Dynamic Gaussian graphical models for modelling genomic networks

Antonio Abbruzzo[1], Clelia di Serio[2] and Ernst Wit[3]

[1] University of Palermo, Palermo, Italy
A.Abbruzzo@unipa.it
[2] Universita Vita-Salute San Raffaele, Milano, Italy
[3] University of Groningen, Johann Bernoulli Institute, Groningen, The Netherlands

Abstract. After sequencing the entire DNA for various organisms, the challenge has become understanding the functional interrelatedness of the genome. Only by understanding the pathways for various complex diseases can we begin to make sense of any type of treatment. Unfortunately, decyphering the genomic network structure is an enormous task. Even with a small number of genes the number of possible networks is very large. This problem becomes even more difficult, when we consider dynamical networks. 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 developed in convex optimization.

1 Introduction

Networks are important models to address specific questions in genomics. Dynamic gene-regulatory networks are complex objects since the number of potential components involved in the system is very large. For example, one important direction in systems biology is to discover gene regulatory networks from microarray data based on the observed mRNA levels of thousands of genes under various conditions. We shall show that one solution to such problem is the use of penalized Gaussian graphical models, which have been extensively used to estimate sparse static graphs.

Proteins are essential parts of the cell that determine the cell's structure and execute nearly all its functions. The production of proteins is carried out by the *ribosomes*, but the information needed for their production is encoded in *genes*

which are the segments of *DNA*. DNA contains valuable genetic information, that must be preserved. Transient *RNA* is used to carry the message from DNA to ribosomes. In all living cells, the flow of genetic information is thought to go in this way

$$DNA \rightarrow RNA \rightarrow \text{PROTEIN}.$$

This fundamental principle in biology is called the *central dogma* of molecular biology. The step from DNA to RNA consists of copying the information from genes to RNA and it is called *transcription*. The step from RNA to protein consists of decoding the information from RNA by ribosomes and it is called *translation*. Together these two processes are known as *gene expression*.

The process of transcription is carried out by special enzymes called *RNA polymerases* (RNAp). RNA polymerase binds to the promoter and then opens up the double helix of the DNA sequence immediately in front of it and slides down the gene producing the RNA molecule. The *promoter* is a region of DNA that facilitates the transaction of a particular gene and contains a sequence of nucleotides indicating the starting point for RNA synthesis. Chain elongation continues until enzyme encounters a second signal in DNA, the *terminator*, where RNAp halts and releases both the DNA chain and the newly made RNA chain. RNA which encodes information for production of a certain protein is called *messenger RNA*(mRNA).

However, to do all of this RNAp needs help from special proteins called *transcription factors*. Transcription factors bind at the promoter and form a transcription initiation complex. They position the RNAp correctly on the promoter and aid in pulling apart the two strands of DNA to allow transcription to begin and to allow RNAp to leave promoter as transcription begins. After RNAp is released from the complex it starts making RNA. Once transcription has begun, most of the transcription factors are released from the DNA so that they are available to initiate another round of transcription with a new RNAp molecule. The synthesis of the next RNA usually starts before the first RNA is completed. There maybe several polymerases moving along a single stretch of DNA and RNAs.

The main goal of gene transcription is to produce mRNA which will be translated by ribosomes to make proteins. Each mRNA can be translated several times by 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, where one regulatory protein controls 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, proteins etc. The directed or undirected edges from one node to another represent the corresponding interaction, for example, activation, repression or translation. Being able to create gene regulatory networks from experimental data and to use them to think about their dynamics is the aim of this paper.

2 Graphical models

An *undirected graphical model* is also called a Markov random field. It is defined as a pair $(G, \mathbb{P})$ that specifies a probability density function f for their joint distribution $\mathbb{P}$ in the form

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

where C is a set of cliques, i.e. complete subsets of V that are maximal, in G, $\psi_c(\mathbf{y}_c)$ 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. If the factorization (F) is possible, then it implies the 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 S separates A from B in G

$$\text{(G)} \qquad \mathbf{Y}_A \perp \mathbf{Y}_B | \mathbf{Y}_S.$$

The global Markov property in turn implies the local and pairwise Markov properties. A probability distribution function is said to obey:

(L) the *local Markov property*, relative to G, if for any vertex $i \in V$

$$Y_i \perp \mathbf{Y}_{V \setminus \{cl(i)\}} | \mathbf{Y}_{bd(i)},$$

(P) the *pairwise Markov property*, relative to G, if for any pair (i, j) of non-adjacent vertices

$$Y_i \perp Y_j | \mathbf{Y}_{V \setminus \{i,j\}},$$

The boundary of i is the set of nodes such that $bd(i) = pa(i) \cup ne(i)$, and the closure of i is the set of nodes such that $cl(i) = i \cup bd(i)$. The expression $V \setminus \{i, j\}$ indicates the set of nodes V except nodes i and j. The expression $Y_i \perp Y_j | \mathbf{Y}_{V \setminus \{i,j\}}$ means that the probability distribution function can be factorized as follows:

$$f_{Y_i, Y_j | \mathbf{Y}_{V \setminus \{i,j\}}}(y_i, y_j | \mathbf{y}_{V \setminus \{i,j\}}) = f_{Y_i | \mathbf{Y}_{V \setminus \{i,j\}}}(y_i | \mathbf{y}_{V \setminus \{i,j\}}) f_{Y_j | \mathbf{y}_{V \setminus \{i,j\}}}(y_j | \mathbf{y}_{V \setminus \{i,j\}}).$$

It can be shown that $(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P)$ (Lauritzen, 1996). Moreover, Hammersley and Clifford's theorem states that:

Theorem 1 (Hammersley and Clifford). *A probability distribution $\mathbb{P}$ with positive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property with respect to an undirected graph G if and only if it factorizes according to G.*

This theorem gives the necessary and sufficient condition for $(P) \Leftrightarrow (F)$, and under this condition we have that all Markov properties are equivalent:

$$(F) \Leftrightarrow (G) \Leftrightarrow (L) \Leftrightarrow (P).$$

Undirected graphical models are useful when random variables can be analysed symmetrically. Specific undirected graphical models are distinguished by the choice of the undirected graph G and the potential functions ψ_c.

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 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 and Wermuth, 1996). 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 $\rho_{ij} = -\theta_{ij}/\sqrt{\theta_{ij}\theta_{ij}}$ 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.

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.

3 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 flowchart 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. 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.

3.1 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*.

3.2 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, it makes sense to restrict the attention to models in which

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

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 $(T-1)p^2$.

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

$$-\log |\Theta| + \text{trace}(\Sigma S) \text{ subject to } \sum_{i \neq j} |\theta_{ij}| \leq t, \tag{2}$$

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

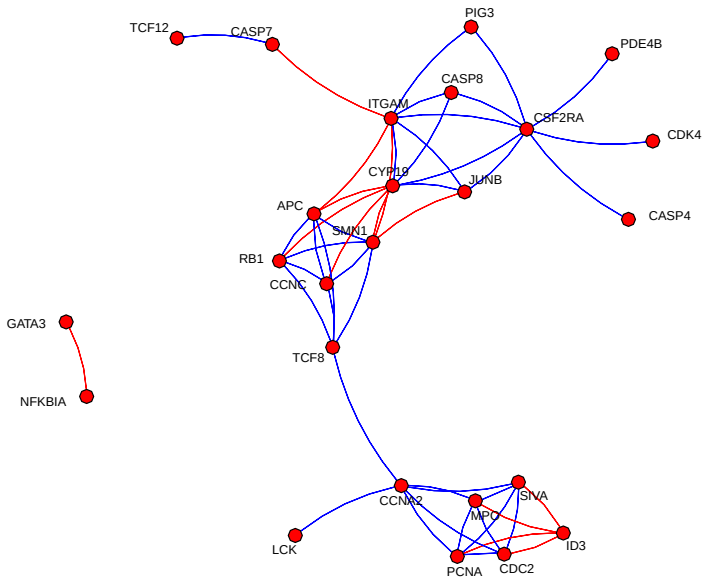

Fig. 1. The lag zero network selected in the case of the T-cell data. It shows two hubs involving the JUNB and CCNC genes, which are well-known for being central regulator. Blue and red links represent positive and negative partial correlations, respectively.

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.

4 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 (Vanderberghe et al., 1996). We make use of the SDPT3 algorithm (Toh, 2006) to manage higher dimensional problems. We consider the optimization problem:

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

$$\text{subject to } \Theta(\beta) \geq 0, \quad F(\beta) \geq 0, \quad L\beta = b;$$

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

5 Application to T-cell data

Tcell dataset is a large time-series experiment to characterize the response of a human T-cell line (Jurkat) to PMA and ionomycin treatment. The data set contains the temporal expression levels of 57 genes for 10 unequally spaced time points. At each time point there are 44 separate measurements. See Rangel et al. (2004) for more details.

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 570×570 inverse covariance matrix Θ. This requires estimating more than 160,000 parameters with only 25,000 observations. However, there is good reason to impose some constraints on Θ.

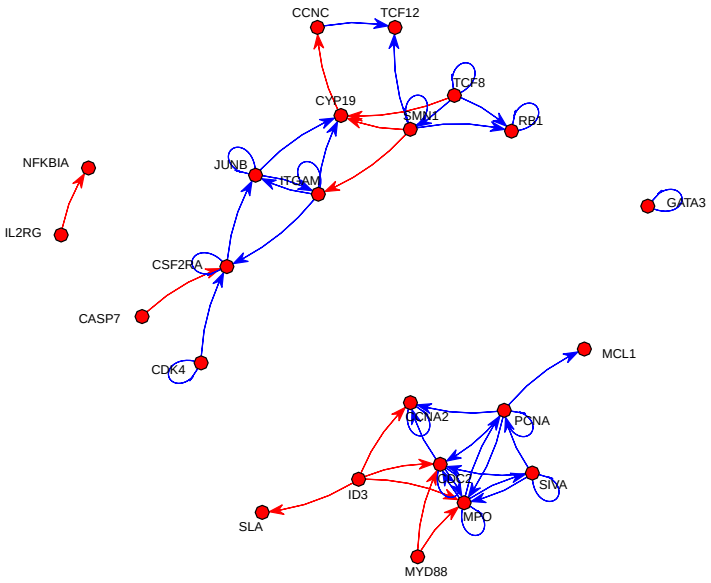

Fig. 2. The lag one network for the T-cell data: the arrows are a semantic interpretation of the graphical model. They are given their direction by pointing from the past to the future, although in the structure of the graphical model they are in fact undirected. Blue and red links represent positive and negative partial correlations, respectively.

1. **Markov assumption:** we assume that except for lag zero and lag one, there are no higher order interactions between the genes, i.e.,

$$Cov(X_{gt}, X_{g't'}) = 0 \text{ for } |t - t'| > 1.$$

2. **Interaction persistence:** For the lag zero and lag one interactions, we assume that the interactions are persistent across all ten time points, i.e.,

$$\text{Lag 0: } \Omega_{gt,g't} = \Omega_{gs,g's},$$
$$\text{Lag 1: } \Omega_{gt,g't+1} = \Omega_{gs,g's+1}.$$

This reduces the number of parameters from over 160,000 to a manageable number less than 5,000. Furthermore, the shrinkage induced by the L_1 penalty further stabilizes the estimates. The application of the above model to the T-cell data, results in the lag zero graph shown in Figure 1 and the lag one graph shown

in Figure 2. Blue and red links represent positive and negative partial correlations, respectively. We see a typicall feature that the majority of links are blue, as it is impossible to have stable networks with *a lot* of negative interactions. Furthermore, the networks we infer seem to have other typical characteristics of genomic networks, such as modularity and small world properties.

6 Conclusions

As more and more large datasets become available, the need for efficient tools to analyse such data has become imperative. In this paper, we have considered sparse dynamic Gaussian graphical models with ℓ_1-norm penalty. This type of modelling offers a straightforward interpretation: the edges of the graph define the partial conditional correlations among the nodes. In particular, under the sparsity assumption, a large part of the precision matrix can be filled with zeroes a priori. Based on the consideration of dynamic and model-oriented definitions, we are able to reduce the number of parameters to be estimated, which allows for more relevant interpretations in real data analysis.

References

Breiman, L.: Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**(6), 2350–2383 (1996)

Cox, D. R. and Wermuth, N.: *Multivariate Dependencies*, Chapman Hall/CRC Press (1996)

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

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

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

Friedman J., Hastie T. and Tibshirani R.: Sparse inverse covariance estimation with the graphical lasso. *Technical Report* (2007)

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

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

Rangel, C. and Angus, J. and Ghahramani, Z. and Lioumi, M. and Sotheran, E. and Gaiba, A. and Wild, D.L. and Falciani, F.: Modeling T-cell activation using gene expression profiling and state-space models, *Bioinformatics*, **20**(9), 1361–1372 (2004)

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

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

Whittaker J.: *Graphical Models in Applied Multivariate Statistics*, Wiley, United Kingdom (1990).