Objective Methods for Graphical Structural Learning

Petrakis Nikolaos, * Peluso Stefano, † Fouskakis Dimitris,‡
Consonni Guido §

Abstract

We propose a theoretically sound Objective Bayes procedure for graphical model selection. Our method is based on the Expected–Posterior Prior (EPP) of Pérez and Berger (2002) and on the Power–Expected–Posterior Prior (PEPP) of Fouskakis et al. (2015). Being the input of the proposed methodology a default improper prior, we do not need subjective prior elicitations and we suggest computationally efficient approximations of Bayes factors and posterior odds. In diverse simulated scenarios with varying number of nodes and sample sizes, we show that we perform equally well or better than benchmarks. Finally, an application to protein-signalling data reveals results in line with the literature.

1 Introduction

Graphical models represent conditional independence relationships among variables by the means of a graph having the variables as nodes. They are widely used in genomic studies (Dobra et al. 2004 and Bhadra and Mallick 2013), finance (Sohn and Kim 2012 and Carvalho and Scott 2009), energy forecasting (Wytock and Kolter 2013), among other fields. Theoretical foundations of graphical models can be found in Lauritzen (1996), Cowell et al. (1999) and Dawid and Lauritzen (1993).

More specifically, for a collection of q random variables with conditional independence structure represented by an Undirected Graph (UG), we assume the underlying graph's structure to be unknown, and we want to infer it from the data, through what is known as *structure learning*. Our approach to the problem will be purely Bayesian, in order to elaborately handle uncertainty: we

*University of Milano-Bicocca, Department of Statistics and Quantitative Methods (n.petrakis@campus.unimib.it)
†Università Cattolica del Sacro Cuore, Department of Statistical Sciences
‡National Technical University of Athens, Department of Mathematics
§Università Cattolica del Sacro Cuore, Department of Statistical Sciences

first assign a prior to the graph and then a prior distribution to the covariance matrix Σ given the graph. Specifying a conditional prior distribution on Σ is non-trivial because each graph induces a different independence structure that affects the shape of the parameter space. In this context, the direct use of an improper prior is infeasible, since is would incur in indeterminate Bayes factors. Thus, we need to carefully elicit a prior distribution under each graph, a task with no hope in high dimensions, or to follow an Objective Bayes (OB) procedure that, starting from a default improper prior, obtains a "usable" prior distribution.

OB contributions to structure learning can be found in Carvalho and Scott (2009), Consonni et al. (2017), Castelletti et al. (2018). They all opt for the Fractional Bayes Factor approach of O'Hagan (1995), which is mathematically convenient but with the important disadvantage of a double use of the data, both for prior specification and model selection. In the current paper we introduce a structural learning approach based on the Expected–Posterior Prior (EPP) approach of Pérez and Berger (2002) and the Power–Expected–Posterior Prior (PEPP) of Fouskakis et al. (2015), which are theoretically sounder than what proposed in the literature in the context of objective approaches, since double usage of data is avoided and compatible prior distributions are provided; more details on prior compatibility can be found in Consonni and Veronese (2008).

The remainder of the paper is as follows. After introducing basic notations and distributions in Section 2, in Section 3 we describe basic notions of Bayesian model selection under OB, with focus on EPP and PEPP (Section 3.1) and on computational challenges (Section 3.2). In Section 4 we face the structural learning problem: in Section 4.1 we estimate the Bayes factor and in Section 4.2 we specify the prior distribution on the graphs. In Section 5 we apply the proposed methodologies to simulated settings and to the protein signalling data of Sachs et al. (2005). We conclude in Section 6, highlighting further directions of investigation.

2 Overview of Graphical Models

Let $G = (V, E)$ denote an undirected graph with a finite set of nodes V and a set of edges $E \subseteq V \times V$. Self–loops are not allowed, so that for any edge $(a, b) \in E$ we have that $a \neq b$. Nodes $a, b \in V$ will be assumed adjacent if the edge $(a, b) \in E$ and if all nodes of G are adjacent then G will be assumed complete. A complete subgraph $C \subset V$ that is maximal with respect to $\subset$ will be called a clique. A triple (A, S, B) of subsets of V forms a decomposition of G if $V = A \cup B$ and $S = A \cap B$ is complete and S separates A from B, where S will be called a separator. Each graph G will be associated with a clique set $\mathscr{C}$ and separator set $\mathscr{S}$. A graph G will be called decomposable if its cliques and separators admit a perfect ordering i.e. $\forall j > 1$ and for $k < j$ we obtain $S_j = C_j \cap H_{j-1} \subset C_k$ where $H_{j-1} = \cap_{k=1}^{j-1} C_k$. Every graph throughout this paper will be assumed to be decomposable.

Each element $a \in V$ will be related to a random variable Y_a with values in a sample space $\mathscr{Y}_a$. For a given set $A \subseteq V$ we define $Y_A = (Y_a)_{a \in A}$ as a collection of random variables $\{Y_a : a \in A\}$ with values in $\mathscr{Y}_A = \times_{a \in A} \mathscr{Y}_a$. A probability distribution over $A \subseteq V$ implies a joint distribution for Y_A over $\mathscr{Y}_A$. Let P be a distribution over $U \subseteq V$ and $A, B \subseteq U$, then P_A will imply the marginal distribution of Y_A and $P_{B|A}$ the conditional distribution of Y_B given $Y_A = y_A$. The distribution P over the vertex set V is Markov w.r.t. to a graph G, if for any decomposition (A, S, B) of G we obtain that $Y_A \perp\!\!\!\perp Y_B \,|\, Y_S$, where $\perp\!\!\!\perp$ implies conditional independence between variables.

Let $(\mathbf{y}_1, \cdots, \mathbf{y}_n)$ be n independent observations from a $N_q(\mathbf{0}, \Sigma)$, and

$$\mathbf{Y} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_q) = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix} \tag{1}$$

be the $n \times p$ data matrix of these observations, where $\mathbf{y}_i = (y_{i1}, \cdots, y_{iq})$ denotes the i-th observation and $\mathbf{Y}_j = (y_{1j}, \cdots, y_{nj})$ denotes the observations on the j-th variable. We assume that the matrix $\mathbf{Y}$ follows a Matrix–Normal distribution with mean parameter $M = 0$, row–covariance matrix I_n and column–covariance matrix Σ; i.e. $\mathbf{Y} \sim MN_{n \times q}(0, I_n, \Sigma)$. The density of $\mathbf{Y}$ given Σ is

$$f(\mathbf{Y}|\Sigma) = \frac{det(\Sigma)^{-n/2}}{(2\pi)^{nq/2}} \exp\Big\{ -\frac{1}{2} tr(\Sigma^{-1}\mathbf{S}) \Big\}, \tag{2}$$

where $\mathbf{S} = \mathbf{Y}^T\mathbf{Y}$ and $tr(\cdot)$ denotes trace of a matrix.

By the term graphical model, we refer to a family of distributions which are Markov with respect to a graph G. A Gaussian graphical model is defined by assuming that $\mathbf{Y} \sim MN_{n \times q}(0, I_n, \Sigma)$, adhered to the Markov property with respect to the graph G, where its structure is represented through non–zero entries on the concentration matrix $K = \Sigma^{-1}$. We write $\Sigma \in M^+(G)$ to denote that Σ is positive definite and coherent with the Markov property to G. The density of $\mathbf{Y}$ given Σ and G follows the graph decomposition in cliques and separators as follows:

$$f(\mathbf{Y}|\Sigma, G) = \frac{\prod_{C \in \mathscr{C}} f(\mathbf{Y}_C|\Sigma_C, G)}{\prod_{S \in \mathscr{S}} f(\mathbf{Y}_S|\Sigma_S, G)}, \tag{3}$$

where under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) the matrix

$$\mathbf{Y}_C = \begin{pmatrix} \mathbf{y}_{1C}^T \\ \vdots \\ \mathbf{y}_{nC}^T \end{pmatrix}$$

follows a Matrix Normal distribution, such that $\mathbf{Y}_C \sim MN_{n \times |C|}(\mathbf{0}, I_n, \Sigma_C)$, with $|\cdot|$ denoting set cardinality. Matrices $\mathbf{Y}_C$ and $\mathbf{Y}_S$ denote submatrices of $\mathbf{Y}$ consisting of the columns indexed by $C \in \mathscr{C}$ or $S \in \mathscr{S}$.

A commonly used conjugate prior on the covariance matrix Σ is the Hyper-Inverse Wishart distribution; see Dawid and Lauritzen (1993). We use the notation $\Sigma \sim HIW_G(b, D)$ to denote that Σ follows a Hyper–Inverse Wishart distribution with respect to a graph G, having $b > 0$ degrees of freedom and scale matrix $D \in M^+(G)$. The density of Σ has a similar factorization as in (3) i.e.

$$\pi(\Sigma|G) = \frac{\prod_{C\in\mathscr{C}} \pi(\Sigma_C|b, D_C, G)}{\prod_{S\in\mathscr{S}} \pi(\Sigma_S|b, D_S, G)}, \tag{4}$$

where under each clique $C \in \mathscr{C}$ (and separator $S \in \mathscr{S}$) the respective matrix Σ_C follows an Inverse Wishart distribution $IW_{|C|}(b, D_C)$ with density

$$\pi(\Sigma_C|b, D_C, G) = K_C\, det(\Sigma_C)^{-(b/2+|C|)} exp\Big\{ -\frac{1}{2} tr(\Sigma_C^{-1} D_C)\Big\}, \tag{5}$$

where

$$K_C = \frac{det(D_C)}{2^{b|C|/2}\Gamma_{|C|}(\frac{b}{2})},$$

having $D_C, \Sigma_C \in M^+(G)$. Matrix Σ_C corresponds to the marginal covariance matrix of $\mathbf{y}_C$ obtained from $\mathbf{y}$ by selecting the elements of $\mathbf{y}$ indexed by $C \in \mathscr{C}$ (similarly for $S \in \mathscr{S}$).

By Roverato (2000), under a graph G, the density of $K = \Sigma^{-1}$ is

$$\pi(K|\delta, A, G) \propto det(K)^{(b-2)/2} \exp\Big\{ -\frac{1}{2} tr(KA^{-1})\Big\},$$

where $A = D^{-1}$ and $K, A \in M^+(G)$. If G is the complete graph, then K follows a Wishart distribution. If G is not the complete graph, then the density of K is proportional to a Wishart distribution, conditioned to the event $K \in M^+(G)$. This distribution is known as *G-conditional Wishart distribution* (or G-Wishart distribution); i.e. $K \sim W_G(b + |V| - 1, A)$. This distribution will be useful in the sequel to generate observations from a Hyper–Inverse Wishart distribution.

3 Objective Bayes Model Selection

We assume the reader is familiar with the basic concepts of model selection from a Bayesian viewpoint, as described for instance in O'Hagan and Forster (2004). In the present section, we provide a background on Objective Bayes model selection (OB), following Consonni et al. (2018) and focusing on the priors of Pérez and Berger (2002) and Fouskakis et al. (2015). The developments in the current section will then be adopted in Section 4 in the context of Gaussian graphical model selection of undirected decomposable graphs.

In the Bayesian model selection framework, our interest lies on a finite set $\mathscr{M} = \{M_1, \cdots, M_k\}$ of statistical models, where we assign a prior model probability $\pi(M_j)$ on each model in $\mathscr{M}$. With $\mathbf{y} = (y_1, \cdots, y_n)^T$ be the data, each model $M_j \in \mathscr{M}$ $(j = 1, \ldots, k)$ is a family of sampling densities $f(\mathbf{y}|\boldsymbol{\theta}_j, M_j)$, indexed by a model–specific parameter $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$, with a prior $\pi(\boldsymbol{\theta}_j|M_j)$. Model comparison of two competing models $M_j, M_i \in \mathscr{M}$ is based on posterior model odds, provided by

$$PO_{M_j:M_i} = \frac{\pi(M_j|\mathbf{y})}{\pi(M_i|\mathbf{y})} = \frac{\pi(M_i)}{\pi(M_j)} \times \frac{m_j(\mathbf{y})}{m_i(\mathbf{y})}; \tag{6}$$

the ratio $m_j(\mathbf{y})/m_i(\mathbf{y})$ represents the Bayes factor of model M_j versus model M_i and is denoted by $BF_{M_j:M_i}(\mathbf{y})$, with $m_j(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta}_j, M_j)\pi(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j$ being the marginal likelihood of the observable data $\mathbf{y}$ under model $M_j \in \mathscr{M}$.

The OB approach is adopted in cases where we either want to express our prior ignorance on parameters or it is infeasible to successfully elicit a prior distribution, especially on high–dimensional parameter spaces. We use a non–informative default prior distributions, denoted by $\pi^N(\boldsymbol{\theta}_j|M_j)$, which are usually improper (i.e. having non–finite total mass). These kind of prior distributions cannot be directly used for model comparison, since the resulting Bayes factors will depend on ratios of arbitrary normalizing constants. An extended review on how to handle improper prior distributions in model selection problems can be found in Consonni et al. (2018), with the most notable of them being the Fractional Bayes factor of O'Hagan (1995), the Intrinsic Bayes factor of Berger and Pericchi (1996), the Expected–Posterior Prior approach of Pérez and Berger (2002) and the Power–Expected Posterior Prior approach of Fouskakis et al. (2015). For the needs of this paper we will briefly describe the latter two in the following subsection.

3.1 Objective Priors for Model Selection

Objective Bayes priors avoid subjective elicitation by only using the information available from the statistical model. An example in this regard is the Expected–Posterior Prior approach (EPP) of Pérez and Berger (2002), conceived as the expectation of the posterior distribution given imaginary observations (Spiegelhalter and Smith 1980, Iwaki 1997) rising from a prior predictive distribution. There are two main advantages in the use of EPPs in model selection: (i) they can expedite the use of improper baseline prior distributions, since there are no issues of indeterminacy in the resulting Bayes factors; (ii) elicitation of the parameters of each prior when the goal is model selection presents specific challenges; the main one being compatibility of priors across models (Consonni and Veronese, 2008) and EPPs provide a valid construction of compatible priors, being related to the same *reference model*.

On the other hand, EPPs also show some limitations, since they rely on features of the imaginary sample, namely to the imaginary sample size and on

the imaginary design matrix. Optimal choices proposed in the literature of the minimal imaginary sample size are not entirely satisfactory, since the resulting priors can still have a significant role in the Bayes factor, especially when the number of parameters is close to the number of observations. These EPP limitations led Fouskakis et al. (2015) to introduce the Power–Expected–Posterior Prior (PEPP) approach, where they combine ideas from the power–prior approach of Ibrahim and Chen (2000) and the unit–information prior approach of Kass and Wasserman (1995). The rationale of the PEPP approach is the construction of minimally–informative prior distributions, in which the effect of the imaginary data on the posterior distribution collapses in one data point. To achieve this result, Fouskakis et al. (2015) raised the likelihood function involved in the calculation of the EPP to the power $1/\delta$, with $\delta = n$, where n denotes the sample size. The choice $\delta = n$ leads to an imaginary design matrix equal to the observed one, and therefore the selection of a training sample of covariates and its effects on the posterior model comparison is avoided. Therefore, with the PEPP approach the sample size of the imaginary data does not have a significant effect on the output of the method and the computational cost can be reduced.

More formally, let $\mathbf{y}^* = (y_1^*, \cdots, y_m^*)^T$ be m independent imaginary observations which rise independently from a random variable $\mathbf{Y}^*$ on sample space $\mathscr{Y}^*$. We will assume that both random variables $\mathbf{Y}^*$ and $\mathbf{Y}$ are i.i.d. random variables on a common sample space $\tilde{\mathscr{Y}}$. Under a given model $M_j \in \mathscr{M}$, starting from a default baseline (usually improper) prior $\pi^N(\boldsymbol{\theta}_j|M_j)$, the posterior distribution of $\boldsymbol{\theta}_j$ given the imaginary data vector $\mathbf{y}^*$ will be provided by $\pi^N(\boldsymbol{\theta}_j|\mathbf{y}^*, M_j) \propto f(\mathbf{y}^*|\boldsymbol{\theta}_j, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)$, where $f(\mathbf{y}^*|\boldsymbol{\theta}_j, M_j)$ denotes the sampling density of $\mathbf{y}^*$ under model M_j. The EPP for the parameter $\boldsymbol{\theta}_j$ under model $M_j \in \mathscr{M}$ is

$$\pi^{EPP}(\boldsymbol{\theta}_j|M_j) = \int \pi^N(\boldsymbol{\theta_j}|\mathbf{y}^*, M_j)m^*(\mathbf{y}^*)d\mathbf{y}^*, \tag{7}$$

where $m^*(\mathbf{y}^*)$ is provided by

$$m^*(\mathbf{y}^*) \equiv m_0^N(\mathbf{y}^*) = \int f(\mathbf{y}^*|\boldsymbol{\theta}_0, M_0)\pi^N(\boldsymbol{\theta}_0|M_0)d\boldsymbol{\theta}_0; \tag{8}$$

i.e. the marginal likelihood, evaluated on $\mathbf{y}^*$, of a *reference model* M_0, using the default baseline prior $\pi^N(\boldsymbol{\theta}_0|M_0)$. In nested cases, M_0 is chosen to be the "simplest" model in $\mathscr{M}$; by this way we *a priori* support parsimony; i.e. in absence of enough evidence from the data, simpler models will be favored.

To define the PEPP of $\boldsymbol{\theta}_j$ under model $M_j \in \mathscr{M}$, Fouskakis et al. (2015) used the normalized power–likelihood

$$f(\mathbf{y}^*|\boldsymbol{\theta}_j, \delta, M_j) = \frac{f(\mathbf{y}^*|\boldsymbol{\theta}_j, \delta, M_j)^{1/\delta}}{\int f(\mathbf{y}^*|\boldsymbol{\theta}_j, \delta, M_j)^{1/\delta}d\mathbf{y}^*}. \tag{9}$$

This form of the likelihood adapts to the variable selection problem of Gaussian linear models, yet it does not hold for all members of the exponential family;

further information provided in Fouskakis et al. (2018). Using (9), the power–posterior distribution of $\boldsymbol{\theta}_j$ given the imaginary data vector $\mathbf{y}^*$ is provided by $\pi^N(\boldsymbol{\theta}_j|\mathbf{y}^*, \delta, M_j) \propto f(\mathbf{y}^*|\boldsymbol{\theta}_j, \delta, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)$. Then the PEPP of $\boldsymbol{\theta}_j$ is given by

$$\pi^{PEPP}(\boldsymbol{\theta}_j|M_j, \delta) = \int \pi^N(\boldsymbol{\theta_j}|\mathbf{y}^*, \delta, M_j)m^*(\mathbf{y}^*|\delta)d\mathbf{y}^*, \tag{10}$$

with

$$m^*(\mathbf{y}^*|\delta) \equiv m_0^N(\mathbf{y}^*|\delta) = \int f(\mathbf{y}^*|\boldsymbol{\theta}_0, \delta, M_0)\pi^N(\boldsymbol{\theta}_0|M_0)d\boldsymbol{\theta}_0. \tag{11}$$

Note that in the above, for $\delta = 1$ we obtain the EPP of $\boldsymbol{\theta}_j$ under $M_j \in \mathscr{M}$.

3.2 Computational Aspects

Pérez and Berger (2002) claimed that the EPP (and thus the PEPP as well) in (7), can be viewed as a two–stage hierarchical prior, whereas the first–stage prior is the posterior distribution $\pi^N(\boldsymbol{\theta}_j|\mathbf{y}^*, M_j)$ and the second–stage prior would be the predictive density $m^*(\mathbf{y}^*)$. Thus, they deduced that the EPP (and PEPP) approach could be integrated using MCMC approaches. They provide computational guidelines to approximate Bayes factors using importance sampling, especially when the predictive density $m^*(\mathbf{y}^*)$ is not proper. Fouskakis et al. (2015) provided four different approximations of Bayes factors as well, where one was aligned with the approach of Pérez and Berger (2002), using the power–likelihood of (9).

Following an importance sampling simulation scheme, the Bayes factor of a model $M_j \in \mathscr{M}$ versus M_0, under the PEPP approach (or the EPP for $\delta = 1$), can be approximated by

$$\widehat{BF}^{PEPP}_{M_j:M_i}(\mathbf{y}|\delta) = BF^N_{M_j:M_0}(\mathbf{y}|\delta)\frac{1}{R}\sum_{r=1}^{R} BF^N_{M_0:M_j}(\mathbf{y}^{*(r)}|\delta), \tag{12}$$

with $\mathbf{y}^{*(r)}, r = 1, \cdots, R$ being R i.i.d. samples from the importance density $g(\mathbf{y}^*|\delta) = m_j^N(\mathbf{y}^*|\mathbf{y},\ \delta)$, where $m_j^N(\mathbf{y}^*|\mathbf{y},\ \delta) = m_j^N(\mathbf{y}^*, \mathbf{y}|\delta)/m_j^N(\mathbf{y}|\delta)$. The marginal likelihoods required for calculating the importance density, are provided by

$$m_j^N(\mathbf{y}^*, \mathbf{y}|\delta) = \int f(\mathbf{y}^*, \mathbf{y}|\boldsymbol{\theta}_j, \delta, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j$$

and

$$m_j^N(\mathbf{y}|\delta) = \int f(\mathbf{y}|\boldsymbol{\theta}_j, \delta, M_j)\pi^N(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j$$

respectively, whereas the Bayes factors included in (12) are provided using the default baseline priors $\pi^N(\boldsymbol{\theta}_j|M_j)$. If the predictive density $m_j^N(\mathbf{y}^*|\mathbf{y},\ \delta)$ is not available in a closed–form expression, Gibbs sampling scheme will be deployed for generating importance samples, which can be performed as follows:

For $r = 1, \cdots, R$:

- Generate $\boldsymbol{\theta}^{(r)}$ from $\pi^N(\boldsymbol{\theta}_j|\mathbf{y}, \delta, M_j)$.
- Generate a sample $\mathbf{y}^{*(r)}$ from $f(\mathbf{y}^*|\boldsymbol{\theta}^{(r)}, \delta, M_j)$.

---

4 Structural Learning using EPP and PEPP

4.1 Bayes Factor Estimation

Consider the sample data matrix $\mathbf{Y}$ in (1) and let $\mathscr{G}$ denote the entire collection of all undirected decomposable Gaussian graphical models on q nodes. Before we proceed with the development of the PEPP (and EPP) approach for the Gaussian graphical model selection procedure, we need to define an *independence graph* $G_0 = (V, E_0)$ where $E_0 = \emptyset$; this is the "simplest" model among all models in $\mathscr{G}$. Given a graph $G \in \mathscr{G}$, following Carvalho and Scott (2009) we will consider the conjugate and computationally convenient improper baseline prior, exploiting the same factorization over cliques and separators as in (3) and (4):

$$\pi^N(\Sigma|G) \propto \frac{\prod_{C\in\mathscr{C}} det(\Sigma_C)^{-|C|}}{\prod_{S\in\mathscr{S}} det(\Sigma_S)^{-|S|}}, \tag{13}$$

where the covariance matrix $\Sigma \in M^+(G)$.

Let $(\mathbf{y}_1^*, \cdots, \mathbf{y}_m^*)$ be m independent imaginary multivariate observations and $\mathbf{Y}^*$ be the $m \times q$ matrix of these observations, similarly as in (1). We let $\mathbf{Y}$ and $\mathbf{Y}^*$ to be independent on a common sample space $\mathscr{Y}$. Following (2), (3) and (9), the power–likelihood of $\mathbf{Y}_C^*$ given $\Sigma_C \in M^+(G)$ under model $G \in \mathscr{G}$, is given by

$$f(\mathbf{Y}_C^*|\Sigma_C, \delta, G) = \frac{\det(\delta\Sigma_C)^{-m/2}}{(2\pi)^{m|C|/2}} \exp\Big\{\frac{1}{2} tr(\delta\Sigma_C)^{-1}\mathbf{S}_C^*\Big\}, \tag{14}$$

which represents a $MN_{m\times|C|}(\mathbf{0}, I_m, \delta\Sigma_C)$, where $\mathbf{S}_C^* = \mathbf{Y}_C^{*T}\mathbf{Y}_C^*$. Then, under each clique $C \in \mathscr{C}$ the power–posterior distribution of Σ_C given $\mathbf{Y}_C^*$, under the baseline prior, is provided by

$$\pi^N(\Sigma_C|\mathbf{Y}_C^*, \delta, G) \propto det(\Sigma_C)^{-(m/2+|C|)} exp\Big\{-\frac{1}{2} tr(\Sigma_C^{-1}\delta^{-1}\mathbf{S}_C^*)\Big\} \tag{15}$$

which represents an $IW_{|C|}(m, \delta^{-1}\mathbf{S}_C^*)$ distribution. Thus, under a graph $G \in \mathscr{G}$, the power–likelihood of $\mathbf{Y}^*$ given $\Sigma \in M^+(G)$ is a $MN_{m\times q}(\mathbf{0}, I_m, \delta\Sigma)$ and the power–posterior of Σ given $\mathbf{Y}^*$, under the baseline prior, is a $HIW_G(m, \mathbf{S}^*/\delta)$. Following the discussion in Section 3.1, the predictive density in (11) is given by

$$m^N(\mathrm{Y}^*|\delta, G_0) = \int f(\mathbf{Y}^*|\Sigma, \delta, G_0)\pi^N(\Sigma|G_0)d\Sigma. \tag{16}$$

Following (14), (15) and (16), the PEPP (EPP for $\delta = 1$) of Σ under a graph $G \in \mathscr{G}$ is

$$\pi^{PEPP}(\Sigma|G, \delta) = \int \pi^N(\Sigma|\mathbf{Y}^*, \delta, G)m^N(\mathbf{Y}^*|\delta, G)d\mathbf{Y}^*. \tag{17}$$

Our guideline to define a minimal size for the number of rows m of the imaginary data matrix $\mathbf{Y}^*$, is based on the posterior distribution included in (17). In order, for this respective posterior distribution to be finite, the number of rows m of the imaginary matrix $\mathbf{Y}^*$ must always satisfy $m \geq |C|, \quad \forall C \in \mathscr{C}, \quad \forall G \in \mathscr{G}$, which implies that $m \geq q$.

4.2 Bayes Factor Computation

As mentioned in Section 3.2, the PEPP and EPP will be improper since they will contain the arbitrary normalizing constant that rises under the reference model. Using the importance sampling scheme of Section 3.2, the Bayes factor of any graph $G \in \mathscr{G}$ versus the independence graph G_0, under the PEPP (or EPP for $\delta = 1$) approach, is approximated by

$$\widehat{BF}^{PEPP}_{G:G_0}(\mathbf{Y}|\delta) = K(\mathbf{Y}, G)\, H(\mathbf{Y}, G|\delta) \sum_{r=1}^{R} K(\mathbf{Y}^*, G)\, H(\mathbf{Y}^*, G|\delta), \tag{18}$$

where for a data matrix $\mathbf{X}_{n\times q}$ and an undirected decomposable graph G we define

$$K(\mathbf{X}, G) = \frac{\prod_{C\in\mathscr{C}} \Gamma_{|C|}(\frac{n+|C|-1}{2})}{\prod_{S\in\mathscr{S}} \Gamma_{|S|}(\frac{n+|S|-1}{2})} \Gamma^{-q}\left(\frac{n}{2}\right) \tag{19}$$

and

$$H(\mathbf{X}, G|\delta) = \prod_{j=1}^{q} det(\frac{1}{2\delta}\mathbf{S}_j)^{\frac{n}{2}} \frac{\prod_{C\in\mathscr{C}} det(\frac{1}{2\delta}\mathbf{S}_C)^{-\frac{n+|C|-1}{2}}}{\prod_{S\in\mathscr{S}} det(\frac{1}{2\delta}\mathbf{S}_S)^{-\frac{n+|S|-1}{2}}}. \tag{20}$$

Using the generation process of the imaginary observations, following the Gibbs scheme provided in Section 3.2, we approximate the Bayes factor of any model $G \in \mathscr{G}$ versus G_0 using the following scheme:

1. Generate imaginary matrices $\mathbf{Y}^{(1)^*}, \cdots, \mathbf{Y}^{(R)^*}$ as follows:
   - Generate $K^{(r)}$ from $W_G(n+q-1, (\mathbf{S}/\delta)^{-1})$.
   - Generate $\mathbf{Y}^{*(r)}$ from $MN_{m\times q}(\mathbf{0}, I_m, \delta {K^{(r)}}^{-1})$.
2. Approximate $BF^{PEPP}_{G:G_0}(\mathbf{Y}|\delta)$ by (18).

The size R of importance samples is defined using a trial-and-error approach, to control the computational cost of the estimation. For both EPP and PEPP approaches, the size of the imaginary data observation will be defined as $\mathbf{Y}^*_{m\times q}$ with $m = q$. The guideline provided by Fouskakis et al. (2015) is to consider $m = n$ and $\delta = n$, for compressing the effect of the imaginary observations to one data point. In this paper, we set the power parameter $\delta = q$ for two reasons: (i) we need to control the computational cost of our approach; (ii) after several trials with different choices of δ ranging from q to n there were not significant gains in terms of performance. Thus the PEPP we obtain here does not have the unit–information principle, but still has reduced effects from the imaginary observations.

Regarding the prior model probabilities, we follow Carvalho and Scott (2009); i.e. we assign a prior probability on each graph $G \in \mathcal{G}$ using

$$\pi(G) \propto \frac{1}{z+1}\binom{z}{k}^{-1},$$

where k denotes the number of edges of graph G and $z = q(q-1)/2$ is the maximum number of edges a graph G can have. Under this prior, each edge has a prior probability of inclusion equal to 0.5.

4.3 Feature–Inclusion Stochastic Search

To apply the EPP and PEPP approaches, we utilize a serial algorithm of Carvalho and Scott (2008), the Feature–Inclusion Stochastic Search (FINCS), which operates using the following principle: it uses on–line estimates of posterior edge–inclusion probabilities to guide the search process in the space of decomposable graphical models. Applications of FINCS algorithm can be found in Fitch et al. (2014), Carvalho and Scott (2009) and Altomare et al. (2013). Furthermore, we introduce two computational improvements: (i) to alleviate the accumulating processing cost arising from the Bayes factor approximations, we perform the global move of FINCS by deterministically selecting the median graph, as in Altomare et al. (2013); (ii) we control the number of iterations through pilot runs, in order to keep them as small as possible, since our experimental studies show that our algorithm can reach the optimal model choice quite fast, rendering superfluous further runs. Our version of FINCS algorithm

is structured as follows:

For $t = 1, \cdots, T$ (iterations):

1. For a given model G_t generate importance samples following the Gibbs scheme of Section 4.2.
2. Estimate Bayes factor under PEPP (or EPP for $\delta = 1$) using (18).
3. Update posterior edge inclusion probabilities.
4. Propose a new model following the FINCS algorithm.

---

5 Experimental Studies

We compare the performance of EPP and PEPP with the Fractional Bayes Factor (FBF) approach of Carvalho and Scott (2009), and with the Birth–Death MCMC (BDMCMC) approach of Mohammadi and Wit (2015), on simulation studies and a on a real data application. The FBF approach was applied by Carvalho and Scott (2009) using the FINCS algorithm, whilst the BDMCMC approach is an MCMC algorithm based on a continuous–time birth–dead process of edges. For the FBF approach, in all simulation scenarios considered we choose the default choice for $b = q/n$ provided by O'Hagan (1995), whilst for BDMCMC approach we follow the standard guidelines provided by Mohammadi and Wit (2015). Note that BDMCMC is a fully Bayesian, transdimensional method that performs structural learning in an explicit Bayesian context, rather than using Bayes factors as in our approach, and it is applicable to all types of graphical models, whereas here we only focus on decomposable graphical models. The BDMCMC approach was applied through the `BDgraph` and `huge` packages of R, following the developers guidelines. For our simulation study, we consider two simulation set–ups with multiple scenarios, while for the real data application we consider the protein-signalling data in Sachs et al. (2005). All codes were written using parallel computing in R and are available upon request.

5.1 Simulation Studies

Each simulation scenario is characterized by the pair (q, n), where $q = \{10, 20, 30\}$ is the number of nodes and $n = \{100, 300, 500\}$ the sample size, resulting in nine scenarios. We consider two simulation set-ups: the *Random Scenario* and the *Star Scenario*. Under the Random Scenario, we generate a total of 40 datasets, corresponding to 40 true UG models, not guaranteed to be decomposable. Following Peters and Buhlmann (2014), we generate an UG with probability of edge inclusion $p_{edge} = 3/(2q - 2)$. Under the Star Scenario, we generate a total of 40 datasets, where the true graph is constructed by setting all nodes adjacent

to the first node. Under both scenarios, we generate i.i.d. observations using the following scheme:

For $1, \cdots, 40$ simulation runs:

1. For the Random Scenario generate an undirected graph G using the `BDgraph` package in R. The adjacency matrix of the true graph under the Star Scenario, is identical for all generated samples.

2. Under both scenarios, generate a matrix K from the G-Wishart distribution having $b = 10$ degrees of freedom and scale matrix $D = I_q$, based on the adjacency matrix provided by each respective graph G_{True}.

3. Generate a data matrix $Y_{n\times q} \sim MN_{n\times q}(\mathbf{0}, I_n, K^{-1})$.

To compare methodologies, we first define the estimated posterior edge inclusion probability of an edge e_{ij} to be

$$\hat{q}_{ij} = \sum_{G\in\mathcal{G}} I_{e_{ij}\in G}\, \hat{\pi}(G|\mathbf{Y}), \tag{21}$$

where $\hat{\pi}(G|\mathbf{Y})$ denotes the estimated posterior probability of the distinctive graphs visited by the algorithm. The *median probability (graphical) model* is defined as the graph that contains edges having posterior edge inclusion probability greater than 0.5 (not guaranteed to be decomposable). This definition is identical to the one in Barbieri and Berger (2004) for variable selection problems in Gaussian linear models.

Each approach under consideration will be evaluated using three performance indicators that provide evidence on the ability of each approach to identify the true underlying UG. The first measure is the Structural Hamming Distance (SHD), that is the number of edge insertions or deletions required to retrieve the true graph's structure. Clearly, lower values correspond to better performances. Another performance measure is the F_1-score (Baldi et al. (2000); Powers (2011), Mohammadi and Wit (2015)), which is defined by

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{22}$$

where TP, FP and FN denote the number of true positives, false positives and false negatives, respectively. Note that $F_1 \in [0, 1]$, where values closer to one correspond to better identification of edges, whereas values near zero correspond to worse edge detections.

Boxplots, over the 40 simulated datasets, for SHDs under the Random Scenario, are presented in Figure 1. As expected, the performances of all methods improve as we increase the sample size, and deteriorate as we increase the number of nodes. Furthermore, PEPP distances are in all cases better the those

recovered by EPP. The BDMCMC method performs worse, but it tends to allineate as the number of observations increases. We should again highlight that the comparison is not fair, since the output provided by BDMCMC is a much richer MCMC output, and therefore it naturally require higher computational time and higher sample sizes. Finally, PEPP reveals a performance comparable to FBF, especially for smaller sample sizes, without suffering from double usage of data as in FBF. Therefore we believe that PEPP is a valid Bayesian alternative to FBF, since it can provide similar accuracy in a more theoretically sound procedure.

Similar is the picture under the Star Scenario; for brevity reasons the plot is not presented here. In Tables 1 and 2 we present the mean and the variance of the F_1 scores under, respectively, the Random and the Star graph simulation setting, for each scenario and method. In the majority of cases, FBF is the best performer, but with its theoretical limitation of using twice the data, to both regularize the prior and the marginal likelihood evaluation. The PEPP approach closely follows, suggesting a theoretically valid alternative for Bayesian structural learning.

In terms of computational speed, Figure 2 shows that, moving from lower to higher number of vertices, the computational cost of the FINCS algorithm is significantly higher. The FBF approach is similarly constrained, but Carvalho and Scott (2008) have developed the FINCS algorithm in C++ environment, reducing the runtime up to 176 seconds for 50000 iterations. We are currently investigating the transition of our approach to C++ aiming to the same computational gains. The average runtime of BDMCMC with $q = 30$ over 40 datasets is 44 seconds, a performance close to 500 iterations of FINCS algorithm under EPP or PEPP approach.

5.2 Real Data Application

We now provide an application to the Sachs et al. (2005) data, which include levels of eleven phosphorylated proteins and phospholipids quantified using flow cytometry. Nine different experimental studies were conducted, differing on certain aspects, with the respective sample size of each experiment to be between 700 and 1000. These data were originally used by Sachs et al. (2005) to infer a single DAG and Friedman et al. (2008) merged these 9 datasets to infer a single UG. Peterson et al. (2015) used the same dataset to infer a different UG under each experimental condition, and we will share the same purpose.

We apply the PEPP approach following the same algorithmic set–up of our simulation studies for $q = 10$, and we compute, for each dataset, the SHD between the benchmarks and PEPP. The resulting estimated posterior edge inclusion probabilities were provided by considering 16 runs of the FINCS algorithm for each dataset and then calculate their average. First, in Figure 3 we present the learnt graphical structures under the PEPP approach for every dataset. In Table 3 we report the SHDs between the PEPP approach and the

benchmarks, whereas in Table 4 the number of edges under PEPP and the alternative methods. We note that there are not significant differences between EPP and PEPP. All proposed methodologies provide the same output for datasets 3 and 8, whilst the most significant differences can be found in dataset 5, where the SHD between PEPP and FBF, as well as BDMCMC, is up to five different edges. Our findings are similar to Peterson et al. (2015), where they identified Dataset 5 as the one that differs the most compared to the others.

A further advantage of EPP and PEPP over FBF is that the formers associate a higher posterior probability to the estimated MAP model, i.e. the model with the highest estimated posterior probability. In Table 5 we present the estimated posterior probability of the MAP model under datasets 3 and 8, the two datasets where EPP, PEPP and FBF returned the same estimated median probability model. Thus, EPP and PEPP can more easily distinct the "optimal" model relative to FBF. The results of BDMCMC are omitted from this Table since the output of respective package returns only estimated posterior edge inclusion probabilities.

6 Conclusions and Further Directions

We have introduced two theoretically sound Objective Bayes approaches for model selection of undirected Gaussian graphical models. The key difference between the proposed methodologies and the Fractional Bayes Factor approach previously proposed in the literature is that the Expected–Posterior Prior (EPP) and Power–Expected–Posterior Prior (PEPP), that we adopt here, rely on imaginary observations, avoiding double use of data. In their core, they utilize improper prior distributions when it is difficult to successfully elicit a subjective one, alleviating the indeterminacy in Bayes factors arising from the existence of arbitrary normalizing constants. The advantage of PEPP over EPP is that the former reduces the effect of imaginary data, leading in some cases to more accurate estimation.

Our studies show that PEPP performs better or equally well to EPP. Furthermore, the performance of PEPP is better than other benchmark methods, for smaller sample sizes and higher number of nodes. In terms of the estimated MAP model, our results indicate that EPP and PEPP perform better than FBF, since they manage to distinguish the "optimal" model with higher posterior probability. Higher estimated posterior weights lead to a more efficient exploration of the space of graphs, which is useful when we are restricted in computational time. This feature of EPP and PEPP led us to consider the FINCS algorithm instead of "traditional" MCMC approaches (for example the MC^3 approach in Fouskakis et al. (2015), or Small World MCMC with modified proposals as in Guan et al. (2006)).

In the current development, EPP and PEPP are not feasible for higher number of nodes, due to the incremental computational cost of the importance

sampling estimation of the Bayes factors. We are currently investigating the extension to C++ routines that will significantly reduce the computational burden for the calculation of the Bayes factor in (12), following the lines suggested in Carvalho and Scott (2008). Finally, we are currently interested in extending our methodology to the covariate-adjusted graphical model selection framework of Consonni et al. (2017).

Acknowledgments

We wish to thank the Associate Editor and the Reviewer for comments that greatly strengthened the paper.

References

Altomare, D., Consonni, G., and La Rocca, L. (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics*, 69:478–487.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32:870–897.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

Bhadra, A. and Mallick, B. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics*, 69:447–457.

Carvalho, C. and Scott, J. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17:790–808.

Carvalho, C. and Scott, J. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96:497–512.

Castelletti, F., Consonni, G., Marco, D. V., and Peluso, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Analysis*, 13:1235–1260.

Consonni, G., Fouskakis, D., Ntzoufras, I., and Liseo, B. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13:627–679.

Consonni, G., La Rocca, L., and Peluso, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics*, 44:741–764.

Consonni, G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23:332–353.

Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer.

Dawid, P. and Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 3:1272–1317.

Dobra, A., Jones, B., Hans, C., Nevins, J., and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Mult. Anal*, 90:126–212.

Fitch, M., Beatrix, J. M., and Massam, H. (2014). The performance of covariance selection methods that consider decomposable models only. *Bayesian Analysis*, 9:659–684.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, 10:75–107.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). Power-expected-posterior priors for generalized linear models. *Bayesian Analysis*, 13:721–748.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Guan, Y., Fleissner, R., Joyce, P., and Krone, S. M. (2006). Markov chain monte carlo in small worlds. *Statistical Computing*, 16:193–202.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *J. Statist. Sci*, 1:46–60.

Iwaki, K. (1997). Posterior expected marginal likelihood for testing hypotheses. *Journal of Economics, Asia University*, 21:105–134.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Mohammadi, R. and Wit, E. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10:109–138.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:99–138.

O'Hagan, A. and Forster, J. (2004). *Bayesian Inference. Kendall's Advanced Theory of Statistics*. Arnold.

Pérez, J. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89:491–511.

Peters, J. and Buhlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228.

Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110:159–174.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2:37–63.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87:99–122.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein signaling networks derived from multiparameter single-cell data. *Science*, 308:523– 529.

Sohn, K.-A. and Kim, S. K. (2012). Joint estimation of structured sparsity and output structure in multiple - output regression via inverse - covariance regularization. *In Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 22:1081 – 1089.

Spiegelhalter, D. J. and Smith, A. F. M. (1980). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44:377–387.

Wytock, M. and Kolter, Z. (2013). Sparse Gaussian conditional random fields. algorithms, theory, and application to energy forecasting. *In Proceedings of the 30th International Conference on Machine Learning*, 28:1265–1273.

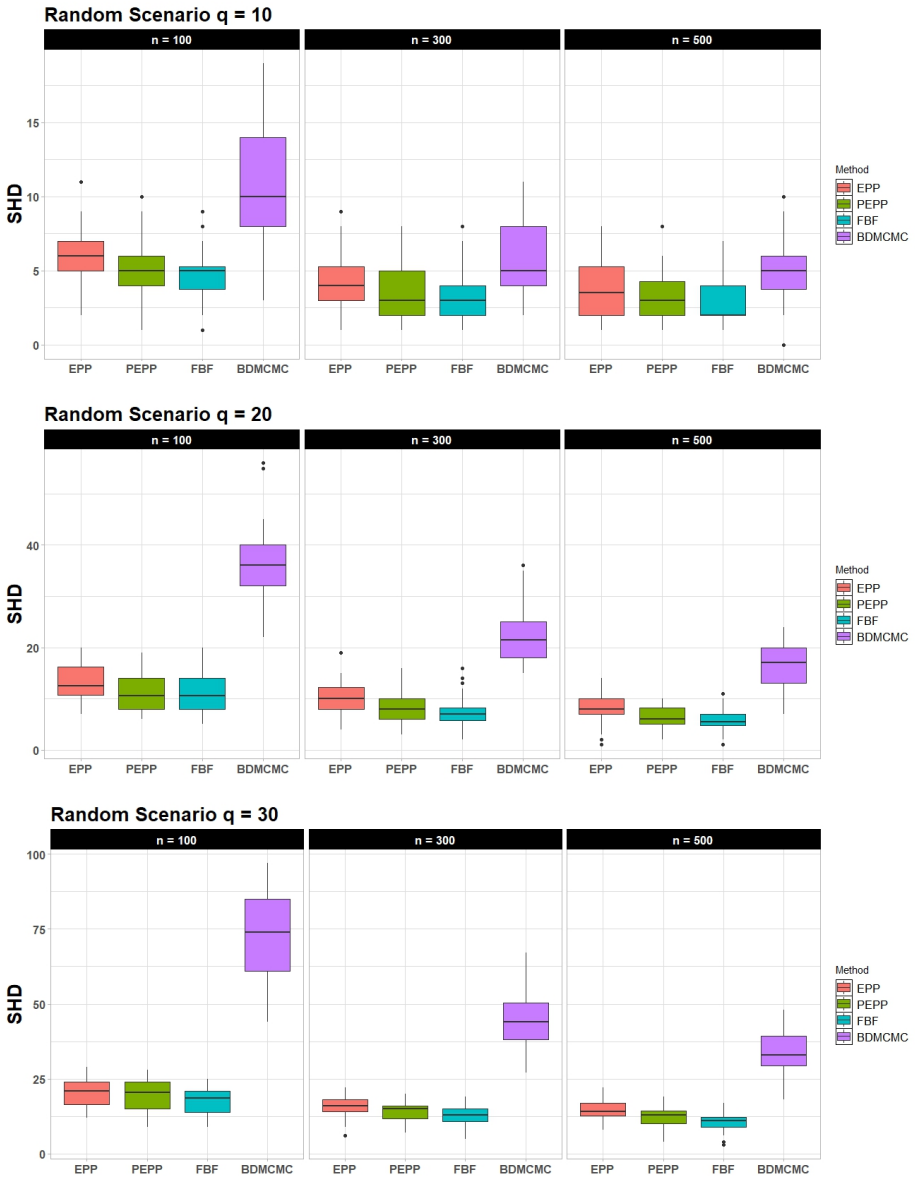


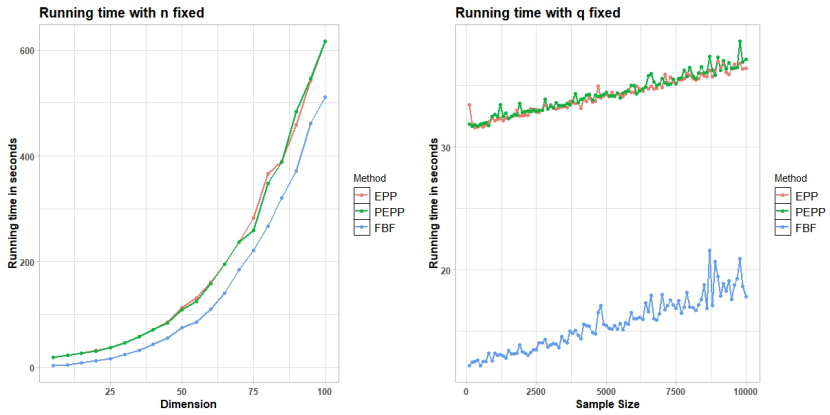
Figure 1: Simulation study under Random Scenario. Structural Hamming distances between estimated and true graphs, over 40 datasets, for number of nodes $q = \{10, 20, 30\}$ and sample size $n = \{100, 300, 500\}$. Performances are measured for EPP, PEPP, FBF and BDMCMC.

Case Studies			Approaches			
Scenario	q	n	EPP	PEPP	FBF	BDMCMC
Random	10	100	0.31 (0.05)	0.47 (0.06)	**0.51 (0.06)**	0.50 (0.02)
		300	0.54 (0.05)	0.65 (0.03)	**0.71 (0.03)**	0.66 (0.02)
		500	0.59 (0.07)	0.68 (0.05)	**0.74 (0.04)**	0.70 (0.02)
	20	100	0.25 (0.02)	**0.48 (0.02)**	0.41 (0.03)	0.36 (0.01)
		300	0.47 (0.02)	0.62 (0.01)	**0.66 (0.02)**	0.50 (0.01)
		500	0.59 (0.02)	0.71 (0.01)	**0.76 (0.01)**	0.59 (0.01)
	30	100	0.23 (0.01)	**0.42 (0.01)**	0.37 (0.02)	0.31 (0.01)
		300	0.44 (0.02)	0.55 (0.01)	**0.59 (0.01)**	0.44 (0.01)
		500	0.51 (0.01)	0.60 (0.01)	**0.68 (0.01)**	0.50 (0.01)

Table 1: Sumulated data. Means of F_1-score (variances in parentheses) under the Random Scenario.

Case Studies			Approaches			
Scenario	q	n	EPP	PEPP	FBF	BDMCMC
Star	10	100	0.33 (0.06)	**0.57 (0.05)**	**0.57 (0.05)**	0.54 (0.01)
		300	0.55 (0.05)	0.64 (0.03)	**0.66 (0.03)**	0.63 (0.01)
		500	0.62 (0.04)	0.71 (0.01)	**0.78 (0.01)**	0.73 (0.01)
	20	100	0.25 (0.03)	**0.47 (0.02)**	0.41 (0.03)	0.37 (0.00)
		300	0.45 (0.04)	0.60 (0.04)	**0.63 (0.04)**	0.51 (0.00)
		500	0.61 (0.03)	0.72 (0.02)	**0.75 (0.02)**	0.61 (0.01)
	30	100	0.22 (0.03)	**0.38 (0.02)**	0.29 (0.04)	0.29 (0.00)
		300	0.48 (0.02)	0.62 (0.01)	0.64 (0.01)	0.45 (0.00)
		500	0.52 (0.01)	0.64 (0.01)	**0.68 (0.01)**	0.50 (0.00)

Table 2: Simulated data. Means of F_1-score (variances in parentheses) under the Star Scenario.



(a) q (b) n

Figure 2: Simulated data. Computational time (in seconds) of 500 iterations of EPP, PEPP and FBF, as a function of q for $n = 500$ (left panel) and as a function of the sample size n for a fixed number of nodes $q = 20$ (right panel).

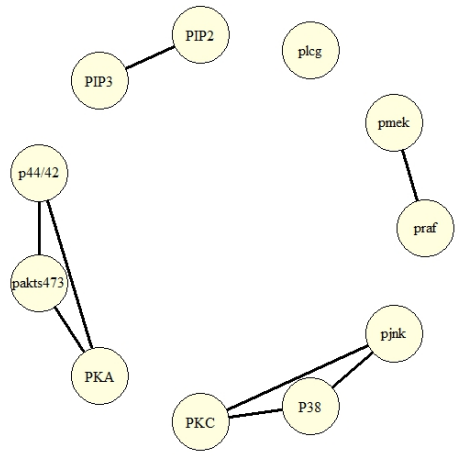


Figure 3: Protein Signalling data. Estimated median probability graphs under the first experimental condition, using PEPP approach.

Dataset	EPP	FBF	BDMCMC
1	2	1	1
2	0	2	3
3	0	0	0
4	2	1	1
5	1	5	5
6	1	1	1
7	2	1	1
8	0	0	0
9	2	1	1

Table 3: Protein Signalling data. Structural Hamming Distances between estimated graphs, under PEPP and every alternative method, for all datasets.

Dataset	PEPP	EPP	FBF	BDMCMC
1	8	6	9	9
2	8	8	10	11
3	9	9	9	9
4	6	6	7	7
5	6	5	11	11
6	8	7	9	9
7	8	6	9	9
8	10	10	10	10
9	9	7	10	10

Table 4: Protein Signalling data. Total Number of Edges under each competing approach for all datasets.

Dataset	EPP	PEPP	FBF
3	0.96	0.97	0.3
8	0.96	0.96	0.3

Table 5: Protein Signalling data. Estimated Posterior Probability of the MAP model for EPP, PEPP and FBF, for datasets 3 and 8.