Model selection for factorial Gaussian graphical models with an application to dynamic regulatory networks

Veronica Vinciotti
Mathematics, Brunel University, Uxbridge UB8 3PH, U.K.

Luigi Augugliaro
Statistics, University of Palermo, Viale delle Scienze, 90128 Palermo, Italy

Antonino Abbruzzo
Statistics, University of Palermo, Viale delle Scienze, 90128 Palermo, Italy

Ernst C. Wit[1]
Johann Bernoulli Institute, University of Groningen, 9747 AG Groningen, Netherlands

Abstract

Factorial Gaussian graphical Models (fGGMs) have recently been proposed for inferring dynamic gene regulatory networks from genomic high-throughput data. In the search for true regulatory relationships amongst the vast space of possible networks, these models allow the imposition of certain restrictions on the dynamic nature of these relationships, such as Markov dependencies of low order – some entries of the precision matrix are a priori zeros – or equal dependency strengths across time lags – some entries of the precision matrix are assumed to be equal. The precision matrix is then estimated by l_1-penalised maximum likelihood, imposing a further constraint on the absolute value of its entries, which results in sparse networks. Selecting the optimal sparsity level is a major challenge for this type of approaches. In this paper, we evaluate the performance of a number of model selection criteria for fGGMs by means of two simulated regulatory networks from realistic biological processes. The analysis reveals a good performance of fGGMs in comparison with other methods for inferring dynamic networks and of the KLCV criterion in particular for model selection. Finally, we present an application on a high-resolution time-course microarray data from the *Neisseria meningitidis* bacterium, a causative agent of life-threatening infections such as meningitis. The methodology described in this paper is implemented in the R package `sglasso`, freely available at CRAN, `http://CRAN.R-project.org/package=sglasso`.

[1]corresponding author: e.c.wit@rug.nl

1 Introduction

Networks are an important paradigm to describe genomic processes. The gene-regulatory system, for example, is a complex and dynamic process with many potential and continuously interacting components. Networks untangle this system in two constituting parts, namely substrates and functional dynamic relationships between those substrates. Decreasing costs of genomic measurement technologies have made it possible to observe genomic systems at high temporal resolutions. This enables investigating organisms different from typical model organisms. In this paper, we focus on the gene-regulatory system of *Neisseria meningitidis*. This bacterium is often referred to as *meningococcus* and can cause meningitis (Ryan et al., 2010). *Neisseria meningitidis* is a major cause of illness and death during childhood in industrialized countries and has been responsible for epidemics in Africa and in Asia (Genco et al., 2010).

One important direction in systems biology is to discover gene regulatory networks from transcriptional data based on the observed mRNA levels of large numbers of genes. The main goal of gene transcription is the production of mRNA that is translated by ribosomes to make proteins. Each mRNA can be translated several times by a ribosome in order to make proteins. This is done until mRNA reaches the end of its life-span. The network of gene regulation can be very complex, with one regulatory protein controlling genes that produce other regulators that in turn control other genes. Gene regulatory network models can be represented as directed or undirected graphs, where nodes are the elements, such as DNA, RNA or proteins, and the directed or undirected edges from one node to another represent the corresponding interaction, such as activation, repression or translation. The process is inherently dynamic, so time-course expression experiments using microarray or RNA-seq technologies are often conducted to infer temporal interactions between genes.

Dynamic Bayesian network models (Grzegorczyk and Husmeier, 2011) have been proposed to model gene-regulatory networks for circadian regulation (Aderhold et al., 2014). The computational complexity of such models prevents their use in an exploratory setting. Simpler and faster approaches have been developed for large scale networks, such as Rhein and Strimmer (2007). Recent works on Gaussian Graphical Models (GGM) constrained with the ℓ_1-penalty function(Meinshausen and Bühlmann, 2006; Friedman et al., 2008) have spurred new developments in fast methods for large genomic network structure learning (Abegaz and Wit, 2013). Most inference methods of graphical models do not allow for borrowing strength across edges. Dynamic networks, however, naturally suggest various forms of "network persistence", which can improve network identification, particularly in the case of small samples. The factorial Gaussian graphical model (fGGM) of Wit and

Abbruzzo (2015) is developed to this aim and is the modelling framework that we consider in this paper.

One of the bottlenecks in current network identification methods is the issue of model selection in such penalized graphical models. Although some knowledge exist on the optimal asymptotic regime of the tuning parameter (Bühlmann and Van De Geer, 2011), little is known for small numbers of observations. Foygel and Drton (2010) proposed an extended BIC for graphical models, which has nice asymptotic consistency properties, but unknown behaviour for small samples. Liu et al. (2010) developed a stability selection method for model identification by means of resampling, which is particularly suitable for moderate numbers of variables. However, for a small number of samples the method is rather unstable, whereas for slightly larger number of variables, it becomes computationally expensive.

In this paper, we address these issues in the more general context of penalized GGM with generic symmetry restrictions on the entries of the concentration matrix (Højsgaard and Lauritzen, 2008). The fGGM can be seen as a special case of this new class of graphical model. In Section 2, we briefly review the fGGM and introduce the ℓ_1-penalized estimator of a GGM with symmetry restrictions. We derive the asymptotic properties of this estimator and discuss the computational aspects of the estimation, for which we propose a new cyclic coordinate descent algorithm. In Section 3, we review the available model selection criteria for this class of models. We propose a suitably re-scaled version of the KLCV criterion of (Vujačić et al., 2015) and a new criterion inspired by the covariance penalty theory of (Efron, 1986, 2004). The model selection criteria are evaluated on simulated data from two realistic biological scenarios and compared with existing ones. Finally, in Section 4, we apply the proposed methodologies to the inference of the underlying dynamic regulatory network in *Neisseria meningitidis*.

2 Gene regulatory network model

2.1 Factorial Gaussian graphical model

The fGGM (Wit and Abbruzzo, 2015) is suitable for modeling longitudinal multivariate data observed at T time points. For example, we may be interested in understanding how the system of interactions among several genomic components evolves in response to environmental or internal stimuli. Let $Y_t = (Y_1^t, \ldots, Y_p^t)^\top$ be a p-dimensional random variable at time point t and assume that the Tp-dimensional random vector $Y = (Y_1^\top, \ldots, Y_T^\top)^\top$ follows a multivariate normal distribution with zero expectation and covariance matrix Σ. Under this setting the concentration ma-

Table 1: Some equality constrains on the elements of the submatrices $\Theta_{t(t+h)}$ usable to specify the fGGM model.

self-self cond. dependence at lag h			network cond. dependence at lag h		
i.	$\theta_{ii}^{th} = 0$	zero effect	i.	$\theta_{ij}^{th} = 0$	zero effect
ii.	$\theta_{ii}^{th} = s^h$	constant effect	ii.	$\theta_{ij}^{th} = n^h$	constant effect
iii.	$\theta_{ii}^{th} = s^{th}$	time effect	iii.	$\theta_{ij}^{th} = n^{th}$	time effect
iv.	$\theta_{ii}^{th} = s_i^h$	unit effect	iv.	$\theta_{ij}^{th} = n_{ij}^h$	unit effect
v.	$\theta_{ii}^{th} = s_i^{th}$	interaction effect	v.	$\theta_{ij}^{th} = n_{ij}^{th}$	interaction effect

trix, i.e. $\Theta = \Sigma^{-1}$, admits the following natural block decomposition,

$$\Theta = \begin{pmatrix} \Theta_{11} & \Theta_{12} & \dots & \Theta_{1T} \\ \Theta_{12}^\top & \Theta_{22} & \dots & \Theta_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \Theta_{1T}^\top & \Theta_{2T}^\top & \dots & \Theta_{TT} \end{pmatrix}, \tag{1}$$

where Θ_{tt} gives information about the conditional dependence structure among the p random variables at the time t, and $\Theta_{t(t+h)}$ gives information about the conditional dependence structure between Y_t and Y_{t+h}. The fGGM model is based on a natural interpretation of the elements of the submatrices $\Theta_{t(t+h)} = (\theta_{ij}^{th})$; the diagonal elements θ_{ii}^{th} are called self-self conditional dependences at temporal lag h and represent the (negative) self-similarity of a given random variable across different time points. The off-diagonal elements θ_{ij}^{th} are the conditional dependencies among the p random variables with time lag h. Similarly to the factorial analysis of variance model, the fGGM is totally specified by imposing a set of equality constraints to the elements of Θ. Table 1 reports some of the possible restrictions that the applied researcher may use to specify $\Theta_{t(t+h)}$ in the context of dynamic networks. Other kinds of restrictions can also be used depending on the researcher's objectives. The algorithm that we propose in Section 2.4, is developed for estimating structured penalized concentration matrices with generic linear equality constraints.

From a modelling point of view, these symmetric restrictions have two important consequences. First, we gain a better comprehension of the dynamic conditional dependence structure among the p underlying substrates. Second, we obtain a more parsimonious representation of the model reducing the total number of parameters that we need to estimate. To gain more insight, let us consider a simple example in which we have only two time points. In this case the block structure (1)

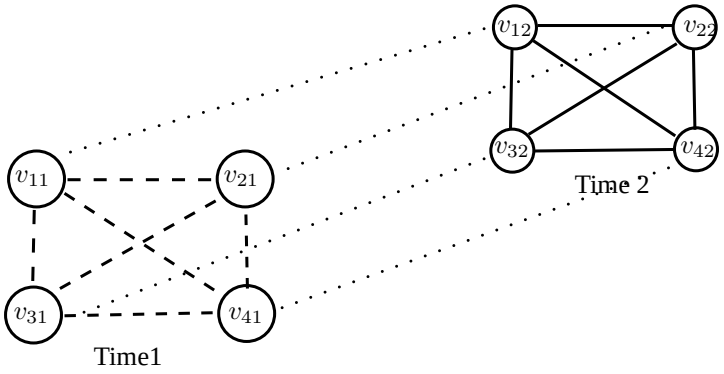


Figure 1: Example of a fGGM with four vertices measured across two time points

is equal to

$$\Theta = \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12}^{\top} & \Theta_{22} \end{pmatrix}.$$

The graphs associated to the diagonal submatrices Θ_{11} and Θ_{22}, denoted as $G_{tt} = (V, E_{tt})$ where $V = \{1, \ldots, p\}$ and $E_{tt} = \{(i, j) \mid \theta_{ij}^{t0} \neq 0\}$, give us information about the conditional independence structure among the p random variables at the t-th time point. The graph $G_{12} = (V \times V, E_{12})$, where $E_{12} = \{(i, j) \mid \theta_{ij}^{11} \neq 0\}$ provides information about the conditional dependence between Y_i^1 and Y_j^2. Figure 1 shows a fGGM specified using time effects for the self-self conditional effects at lag zero and constant effect for the conditional dependence at lag zero. Graph G_{12} is modeled using a constant effect for the self-self conditional effects at lag one and zero effect for the conditional dependence. The fGGM can be seen as a special case of the model proposed in Højsgaard and Lauritzen (2008), which is a Gaussian graphical model with generic symmetric restrictions on the elements of the concentration matrix.

2.2 ℓ_1-penalized GGM with symmetry restrictions

In the previous section, we defined an important class of graphical models by considering specific equality constraints on the concentration matrix. These constraints lead to a considerable reduction in the number of parameters to be estimated. However, dynamic genetic graphs are usually sparse, which means that few vertices will be connected. For this reason we focus our statistical inference on the maximum likelihood estimation subject to an ℓ_1-norm penalty on the concentration matrix to induce sparsity. The advantage of the ℓ_1-norm is that it is the only convex ℓ_q norm

that induces sparsity. Exact zeros will be induced for $q \leq 1$ only, while the optimization problem is convex for $q \geq 1$, which makes it feasible for high-dimensional problems (Banerjee et al., 2008).

We now describe the statistical inference for sparse Gaussian graphical models with generic equality constraints on the concentration matrix. This class of model contains, as special cases, the ℓ_1-penalized Gaussian graphical models, obtained when no restrictions are imposed on the concentration matrix, and the fGGM previously described. Following the approach proposed in Højsgaard and Lauritzen (2008), the GGM with symmetry restrictions is defined specifying the concentration matrix as follow

$$\Theta = \sum_{m=1}^{M} \theta_m T_m, \tag{2}$$

where Θ is a matrix of dimension $K \times K$, with $K = pT$, and θ_m, $m = 1, \ldots, M$, are the model parameters (e.g. $M = 1$ if all the free parameters in the matrix are set to be equal) and $T_m = (t_{ij}^m)$ are matrices with entries t_{ij}^m equal to zero or one, identifying the location of the θ_m parameters in the precision matrix. To gain more insight into the restriction (2), we consider the following simple example. Suppose that we have two genes and two time points, i.e. $K = 4$, and that the concentration matrix is specified as follows

$$\begin{aligned}
\Theta &= \theta_1 \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) + \theta_2 \left(\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right) + \theta_3 \left(\begin{array}{cc|cc} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right) = \\
&= \theta_1 T_1 + \theta_2 T_2 + \theta_3 T_3.
\end{aligned}$$

In other words, we are modeling the two submatrices at lag zero, i.e. Θ_{11} and Θ_{22}, in the same way: constant effect for both diagonal and off-diagonal entries. These equality constraints are introduced in our model by T_1 and T_2. The sub matrix at lag one is modeled using the constant effect for the diagonal entries and the zero effect for the off-diagonal entries. This equality constraint is specified by T_3. The specification (2) shows that the restricted concentration matrix lies in a linear subspace of the space of the symmetric positive-definite matrices. Denoting by S the empirical covariance matrix, the log-likelihood function can be written as

$$\ell(\theta) = \frac{N}{2}\{r^T \theta - b(\theta)\}, \tag{3}$$

where $r = (r_1, \ldots, r_M)^T$, $r_m = -\mathrm{tr}\,(ST_m)$, $\theta = (\theta_1, \ldots, \theta_M)$, $b(\theta) = -\log|\Theta|$ and N is the sample size. Using expression (3) the proposed ℓ_1-penalized estimator can be

formally defined as

$$\hat{\theta}^{\rho} = \text{argmax}_{\theta}\, \ell(\theta) - \rho \sum_{m=1}^{M} w_m |\theta_m|, \tag{4}$$

where $w_m = \sum_{ij} t_{ij}^m$ are fixed weights used to consider the effects coming from the number of non-zero entries in each matrix T_m. As a consequence of the restriction (2), we have that $\widehat{\Theta}_{\rho} = \sum_{m=1}^{M} \hat{\theta}_m^{\rho} T_m$. In the following of this paper we call the estimator (4) the structured graphical lasso (sglasso) in order to emphasize that it is equal to a graphical lasso estimator applied to estimate the structured precision matrix (2).

2.3 Asymptotic properties

In this section, we study the asymptotic behaviour of the proposed estimator (4). Before dealing with the technical details, we fix the necessary notation. First of all, in this section we shall drop the dependence of the estimator on the tuning parameter ρ. Furthermore, the true parameter vector is denoted by $\theta^{\star}$ and the corresponding concentration and covariance matrices are denoted as $\Theta^{\star}$ and $\Sigma^{\star}$, respectively. By $\mathscr{S} = \{m : \theta_m^{\star} \neq 0\}$ we denote the set of indices identifying the true parameters different from zero and we let $s = \sum_{m \in \mathscr{S}} w_m$. The set of M equality constraints used in definition (2) can be divided into two disjoint subsets denoted as D and O. The first one contains the indices identifying the equality constraints on the diagonal entries of the matrix Θ, i.e. $n \in D$ if and only if T_n is a diagonal matrix. The set O contains the remaining indices, formally $m \in O$ if and only if T_m is not a diagonal matrix. For each $n \in D$ we define the set $V_n = \{i : t_{ii}^n = 1\}$. It easy to see that the cardinality of the set V_n is equal to w_n and that $K = \sum_{n \in D} w_n$. In the same way, for each $m \in O$ we define the sets $E_m = \{(i,j) : t_{ij}^m = 1\}$ and $V_m = \{i :$ exists a j such that $(i,j) \in E_m\}$. Finally, we let $\overline{K} = \max_{n \in D} w_n \vee \max_{m \in O} |V_m|$.

Theorem 1 is based on the assumption that exist two positive constants, denoted as $\underline{k}$ and $\overline{k}$, such that

$$\underline{k} \leq \underline{\lambda}(\Sigma^{\star}) \leq \overline{\lambda}(\Sigma^{\star}) \leq \overline{k}, \tag{5}$$

where $\underline{\lambda}(\Sigma^{\star})$ and $\overline{\lambda}(\Sigma^{\star})$ are the smallest and largest eigenvalues of $\Sigma^{\star}$, respectively. This assumption is standard and used in Bickel and Levina (2008), Rothman et al. (2008) and Guo et al. (2011). It guarantees that the true concentration matrix exists and is well-conditioned.

Theorem 1. *If* $\rho \asymp \sqrt{\log \max_{m \in O} |V_m|/N}$ *and under the assumption (5), the sglasso estimator is such that*

$$\|\hat{\theta} - \theta^\star\|_2 = O_p\left(\sqrt{\frac{(K+s)\log \overline{K}}{N}}\right).$$

Proof. The proof is based on the approach originally proposed in Rothman et al. (2008). First observe that $\hat{\Delta} = \hat{\theta} - \theta^\star$ can be defined as minimizer of the following function

$$D(\Delta) = \sum_{m=1}^{M} \Delta_m \text{tr}\,(ST_m) - (\log\det\Theta - \log\det\Theta^\star) + \rho \sum_{m=1}^{M} w_m(|\theta_m^\star + \Delta_m| - |\theta_m^\star|).$$

Let $r_N = (K+s)\log\overline{K}/N$ and consider the closed ball $B_C = \{\Delta : (\sum_{m=1}^{M} w_m \Delta_m^2)^{1/2} \leq C\sqrt{r_N}\}$, where C is a positive constant. Since $D(\Delta)$ is a convex function and $D(\hat{\Delta}) \leq D(0) = 0$, if we show that the function $D(\Delta)$ restricted to the boundary ∂B_C is positive, we can conclude that $\hat{\Delta} \in B_C$, which means that

$$\|\hat{\theta} - \theta^\star\|_2 \leq \|\hat{\Theta} - \Theta^\star\|_F = \left(\sum_{m=1}^{M} w_m \hat{\Delta}_m^2\right)^{1/2} \leq C\sqrt{r_N} \to 0.$$

As shown in Rothman et al. (2008), using assumption (5) the second term in $D(\Delta)$ can be bounded as follows

$$\begin{aligned} \log\det\Theta - \log\det\Theta^\star &\leq \text{tr}\,\{\Sigma^\star(\Theta - \Theta^\star)\} - \frac{1}{4}\underline{k}^2\|\Theta - \Theta^\star\|_F^2 \\ &= \sum_{m=1}^{M} \Delta_m \text{tr}\,(\Sigma^\star T_m) - \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2, \end{aligned}$$

then we have that

$$D(\Delta) \geq \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2 + \sum_{m=1}^{M} \Delta_m \text{tr}\,\{(S - \Sigma^\star)T_m\} + \rho \sum_{m=1}^{M} w_m(|\theta_m^\star + \Delta_m| - |\theta_m^\star|).$$

To handle the term $\sum_{m=1}^{M} \Delta_m \text{tr}\,\{(S - \Sigma^\star)T_m\}$, observe that

$$\sum_{m=1}^{M} \Delta_m \text{tr}\,\{(S - \Sigma^\star)T_m\} = \sum_{m \in O} \Delta_m \text{tr}\,\{(S - \Sigma^\star)T_m\} + \sum_{n \in D} \Delta_n \text{tr}\,\{(S - \Sigma^\star)T_n\}. \quad (6)$$

Using Lemma A.3. in Bickel and Levina (2008) the two terms in (6) can be easily bounded as follows

$$
\begin{aligned}
|\sum_{m\in O}\Delta_m \mathrm{tr}\left\{(S-\Sigma^\star)T_m\right\}| &\leq \sum_{m\in O}|\Delta_m|\sum_{(i,j)\in E_m}|s_{ij}-\sigma^\star_{ij}| \\
&\leq \sum_{m\in O}|\Delta_m|\left(w_m \max_{(i,j)\in E_m}|s_{ij}-\sigma^\star_{ij}|\right) \\
&\leq \sum_{m\in O} w_m|\Delta_m| C_1\sqrt{\frac{\log|V_m|}{N}} \\
&\leq C_1\sqrt{\frac{\log\max_{m\in O}|V_m|}{N}}\sum_{m\in O} w_m|\Delta_m| \\
&\leq C_1\sqrt{\frac{\log\max_{m\in O}|V_m|}{N}}\sum_{m=1}^{M} w_m|\Delta_m| \\
&= C_1\sqrt{\frac{\log\max_{m\in O}|V_m|}{N}}\left(\sum_{m\in\mathscr{S}} w_m|\Delta_m|+\sum_{m\notin\mathscr{S}} w_m|\Delta_m|\right),
\end{aligned}
$$

and

$$
\begin{aligned}
|\sum_{n\in D}\Delta_n \mathrm{tr}\left\{(S-\Sigma^\star)T_n\right\}| &\leq \sum_{n\in D}|\Delta_n|\sum_{i\in V_n}|s_{ii}-\sigma^\star_{ii}| \leq \sum_{n\in D}|\Delta_n|\left(w_n\max_{i\in V_n}|s_{ii}-\sigma^\star_{ii}|\right) \\
&\leq \sum_{n\in D} w_n|\Delta_n| C_2\sqrt{\frac{\log w_n}{N}} \leq C_2\sqrt{\frac{\log\max_{n\in D} w_n}{N}}\sum_{n\in D} w_n|\Delta_n|.
\end{aligned}
$$

Using the reverse triangular inequality, the last term in $D(\Delta)$ can be easily bounded as follows

$$
\sum_{m=1}^{M} w_m(|\theta_m+\Delta_m|-|\theta^\star_m|) \geq \sum_{m\notin\mathscr{S}} w_m|\Delta_m| - \sum_{m\in\mathscr{S}} w_m|\Delta_m|.
$$

Combining all the previous inequalities we have

$$
\begin{aligned}
D(\Delta) \geq\ & \frac{1}{4}\underline{k}^2\sum_{m=1}^{M} w_m\Delta_m^2 - \left(C_1\sqrt{\frac{\log\max_{m\in O}|V_m|}{N}}-\rho\right)\sum_{m\notin\mathscr{S}} w_m|\Delta_m| \\
& - \left(C_1\sqrt{\frac{\log\max_{m\in O}|V_m|}{N}}+\rho\right)\sum_{m\in\mathscr{S}} w_m|\Delta_m| - C_2\sqrt{\frac{\log\max_{n\in D} w_n}{N}}\sum_{n\in D} w_n|\Delta_n|.
\end{aligned}
$$

Letting $\rho = \varepsilon C_1 \sqrt{\log \max_{m\in O} |V_m|/N}$, with $\varepsilon \geq 1$, the second term in the previous decomposition is non-negative then it can be removed obtaining

$$\begin{aligned} D(\Delta) &\geq \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2 - C_1 \sqrt{\frac{\log \max_{m\in O} |V_m|}{N}} (1+\varepsilon) \sum_{m\in\mathscr{S}} w_m |\Delta_m| \\ &\quad - C_2 \sqrt{\frac{\log \max_{n\in D} w_n}{N}} \sum_{n\in D} w_n |\Delta_n| \\ &\geq \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2 - C_1 \sqrt{\frac{\log \overline{K}}{N}} (1+\varepsilon) \sum_{m\in\mathscr{S}} w_m |\Delta_m| - C_2 \sqrt{\frac{\log \overline{K}}{N}} \sum_{n\in D} w_n |\Delta_n|. \end{aligned}$$

Noting that

$$\begin{aligned} \sum_{m\in\mathscr{S}} w_m |\Delta_m| &\leq \sqrt{\sum_{m\in\mathscr{S}} w_m} \sqrt{\sum_{m\in\mathscr{S}} w_m \Delta_m^2} \leq \sqrt{s} \sqrt{\sum_{m=1}^{M} w_m \Delta_m^2} \\ \sum_{n\in D} w_n |\Delta_n| &\leq \sqrt{\sum_{n\in D} w_n} \sqrt{\sum_{n\in D} w_n \Delta_n^2} \leq \sqrt{K} \sqrt{\sum_{m=1}^{M} w_m \Delta_m^2}, \end{aligned}$$

we can write that

$$\begin{aligned} D(\Delta) &\geq \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2 - C_1 \sqrt{\frac{s \log \overline{K}}{N}} (1+\varepsilon) \sqrt{\sum_{m=1}^{M} w_m \Delta_m^2} - C_2 \sqrt{\frac{K \log \overline{K}}{N}} \sqrt{\sum_{m=1}^{M} w_m \Delta_m^2} \\ &\geq \frac{1}{4}\underline{k}^2 \sum_{m=1}^{M} w_m \Delta_m^2 - \sqrt{\frac{(K+s)\log \overline{K}}{N}} \{C_1(1+\varepsilon) + C_2\} \sqrt{\sum_{m=1}^{M} w_m \Delta_m^2}. \end{aligned}$$

When we evaluate the function $D(\Delta)$ on ∂B_C, the previous inequality can be written as follows

$$D(\Delta) \geq C^2 r_N \frac{1}{4}\underline{k}^2 - C r_N \{C_1(1+\varepsilon) + C_2\} = C^2 r_N \left(\frac{1}{4}\underline{k}^2 - \frac{C_1(1+\varepsilon) + C_2}{C} \right) > 0,$$

for C sufficiently large. □

The importance of this theorem is that on some mild assumptions, the sglasso estimator is consistent.

2.4 A cyclic coordinate descent algorithm

From a computational point of view, the maximization problem used in definition (4) is a challenging task. One method is the LogdetPPA algorithm, which

combines a proximal point algorithm (PPA) inside a preconditioned conjugate gradient solver needed for Newton's method (Wang et al., 2010). This solver was used in Wit and Abbruzzo (2015) to deal with fGGM. For a single tuning parameter, this method can be quite efficient, but solving an entire solution path for a range of tuning parameters is non-trivial. Instead, we propose a cyclic coordinate descent method. The main idea underlying this family of algorithms is to choose, at each iteration, an index and then optimize the objective function with respect to the corresponding parameter keeping all the remaining indexes fixed. This kind of algorithm was originally proposed for lasso regression models in Wu and Lange (2008) and studied in more detail in Friedman et al. (2007). In Friedman et al. (2010), the cyclic coordinate descent algorithm is extended to ℓ_1-penalized generalized linear models.

In order to simplify our notation, in what follows we denote by $\ell(\theta_m)$ the log-likelihood function (3) seen as a function of the parameter θ_m while the remaining parameters are kept fixed at the current values. Suppose that we have the sglasso estimator $\hat{\theta}^{\rho'}$ for a given value of the tuning parameter, say ρ', and we want to compute a new estimate for $\rho < \rho'$. If ρ is close enough to ρ', then $\hat{\theta}^{\rho}$ is in a neighborhood of $\hat{\theta}_m^{\rho'}$ and, consequently, we can approximate $\ell(\theta_m)$ by a standard Taylor expansion. Formally

$$\begin{aligned}
\ell(\theta_m) &\approx \ell(\hat{\theta}^{\rho'}) - \rho \sum_{n \neq m}^{M} w_n |\hat{\theta}_n^{\rho'}| + \\
&\quad + \frac{\partial \ell(\hat{\theta}^{\rho'})}{\partial \theta_m}(\theta_m - \hat{\theta}_m^{\rho'}) + \frac{1}{2}\frac{\partial^2 \ell(\hat{\theta}^{\rho'})}{\partial \theta_m^2}(\theta_m - \hat{\theta}_m^{\rho'})^2 - \rho w_m |\theta_m| \\
&= C(\hat{\theta}^{\rho'}) + \frac{1}{2}\frac{\partial^2 \ell(\hat{\theta}^{\rho'})}{\partial \theta_m^2}(\theta_m - \hat{\vartheta}_m^{\rho'})^2 - \rho w_m |\theta_m|, \qquad (7)
\end{aligned}$$

where $C(\hat{\theta}^{\rho'}) = \ell(\hat{\theta}^{\rho'}) - \rho \sum_{n \neq m}^{M} w_n |\hat{\theta}_n^{\rho'}| - \frac{1}{2}\{\partial^2 \ell(\hat{\theta}^{\rho'})/\partial \theta_m^2\}^{-1}\{\partial \ell(\hat{\theta}^{\rho'})/\partial \theta_m\}^2$ is a constant with respect to θ_m and $\hat{\vartheta}_m^{\rho'} = \hat{\theta}_m - \{\partial^2 \ell(\hat{\theta}^{\rho'})/\partial \theta_m^2\}^{-1}\partial \ell(\hat{\theta}^{\rho'})/\partial \theta_m$. Using approximation (7), the original maximization problem can be locally substituted by the simpler problem

$$\min_{\theta_m \in \mathbb{R}} \frac{1}{2} I_m(\hat{\theta}^{\rho'})(\theta_m - \hat{\vartheta}_m^{\rho'})^2 + \rho w_m |\theta_m|, \qquad (8)$$

where $I_m(\hat{\theta}^{\rho'}) = -\partial^2 \ell(\hat{\theta}^{\rho'})/\partial \theta_m^2$ is the Fisher information for θ_m evaluated at $\hat{\theta}^{\rho'}$. The problem (8) can be solved in closed form, i.e. $\hat{\theta}_m^{\rho} = S(\hat{\vartheta}_m^{\rho'}; w_m I_m^{-1}(\hat{\theta}^{\rho'})\rho)$, where $S(x;\lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-threshold operator (Friedman et al., 2007). Using this result, the proposed cyclic coordinate descent algorithm can be

Algorithm 1 Pseudo-code of the proposed cyclic coordinate descent algorithm

1: inizialize $\hat{\theta}^{\rho}$ to a previous estimate
2: Compute $\widehat{\Theta}_{\rho} = \sum_{m=1}^{M} \hat{\theta}_m^{\rho} T_m$ and $\widehat{\Sigma}_{\rho} = \widehat{\Theta}_{\rho}^{-1}$
3: **repeat**
4: **for** $m = 1$ to M **do**
5: $\partial \ell(\hat{\theta}^{\rho}) / \partial \theta_m = \text{tr}\,\{T_m(\widehat{\Sigma}^{\rho} - S)\}$ ▷ $w_m/2$ operations
6: $I_m(\hat{\theta}^{\rho}) = \text{tr}\,\{T_m \widehat{\Sigma}^{\rho} T_m \widehat{\Sigma}^{\rho}\}$ ▷ $O(w_m^2)$ operations
7: $\vartheta_m^{\rho} = \hat{\theta}_m^{\rho} + I_m^{-1}(\hat{\theta}^{\rho})\{\partial \ell(\hat{\theta}^{\rho}) / \partial \theta_m\}$
8: $\hat{\theta}_m^{\rho} = S(\vartheta_m^{\rho}; w_m I_m^{-1}(\hat{\theta}^{\rho})\rho)$
9: $\widehat{\Sigma}_{\rho} \leftarrow \{\widehat{\Theta}_{\rho} + \hat{\theta}_m^{\rho} T_m\}^{-1}$ ▷ $O(K^2)$ operations
10: **end for**
11: **until** a convergence criterion is met

described by the pseudo-code reported in Algorithm 1. A closer look at the pseudo-code reveals that in each inner loop we need to compute the inverse of the concentration matrix in step 9. Standard algorithms require $O(K^3)$ operations, which can be prohibitive when p or T are large. However, as the inversion involves a sum of two matrices, it is possible to reduce the computational burden to only $O(K^2)$ operations, using the iterative algorithm proposed in Miller (1981). We have implemented the solver in our `R`-package `sglasso`. In particular, the fGGM described in Section 2.1, with time, unit and interaction effects, can be inferred using the function `fglasso`. At the moment, the package can handle efficiently covariance matrices of up to size 1000×1000, e.g. 100 genes across 10 time points.

3 Model selection for fGGMs

The behaviour of the factorial Gaussian graphical model estimator (4) and its corresponding precision matrix $\widehat{\Theta}_{\rho}$ is closely related to selecting the optimal value of the tuning parameter. Although we are guaranteed consistency for some asymptotic regime of ρ, as we have shown in the previous section, the choice of ρ for finite samples is less clear. The aim of this section is to provide an overview of available methods in these scenarios and to make recommendations about sensible choices in practical circumstances. We will consider the behaviour of various model selection criteria in two realistic biological scenarios to evaluate their usefulness.

Model selection ideas have centered around two main ideas (Wit et al., 2012): either minimizing the distance of the selected model to the true model or maximizing the probability of selecting the true model. The first idea centers around

the concept of Kullback-Leibler divergence,

$$KL(\hat{\Theta}_\rho|\Theta^\star) = E_{\Theta^\star}\frac{l_Y(\hat{\Theta}_\rho)}{l_Y(\Theta^\star)},$$

where the expectation is taken over Y from the true model with parameters $\Theta^\star$. The second idea is based on the probability of the true model, here interpreted as the underlying graph G, integrating out all the parameters

$$P(\hat{G}_\rho|\text{data}) = \int_\Theta \frac{P(\text{data}|\Theta,\hat{G}_\rho)P(\hat{G}_\rho)}{P(\text{data})}\, d\Theta.$$

The aim of minimizing the first Kullback-Leibler divergence is related to getting models that have good predictive properties. It is closely related to minimizing prediction errors, cross-validation and the covariance penalty theory (Efron, 1986, 2004). Maximizing the posterior model probability is related to obtaining the true underlying structure of the model, focussing on having the right features, in our case the correct links of the graph.

The above quantities involve population parameters and in order to use them in practice, they need to be estimated. The estimates are typically referred to as information criteria. As a result from a Laplace expansion in the integral in either the Kullback-Leibler or the posterior model probability, a bias correction quantity to the observed likelihood is derived, which involves a quantity that is typically referred to as the *degrees of freedom* of the model. In simple regression models, the quantity can be shown to be equal in expectation to the number of covariates in the model. This has been a powerful concept, both practically and philosophically. The immediacy of the trade-off between model fit and model complexity has made the degrees of freedom a key concept in the model selection literature.

Although it easy to define the degrees of freedom in a graphical model as the non-zero entries of the concentration matrix,

$$\text{df}_1(\hat{\Theta}_\rho) = |\{\hat{\theta}^\rho_m \mid \hat{\theta}^\rho \neq 0\}|,$$

there are several problems with this idea. Within the classical theory of linear regression models, it is well known that the degrees-of-freedom are equal to the number of covariates but for non-linear modelling procedures this equivalence is not satisfied. Furthermore, the number of parameters is an asymptotic concept and in modern applications, such as in high-dimensional graphical models, we are rarely

in this regime. In fact, the degrees of freedom in a graphical model, i.e. the difference between the observed likelihood and the expected likelihood, has the following form,

$$\begin{aligned} \mathrm{df}_2(\hat{\Theta}_\rho) &= E_{\Theta^\star}\ell_Y(\hat{\Theta}_\rho) - \ell_y(\hat{\Theta}_\rho) \\ &= -\frac{N}{2}\mathrm{tr}\,\{\hat{\Theta}_\rho(S-\Theta^{\star-1})\} \qquad (9) \\ &= -\frac{N}{2}\sum_{m=1}^{M}\hat{\theta}_m^\rho \mathrm{tr}\,\{T_m(S-\Sigma^\star)\}. \qquad (10) \end{aligned}$$

where $\ell_y(\hat{\Theta}_\rho)$ is the observed likelihood as an estimate of $E_{\Theta^\star}\ell_Y(\hat{\Theta}_\rho)$ and $\Sigma^\star$ denotes the true covariance matrix. In recent years, several authors have studied the problem of how to generalize the classical notion of degrees-of-freedom for penalized regression models. For the lasso estimator, Zou et al. (2007) developed an adaptive model selection criterion to select the regularization parameter based on a rigorous definition of degrees of freedom based on the covariance penalty theory (Efron, 1986, 2004). This approach was also used in Augugliaro et al. (2013) to derive the notion of degree-of-freedom for a differential-geometric extension of the least angle regression method (Efron et al., 2004). Since expression (3) shows that the probability density function of the Gaussian distribution with equality constraints on the precision matrix belongs to the exponential family, we can easily use Theorem 2 in Efron (1986) to derive a genuine extension of the notion of degrees-of-freedom for the proposed estimator, i.e.

$$\mathrm{gdf}(\rho) = \frac{N}{2}\sum_{m=1}^{M}\mathrm{cov}(r_m,\hat{\theta}_m^\rho) = -\frac{N}{2}\sum_{m=1}^{M}E_Y[\hat{\theta}_m^\rho \mathrm{tr}\,\{T_m(S-\Sigma^\star)\}]. \qquad (11)$$

We refer to the left-hand-side of (11) as the *generalized degrees-of-freedom* (gdf) of the proposed estimator, since it generalizes the classical notion of degree-of-freedom. It is effectively the expected value of the expression (10) and rather than expressing the quality of a particular estimator $\hat{\Theta}_\rho$, it evaluates the choice of the value ρ on average.

3.1 Information criteria for fGGM

The Kullback-Leibler divergence and the posterior model probability, as well as the concepts of degrees-of-freedom, that we introduced in this section, involve population parameters and therefore cannot be calculated on a particular dataset. Given

such dataset, so-called information criteria are estimates of either the Kullback-Leibler divergence or the posterior model probability. Among the standard approaches, the Akaike Information Criterion (AIC) (Akaike, 1973),

$$AIC(\rho) = -N\left(\log|\widehat{\Theta}_\rho| - \mathrm{tr}(\widehat{\Theta}_\rho S)\right) + 2\mathrm{df}(\widehat{\Theta}_\rho),$$

is an estimate for the former, whereas the Bayesian Information Criterion (BIC) (Schwarz, 1978),

$$BIC(\rho) = -N\left(\log|\widehat{\Theta}_\rho| - \mathrm{tr}(\widehat{\Theta}_\rho S)\right) + \mathrm{df}(\widehat{\Theta}_\rho)\log N,$$

is an estimate for the latter. Modifications and extensions of these criteria have been suggested in the literature, such as the extended BIC (Foygel and Drton, 2010), modified BIC (Gao et al., 2012) and RIC (Lysen, 2009). There are two main issues facing the use of these measures. On the one hand, the traditional definition of the degrees of freedom as the number of non-zero parameters in penalized inference is not very well defined. Shrinked parameters should probably count as fewer than a real parameter, but it is not clear by how much. On the other hand, the number of observations in genomic experiments are typically so small that the asymptotic assumptions on which the AIC and BIC are based are equally suspect.

Other estimators have been suggested in the literature, in an attempt to overcome the limitations of AIC and BIC measures. StARS (Liu et al., 2010) is based on a resampling approach, which is particularly suitable for moderate numbers of variables but becomes computationally expensive for a large number of variables and rather instable for small sample sizes. Vujačić et al. (2015) derived the following estimator of the cross-validated Kullback-Leibler divergence:

$$KLCV(\rho) = -N\left(\log|\widehat{\Theta}_\rho| - \mathrm{tr}(\widehat{\Theta}_\rho S)\right) + 2\sum_{i=1}^{N} p_i(\widehat{\Theta}_\rho, S),$$

where the cross-validated estimate of the degrees of freedom, p_i, is given by

$$p_i(\rho) = \frac{\mathrm{vec}[(\widehat{\Theta}_\rho^{-1} - S_i)\circ I_\rho]^t \mathrm{vec}[\widehat{\Theta}_\rho\{(S-S_i)\circ I_\rho\}\widehat{\Theta}_\rho]}{2N-2}, \tag{12}$$

where I_ρ is the indicator matrix, whose entry is 1 if the corresponding entry in the precision matrix $\widehat{\Theta}_\rho$ is nonzero and zero if the corresponding entry in the precision matrix is zero. This estimator is not only computationally efficient, but it takes also parameter shrinkage into account.

The original KLCV was proposed for graphical models without constraints. The equality constraints and the structural zero constraints reduce the number of parameters in the graphical model and the bias correction term can be adjusted

accordingly. The bias correction term we use in this paper is p_i^*, which is defined as

$$p_i^* = cp_i,$$

where c is the ratio of M, the number of free parameters used in the graphical model, and $K(K+1)/2$, the total number of parameters of the unconstraint graphical model.

In addition to the estimators described above, we propose a new estimator which is motivated by the covariance penalty theory of Efron (1986). In particular, we propose

$$\text{CovPen}(\rho) = -\text{N}\left(\log|\widehat{\Theta}_\rho| - \text{tr}(\widehat{\Theta}_\rho \text{S})\right) + 2\widehat{\text{gdf}}(\rho),$$

with

$$\widehat{\text{gdf}}(\rho) = \frac{N}{2}\left(\rho \sum_{m=1}^{M} w_m|\hat{\theta}_m^\rho| + \frac{\sum_{i=1}^{N} y_i^\top \widehat{\Theta}_\rho^{(-i)} y_i}{N} - K\right), \tag{13}$$

and using the notation previously introduced. The technical details of the derivation of $\widehat{\text{gdf}}(\rho)$ are placed in the Appendix.

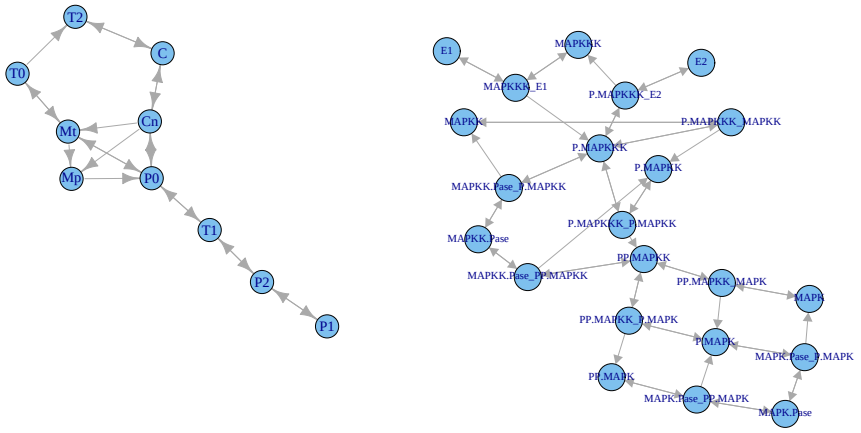
3.2 A simulation study

We perform a simulation study to compare the different model selection criteria. We consider two simulated data from realistic biological processes. In particular, we use the software COPASI 4.16 (Hoops et al., 2006) to simulate data from two biochemical networks: the Drosophila circadian clock based on the PER and TIM genes, as published in (Leloup and Goldbeter, 1999), and the MAPK cascade model from (Huang and Ferrell, 1996). The circadian network has 10 proteins, for which we simulate data over 10 hours every 6 minutes, whereas the MAPK network has 22 proteins, for which we simulate data over 17 milliseconds every 0.17 milliseconds. Figure 2 shows the two real networks, from which the data are simulated using a Gillespie algorithm (direct method).

We use the fGGM model described before, where:

i. lag zero submatrices, i.e. Θ_{tt}, are modelled using model (iv) according to Table 1;
ii. lag one submatrices, i.e. $\Theta_{t(t+1)}$, are also modelled by unit effects;
iii. the entries of the remaining submatrices are equal to zero.

We compare this model with a Time Series Chain Graphical Model (TSCGM), as proposed in Abegaz and Wit (2013) and implemented in the `SparseTSCGM` R package, and the graphical lasso (Glasso), as proposed in (Friedman et al., 2008) and



(a) (b)

Figure 2: True biochemical networks of (a) the circadian clock and (b) MAPK processes.

implemented in the `huge` R package. We select the tuning parameter using various model selection criteria and for the selected network we report: recall $\left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right)$, precision $\left(\frac{\text{TP}}{\text{TP}+\text{FP}}\right)$, True Negative Rate (TNR) and F_1 score $\left(\frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}}\right)$. For a fair comparison with `Glasso`, Table 2 reports these measures only on the lag 0 networks. The dynamic nature is not captured in the Glasso method and under all kinds of model selection methods it performs poorly (F1 scores 0-0.077 on the Circadian Clock and 0-0.262 on MAPK). The TSCGM method does account for the dynamic nature of the data and performs much better (F1 scores 0.216 - 0.242 on the Circadian Clock and 0.323-0.387 on MAPK). The fGGM performs comparably to TSCGM on the MAPK network (F1 scores 0.321 - 0.337) but better on the Circadian Clock example (F1 scores 0.364 - 0.368).

Figure 3 shows the behaviour of the model selection criteria for the fGGM model on the two datasets. KLCV shows a clear minimum and tends to lead to the selection of sparser models, compared with AIC, BIC and CovPen which lead to similar selections. On the MAPK network, KLCV and CovPen lead to the same selection, but KLCV is considerably faster as it does not need the re-estimation of the precision matrix at each fold (Vujačić et al., 2015). Overall, the simulation shows a comparable performance between fGGM and TSCGM and a good performance of KLCV as a model selection criterion for fGGM.

	Circadian Clock				MAPK			
	Recall	Precision	TNR	F_1	Recall	Precision	TNR	F_1
fGGM								
AIC	0.400	0.333	0.800	0.364	0.520	0.232	0.802	0.321
BIC	0.400	0.333	0.800	0.364	0.520	0.232	0.802	0.321
KLCV	0350	0.389	0.863	**0.368**	0.560	0.241	0.797	0.337
CovPen	0.400	0.333	0.800	0.364	0.560	0.241	0.797	0.337
TSCGM								
AIC	0.190	0.250	0.848	0.216	0.240	1.000	1.000	**0.387**
BIC	0.190	0.333	0.899	0.242	0.200	0.833	0.995	0.323
eBIC	0.190	0.286	0.873	0.229	0.200	0.833	0.995	0.323
BIC_{mod}	0.190	0.286	0.873	0.229	0.200	0.833	0.995	0.323
GIC	0.190	0.333	0.899	0.242	0.200	0.833	0.995	0.323
Glasso								
AIC	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
BIC	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
eBIC	0.000	0.000	1.000	0.000	0.780	0.152	0.500	0.255
RIC	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000
StARS	0.050	0.167	0.938	0.077	0.440	0.186	0.779	0.262

Table 2: Reconstruction performance of various model selection criteria applied to three main network reconstruction models: factorial gaussian graphical models (fGGM) used in this paper and originally proposed in Wit and Abbruzzo (2015), time series chain graphical models (TSCGM) as proposed in Abegaz and Wit (2013) and the graphical lasso (Glasso) as proposed in (Friedman et al., 2008).

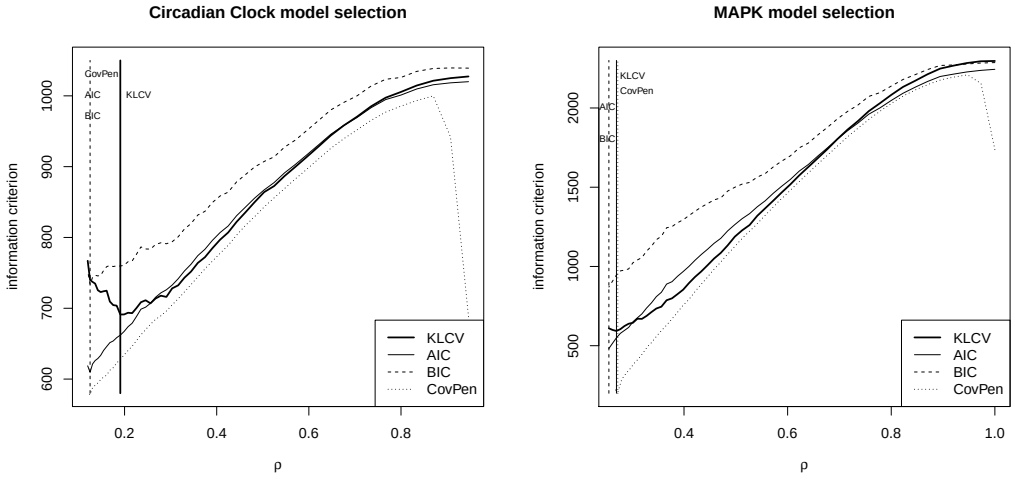


Figure 3: Model selection criteria for fGGM on data simulated from (a) the circadian clock and (b) MAPK processes.

4 Regulatory network of *Neisseria meningitidis*

We apply the methodology to microarray data from a high-resolution time-course experiment using the sequenced *Neisseria meningitidis* serogroup B strain MC58 (Tettelin et al., 2000). The expression of 2129 transcripts was determined using dendrimer labelling of the parent of the sequenced strain with established methods (Jordan and Saunders, 2009; Saunders and Davies, 2012), in rapidly growing liquid cultures at 10 minute intervals in the early and log phases of growth (0 to 130 minutes) and at 20 minute intervals thereafter (to 250 minutes). Two biological replicate cultures, grown in parallel, were sampled. In this study, we focus on 60 transcripts that have been found to be differentially expressed upon deletion of the master regulator FarR in highly replicated microarrays studies and have been validated by qPCR and gel shift assays (Nigel Saunders, unpublished), but whose regulatory mechanisms are largely unknown (See table in the supplementary material for the list of genes). For the analysis, we combine two consecutive time points into one time point, in order to increase the number of observations per time point to four. We finally scale the data to have mean zero and variance one for each protein and across all time points.

The resulting precision matrix has dimension 600×600, thus about 180,000 parameters to be estimated. To improve the comprehension of the model used to study the data set and reduce the number of parameter, we fit a fGGM specified as

follows:

1. **lag zero submatrices:** we assume that diagonal and off-diagonal entries of the lag zero submatrices are modeled by unit effects, i.e. the conditional dependence structure is persistent across all ten time points;
2. **lag one submatrices:** the diagonal and off-diagonal entries of the lag one submatrices are also modeled by unit effects;
3. the entries of the remaining submatrices are set equal to zero.

This reduces the number of parameters from about 180,000 to a manageable number less than 5,500. Furthermore, the shrinkage induced by the l_1-penalty further stabilizes the estimates. Using the KLCV measure for model selection, the optimal value of ρ is found to be 0.453

A number of analyses were conducted to evaluate the robustness of the selected network on the 60 proteins. Firstly, we performed a bootstrap analysis: for each protein, we simulated 100 bootstrap samples by adding noise to the real data at a level of variability estimated by fitting a smoothing spline to the time series data. We then fitted a factorial graphical model to the bootstrap data with $\rho = \rho^*$. Effectively, this post-analysis allowed us to explore the space of precision matrices around $\hat{\Theta}_{\rho^*}$, by showing how robust the inference is to obtaining slightly different, but equally plausible data. Overall, we found that 70.59% of the links in the lag 0 network and 21.35% of the links in the lag 1 network were found consistently in more than 50% of bootstrap samples. We have also tested the robustness of the network, by repeating the analysis on 8 replicates, i.e. aggregating the data by combining four (rather than two) consecutive time points. The precision matrix is now 300×300, thus reducing significantly the number of model parameters. The new criterion selected a network of similar sparsity and there was full agreement with the earlier analysis: the 238 lag 0 links detected from the dataset with 4 replicates were all detected by the analysis on the dataset with 8 replicates, where 53 additional links were found. Finally, we have compared the network inferred by the proposed fglasso with the network inferred by the sparse vector autoregressive approach of Abegaz and Wit (2013). From a modelling point of view, this is the closest approach to ours, as it fits an autoregressive process of order 1 under sparsity constraints. The main difference is that no equality constraints are imposed in the lag 0 and the lag 1 network. We selected the two tuning parameters so as to have the same sparsity level of our inferred lag 0 and lag 1 networks, respectively. The SparseTSCGM lag 0 network contained 24.3% of the fglasso lag 0 links and the SparseTSCGM lag 1 network contained 41.57% of the fglasso lag 1 links.

Figure 4 shows the lag zero and the lag one fglasso inferred graphs on the 60 proteins (from 4 replicates), where links are found in at least 50% of bootstrap

samples. The lag 0 network is generally more connected than the lag 1 network. In particular, NMB0035, NMB0913 and NMB1870 are the most connected nodes, with 18, 16 and 15 connections respectively. NMB0913 further interacts with the meningococcus-specific Neisseria adhesin A (NMB1994, NadA), which has been highlighted by a number of studies for its role in the regulation of *Neisseria meningitidis* and which is one of the components of a recombinant vaccine against meningococcal serogroup B (Giuliani et al., 2006; Schielke et al., 2009; Pizza and Rappuoli, 2015). In particular, nadA has been found to interact with the meningococcal FarR homologue NMB1843, which was recently renamed NadR, due to its main role in the regulation of NadA repression (Fagnocchi et al., 2012). The fglasso network inference detects NMB1843-NMB1994 in at least 60% of bootstrap samples, whereas this interaction was not found by the SparseTSCGM method. The fact that farR also represses itself (Fagnocchi et al., 2012) is most likely the reason why the partial correlation value associated to this link has a positive sign. Figure 5 shows a number of other proteins that were found linked with farR and nadA. Of particular notice is NMB0788, an amino acid ABC transporter that is found to be repressed both by farR and nadA. This protein is further detected to repress NMB1476 (gdhB), a NAD-specific glutamate dehydrogenase that is found to be direct target of farR by Fagnocchi et al. (2012). In the lag 1 network, NMB1476 appears to repress NMB0888 (a pilus assembly protein pilW), which is the most connected node in the lag 1 network (7 connections), but whose role in regulation is largely unknown.

In order to shed light into regulatory mechanisms, we have visualised the network at the level of pathways and Gene Ontology (GO) terms. The GO terms were downloaded from `https://www.ebi.ac.uk/interpro/`. Focussing on biological processes, 37 terms were found associated to the 60 proteins. We considered the four most common groups, namely `transport` (NMB0788, NMB1540, NMB0881,NMB1122), `oxidation-reduction process` (NMB0401, NMB1044, NMB0251, NMB0955, NMB1476, NMB1676, NMB2068), `transmembrane transport` (NMB0213, NMB1122, NMB0318) and `metabolic process` (NMB0401, NMB0955, NMB1576), and used the package `pnea` (Signorelli et al., 2015) to detect enrichment at the network level between any pair of these groups. A moderate over-enrichment was detected between the `oxidation-reduction process` and `transmembrane transport` groups in the lag 0 network (p-value 0.133, 7 links between the two groups). As for pathways, we have selected 14 out of the 98 KEGG pathways in *Neisseria meningitidis*, available from the bioconductor package `KEGGREST`. The selection was made to make sure that each pathway was uniquely identified by at least one of the 60 proteins in our study. In particular, we found that 15 out of the 60 proteins (25%) belonged uniquely to one of the 14 pathways. However, only one or two genes identified each of these pathways. As the groups are too small

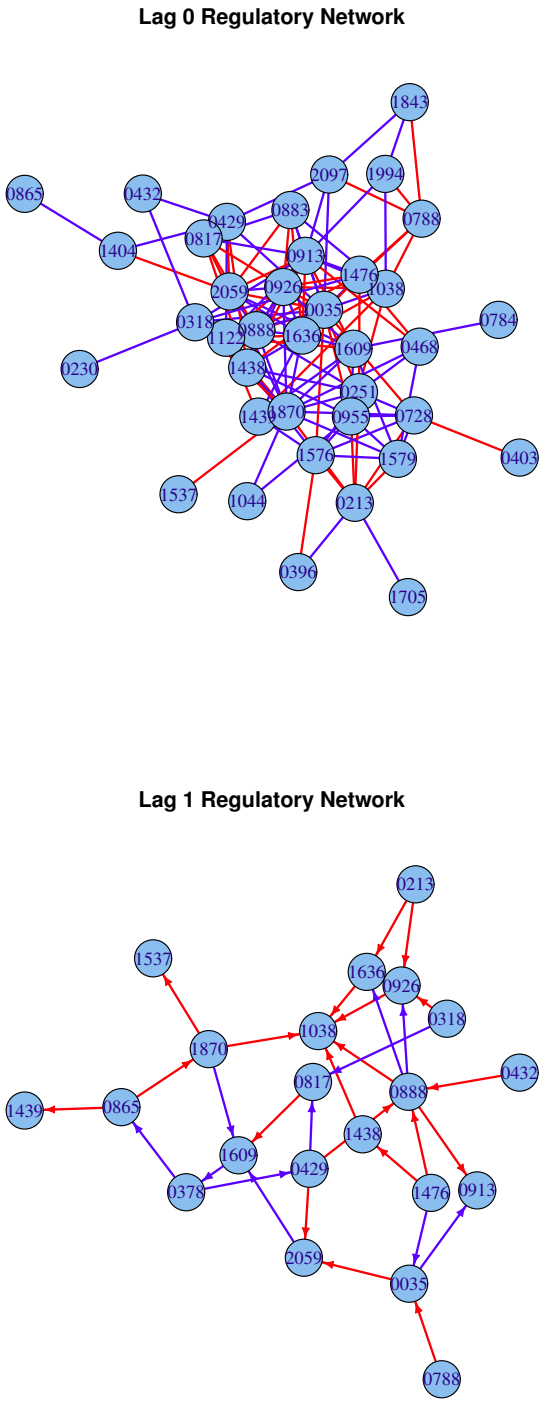


Figure 4: Static (lag 0) and dynamic (lag 1) regulatory network of 60 proteins in *Neisseria meningitidis*. The links were found in at least 50% of bootstrap samples. Red links correspond to negative partial correlations, blue links to positive partial correlations.

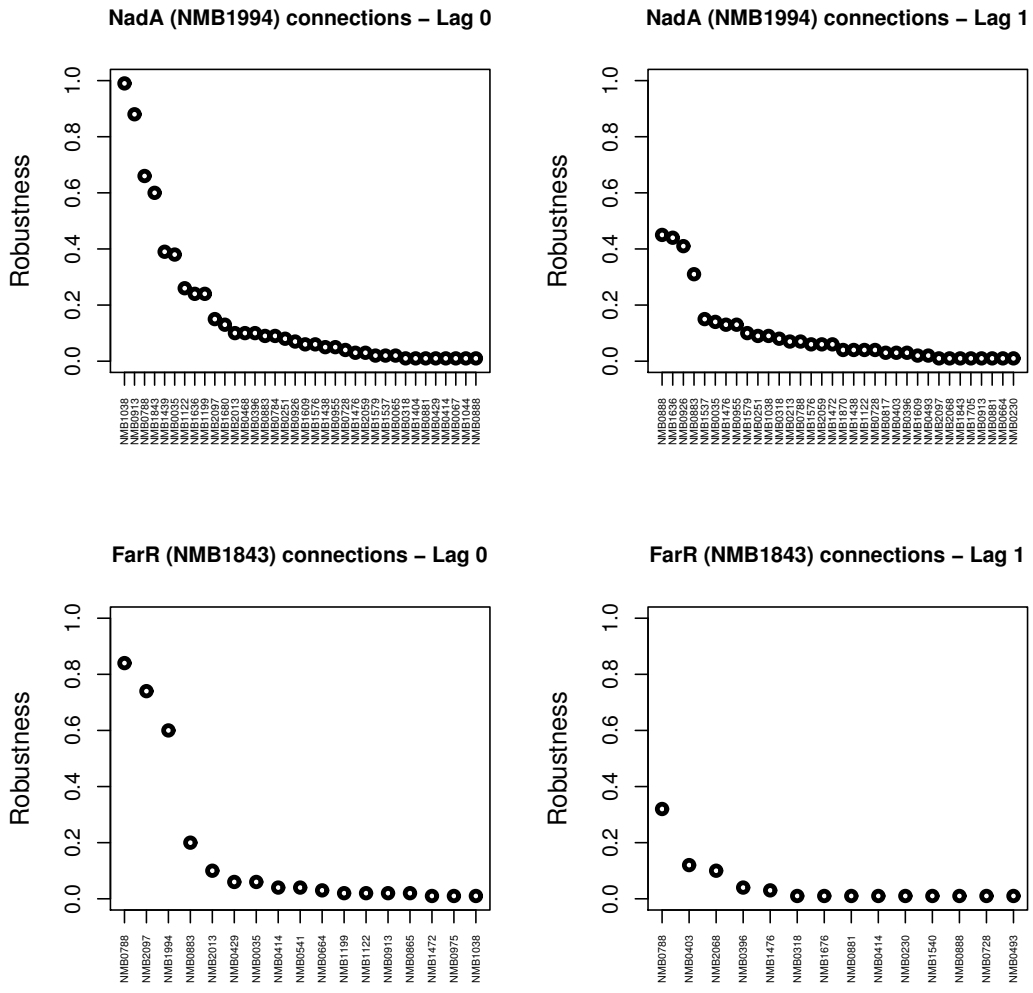
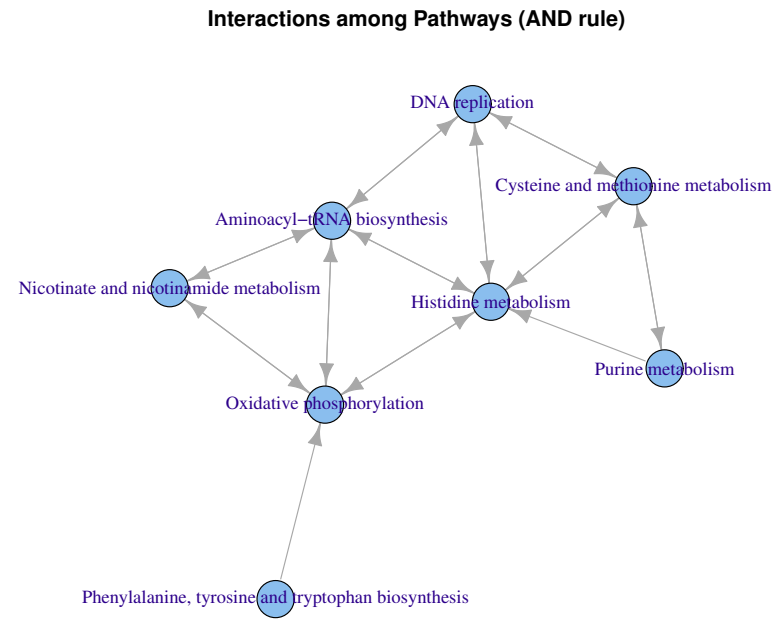


Figure 5: Robustness values of links connected to farR (top) and nadA (bottom) in the lag 0 (left) and lag 1 (right) networks.

for the network enrichment analysis, Figure 6 simply shows the networks obtained using an AND (top) and OR (bottom) rule, respectively. In the AND case, a link between two pathways is present if all proteins in one pathway are connected with all proteins in the other pathway, in either the lag 0 or lag 1 network. In the OR case, a link between two pathways is present if at least one protein in one pathway is connected with at least one (different) protein in the other pathway, in either the lag 0 or lag 1 network. The figure shows that the inferred networks contain many connections among the proteins that belong to known pathways. These connections show that core functions associated with growth are coordinated: DNA replication, amino acid metabolism (several types, particularly in the bottom figure), DNA precursor manufacture, and redox determinants. This is to be expected for a high-resolution time expression data involving the first 210 minutes of growth of the bacteria. Some closely linked biological processes are also found connected, such as the metabolism of nicotinate, nicotinamide, alanine, aspartate and glutamate, and that of cystine and sulfur (Schoen et al., 2014).

5 Conclusions

In this paper, we have proposed a new ℓ_1-penalized estimator for GGM with equality constraints on the precision matrix, called sglasso estimator. This new estimator allows us to apply the fGGM of Wit and Abbruzzo (2015) also when we have a limited amount of data. This model allows to borrow strength across time by imposing suitable equality constraints, and to restrict the possible class of models by setting many entries of the precision matrix to zero a priori or to equal values. We present a cyclic coordinate descent algorithm for the likelihood optimization, which is more efficient than that proposed by Wit and Abbruzzo (2015) and which is now available in the R package `sglasso`. We give the necessary condition for the consistency of the proposed estimator. Then we evaluate different model selection criteria on simulated data from two biochemical networks. The simulation study shows a comparable performance of the fGGM model with other dynamic network models and a good performance of the KLCV criterion for model selection in particular. For the latter, we use a re-scaled version of the criterion developed by (Vujačić et al., 2015) for sparse unstructured Gaussian graphical models. The analysis on real data leads to the selection of a sparse dynamic regulatory network of *Neisseria meningitidis*, which is shown to be robust against sample size and re-sampling, and which features some links that are supported by existing - but rather limited - biological knowledge on the regulatory mechanisms of this biological system.



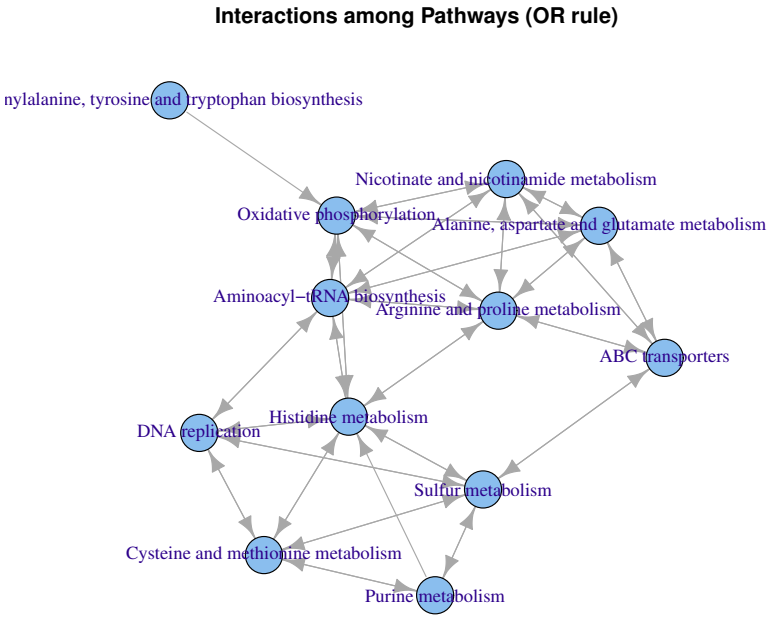


Figure 6: Enrichment analysis of the Neisseria inferred network. In the top network, there is a link between two pathways if all genes in one pathways are connected with all genes in the other pathways. In the bottom network, there is a link if at least one gene in one pathway is connected with at least one gene of the other pathway.

Acknowledgement

We thank Prof. Nigel Saunders from Brunel University London for providing the data used in this analysis.

Appendix

Derivation of $\widehat{\mathbf{gdf}}(\rho)$

In this section, we derive the estimator of equation 13. Definition (11) can be further simplified using the Karush-Kuhn-Tucker conditions, i.e. $\hat{\theta}_m^\rho$ is different from zero if and only if

$$\text{tr}\,\{T_m S\} - \text{tr}\,\{T_m \widehat{\Sigma}_\rho\} + \rho w_m \text{sign}\hat{\theta}_m^\rho = 0, \tag{14}$$

where $w_m = \sum_{ij} T_{ij}^m$. By equation (14) we have that

$$\begin{aligned}
\sum_{m=1}^{M} \hat{\theta}_m^\rho \text{tr}\,\{T_m(S-\Sigma)\} &= \sum_{m=1}^{M} \hat{\theta}_m^\rho \text{tr}\,\{T_m S\} - \sum_{m=1}^{M} \hat{\theta}_m^\rho \text{tr}\,\{T_m \Sigma\} = \\
&= \sum_{m=1}^{M} \hat{\theta}_m^\rho \text{tr}\,\{T_m \widehat{\Sigma}_\rho\} - \rho \sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho| - \sum_{m=1}^{M} \hat{\theta}_m^\rho \text{tr}\,\{T_m \Sigma\} = \\
&= \text{tr}\,\widehat{\Theta}_\rho \widehat{\Sigma}_\rho - \rho \sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho| - \text{tr}\,\widehat{\Theta}_\rho \Sigma \\
&= K - \rho \sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho| - \text{tr}\,\widehat{\Theta}_\rho \Sigma,
\end{aligned}$$

and consequently, the generalized degrees-of-freedoms can be defined as

$$\text{gdf}(\rho) = \frac{N}{2}[\rho E_Y(\sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho|) + E_Y(\text{tr}\,\widehat{\Theta}_\rho \Sigma) - K]. \tag{15}$$

Definition (15) shows that $\text{gdf}(\rho)$ depends on two distinct expected values, i.e. $E_Y(\sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho|)$ and $E_Y(\text{tr}\,\widehat{\Theta}_\rho \Sigma)$. The first one can be estimated by $\sum_{m=1}^{M} w_m |\hat{\theta}_m^\rho|$, since it is an unbiased estimator, while to develop an unbiased estimator of the second expected value observe that

$$E_Y(\text{tr}\,\widehat{\Theta}_\rho \Sigma) = E_Y\{\text{tr}\,\widehat{\Theta}_\rho E_{\bar{Y}}(\bar{Y}\bar{Y}^\top)\} = E_{\bar{Y}} E_Y(\bar{Y}^\top \widehat{\Theta}_\rho \bar{Y}), \tag{16}$$

where $\bar{Y}$ is an independent copy of Y. The identity (16) suggests that the second expected value can be estimated by leave-one-out cross-validation method, i.e.

$$\hat{E}_Y(\text{tr}\,\widehat{\Theta}_\rho \Sigma) = \sum_{i=1}^{N} y_i^\top \widehat{\Theta}_\rho^{(-i)} y_i / N,$$

where $\widehat{\Theta}_{\rho}^{(-i)}$ denotes the sglasso estimate obtained after removing the ith observation from the data. This leads to the estimator

$$\widehat{\mathrm{gdf}}(\rho) = \frac{N}{2}\left(\rho \sum_{m=1}^{M} w_m |\hat{\theta}_m^{\rho}| + \frac{\sum_{i=1}^{N} y_i^{\top} \widehat{\Theta}_{\rho}^{(-i)} y_i}{N} - K\right).$$

References

Abegaz, F. and E. Wit (2013): "Sparse time series chain graphical models for reconstructing genetic networks." *Biostatistics*, 14, 586–599.

Aderhold, A., D. Husmeier, and M. Grzegorczyk (2014): "Statistical inference of regulatory networks for circadian regulation," *Statistical applications in genetics and molecular biology*, 13, 227–273.

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," in B. N. Petrov and F. Czaki, eds., *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267–281.

Augugliaro, L., A. M. Mineo, and E. C. Wit (2013): "Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models," *J. Roy. Statist. Soc. Ser. B*, 75, 471–498.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008): "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *The Journal of Machine Learning Research*, 9, 485–516.

Bickel, P. J. and E. Levina (2008): "Regularized estimation of large covariance matrices," *Ann. Statist.*, 36, 199–227.

Bühlmann, P. and S. Van De Geer (2011): *Statistics for high-dimensional data: methods, theory and applications*, Springer.

Efron, B. (1986): "How biased is the apparent error rate of a prediction rule?" *J. Amer. Statist. Assoc.*, 81, 461–470.

Efron, B. (2004): "The estimation of prediction error: Covariance penalties and cross-validation," *J. Amer. Statist. Assoc.*, 99, 619–632.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004): "Least angle regression," *Ann. Statist.*, 32, 407–499.

Fagnocchi, L., E. Pigozzi, V. Scarlato, and I. Delany (2012): "In the NadR regulon, adhesins and diverse meningococcal functions are regulated in response to signals in human saliva," *J Bacteriol*, 194, 460–474.

Foygel, R. and M. Drton (2010): "Extended Bayesian information criteria for gaussian graphical models," in *Advances in Neural Information Processing Systems*, 604–612.

Friedman, J., T. Hastie, H. Höfling, R. Tibshirani, et al. (2007): "Pathwise coordinate optimization," *The Annals of Applied Statistics*, 1, 302–332.

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

Friedman, J. H., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1–22.

Gao, X., D. Q. Pu, Y. Wu, and X. Xu (2012): "Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model," *Statistica Sinica*, 22, 1123–1146.

Genco, C. A., L. Wetzler, and L. M. Wetzler (2010): *Neisseria: molecular mechanisms of pathogenesis*, Horizon Scientific Press.

Giuliani, M. M., J. Adu-Bobie, M. Comanducci, B. Aricò, S. Savino, L. Santini, B. Brunelli, S. Bambini, A. Biolchi, B. Capecchi, E. Cartocci, L. Ciucchi, F. Di Marcello, F. Ferlicca, B. Galli, E. Luzzi, V. Masignani, D. Serruto, D. Veggi, M. Contorni, M. Morandi, A. Bartalesi, V. Cinotti, D. Mannucci, F. Titta, E. Ovidi, J. A. Welsch, D. Granoff, R. Rappuoli, and M. Pizza (2006): "A universal vaccine for serogroup B meningococcus," *Proceedings of the National Academy of Sciences*, 103, 10834–10839.

Grzegorczyk, M. and D. Husmeier (2011): "Non-homogeneous dynamic Bayesian networks for continuous data," *Machine Learning*, 83, 355–419.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011): "Joint estimation of multiple graphical models," *Biometrika*, 98, 1–15.

Højsgaard, S. and S. Lauritzen (2008): "Graphical Gaussian models with edge and vertex symmetries," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 1005–1027.

Hoops, S., S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer (2006): "Copasia complex pathway simulator," *Bioinformatics*, 22, 3067–3074.

Huang, C. Y. and J. E. Ferrell (1996): "Ultrasensitivity in the mitogen-activated protein kinase cascade," *Proceedings of the National Academy of Sciences*, 93, 10078–10083.

Jordan, P. and N. Saunders (2009): "Host iron binding proteins acting as niche indicators for Neisseria meningitidis." *PLoS ONE*, 4, e5198.

Leloup, J.-C. and A. Goldbeter (1999): "Chaos and birhythmicity in a model for circadian oscillations of the {PER} and {TIM} proteins in drosophila," *Journal of Theoretical Biology*, 198, 445 – 459.

Liu, H., K. Roeder, and L. Wasserman (2010): "Stability approach to regularization selection (stars) for high dimensional graphical models," in J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds., *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., 1432–1440.

Lysen, S. (2009): *Permuted Inclusion Criterion: A Variable Selection Technique.*, PhD thesis, University of Pennsylvania.

Meinshausen, N. and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, 1436–1462.

Miller, K. S. (1981): "On the inverse of the sum of matrices," *Mathematics Magazine*, 54.

Pizza, M. and R. Rappuoli (2015): "Neisseria meningitidis: pathogenesis and immunity," *Current Opinion in Microbiology*, 23, 68 – 72.

Rhein, R. O. and K. Strimmer (2007): "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data," *BMC Systems Biology*, 1, 37+.

Rothman, A., P. J. Bickel, E. Levina, and J. Zhu (2008): "Sparse permutation invariant covariance estimation," *Electron. J. Stat.*, 2, 494–515.

Ryan, K. J., C. G. Ray, et al. (2010): *Sherris medical microbiology*, McGraw Hill Medical New York.

Saunders, N. and J. Davies (2012): "The use of the pan-Neisseria microarray and experimental design for transcriptomics studies of neisseria." *Methods Mol Biol.*, 799, 295–317.

Schielke, S., C. Huebner, C. Spatz, V. Nägele, N. Ackermann, M. Frosch, O. Kurzai, and A. Schubert-Unkmeir (2009): "Expression of the meningococcal adhesin NadA is controlled by a transcriptional regulator of the MarR family," *Molecular Microbiology*, 72, 1054–1067.

Schoen, C., L. Kischkies, J. Elias, and B. J. Ampattuu (2014): "Metabolism and virulence in Neisseria meningitidis," *Frontiers in Cellular and Infection Microbiology*, 4.

Schwarz, G. (1978): "Estimating the dimension of a model," *Ann. Statist.*, 6, 461–464.

Signorelli, M., V. Vinciotti, and E. C. Wit (2015): *pnea: Parametric Network Enrichment Analysis*, URL `http://CRAN.R-project.org/package=pnea`, `r` package version 1.2.0.

Tettelin, H., N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, W. C. Nelson, M. L. Gwinn, R. DeBoy, J. D. Peterson, E. K. Hickey, D. H. Haft, S. L. Salzberg, O. White, R. D. Fleischmann, B. A. Dougherty, T. Mason, A. Ciecko, D. S. Parksey, E. Blair, H. Cittone, E. B. Clark, M. D. Cotton, T. R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H. O. Smith, C. M. Fraser, E. R. Moxon, R. Rappuoli, and J. Craig Venter (2000): "Complete genome sequence of neisseria meningitidis serogroup B strain MC58," *Science*, 287, 1809–1815.

Vujačić, I., A. Abbruzzo, and E. Wit (2015): "A computationally fast alternative to

cross-validation in penalized Gaussian graphical models," *Journal of Statistical Computation and Simulation*, 1–13.

Wang, C., D. Sun, and K. Toh (2010): "Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm," *SIAM Journal on Optimization*, 20, 2994.

Wit, E. and A. Abbruzzo (2015): "Factorial graphical models for dynamic networks," *Network Science*, 3, 37–57.

Wit, E., E. v. d. Heuvel, and J.-W. Romeijn (2012): "all models are wrong...: an introduction to model uncertainty," *Statistica Neerlandica*, 66, 217–236.

Wu, T. T. and K. Lange (2008): "Coordinate descent algorithms for lasso penalized regression," *Ann. Appl. Statist.*, 2, 224–244.

Zou, H., T. Hastie, and R. Tibshirani (2007): "On the "degrees of freedom" of the lasso," *Ann. Statist.*, 35, 2173–2192.