Sparse model-based network inference using Gaussian graphical models

Ernst Wit[1] and Antonio Abbruzzo [2]

[1] Johann Bernoulli Institute, University of Groningen
[2] Statistics Department, University of Palermo

Abstract

We consider the problem of estimating a sparse dynamic Gaussian graphical model with L_1 penalized maximum likelihood of structured precision matrix. The structure can consist of specific time dynamics, known presence or absence of links in the graphical model or equality constraints on the parameters. The model is defined on the basis of partial correlations, which results in a specific class precision matrices. A priori L_1 penalized maximum likelihood estimation in this class is extremely difficult, because of the above mentioned constraints, the computational complexity of the L_1 constraint on the side of the usual positive-definite constraint. The implementation is non-trivial, but we show that the computation can be done effectively by taking advantage of an efficient maximum determinant algorithm (SDPT3) developed in convex optimization. For selecting the tuning parameter, we compare several selection criteria and argue that the traditional AIC and BIC should not expect to work. We compare our method with related methods, such as *glasso* (Friedman et al. 2007).

Key words: Covariance Selection; Lasso; $SDPT3$ Algorithm; Penalized likelihood; Gaussian Graphical Model; Structured Correlation Matrix.

1 Introduction

A multivariate Gaussian graphical model (GGM) for an undirected graph G is defined in terms of its Markov properties. Variables, i.e. nodes in the graph, are independent conditional on a separating set. In other words, let $X = (X_1, X_2, \ldots, X_p)^T$ be a multivariate Gaussian vector, then an undirected edge is drawn between two nodes i and j, if and only if the corresponding variables X_i and X_j are conditionally dependent given the remaining variables. Let $G = (X, E)$ be an undirected graph with vertex set $X = \{X_1, ..., X_p\}$ and edge set $E = \{e_{ij}\}$, where $e_{ij} = 1$ or 0 according to whether vertices i and j are adjacent in G or not. The GGM model $N(G)$ consists of all p-variate normal distributions $N_p(\mu, \Sigma)$, for arbitrary mean vectors μ and covariance matrices Σ, assumed nonsingular, for which the concentration or precision matrix $\Theta = \Sigma^{-1}$ satisfies the following linear restriction

$$e_{ij} = 0 \Leftrightarrow \theta_{ij} = 0.$$

The model $N(G)$ has also been called a covariance selection model (Dempster 1972) and a concentration graph model (Cox 1996, Chapter 2). The reader is referred to Whittaker (1990, Chapter 6) for statistical properties of these models, including methods for parameter estimation, model testing and model selection. The model $N(G)$ also can be defined in terms of pairwise conditional independence. If $X = (X_1, \ldots, X_p)^T \sim N_p(\mu, \Sigma)$, then

$$\theta_{ij} = 0 \Leftrightarrow X_j \perp X_i | X_{\{-(i,j)\}} \Leftrightarrow \rho_{ij} = 0$$

where

$$\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ij}\theta_{ij}}}$$

denotes the partial correlation between X_i and X_j, i.e. the correlation between X_i and X_j given $X_{\{-(i,j)\}}$. This suggests that the determination of the graph G, can be based on the set of sample partial correlations $\hat{\rho}_{ij}$ arising from independent and identically distributed observations $X \sim N_p(\mu, \Sigma)$, where $n >> p$ is assumed in order to guarantee positive definiteness of the sample covariance matrix. In other words, given a random sample X we wish to estimate the concentration matrix Θ. Of particular interest is the identification of zero entries in the concentration matrix $\Theta = \{\theta_{ij}\}$, since a zero entry $\theta_{ij} = 0$ indicates the conditional independence between the two variables X_i and X_j given all other variables.

Graphical models are probability models for multivariate random variables whose independence structure is characterized by a conditional independence graph. The standard theory of estimating GGMs can be exploited only when the number of measurements n is much higher than the number of variables p. This ensures that the sample covariance matrix is positive definite with probability one. Instead, in most application, such as microarray gene expression data sets, we have to cope with the opposite situation $(n \ll p)$. Thus, the growing interest in "small n, large p" problems, requires an alternative approach. In problems where the number of nodes is large, but the number of links are relatively few per node, sparse inference of Θ in the framework of a GGM is useful, because:

- it reduces the complex high dimensional object into simpler, low dimensional objects;
- it groups the variables into several sets;
- it highlights that some variables are crucial;
- it asserts that some variables will be sufficient to predict other ones.

Estimating the dimensionality of the GGM model is complicated issue. The standard approach is greedy stepwise forward-selection or backward-deletion, and parameter estimation is based on the selected model. In each step the edge selection or deletion is typically done through hypothesis testing at some level α. It has long been recognized that this procedure does not correctly take account of the multiple comparisons involved (Edwards, 2000). Another drawback of the common stepwise procedure is its computational complexity. To remedy these problems, Drton and Perlman (2004) proposed a method that produces conservative simultaneous $1 - \alpha$ confidence intervals, and use these confidence intervals to do model selection in a single step. The method is based on asymptotic considerations. Meinshausen and Buehlmann (2006) proposed a computationally attractive method for covariance selection that can be used for very large Gaussian graphs. They perform neighbourhood selection for each node in the graph and combine the results to learn the structure of a Gaussian concentration graph model. They showed that their method is consistent for sparse high-dimensional graphs. However, in all of the above mentioned methods, model selection and parameter estimation are done separately. The parameters in the concentration matrix are typically estimated based on the model selected. As demonstrated by Breiman (1996), the discrete nature of such procedures often leads to instability of the estimator: small changes in the data may result in very different estimates.

Here, we propose a sparse dynamic Gaussian graphical model with L_1 penalty of structured correlation matrix that does model selection and parameter estimation simultaneously in the Gaussian concentration graph model. We employ an L_1 penalty on the off-diagonal elements of the correlation matrix. This is similar to the idea of the *glasso* (Friedman, 2007). The L_1 penalty encourages sparsity and at the same time gives shrinkage estimates. In addition, we can model arbitrary, locally additive models for the precision matrix, while explicitly ensuring that the estimator of the concentration matrix is positive define. This is achieved via an efficient semi-definite programming algorithm $(SDPT3)$.

The remainder of the paper is organized as follows. In Section 2, we provide the problem of gaussian profile likelihood model and the linked with the maximum determinant problem. The type of model will be specified. In Section 3, we describe the algorithm $SDPT3$ to solve the optimization problem, the general idea and the implementation of the method as well.

2 Dynamic Gaussian graphical model for networks

The graph structure of the Gaussian graphical model describes the conditional independence structure between the variables. The two main applications of this conditional independence are either (i) modular dependency structures and (ii) Markovian dependency structures. The former are used in expert systems or flow-chart descriptions of causal structures, whereas the latter is typical for spatio-temporal forms of (in)dependence. A dynamic gaussian graphical model for a network contains both types of conditional dependence: a Markovian dependence structure would capture that temporal relatedness of nearby observations, which is broken by one (or more) conditioning, intervening observations. The network itself has an internal relatedness due to the modular structure of the network: the results of the observed outcomes at the nodes flow through the links to the other nodes, thereby affecting neighbouring vertices. Due to its computational tractibility is the multivariate normal distribution uniquely suited as an initial model for a dynamic graphical model. For example, its conditional independence structure is simply characterized in terms of zeros in the inverse of its covariance matrix,

$$\Theta = \Sigma^{-1}.$$

If we measure a univariate outcome at p nodes across T discrete time-points, then initially we describe the data X as coming from a multivariate normal distribution:

$$X \sim N_{pT}(\mu, \Theta^{-1}).$$

In many practical example, it may be the case that only a single replicate X has been observed. Estimation will only be possible if we are willing to impose restrictions on the parameters. There are two types of restrictions that we will consider: sparsity restrictions and model definitions.

2.1 Sparsity restrictions of the precision matrix

The arrival of the high-throughput era in genomics has seen an explosion of data gathering: for a fraction of the amount of time and money it used to cost to monitor the level of a particular gene or protein, now thousands are monitored. Nevertheless, the underlying physical reality will not have changed as a result of our data-gathering. The particular protein that used to bind to the promotor region of the particular gene will still do so: the fact that we monitor thousands of genomic variables has not made the genomic reality itself any more difficult. Obviously, this reality is certainly highly complex, but at the same time it is also highly structured as DNA sequences are highly specific for binding to particular proteins. Therefore, the genomic network can be thought to be highly sparse set of relations between thousands of genomic players, such as DNA, mRNA and proteins. Obviously, we don't know exactly which links should be assumed to be zero, but we want to create a model that encourages zeroes between the vertices.

Furthermore, the fact that we are considering dynamic models with observations of the genomic system spaced in time, it is probably sufficient to assume – especially given the usual spacing of genomic observations – the existence of first or at most second order Markov dependence. This means that large part of the precision matrix can be filled with zeroes *a priori*.

2.2 Model restrictions of the precision matrix

Given the sparsity of the data, it is essential to define models that are finely tuned to be able to estimate interesting quantities of interest. For example, we have seen in the previous paragraph that Markov assumptions are sensible ways to reduce the dimensionality of the estimation problem. Additionally, given that the temporal correlation is probably not particularly important, it makes sense to compromise a little on the amount of variables we use to model it. For example, makes sense to restrict the attention to models in which

$$\forall i, t : \text{cor}(x_{i,t}, x_{i,t-1}|x_{-i}) = \rho.$$

This reduces the number of parameters in Θ by $pT - 1$. Moreover, it may, in certain circumstances, be sensible to assume that the genomic network at each time-point is the same. This reduces the number of parameters by $(T-1)p^2$.

2.3 Maximum Likelihood

The most simple model is the unconstrained Θ with no penalty on the elements θ_{ij} on the precision matrix Θ. The log-likelihood for μ and $\Theta = \Sigma^{-1}$ based on a random sample $X = (X^{(1)}, \ldots, X^{(n)})$ is

$$l(\mu, \Sigma; X) \cong \frac{n}{2} \log|\Theta| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T\Theta\ \ (X_i - \mu) \tag{1}$$

up to a constant not depending on μ and Θ. Even if $S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$ is of full rank (only if $n > pT$), the matrix S^{-1} will not be 'sparse'. To achieve 'sparse' graph structure and to obtain a better estimator of the concentration matrix, we introduce an L_1 penalty on the likelihood, i.e. we want a minimizer Θ of

$$-\log|\Theta| + \text{trace}(\Sigma S) \tag{2}$$

$$\text{subject to} \sum_{i \neq j} |\theta_{ij}| \leq t,$$

over the set of positive definite matrices Θ. Here $t \geq 0$ is a tuning parameter.

The constraint as formulated above does not penalize the diagonal of Θ. We could also choose not to penalize links that we know are there or time-dependencies which are so low-dimensional that it is not worth penalizing.

3 Max Determinant optimization problem

The non-linearity of the objective function, the positive definiteness constraint and the structured correlation make the optimization problem non-trivial. We take advantage of the connection of the penalized likelihood and the the the max-determinant optimization problem (Vandenberghe et al. 1996). We make use of the SDPT3 algorithm (Toh) to manage higher dimensional problems. We consider the optimization problem:

$$\min\ \ c^T\beta + log|\Theta(\beta)| \tag{3}$$

$$\text{subject to}\ \ \Theta(\beta) \geq 0$$

$$F(\beta) \geq 0$$

$$L\beta = b;$$

where the optimization variable is the vector $\beta \in R^m$. The functions $\Theta : R^m \rightarrow R^{l\times l}$ and $F : R^m \rightarrow R^{n\times n}$ are affine:

$$\Theta(\beta) = \Theta_0 + \beta_1\Theta_1 + \ldots + \beta_m\Theta_m$$

$$F(\beta) = F_0 + \beta_1 F_1 + \ldots + \beta_m F_m,$$

where $\Theta_i = \Theta^T$ and $F_i = F_i^T$. The inequality signs in (3) denote matrix inequalities, *i.e.*, $\Theta(\beta) > 0$ means $z^T\Theta(\beta)z \geq 0$ for all nonzero z and $F(\beta) \geq 0$ means $z^T F(\beta)z \geq 0$ for all z. We will refer to problem (3) as a $maxdet$ problem.

The $maxdet$ problem is a convex optimization problem, i.e. the objective function $c^T\beta + log|\Theta(\beta)|$, is convex (on $\{x : \Theta(\beta) \geq 0\}$, and the constraint set is convex. The current version of $SDPT3$, version 4.0, is designed to solve conic programming problems whose constraint cone is a product of semidefinite cones, second-order cones, nonnegative orthants and Euclidean spaces; and whose objective function is the sum of linear functions and log-barrier terms associated with the constraint cones. This means that it is possible to solve more general problems then $maxdet$ algorithm. The algorithm implemented in SDPT3 is an infeasible primal-dual path-following algorithm, at each iteration, it first computes a predictor search direction aimed at decreasing the duality gap as much as possible. After that, the algorithm generates a Mehrotra-type corrector step with the intention of keeping the iterates close to the central path. However, it does not impose any neighbourhood restrictions. Initial iterates need not be feasible the algorithm tries to achieve feasibility and optimality of its iterates simultaneously. The algorithms can start with an infeasible starting point. However, the performance of these algorithms is quite sensitive to the choice of the initial iterate so it is desirable to choose an initial iterate that at least has the same order of magnitude as an optimal solution of the $SQLP$.

4 Simulation Study

We consider a small simulation study with $n = 10$ replicates, whereby we sample $p = 5$ independent vertices with time-correlation across $T = 2$ time points. The true model for Θ, therefore, is matrix with 2 identical 5×5 diagonal matrices on its diagonal and with 2 identical 5×5 diagonal matrices on its off-diagonal. The number of true parameters in Θ is therefore 2, one for the diagonal term and one or the off-diagonal term. We fit 5 different models. Three are versions of the glasso model (Friedman et al. 2007): the default glasso considers the full 10×10 matrix, glasso1 considers the 15 parameter model with a free diagonal and free off-diagonal, and glasso2 considers the 35 parameter model with two free 5×5 diagonal blocks and a off-diagonal matrix with only entries on its diagonal. Furthermore, we consider 2 constrained dynamic models: MaxDet3 has tree parameters, a diagonal, the off-diagonal entries of the diagonal blocks and the diagonal of the off-diagonal blocks, and MaxDet400 considers a model with 20 parameters, consisting of two identical, but free, matrices on the diagonal and an off-diagonal matrix with a free diagonal and the rest zeroes.

References

Bozdogan, H. Haughton, D.M.A (1998). Information complexity criteria for regression models. Computation Statistics & Data Analysis 28: 51-76.

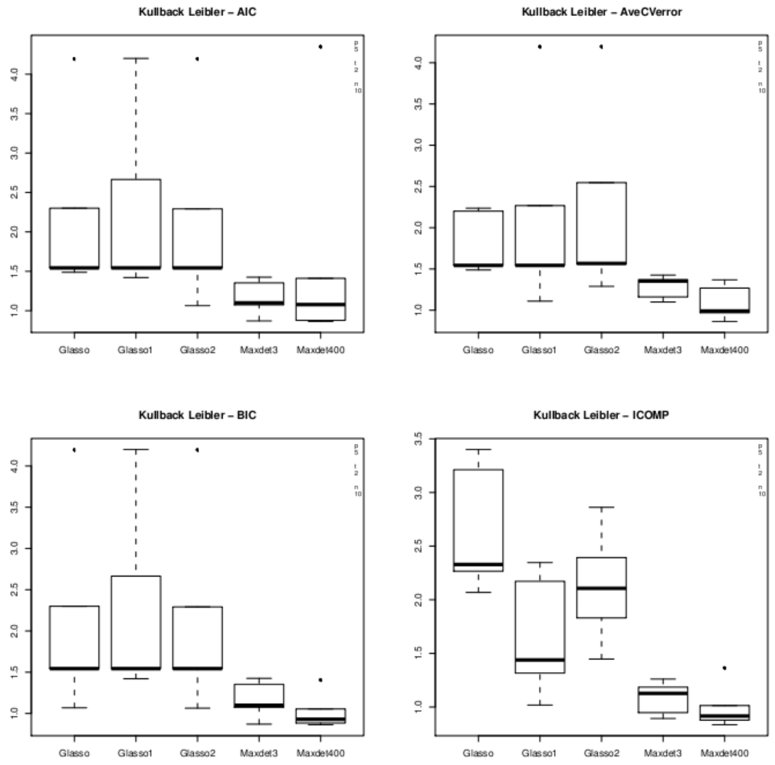

Figure 1: Average Kullback-Leibler divergence across 5 simulation for different selection criteria of the tuning parameter, namely AIC, Average cross-validation error, BIC and ICOMP (Bozdogan and Haughton 1998).

Cox D. R. and Nanny Wermuth, 1996. *Multivariate Dependencies*, Chapman Hall/CRC Press.

Dempster D.P., 1972. Covariance selection. *Biometrika*, 91, 591-602.

Drton, M. and Perlman, M. D., 2004. Model selection for Gaussian concentration graphs. *Biometrika*, 91, 591-602.

Edwards D.M., 2000. *Introduction to Graphical Modelling*, New York: Springer.

Friedman Jerome, Hastie Trevor and Tibshirani Robert, 2007. Sparse inverse covariance estimation with the graphical lasso. *Technical Report*.

Meinshausen N. and Buhlmann, P., 2006. High-dimensional graphs with the Lasso. *Ann. Statist.*, 34, 1436-62.

Toh K.C., Tutuncu R.H. and Todd M.J., 2006. *On the implementation and usage of* $SDPT3$ *- a MATLAB software package for semidefinite quadratic linear programming, version 4.0*.

Vanderberghe, L., Boyd, S. and Wu, S., 1996. Determinant maximization with linear inequality constraints. *Journal on Matrix Analysis Application*.

Whittaker Joe, 1990. *Graphical Models in Applied Multivariate Statistics*, Wiley, United Kingdom.