DE GRUYTER Statistical Applications in Genetics and Molecular Biology

Model selection for structured dynamic gene regulatory networks of \emph{Neisseria meningitidis}

Journal:	*Statistical Applications in Genetics and Molecular Biology*
Manuscript ID:	SAGMB.2014.0075
Manuscript Type:	Research Article
Date Submitted by the Author:	02-Oct-2014
Complete List of Authors:	Vinciotti, Veronica; Brunel University London, Mathematics Abbruzzo, Antonino; University of Palermo, Statistics Augugliaro, Luigi; University of Palermo, Statistics Saunders, Nigel; Brunel University London, Biosciences Wit, Ernst ; Rijksuniversiteit Groningen, NL,
Classifications:	62F30, 62H20
Keywords:	Graphical models, penalized inference, sparse networks, gene-regulatory systems
Abstract:	Factorial graphical models have recently been proposed for inferring dynamic regulatory networks from high-throughput data. In the search of true regulatory relationships amongst the vast space of possible networks, these models allow to impose certain restrictions on the dynamic nature of these relationships, such as that Markov dependencies are of low order, i.e. some entries of the precision matrix are a priori zeros, or that the strength of the dependencies depend only on time lags, i.e. some entries of the precision matrix are assumed to be equal. The precision matrix is then estimated by l_1 penalised likelihood, imposing a further constraint on the absolute value of its entries, which results in sparse networks. The problem of selecting the optimal sparsity level is traditionally framed in terms of the Kullback-Leibler (KL) divergence. In this paper, we present a KL-motivated model selection criterion for factorial graphical models, by taking into account the a priori structural constraints. We test the performance of this method on simulated data and compare it with existing approaches. Finally, we present an application on a detailed time-course microarray data from the \textit{Neisseria meningitidis} bacterium, a causative agent of life-threatening infections such as meningitis.

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 be able to observe genomic systems at high temporal resolution. This enables investigation into 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 microarray 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. 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 prevent their use in an exploratory setting. Recent work in penalized Gaussian graphical models (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. 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 slightly less

known 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 develop a dynamic graphical model for gene expression in *Neisseria meningitidis* that borrows strength across time by imposing suitable equality constraints and whose sparsity is selected automatically via a computationally efficient and accurate model selection algorithm. In section 2 we introduce the structured dynamic gene-regulatory network model by means of undirected graphical models. In section 3 we derive an efficient estimator of the Kullback-Leibler divergence, which can be used directly to perform model selection. In section 4, we test the performance of this method on simulated data. Finally, in section 5, we apply the methodology to the inference of a dynamic regulatory network in *Neisseria meningitidis*.

2 Structured dynamic gene regulatory network model

An important issue in system biology is to understand the system of interactions among several biological components, such as protein-protein interaction and gene regulatory networks. Various genome-wide measurement techniques have opened up the possibility of achieving such ambitious goals. For instance, RNA-seq chips or microarrays measure simultaneously thousands of gene expression levels, i.e., concentrations of messenger RNA produced when genes are transcribed. Gene expression is a temporal process, which evolves dynamically in response to internal, genomic, and external, environmental, cues. Even under stable conditions, mRNA is transcribed continuously and new proteins are generated. This process is highly regulated. In many cases, the expression program starts by activating a few transcription factors, which in turn activate other genes. Transcription factors are proteins that bind to specific DNA sequences, thereby controlling the flow of genetic information from DNA to mRNA. Taking a snapshot of the expression profile following some intervention may reveal which genes have specifically changed. But rather than determining the set of differentially expressed genes, such as in the early days of microarray analysis, biologists have become more interested in determining the transcriptional program, i.e., determining the functional pathways in the genomic network. In order to infer the temporal interaction between the genes, it is necessary to perform time-course expression experiments.

2.1 Gaussian graphical models

The presence of a link or undirected edge in a network can mean, in general, a variety of things. In edge-based network models, it refers to a dyadic relationship between the vertices, such as the presence of friendship relationship. In vertex-based network model, an edge refers to a relationship between some properties of the vertices themselves. We will define in this section the absence of a link in a network as the conditional independence of some quantity of interest measured on the vertices, when fixing all other vertex levels. This corresponds roughly to the definition of a graphical model. For the purposes of this paper, we will consider exclusively Gaussian graphical models. Let $G = (V, E)$ be a graph with finite vertex set $V = \{1, \ldots, p\}$ and undirected edge set $E \subset V \times V$.

A Gaussian Graphical Model (GGM) with respect to an undirected graph $G = (V, E)$ is a random vector $Y = (Y_1, \ldots, Y_p)$ with multivariate normal distribution $N(\mu, \Sigma)$, such that Y_i is independent from Y_j when fixing the rest if and only if the edge (i, j) is not in the edge set E,

$$(i, j) \notin E \Leftrightarrow Y_i \perp Y_j \mid Y_{V \setminus \{i,j\}}.$$

This property is known as the pairwise Markov property. As a Gaussian density is strictly positive, it can be shown (Lauritzen, 1996) that the pairwise Markov property is equivalent with the following two properties:

- **Global Markov property**: A probability distribution $\mathbb{P}$ is said to obey the *global Markov property* relative to G, if for any triple (A,B,S) of disjoint subsets of V such that all paths from A to B in G pass through S, we have that the measurements on the nodes in A are conditionally independent from those on B, when fixing the measurements on the nodes in S, i.e.,

$$\mathbf{Y}_A \perp \mathbf{Y}_B | \mathbf{Y}_S.$$

- **Factorization property**: A probability distribution $\mathbb{P}$ is said to satisfy the *factorization property* relative to G, if the density function f of the joint distribution $\mathbb{P}$ can be written in the form

$$f(y_1, \ldots, y_p) = \frac{1}{z} \prod_{c \in C} \psi_c(\mathbf{y}_c),$$

where C is a set of cliques, i.e. the largest subsets of V that form complete graphs in G, the function $\psi_c(\mathbf{y}_c)$ is a potential function, which is a positive function of the variables $\{y_i\}_{i \in C}$, and $z = \sum_{\mathbf{y}} \prod_{c \in C} \psi_c(\mathbf{y}_c)$ is a normalization factor.

The reason why these various characterizations of a graphical model are relevant is that the factorization property will enable us to write down a convenient likelihood, whereas the global Markov property will allow for an intuitive interpretation of the graph as a quasi-modular network. The local Markov property can be useful to infer the graph structure from sparse data, although in this paper we will pursue a different route.

The normal density is strictly positive and it can be written as

$$f(y) = (2p)^{-p/2}|\Theta|^{1/2}\prod_{i,j} -\theta_{ij}(y_i-\mu_i)(y_j-\mu_j),$$

where $\Theta = \Sigma^{-1}$ is the precision matrix. From the equivalence of all the Markov and factorization properties, one can see that

$$\theta_{ij} = 0 \Leftrightarrow Y_j \perp Y_i | Y_{V\setminus\{(i,j)\}} \Leftrightarrow (i,j) \notin E,$$

This suggests that the determination of the graph G, can be based on the set of sample precisions $\hat{\theta}_{ij}$ estimated from a set of observations. Of particular interest will be the identification of zero entries in the concentration matrix $\Theta = \{\theta_{ij}\}$, since a zero entry $\theta_{ij} = 0$ indicates the absence of the link in the conditional independence graph G.

2.2 Dynamic Gaussian graphical model

In this section, we consider a special case of the Gaussian graphical model, in order to represent a dynamic network with particular symmetry constraints. We consider a set of arbitrary units $\Gamma = \{1,\ldots,p\}$ and an ordered set $T = (1,\ldots,\tau\}$, typically describing time points. This allows us to formally define a dynamic Gaussian graphical model, in which each unit at each time point is considered a, not necessarily independent, random variable.

Consider Γ a finite set of unordered units, and T is a finite set of ordered time points. A dynamic Gaussian graphical model is a pair $\left(G = (V,E), N(\mu,\Theta^{-1})\right)$, where $V = \Gamma \times T = \left\{v_{ij}\right\}_{i\in\Gamma, j\in T}$ is a set of vertices and $E \subseteq V \times V$ is a set of pairs of vertices. An observation $Y \in \mathbb{R}^{n_\Gamma n_T}$ from the dynamic Gaussian graphical model is normally distributed set of observations across all units and time points,

$$Y \sim N(\mu,\Theta^{-1}),$$

such that the conditional independence relationships satisfy G, $\theta_{it,js} = 0 \Leftrightarrow (v_{it},v_{js}) \notin E$. Links in the dynamic Gaussian graphical models represent conditional dependencies between the units within the same time point or across time. This class of

dynamic graphical models is, in principle completely, general. We will consider various subclasses of these model. First consider the dynamic graphical model associated with the graph in Figure 1. This represent an 8 dimensional random vector Y, whose elements $(Y_{11}, Y_{21}, Y_{31}, Y_{41})$, are those associated with time point 1, whereas the others are connected to time point 2. Edges or links within these two groups, excluding self-links, represent the instantaneous or lag zero networks, N_0. Edges between these two groups, excluding links between the same unit, represent the lag 1 network, or the time-delay interactions, N_1. Similarly, self-self interaction with lag 0, S_0, or lag 1, S_1, can be identified. To make the concept of networks and self-self interactions more precise, we introduce the concept of a natural partition of the precision matrix Θ.

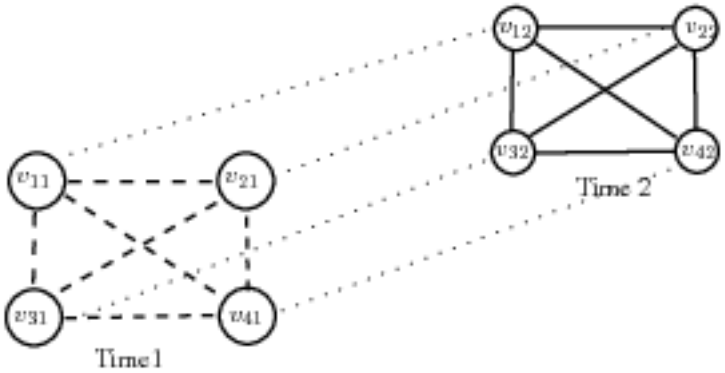

Figure 1: Example of a dynamic graph with four vertices measured across two time points. The graph corresponds to the dynamic graphical model $S_0 \prec 1, N_0 \prec T, S_1 \prec 1, N_1 \prec 0$.

For a dynamic Gaussian graphical model $\left(G, N(\mu, \Theta^{-1})\right)$ the natural partition of Θ given by $\{S_i\}_{i=0}^{\tau-1}$ of self interactions and $\{N_i\}_{i=0}^{\tau-1}$ of network interactions, defined as subsets of Θ,

$$S_i = \{\theta_{jt,jt+i} \in \Theta \mid j \in \Gamma, t \in T\},$$

and

$$N_i = \{\theta_{jt,kt+i} \in \Theta \mid \forall j \neq k \in \Gamma, t = 1, \ldots, n_T - i\}.$$

Each element of the natural partition has a natural interpretation: S_i are the lag i self-self interactions of the units Γ, whereas N_i are the time lag i interactions between the units of Γ. The next step is to define a set of models that can be applied to each of the elements of the natural partition. Consider an element $B \subset \Theta$ of the natural partition of Θ, i.e., $B = S_i$ or $B = N_i$ for some i, then we define the following models with respect to the unit set Γ and time ordering T,

- Constant model 1: for all $\theta_{it,js} \in B$, $\theta_{it,js} = c$.
- Main time effect model T: for all $\theta_{it,js} \in B$, $\theta_{it,js} = c_{ts}$.
- Main unit effect model Γ: for all $\theta_{it,js} \in B$, $\theta_{it,js} = c_{ij}$.
- Interaction effect model ΓT: for all $\theta_{it,js} \in B$, $\theta_{it,js} = c_{it\,js}$.

The constant model means that all the edges represent the same concentration or partial correlation. The main effect time model means that the edge from unit i at time point t to unit j at time point s has the same concentration or partial correlation as the edge from unit k at time point t to unit l at time point s. Conversely, the main effect unit model means that the edge from unit i at time point t to unit j at time point s has the same concentration or partial correlation as the edge from unit i at time point t' to unit j at time point s'. The interaction effect model means that the concentrations or partial correlations can vary freely across time and units.

The models provide structure to the dynamic Gaussian graphical models with a consistent interpretation. We will informally also speak of the "zero model: 0", which corresponds to the absence of all edges. The dynamic models can be applied to each of the elements of the natural partition separately. For example, in Figure 1 we consider the constant model on S_1 and main effect time on N_0 and the zero model on N_1, which we will write, respectively, as $S_1 \prec 1$, $N_0 \prec T$ and $N_1 \prec 0$. By collecting all the models on the elements of the natural partition, we can now define a dynamic graphical model as a Gaussian graphical model $(G, N(\mu, \Theta^{-1}))$ with respect to the equality constraints defined on the elements of the natural partition of Θ.

3 Model selection via efficient KL estimation

In the previous section, we defined an important class of network models by considering dynamic constraints on the concentration or conditional correlation matrix. These constraints lead to a considerable reduction of the number of parameters to be estimated. Each sub-network can be interpreted according to its corresponding natural partition. However, dynamic genetic graphs are usually sparse, which means that few vertices will be connected. In this section, we consider maximum likelihood estimation subject to an ℓ_1-norm penalty on the concentration or conditional correlation 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).

3.1 Penalized likelihood for dynamic graphical models

In this section, we describe the statistical inference for sparse dynamic Gaussian graphical models with equality constraints on the concentration matrix. In the case of high-dimensional data, such constraint is also necessary to guarantee existence of the estimate. The ℓ_1-penalized maximum likelihood estimator, therefore, is defined as

$$\begin{aligned} \widehat{\Theta}_\rho &= \text{argmax}_\Theta \log(\Theta) - \text{tr}(\Theta S) - \rho||\Theta||_1, \\ &\quad \text{subject to dynamic constraints on } \Theta, \end{aligned} \tag{1}$$

where $S \in \mathbb{R}^{p\tau \times p\tau}$ is the sample covariance matrix of the time course data. Maximization of (1) is a challenging task. One way 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). 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 to optimize the objective function with respect to the corresponding parameter keeping all the remaining indexes fixed.

Suppose that we have computed the estimator $\hat{\psi}$ for a given value of the tuning parameter, say ρ', and we want to compute a new estimate for a value of the tuning parameter, say ρ, with $\rho < \rho'$. If ρ is close enough to ρ', the one-dimensional log-likelihood function $\ell(\theta_m)$ can be approximated by standard Taylor expansion, with respect to θ_m, around the old estimate $\hat{\psi}$. By straightforward algebra, it is easy to see that $\ell(\theta_m)$ can be approximated as follows

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

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

$$\min_{\theta_m \in \mathbb{R}} \frac{1}{2} I_m(\hat{\psi})(\theta_m - \hat{\vartheta}_m)^2 + \rho w_m |\theta_m|, \tag{3}$$

where $I_m(\hat{\psi}) = -\partial^2 \ell(\hat{\psi})/\partial \theta_m^2$ is the Fisher information for θ_m evaluated at $\hat{\psi}$. Problem (3) can be solved in closed form (Friedman et al., 2007), i.e. $\hat{\theta}_m = S(\hat{\vartheta}_m; w_m I_m^{-1}(\hat{\theta})\rho)$,

where $S(x;\lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator. We have implemented the solver in `R` in our package `sglasso`.

3.2 Model selection through Kullback-Leibler cross-validation

In this section we address the issue of choosing the best dynamic model, and what should be a good compromise between a sparse and a dense graph. Having selected the factorial model and the tuning parameter we can increase the precision of our estimates by using stability selection.

An information criterion such as AIC or BIC can be used to compare different factorial graphical models. The AIC and BIC are, respectively, given by

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

and

$$BIC(\rho) = -n\left(\log(\widehat{\Theta}_\rho) - \text{tr}(\widehat{\Theta}_\rho S)\right) + \text{df}(\widehat{\Theta}_\rho)\log n.$$

There are two main issues facing the use of such methods. On the one hand, the traditional definition of the degrees of freedom as the number of non-zero parameters in the model is somewhat problematic in penalized inference. Moreover, 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.

Instead, we define a computationally efficient estimator of the Kullback-Leibler divergence based on cross-validation. This should be equally suitable in the small sample as in the large sample scenario. This socalled Kullback-Leiber cross-validation estimator, or KLCV for short, is defined on the same scale as the AIC and therefore an estimate of the corresponding KL divergence can be obtained by dividing by $2n$.

$$KLCV(\rho) = -n\left(\log(\widehat{\Theta}_\rho) - \text{tr}(\widehat{\Theta}_\rho S)\right) + 2\sum_{k=1}^{n} p_k(\hat{\Theta}_\rho, S),$$

where the cross-validated estimated of the degrees of freedom, p_k, is given as

$$p_k(\rho) = \frac{\text{vec}[(\hat{\Theta}_\rho^{-1} - S_k)\circ I_\rho]^t C[C^t(\hat{\Theta}_\rho^{-1}\otimes\hat{\Theta}_\rho^{-1})C]^{-1}C^t\text{vec}[(S - S_k)\circ I_\rho]}{2n-2}, \qquad (4)$$

where $C = \frac{\partial\Theta}{\partial\theta}$ a $q^2 \times m$ matrix, for number of nodes $q = p\tau$ and number of model parameters m, and where I_ρ is the indicator matrix, whose entry is 1 if the corresponding entry in the precision matrix $\hat{\Theta}_\rho$ is nonzero and zero if the corresponding entry in the precision matrix is zero. The original KLCV was proposed in Vujacic

et al. (2013) for unstructured Gaussian graphical models and we refer to this method as KLCV(PMLE). Using the original KLCV, we can also define a fast approximation of (4), by replacing it by

$$p_k(\rho) = \frac{m}{q} \times \frac{\text{vec}[(\hat{\Theta}_\rho^{-1} - S_k) \circ I_\rho]^t M_q (\hat{\Theta}_\rho \otimes \hat{\Theta}_\rho) \text{vec}[(S - S_k) \circ I_\rho]}{2n-2}. \qquad (5)$$

We refer to this method as the KLCV(SPMLE). In section 4 we shall show how the KLCV works in practice. Below, we give a brief derivation of the KLCV for structured dynamic models.

3.3 Derivation of KLCV for structured dynamic models

Consider the scaled log-likelihood for a Gaussian graphical model,

$$l(\Theta; S) = \log|\Theta| - \text{tr}(S\Theta).$$

The Kullback-Leibler divergence up to an arbitrary scale factor and constant is defined as,

$$KL(\hat{\Theta}) = -E_S l(\hat{\Theta}; S),$$

where the expectation is taken with respect to the sample covariance matrix. Using the fact that cross-validation can be used to estimate this expectation, we derive

$$\begin{aligned} KLCV(\hat{\Theta}) &= -\sum_{i=1}^{n} l(\hat{\Theta}^{(-i)}; S_i) \\ &= -l(\hat{\Theta}; S) - \sum_{i=1}^{n} [l(\hat{\Theta}^{(-i)}; S_i) - l(\hat{\Theta}; S_i)] \\ &\approx -l(\hat{\Theta}; S) - \sum_{i=1}^{n} \left[\frac{dl(\hat{\Theta}; S_i)}{d\theta}\right]^t \text{vec}(\hat{\theta}^{(-i)} - \hat{\theta}). \end{aligned}$$

Using matrix differential calculus we have $\frac{\partial l(\hat{\Theta};S_i)}{\partial \theta} = \frac{\partial l}{\partial \Theta}\frac{\partial \Theta}{\partial \theta} = \text{vec}(\hat{\Theta}^{-1} - S_i)C$. The term $\text{vec}(\hat{\theta}^{(-i)} - \hat{\theta})$ is obtained via the Taylor expansion

$$0 = \frac{dl(\hat{\Theta}^{(-i)}; S^{(-i)})}{d\theta} \approx \frac{dl(\hat{\Theta}; S)}{d\theta} + \frac{d^2 l(\hat{\Theta}; S)}{d\theta^2}\text{vec}(\hat{\theta}^{(-i)} - \hat{\theta}) + \frac{d^2 l(\hat{\Theta}; S)}{d\theta dS}\text{vec}(S^{(-i)} - S).$$

From here it follows that

$$\text{vec}(\hat{\theta}^{(-i)} - \hat{\theta}) = -\left(\frac{d^2 l(\hat{\Theta}; S)}{d\theta^2}\right)^{-1} \frac{d^2 l(\hat{\Theta}; S)}{d\theta dS}\text{vec}(S^{(-i)} - S).$$

We have $dl(\hat{\Theta};S)/d\theta = \text{vec}(\hat{\Theta}^{-1} - S)C$, and so $d^2l(\hat{\Theta};S)/d\theta dS = -C^t$, as well as $d^2l(\hat{\Theta};S)/d\theta^2 = -C^t(\hat{\Theta}^{-1} \otimes \hat{\Theta}^{-1})C$ and consequently

$$\text{vec}(\hat{\theta}^{(-i)} - \hat{\theta}) = -[C^t(\hat{\Theta}^{-1} \otimes \hat{\Theta}^{-1})C]^{-1}C^t\text{vec}(S^{(-i)} - S).$$

This results into the general formula for an unpenalized estimator $\hat{\Theta}$ with equality constraints,

$$KLCV = -l(\hat{\Theta};S) + \sum_{i=1}^{n} \frac{\text{vec}(\hat{\Theta}^{-1} - S_i)^t C[C^t(\hat{\Theta}^{-1} \otimes \hat{\Theta}^{-1})C]^{-1}C^t\text{vec}(S - S_i)}{n-1}, \quad (6)$$

where we use the fact that $S^{(-i)} - S = \frac{S - S_i}{n-1}$. To obtain the formula for the penalized estimator $\rho > 0$ we note that asymptotically the covariances between zero elements and nonzero elements are equal to zero. Thus, to obtain the term p_i for the shrinkage estimator we do not only plug in the expression $\hat{\Theta}_\rho$ in formula (6), but we also set the elements of the covariance matrix $[C^t(\hat{\Theta}_\rho^{-1} \otimes \hat{\Theta}_\rho^{-1})C]^{-1}$ that correspond to covariances between zero and nonzero elements to zero.

4 Simulation study

We compare the different model selection criteria on simulated data. We simulate multivariate Gaussian data, with $n = 10$, $p = 40$ and a precision matrix with 20% structural zeros and about 55% of zeros (sparsity). The remaining non-zero parameters are drawn randomly amongst a limited number of possible values between 0.4 and 0.6. This results in a structured precision matrix, with a priori zeros, sparsity and equality constraints. Figure 2 shows the true Kullback-Leibler loss for different penalties ρ, together with the KL estimates given by the different methods. In particular, KLCV(SPMLE) is the criterion proposed in this paper for structured penalised graphical models (5). The plot shows how KLCV(SPMLE) and AIC reach a minimum close to the true KL minimum.

To further compare the different methods, Table 1 reports the Kullback-Leibler minimum, averaged over 50 iterations and with standard errors in brackets, for the methods considered. We consider the case of small sample size, $n = 10$, and vary the number of free parameters in the model, p_{model}, between 10 and 100. In all cases, the model selection criteria for structured graphs have a KL loss close to the oracle value, with KLCV(SPMLE) outperforming AIC. Not considering structural and equality constraints results in poorer selection of the model, as shown by the KLCV(PMLE) results.

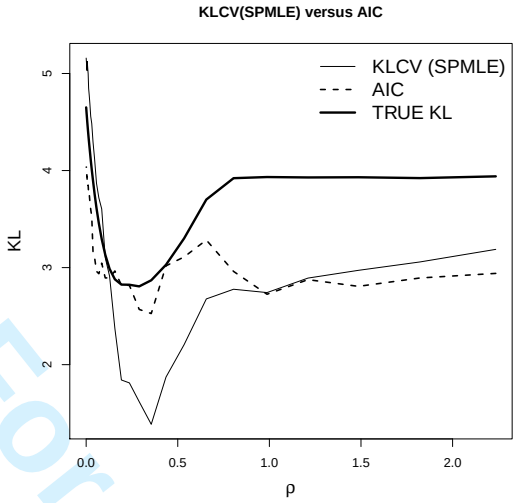

Figure 2: KLCV and AIC estimators on simulated data with $n = 10$, $p = 40$, $p_{\text{model}} = 50$, about 20% structural zeros and 55% sparsity.

Table 1: Simulated data with n=10, p=40, about 20% structural zeros and 55% sparsity. For each method, the table reports the Kullback-Leibler loss, averaged over 50 iterations and with standard errors in brackets. A number of cases are considered by varying the number of model parameters, p_{model}. The best results are highlighted in bold.

p_{model}	Oracle	KLCV(SPMLE)	AIC	KLCV(PMLE)
10	1.03 (0.07)	**1.11 (0.17)**	**1.11 (0.16)**	3.95 (0.65)
20	1.66 (0.07)	**1.74 (0.16)**	1.76 (0.17)	4.06 (0.3)
30	1.93 (0.06)	**1.99 (0.13)**	2.12 (0.17)	4.10 (0.34)
50	2.90 (0.08)	**3.00 (0.17)**	3.16 (0.17)	4.04 (0.22)
100	3.68 (0.09)	**3.93 (0.23)**	4.08 (0.25)	4.16 (0.28)

5 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 upon 60 transcripts that have been highly characterized to be within the FarR regulatory network in highly replicated microarrays studies and validated by qPCR and gel shift assays (NS - unpublished), and 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).

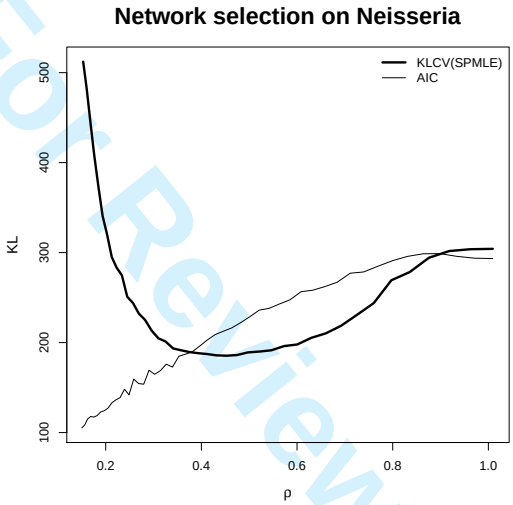

Figure 3: KLCV(PMLE) and AIC on real data with 60 proteins from Neisseria meningitidis.

We consider a particular structure to the graphical model. We define the nodes of the graph to be the genes at a particular time point. This results in a 600×600 inverse covariance matrix Θ, thus about 180,000 parameters to be estimated with only 2,400 observations. However, there is good reason to impose some constraints on Θ. In particular, we make the following two assumptions:

1. **Markov assumption:** we assume that except for lag zero and lag one, there are no higher order interactions between the genes, i.e. $S_i, N_i \prec 0$ for all $i > 1$.
2. **Interaction persistence:** we assume that the lag zero and lag one interactions are persistent across all ten time points, i.e. $S_i, N_i \prec \Gamma$ for $i = 0, 1$.

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. We use a factorial graphical model with the above constraints on the Neisseria data.

Figure 3 shows the KLCV(SPMLE) and AIC criteria on the Neisseria gene expression data. In contrast to AIC, which tends to favour denser networks, the KLCV(SPMLE) criterion selects the optimal network corresponding to the value of $\rho^* = 0.453$.

We further perform a bootstrap analysis to test the robustness of the inferred networks. For each protein, we simulate 100 boostrap 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 fit a factorial graphical model to the bootstrap data with $\rho = \rho^*$. Effectively, this post-analysis allows 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. Figure 4 shows the lag zero and the lag one graphs, where links are found in at least 50% of bootstrap samples. In the lag 0 network, NMB0035, NMB0913 and NMB1636 are the most connected nodes, each with 11 connections and all connected with each other. NMB0913 further interacts with NMB1994 (NadA) which is known from previous studies to respond to FarA deletion (Schielke et al., 2009). In the lag 1 network, NMB0888 is a central node, with 7 connections. This node appears also in the lag 0 network and is connected with 6 other proteins. These interactions will be further validated in future research.

6 Conclusions

In this paper, we have introduced structured dynamic graphical models for inferring regulatory networks from gene expression data. Given the limited amount of data available in typical genomic studies, we propose 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. We further impose a sparsity constraint which stabilizes the parameter estimates. A computationally efficient and accurate model selection criterion is used for choosing automatically the level of sparsity. We show an application of this methodology to a high-resolution time-course gene expression data of *Neisseria meningitidis*.

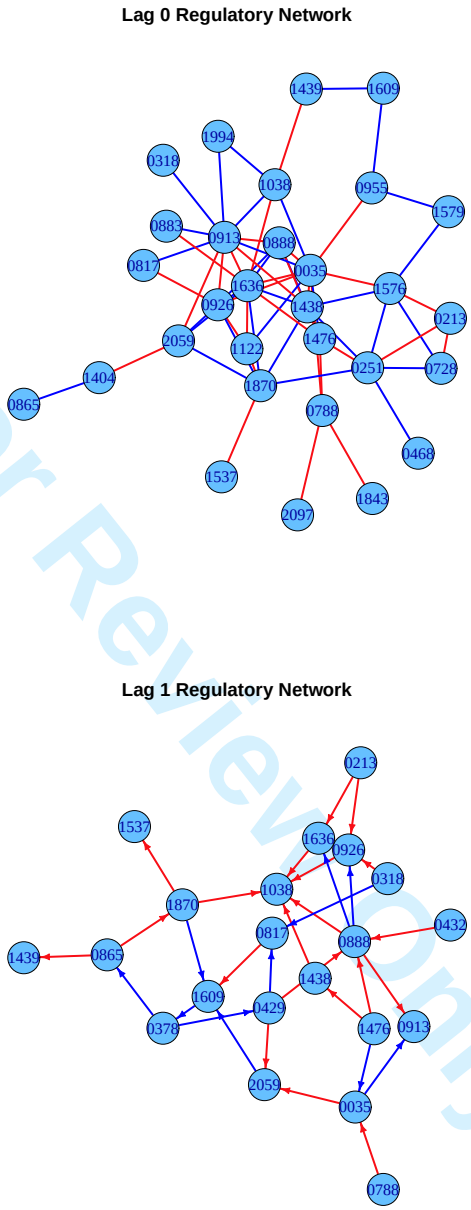

Figure 4: Static (lag 0) and dynamic (lag 1) regulatory network of 60 proteins in Neisseria meningitidis. The links were found in at least 50% of bootstrap samples with the optimal ρ chosen using KLCV(SPMLE). Red links correspond to positive partial correlations, blue links to negative partial correlations.

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.

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.

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

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.

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

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

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

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

Liu, H., K. Roeder, and L. Wasserman (2010): "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Advances in Neural Information Processing Systems*, 1432–1440.

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

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. Ngele, 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.

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.

Vujacic, I., A. Abbruzzo, and E. Wit (2013): "Kullback-leibler loss in gaussian graphical models," *arXiv preprint arXiv:1309.6216*.

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.