcomes of  $\delta \in \Delta$ . On the set  $\mathcal{I}_\Delta = \times_{\delta \in \Delta} \mathcal{I}_\delta$  we assume a multinomial distribution with the probabilities  $p_\Delta$ . For any  $a \subset \Delta$ , the corresponding marginal distribution with probabilities  $p_a$  is defined on  $\mathcal{I}_a = \times_{\delta \in a} \mathcal{I}_\delta$ . We write  $n_\Delta = (n(x) = \sum_{i=1}^n I(x_i = x), x \in \mathcal{I}_\Delta)$  for the observed counts of the different outcomes.

Let  $\lambda_\Delta$  be a vector of constants  $(\lambda(x), x \in \mathcal{I}_\Delta)$ . Assuming the prior distribution for  $p_\Delta$  is the Dirichlet  $(\lambda_\Delta)$  distribution, the corresponding posterior is Dirichlet  $(\lambda_\Delta + n_\Delta)$ . If the Jeffreys' *reference* prior is used, then  $\lambda(x) = 1/2, x \in \mathcal{I}_\Delta$ . However, as discussed in [9], for a large  $\mathcal{I}_\Delta$  this choice leads to marginal priors that are not vague. A more reasonable, symmetric prior is obtained by setting  $\lambda(x)$  equal to  $1/|\mathcal{I}_\Delta|$  for all  $x \in \mathcal{I}_\Delta$ . This prior was originally suggested in [24]. Given the Dirichlet-posterior for  $p_\Delta$ , marginal posteriors for any subset  $a \subset \Delta$  are straightforward to obtain.

For decomposable graphical models, the expected logarithmic utility involves posterior expectations of the entropies  $\{h(f_c(\cdot|\boldsymbol{\theta}_c)), h(f_s(\cdot|\boldsymbol{\theta}_s)), c \in \mathcal{C}(G), s \in \mathcal{S}(G)\}$ . When the posterior is Dirichlet  $(\lambda_\Delta + n_\Delta)$ , the expected entropy  $h(f(\cdot|\boldsymbol{\theta}))$  has the



following expression, see, [31],

$$- \sum_{x \in \mathcal{I}_A} \frac{\lambda(x) + n(x)}{n + \lambda(x)|\mathcal{I}_A|} [\psi(\lambda(x) + n(x) + 1) - \psi(n + \lambda(x)|\mathcal{I}_A| + 1)], \tag{4.1}$$

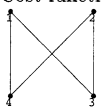
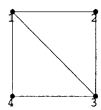

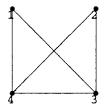
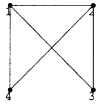
where  $\psi(\cdot)$  is the digamma function. As illustrated in [31], when  $n$  is relatively small and the cardinality of  $\mathcal{I}_A$  is large, the above Bayesian estimator of the entropy tends to be more stable and yield larger values than the maximum likelihood estimator, which is obtained by replacing the probabilities by relative frequencies in the definition of entropy.

We now illustrate the model determination methodology using the risk factor data from Whittaker [30, p. 268]. In the present example and in the next section, the variables are represented by the integers 1, ..., 4 in the given order. The risk factor data consists of a cross-classification of 1841 car factory employees with respect to four binary variables: smoking, strenuous mental work, strenuous physical work, and ratio of lipoproteins.

For  $k = 4$  there are 64 models in  $\mathcal{G}$ , of which 3 are non-decomposable, so it is feasible to compute the expected logarithmic utility for all models. To estimate the expected entropy under a non-decomposable model, we experimented with various numbers of parameter values, and stable estimates were already obtained by using 1000 simulated values from the posterior. All estimates used in the present paper are based on 10,000 values from the posterior.

The results of model determination are presented in Table 1, using the relative utilities (3.4). There is no ambiguity concerning the optimal model, which could be expected bearing the large sample size in mind. The optimal model appears to be non-decomposable, which illustrates the disadvantage of model determination strategies which can be applied only to decomposable models. We note that the current model determination approach for multinomial distributions is expected to produce very similar results as that of Dellaportas and Forster [9].

Table 1  
The models with largest relative expected utilities (3.4) for the risk factor

Cost function	$q(G)(\log n)/2$	$q(G) \log \log n$	$q(G)(\log n)/2$	$q(G) \log \log n$	
	0.960	0.697		0.000	0.052
	0.036	0.148		0.000	0.065
	0.000	0.036			

## 5. Multinormal case

We now consider model determination in the class of graphical Gaussian, or covariance selection models [10,30]. In this class, the joint distribution of  $\Delta$  equals a multinormal distribution  $N_k(\mathbf{0}, \Sigma)$ . Let  $\mathbf{S}$  denote the observed covariance matrix based on a sample of size  $n$ .

The entropy of  $N_k(\mathbf{0}, \Sigma)$  is given by

$$h(\Sigma) = \frac{k}{2}(1 + \log(2\pi)) + \frac{1}{2}\log |\Sigma|. \quad (5.1)$$

Since the maximum likelihood estimate of  $h(\Sigma)$  is provided by replacing  $\Sigma$  with the sample covariance matrix  $\mathbf{S}$ , it tends to be lower than the true value, as  $E|\mathbf{S}| < |\Sigma|$  for  $k > 1$ , see, for instance, [22]. To derive an expression for the Bayesian estimate, we need the following lemma, which follows by an obvious modification of the proof of Theorem 3.4.8 in [21].

**Lemma 5.1.** *Let  $\mathbf{M}$  follow the inverse Wishart distribution  $W_k^{-1}(\Phi, m)$ , and let  $\chi^{-2}(v)$  denote an inverted chi-squared random variable with  $v$  degrees of freedom. Then,  $|\mathbf{M}|$  is distributed as  $|\Phi| \prod_{i=0}^{k-1} \chi^{-2}(m - i)$ , where  $\chi^{-2}(\cdot)$  are all independent.*

Given Lemma 5.1, and the expectations of the logarithms of  $\chi^2(\cdot)$  and  $\chi^{-2}(\cdot)$ , we may state the following result.

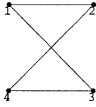

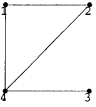
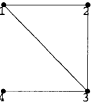
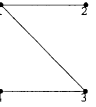
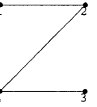
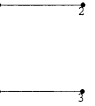
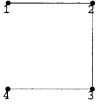
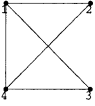
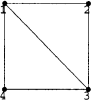
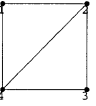
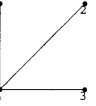
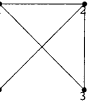
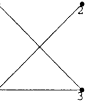
**Proposition 5.2.** *Let  $\mathbf{S}$  be a  $k \times k$  sample covariance matrix under  $N_k(\mathbf{0}, \Sigma)$ . Under the reference prior  $\pi(\Sigma) \propto |\Sigma|^{-(q+1)/2}$  and quadratic loss, the Bayesian estimate of the entropy  $h(\Sigma)$  is unbiased and equals  $\bar{h}(\Sigma) = \frac{k}{2}(1 + \log(2\pi) + \log n) + \frac{1}{2}(\log |\mathbf{S}| - \log 2 - \sum_{i=0}^{k-1} (\psi((n-1-i)/2)))$ .*

The reference posterior based on the prior  $\pi(\Sigma) \propto |\Sigma|^{-(q+1)/2}$  is the inverse Wishart distribution  $W^{-1}(n\mathbf{S}, n-1)$ , with the parameterization of Mardia et al. [21]. As in the multinomial case, the above formula cannot be used for a non-decomposable part of the investigated graph, but the expected entropy has to be found by simulating covariance matrices from the posterior, and projecting them to the corresponding affine subspace.

To illustrate graphical model determination for multinormal data we consider the Fret's heads data set, analyzed previously in [11,21,30]. The data consists of 25 measurements of the head length and breadth of the first and second son, respectively. The models with largest relative expected utilities (3.4) are given in Table 2. For comparison, the posterior model probabilities obtained in [11] are also given.

Situation is quite opposite to the previous example, as there is a considerable uncertainty about the dependence structure. The relative expected utilities differ from the posterior probabilities obtained in [11], since their analysis was restricted to decomposable models only, and favors thus unnecessary complex models according

Table 2  
The models with largest relative expected utilities (3.4) for the Fret’s heads data

						
(1) 0.280	0.111	0.069	0.057	0.051	0.047	0.040
(2) 0.273	0.108	0.067	0.055	0.030	0.030	0.025
(3) 0.000	0.000	0.088	0.073	0.045	0.038	< 0.027
						
(1) 0.028	0.025	0.021	0.015	0.015	0.011	0.005
(2) 0.017	0.074	0.069	0.049	0.030	0.048	0.012
(3) < 0.027	0.153	0.097	0.080	< 0.027	< 0.027	0.027

The used cost functions  $q(G)(\log n)/2$  (1),  $q(G) \log \log n$  (2) are indicated at the beginning of each row. For comparison, the posterior model probabilities (3) obtained in [11] are given.

to our results. However, the ordering of the decomposable models is roughly the same according to the posterior probabilities and the relative utilities based on the cost function  $q(G) \log \log n$ .

To investigate performance of the model determination procedure based on different cost functions more systematically, we generated data sets from multi-normal distributions at various parameter values. The entropy criteria with the cost functions  $q(G)(\log n)/2$  and  $q(G) \log \log n$  are in the sequel referred to as EC1 and EC2, respectively. The SBC criterion of Schwarz [26] was also included in the comparison.

Four different trivariate normal distributions, which were identical except for the covariance between two of the variables, were investigated. The means and variances were chosen to be 0 and 10, respectively. The two fixed covariances were set to 5 and the third ( $\theta$ ) was given the values 2.5, 3, 4, 5, where the first value corresponds to a model with conditional independence between two of the variables. These models are referred to as model 1, ..., 4, respectively. The permutation of variable labels corresponding to the covariance matrix

$$\begin{pmatrix} 10 & 5 & \theta \\ & 10 & 5 \\ & & 10 \end{pmatrix}$$

was chosen. With three variables, there are 8 models in the class  $\mathcal{G}$ . The sample sizes 20, 30, 40, 50, 75, 100, 200 and 300, with 50,000 replications of each, were used. The proportions of replicates for which the correct model was indicated as the optimal

Table 3

Proportions of replicates for which the correct models were indicated as optimal using different criteria

Model	$n$	20	30	40	50	75	100	200	300
1	EC1	0.4050	0.6046	0.7595	0.8537	0.9532	0.9760	0.9871	0.9898
1	EC2	0.4724	0.6669	0.7993	0.8698	0.9350	0.9497	0.9614	0.9648
1	SBC	0.4830	0.6692	0.7976	0.8705	0.9444	0.9626	0.9780	0.9827
2	EC1	0.0139	0.0094	0.0091	0.0140	0.0281	0.0401	0.0610	0.0746
2	EC2	0.0251	0.0220	0.0308	0.0453	0.0729	0.0919	0.1317	0.1667
2	SBC	0.0266	0.0214	0.0266	0.0360	0.0549	0.0671	0.0901	0.1077
3	EC1	0.0041	0.0032	0.0137	0.0395	0.1523	0.2887	0.6332	0.8216
3	EC2	0.0095	0.0185	0.0592	0.1207	0.3070	0.4619	0.7813	0.9118
3	SBC	0.0011	0.0132	0.0520	0.0984	0.2350	0.3932	0.7089	0.8680
4	EC1	0.0021	0.0051	0.0290	0.0842	0.3325	0.5972	0.9757	0.9990
4	EC2	0.0070	0.0336	0.1130	0.2268	0.5472	0.7744	0.9928	0.9999
4	SBC	0.0065	0.0311	0.0940	0.1916	0.4783	0.7138	0.9860	0.9995

one are given in Table 3. For the investigated models EC2 is nearly uniformly better in indicating the correct model than the other two criteria, the exception being model 1 for which SBC performs best. On the other hand, EC1 performs nearly uniformly worst, as might be expected on the basis of the discussion in Section 3. The results thus strongly discourage the use of the cost function  $q(G)(\log n)/2$  in the current model determination procedure.

## 6. Discussion

We have demonstrated the importance of considering both decomposable and non-decomposable models in graphical model determination. Non-decomposable models arise frequently in applications, and may represent a subject-matter theory about a dependence structure equally well as decomposable models. Therefore, learning about dependence structures should not be restricted to decomposable models on the basis of mathematical and computational convenience.

In addition to the identification of the “optimal” model in the light of data, it is important to consider the degrees of optimality of various competing models, as illustrated in our examples. When the relative utilities are not strongly supporting a single model, they are useful in communicating the model determination results and facilitate a weighted posterior inference about any model-dependent quantity of interest.

Although we have presented a unified approach to graphical model determination within the exponential family, some substantial matters remain for further research. First, a practically applicable method for the calculation of expected utilities has still to be developed for the CG distributions of Lauritzen and Wermuth [18]. Second, it would be useful to generalize the currently considered model class to allow for graphs with directed edges. Finally, the current approach should be extended to handle situations with missing data. It is worth to notice that this important issue has

not so far been treated in any of the papers concerning Bayesian graphical model determination.

## References

- [1] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, Wiley, New York, 1978.
- [2] J. Berger, L. Pericchi, The intrinsic Bayes factor for model selection and prediction, *J. Amer. Statist. Assoc.* 91 (1996) 109–122.
- [3] J. Bernardo, Nested hypothesis testing: the Bayesian reference criterion, in: J. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), *Bayesian Statistics*, Vol. 6, Oxford University Press, Oxford, 1999, pp. 101–130.
- [4] J. Bernardo, A. Smith, *Bayesian Theory*, Wiley, Chichester, 1994.
- [5] L. Brown, *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics, Hayward, 1986.
- [6] J. Corander, Graphical model selection for multinomial data using information divergence, Research Report, Department of Statistics, Stockholm University, 1998.
- [7] I. Csiszar, I-divergence geometry of probability distributions and minimization problems, *Ann. Probab.* 3 (1975) 146–158.
- [8] A. Dawid, S. Lauritzen, Hyper-Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.* 21 (1993) 1272–1317.
- [9] P. Dellaportas, J. Forster, Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, *Biometrika* 86 (1999) 615–633.
- [10] A. Dempster, Covariance selection, *Biometrics* 28 (1972) 157–175.
- [11] P. Giudici, P. Green, Decomposable graphical Gaussian model determination, *Biometrika* 86 (1999) 785–801.
- [12] P. Giudici, P. Green, C. Tarantola, Efficient model determination for discrete graphical models, Technical Report, 1999, available at [www.statslab.cam.ac.uk/MCMC/](http://www.statslab.cam.ac.uk/MCMC/).
- [13] E. Hannan, B. Quinn, The determination of the order of an autoregression, *J. Roy. Statist. Soc. B* 41 (1979) 190–195.
- [14] R. Kass, A. Raftery, Bayes factors, *J. Amer. Statist. Assoc.* 90 (1995) 773–795.
- [15] R. Kass, L. Wasserman, The selection of prior distributions by formal rules, *J. Amer. Statist. Assoc.* 91 (1996) 1343–1370.
- [16] J. Key, L. Pericchi, A. Smith, Bayesian model choice: what and why? in: J. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), *Bayesian Statistics*, Vol. 6, Oxford University Press, Oxford, 1999, pp. 343–370.
- [17] S. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996.
- [18] S. Lauritzen, N. Wermuth, Graphical models for associations between variables, some of which are qualitative and some quantitative, *Ann. Statist.* 17 (1989) 31–54.
- [19] D. Madigan, A. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window, *J. Amer. Statist. Assoc.* 89 (1994) 1535–1546.
- [20] D. Madigan, J. York, Bayesian graphical models for discrete data, *Internat. Statist. Rev.* 63 (1995) 215–232.
- [21] K. Mardia, J. Kent, J. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
- [22] R. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
- [23] A. O'Hagan, Fractional Bayes factors for model comparisons, *J. Roy. Statist. Soc. B* 57 (1995) 99–138.
- [24] W. Perks, Some observations on inverse probability including a new indifference rule, *J. Inst. Actuaries* 73 (1947) 285–334.
- [25] T. Rudas, A new algorithm for the maximum likelihood estimation of graphical log-linear models, *Comput. Statist.* 13 (1998) 529–537.

- [26] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* 6 (1978) 461–464.
- [27] A. Smith, D. Spiegelhalter, Bayes factors and choice criteria for linear models, *J. Roy. Statist. Soc. B* 42 (1980) 213–220.
- [28] T. Speed, H. Kiiveri, Gaussian Markov distributions over finite graphs, *Ann. Statist.* 14 (1986) 138–150.
- [29] N. Wermuth, Graphical Markov models, In: S. Kotz, C. Read, D. Banks (Eds.), *Encyclopedia of Statistical Science, Update, Vol. 2*, Wiley, New York, 1998, pp. 284–300.
- [30] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester, 1990.
- [31] L. Yuan, H. Kesavan, Bayesian estimation of Shannon entropy, *Comm. Statist. Theory Methods* 26 (1997) 139–148.