Estimation of Large Covariance Matrices and their Inverses by Graphical Modelling

May 5, 2017

Abstract

This chapter reviews graphical modelling techniques for estimating large covariance matrices and their inverse. The chapter provides a selective survey of different models and estimators proposed by the graphical modelling literature and offers some practical examples where these methods could be applied in health economics.

1 Introduction

In the last few years, there has been a growing literature, both empirical and theoretical, on methods for estimating large covariance matrices, and their inverse, in a context where the number of variables largely exceeds the number of observations. Under this case, it is well-known that the sample covariance matrix is singular, and the aggregation of massive amount of estimation errors can have considerable adverse impacts on the estimation accuracy (Stein 1956). Estimators of the covariance matrix and its inverse that are more accurate and well-conditioned than the sample covariance matrix have been proposed in various different areas, ranging from economics and finance to health, biology, computer science and engineering. In economics and finance, a growing literature has been developing econometric methods for large, cross sectionally correlated panel data, in order to improve the explanatory and predictive power of conventional models. A variety of econometrics models have been proposed to capture contemporaneous correlations, such as the spatial autoregressive or the conditional autoregressive model (see Cressie (1993); Anselin (2010)), and the common factors specification (Bai 2003). These models have been widely adopted for studying many empirical problems, ranging from the analysis of the propagation of consumer's behaviour across a population, the study of technology diffusion, the analysis of co-movements of economic and financial time series, and how linkages between countries explain regional income variations. Moreover, spatial statistical models have been adopted in public health and epidemiology, for modelling small-area disease incidence and for disease mapping in order to investigate the spread of a disease and estimate local disease risk (see Stern and Cressie (2000); Lee, Rushworth, and Sahu (2014) and Chapter XX of this book).

Methods for estimating large covariance matrices have been proposed by the statistical and computer science literature and applied in many areas such as machine learning, biology and medicine. The overall framework for the methodology is that of graphical modelling, which aims at exploring the relationships among a set of random variables through their joint distribution. Under this framework, the Gaussian distribution is often assumed so that the dependence structure is completely determined by the covariance matrix, whereas the conditional dependence structure (the network) is completely determined by the inverse of the covariance matrix, whose off-diagonal elements are proportional to partial correlations (Lauritzen 1996). A number of methods for tackling the challenges of high-dimensional data have resorted to the use of regularization methods, whereby the model likelihood is appropriately penalised in order to achieve sparsity as well as efficient and stable inference. Two seminal papers in the context of graphical modelling are those of (ADD Banerjee et al 2008), who first introduced penalised likelihood for Gaussian graphical models, and ADD Friedman et al (2008), who provided a lasso-based coordinate descent algorithm for efficient estimation. A wide range of extensions and alternative regularisation techniques have been proposed and applied in many different areas. In machine learning, they have been adopted for image and vision processing or magnetic resonance imaging, as well as for speech recognition (Bach et al., 2012; Mardia (1988)), while in biology and medicine these methods have been employed to infer the interactions between biological entities, such as proteins and metabolites, in a biological system under specific conditions (such as disease and time) from microarray and next-generation sequencing data (ADD Abegaz and Wit 2013, Vinciotti et al 2016) or for the inference of brain networks from time-series fMRI data (ADD Cribben and Yu, 2017, Estimating whole brain dynamics using spectral clustering).

This chapter provides a selective review of recent methods for estimating large covariances and their inverses using graphical modelling techniques, from both a theoretical and applied perspective. The outline of the paper is as follows. Section 2 introduces the notion of Gaussian graphical model, while Section 3 provides a bridge between graphical models and models from the spatial econometrics literature. Section 4 reviews the graphical modelling approach to estimation of large precision matrices, and illustrates methods for model selection. Section 5 introduces discrete models and reviews approaches to model and estimate graphical models with discrete variables. Finally, Section 6 concludes with a discussion of existing and potential application of these approaches to tackle health economics issues.

2 Gaussian graphical models

Let y_{it} be the observation on the ith cross section unit at time t, with $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$, and assume that $\mathbf{y}_t = (y_{1t}, y_{2t}, ., , , .y_{Nt})' \sim N(\mu_t, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a $N \times N$ symmetric and positive definite matrix, independent of t, having inverse $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, known as the*precision* matrix. In many applications it is convenient to set the mean of the joint distribution $\mu_t = \mathbf{X}_t\beta$,where β is a k-dimensional vector of unknown parameters, and $\mathbf{X}_t = (\mathbf{x}'_{1t}, \mathbf{x}'_{2t}, ..., \mathbf{x}'_{Nt})'$ is a $N \times k$ matrix of individual-specific, observed regressors.

One key result in the Gaussian graphical modelling literature is that there exists a one-to-one correspondence between the joint Gaussian distribution of a vector of random variables and its conditional Gaussian distribution. More specifically, letting $\mathbf{y}_{t|-(j,k)} = \{y_{\ell t} : \ell \neq j, k\}$, it is possible to show that the (i, j)th element of the precision matrix, θ_{ij}, is zero if and only if y_{jt} is independent of y_{kt} given $\mathbf{y}_{t|-(j,k)}$ (Lauritzen 1996). This result offers an appealing graphical interpretation of the distribution of $\mathbf{y}_t$ as a Gaussian graphical model. In particular, the dependency structure of $\mathbf{y}_t$ can be represented by an undirected graph, namely a collection of nodes and edges, $G = (V; E)$, where V contains the nodes corresponding to the N variables in $\mathbf{y}_t$ and E is the set of edges which represent the conditional dependency relationship between variables. The distribution of a variable observed in a certain node, given values observed in all other nodes, depends only on the observations in its neighborhood, and estimating θ_{jk} is equivalent to estimating whether or not there exists an edge between units j and k. Thus, estimating parameters and identifying zeros in the concentration matrix are equivalent to parameter estimation and model selection in the corresponding Gaussian graphical model. We next provide a bridge between graphical models and the range of models proposed by the spatial econometrics literature. While the spatial econometrics literature has been largely immune to the developments in Gaussian graphical modelling, these methods may be useful for a large number of application in the social and life sciences.

3 Link with the spatial econometrics approach

Gaussian graphical models are known in the spatial econometrics literature as Conditional Autoregressive (CAR) models. CAR models represent data from a given spatial location as a function of data in neighboring locations, and are often seen as an alternative to the well known simultaneous autoregressive (SAR) processes (Whittle 1954). Both CAR and SAR models represent data from a given spatial location as a function of data in neighboring locations, and are used to study how a particular area is influenced by neighboring areas. The neighbourhood structure is represented by the means of the so-called spatial weights matrix, usually assumed to be known a-priori using information on distance between units, such as the geographic, economic, policy, or social distance. However, in many cases the network is not known (or only partly known), so the interest is not only in quantifying the strength/sign of interactions, but also in the detection of interactions. It is possible to show that the problem of estimating the spatial weights matrix in a Conditional Autoregressive (CAR) model is equivalent to a neighbourhood selection problem in a graphical model. Under the CAR specification, y_{it} has a Gaussian conditional distribution with conditional mean and variance given by

$$E\left(y_{it}|y_{jt}, j=1,2,...,N, j\neq i\right) = \beta'\mathbf{x}_i + \sum_{j=1, j\neq i}^{N} w_{ij}\left(y_{jt}-\beta'\mathbf{x}_j\right),\ i=1,2,...,N, \tag{1}$$

$$Var\left(y_{it}|y_{jt}, j=1,2,...,N, j\neq i\right) = \sigma_i^2. \tag{2}$$

In (1), w_{ij} belong to a $N \times N$ matrix, $\mathbf{W}$, known as spatial weights matrix such that $w_{ii} = 0$. In a spatial weights matrix the rows and columns correspond to the cross section observations, and the generic element, w_{ij}, can be interpreted as the strength of potential interaction between units i and j. $\mathbf{W}$ is often written as $\delta\mathbf{W}^*$ where $\mathbf{W}^*$ is a matrix pre-specified by the user, while δ is an unknown parameter that needs to be estimated, measuring the amount of spatial dependence in the data. Estimation of δ, β and σ_i^2 is usually carried by maximum likelihood, exploiting the link existing between conditional and joint distribution, or by the generalised method of moments (Cressie 1993). Besag (1974) has shown that (1)-(2) for the conditional distribution imply the following *joint* normal distribution of $\mathbf{y}$

$$\mathbf{y}_t \sim N\left(\mathbf{X}_t\beta, \left(\mathbf{I}_N - \mathbf{W}\right)^{-1}\mathbf{\Lambda}\right), \tag{3}$$

where $\mathbf{\Lambda} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_N^2)$, provided that $(\mathbf{I}_N - \mathbf{W})$ is invertible and $\mathbf{\Sigma} = \left(\mathbf{I}_N - \mathbf{W}\right)^{-1}\mathbf{\Lambda}$ is symmetric and positive-definite. It is interesting to observe that the reverse also holds. Let

$$\mathbf{y}_t \sim N\left(\mu_t, \mathbf{\Sigma}\right), \tag{4}$$

where $\mathbf{\Sigma}$ is a $N \times N$, positive definite matrix. Then (1)-(2) holds, with (see Mardia (1988); Meinshausen and Buhlmann (2006))

$$w_{ij} = -\frac{\theta_{ij}}{\theta_{ii}}, \tag{5}$$

$$Var\left(y_i | y_j, j = 1, 2, ..., n, j \neq i\right) = \theta_{ii}^{-1}. \tag{6}$$

It follows that the problem of estimating w_{ij} in the CAR model (1)-(2) is equivalent to determining whether y_i and y_j are conditionally independent, i.e., $\theta_{ij} = 0$. The above derivation of the conditional distribution from the joint distribution of a Gaussian random vector has been widely exploited by the statistical literature to propose methods for estimating sparse graphical models. However, the spatial econometrics literature has been largely immune to these developments. From (5)-(6) it is evident that $\mathbf{W}$ can be inferred from $\mathbf{\Sigma}^{-1}$, and vice-versa. Therefore, using (5) the problem of estimating the spatial weights in a CAR model can be seen as the problem of estimating a sparse covariance (Friedman, Hastie, and Tibshirani 2010), or, equivalently, a neighbourhood selection problem (see Section 4.1). Hence, the spatial weights matrix for CAR models can be estimated by using methods from the Gaussian graphical modelling literature for estimating inverse covariance matrices.

4 Estimation

Early graphical approaches for estimating large covariance and precision matrices consisted of adopting a stepwise backward-deletion technique, which starts by removing the least significant edges from a fully connected graph, and continues removing edges until all remaining edges are significant according to some criterion, such as the p-value of a test for vanishing partial correlations. However, this procedure is computationally infeasible for data with even a moderate number of variables, and does not correctly take account of the multiple comparisons involved (Edwards 2000). More recently, researchers have proposed various regularisation techniques to consistently estimate large covariance and precision matrices. Some authors have proposed carrying element-wise transformations on the covariance matrix, such as "banding", namely replacing all entries outside a band around the main diagonal by zeros, "tapering", that is, gradually shrinking the off-diagonal elements of the sample covariance toward zero, or "thresholding", which sets small estimated elements of the covariance matrix to zero (Bickel and Levina 2008). Recently, there has been a surge of interest in regression-based approaches to sparse estimation of precision matrices, featuring methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) techniques by Tibshirani (1996). These approaches reparameterise the precision matrix in a manner that its estimation can be recast as a linear least-squares regression problem. Among these methods, the column-by-column method and the Graphical LASSO approaches have been widely applied. These are reviewed next.

4.1 Column-by-column approach

The column-by-column approach exploits a regression interpretation of the entries of the precision matrix, that stems from the one-to-one relationship between the joint normal distribution and the conditional distribution (see Section 3). Starting from this result, this approach considers the regression equation:

$$y_{it} = \sum_{j=1, j \neq i}^{N} \gamma_{ij} y_{jt} + \varepsilon_{it}.$$

As also explained in Section 3 (see, in particular, equations (1)-(2)) the regression coefficients of y_{it} on $\mathbf{y}_{-i,t}$ are given by $\gamma_{ij} = -\theta_{ij}/\theta_{ii}$, for $j = 1, 2, \ldots, N-1$. It follows that the (i,j)th entry of the precision matrix is, up to a scalar, the regression coefficients of y_{it} on $\mathbf{y}_{-i,t}$ in the multiple regression of variable y_{it} on the rest. On the basis of this result, Meinshausen and Buhlmann (2006) propose to select edges for each node in the graph by regressing the variable observed on one node on that observed on all other nodes, using L_1-penalized regression. In particular, for each node, y_{it}, the regression parameters against all other nodes $\mathbf{y}_{-i,t}$ are found by solving, for $i = 1, 2, ..., N$,

$$\widehat{\gamma}_i = \arg\min_{\beta} [||\mathbf{y}_t - \gamma_i' \mathbf{y}_{-i,t}||_2^2 + \lambda ||\gamma_i||_1], \tag{7}$$

with λ being a tuning parameter. The larger the value of λ, the stronger the constraint on the γ_{ij} coefficients.

Once $\widehat{\gamma}_i$ is obtained, we get the neighborhood edges by reading out the nonzero coefficients of $\widehat{\gamma}_i$. Under a number of regularity conditions, and sparsity of the precision matrix, Meinshausen and Buhlmann (2006) have showed that the above problem can recover the true structure of the graph. However, one problem in the above approach is the absence of symmetry. In fact, it can happen that $\widehat{\gamma}_{ij} = 0$ when predicting y_{jt}, while $\widehat{\gamma}_{ji} \neq 0$ when predicting y_{it}, or vice-versa. The final graph estimate is obtained by either an "and"-type rule, according to which $(i,j) \notin E$ if $\widehat{\gamma}_{ij} = \widehat{\gamma}_{ji} = 0$, or an "or"-type rule stating that $(i,j) \notin E$ if $\widehat{\gamma}_{ij} = 0$ or $\widehat{\gamma}_{ji} = 0$, leading to a sparser network.

A number of extensions of the above simple approach have been proposed in the literature. To deal with the lack of symmetry problem, Friedman et al. (2010) and Peng, Wang, Zhou, and Zhu (2009) propose improved versions of the Meinshausen and Buhlmann (2006) method that preserve the symmetry of the estimation problem. Sun and Zhang (2013) suggest to estimate γ_i by solving a scaled version of the LASSO problem (7). This approach has the advantage that the penalty level for each column is completely determined by the data via convex minimisation, without the need of using model selection criteria for determining the tuning parameter, λ (see Section 4.3). NOT SURE ABOUT THIS LAST SENTENCE i NEED TO CHECK.

4.2 Penalised log-likelihood approach

One of the most commonly used techniques for estimating sparse precision matrices is the penalised maximum likelihood approach, originally advanced by Li (2001). Motivated by the success of LASSO estimators in the context of linear regression with a large number of covariates, this approach can achieve model selection and parameter estimation simultaneously, by inducing sparsity in the precision matrix estimation. This is achieved by adding to a penalty on the off-diagonal entries of the precision matrix to the usual log-likelihood function, i.e.

$$\widehat{\boldsymbol{\Theta}}_\lambda = \arg\min_{\boldsymbol{\Theta}} \left[Tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \sum_{i<j} P_{\lambda_{ij}}\left(|\theta_{ij}|\right) \right], \tag{8}$$

where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix and θ_{ij} is the (i,j)th element of the precision matrix $\boldsymbol{\Theta}$, and $P_{\lambda_{ij}}(.)$ is a penalty function that depends on a tuning parameter, λ_{ij}, which regulates the weights on the corresponding θ_{ij} term. One of the commonly used convex penalty is the L_1 penalty (see, among others, Yuan and Lin (2007); Friedman, Hastie, and Tibshirani (2008); Banerjee, El Ghaoui, and d'Aspremont (2008)), leading to the objective function:

$$\widehat{\boldsymbol{\Theta}}_\lambda = \arg\min_{\boldsymbol{\Theta}} \left[Tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log|\boldsymbol{\Theta}| + \lambda \sum_{i<j} |\theta_{ij}| \right], \tag{9}$$

Banerjee, El Ghaoui, and d'Aspremont (2008) proved that the solution, $\hat{\boldsymbol{\Theta}}_\lambda$, is, by definition, symmetric and for any $\lambda > 0$ is full rank and invertible. It is interesting to observe that larger values of λ lead to stronger penalisation and hence sparser graphs, while smaller values of λ yield denser graphs. Friedman, Hastie, and Tibshirani (2010) propose to solve the minimisation problem in (9) by adopting an iterative procedure whereby at each step the coordinate descent algorithm proposed by Friedman, Hastie, Hoefling, and Tibshirani (2007) is applied. The resulting method, known as Graphical LASSO, is very simple and extremely fast, and has been widely used in empirical applications thanks also to a number of efficient implementations, e.g. the R packages `glasso` and `huge`.

The method provides a stable estimate of the precision matrix, especially in high-dimensional problems where the number of variables is far greater than the number of observations. Theoretical properties of the L_1-penalized maximum likelihood estimator in a large N scenario were derived by Rothman, Bickel, Levina, and Zhu (2008), who showed that the rate of convergence to the true precision matrix is $O_p\left(\sqrt{[(N+s)/T]\ln N}\right)$ in the Frobenious norm and $O_p\left(\sqrt{(s/T)\ln N}\right)$ in the spectrum norm where s is the total number of non-zero elements in the precision matrix.

Lam and Fan (2009) also established the so-called sparsistency property of the penalized likelihood estimator, implying that true zeros are estimated correctly with probability tending to 1. The authors also prove that the LASSO penalty produces a bias even in the simple regression setting due to the linear increase of the penalty on regression coefficients. A number of authors have proposed remedies to this bias issue. In particular, Shen, Pan, and Zhu (2012) extended the above penalized maximum likelihood approach to general non-convex penalties, such as the smoothly clipped absolute deviation (SCAD) penalty. Fan and Li (2001) show that the SCAD penalty, which corresponds to a quadratic spline function, leads to estimators with desirable statistical properties and has what is called the oracle property, which means that the estimator performs similarly to the estimator when the true model is known in advance. Finally, Fan, Feng, and Wu (2009) have proposed the so-called adaptive LASSO penalty in the context of graphical models, following the development and successful use of this penalty in the regression context Zou (2006).

Computational costs The computational cost associated to a coordinate descendent update in the Friedman, Hastie, and Tibshirani (2008) Graphical LASSO is $O(N^2)$. To decrease the computational cost, Hsieh, Sustik, Dhillon, and Ravikumar (2014) propose an algorithm called QUIC that uses Newton's method and employs a quadratic approximation which reaches a $O(N)$ computational cost. However, when N is very large, this could still be prohibitive. The last decades have witnessed the dramatic development of new data acquisition technologies which allow to collect massive amount of data with relatively low cost. When dealing with huge networks, a number of authors have proposed to exploit a-priori information on group membership of observations, or equality constrains, to propose fast, sparse estimation algorithms. These methods often allow to split a large Graphical LASSO problem into many, smaller tractable problems. Guo, Levina, Michailidis, and Zhu (2011) consider an heterogeneous data set where variables, while independent across groups, have a sparse dependency structure within group. The corresponding precision matrix has a block diagonal structure, and the authors propose joint estimation of various blocks by maximising the corresponding penalized log-likelihood functions. A similar approach is taken by Mazumder and Hastie (2012), who propose thresholding estimation of a sparse inverse covariance that is a block diagonal matrix of connected components. Wit and Abbruzzo (2015) impose block equality constraints on the parameters of an undirected graphical model to reduce the number of parameters to be estimated. Vinciotti et al. (2016) discuss various forms of block structures for dynamic networks and propose estimation of the associated precision matrix under sparsity and equality constraints on parameters (also known as parameter tying). Parameter tying does not generally reduce computational costs. However, in some cases, it means that the precision matrix can be inferred from a matrix/network at a latent lower-dimensional space ADD OUR PAPER + SEE Chapter XX.

4.3 Model selection

An important issue with the estimation methods reviewed above is the choice of the regularisation parameter λ, which controls the sparsity pattern of

$\widehat{\mathbf{\Theta}}$. Because each tuning parameter value corresponds to a fitted model, the choice of the optimal regularisation parameter can be seen as a model selection problem. The literature has proposed a wide range of techniques for selecting λ, which can be broadly divided into two groups, depending on their objective. A first class of methods optimizes the posterior probability of the model, which involves an integration over the parameters θ. Thus, these methods aims at recovering the true structure of the network. Amongst these, popular choices are the Bayesian information criterion (BIC) and the Extended Bayesian Information Criterion (EBIC). Let E be the set of edges of a candidate graph, and $l(\vartheta)$ be the log-likelihood function of the associated model. The EBIC for a particular regularisation parameter is given by (Yuan and Lin,2007):

$$EBIC_{\lambda} = -2l(\vartheta) + |E| \ln T + 4 |E| \lambda \ln N, \tag{10}$$

where $|E|$ is the number of nonzero elements of the estimated $\hat{\mathbf{\Theta}}$. If $\lambda = 0$, then the above formula reduces to the conventional BIC, which is well known to lead to (asymptotically) consistent model selection in a setting where the number of variables, N, is fixed and the sample size, T, is growing. Foygel and Drton (2010) establish the consistency of the EBIC, namely, the ability to select the smallest true graph, in a setting where both N and T grow to infinity. We refer to Wang, Li, and Tsai (2007) for an investigation of the use of BIC in the context of penalized likelihood method under a SCAD penalty.

A further class of methods aims to maximize the prediction power of the model, namely, by minimizing the Kulback-Leibler divergence between the estimated $\hat{\mathbf{\Theta}}$ and the true $\mathbf{\Theta}$. Here the values of Θ are of importance, more than the actual structure of the network (i.e. zeros and non-zeros). Popular choices amongst this class of methods are the Aikaike information criterion (AIC), Cross-Validation (CV), and the Generalized Cross-Validation (GACV). The AIC is given by:

$$AIC = -2l(\vartheta) + 2 |E| ,$$

and, as in the case of the BIC, it is has been designed for situations where the dimension N is fixed as T increases.

Cross validation consists of randomly dividing the sample into K subsets of rounghly equals size, $n_1, n_2, ..., n_K$, of which the first $K-1$ are used as the "training" sets and the last as the "validation" set. For each value of the tuning parameter a cross a sequence of values, $\lambda_1, \lambda_2, ..., \lambda_n$, estimation of the graph is carried on the $K-1$ training sets, and then the negative log-likelihood is evaluated on the retained validation set. The results are then averaged over all K groups to obtain a single CV score for each λ. Hence, the optimal parameter is selected as the value of λ that minimises the CV score, achieving the minimal prediction error in the validation sample. The advantage of CV techniques is that they make greater use of the available data, but the disadvantage is that they are computationally very intensive, although recent proposals have been put forward for computationally efficient approximations (ADD VUJACIC PAPER). Another disadvantage, however, is that CV assumes that the training pairs are exchangeable, which is not true in many real problems, in which there is a structure in the training inputs (for example, when data refer to points in time or correspond to locations on a line). Finally, Liu, Roeder, and Wasserman (2010) have shown that CV tends to overfit in graph estimation. We refer to Wit, van den Heuvel, and Romeijn (2012) for a discussion on model selection from a model uncertainty perspective, and a review of existing approaches.

The choice of the tuning parameter is still an open issue and existing criteria are not completely satisfactory. To handle the challenge of tuning parameter selection, a number of tuning insensitive estimation techniques have been proposed. Among these, it is worth citing the approach advanced by Liu and Wang (2012), known as TIGER, which can be viewed as a tuning-insensitive extension of the nodewise LASSO method proposed byMeinshausen and Buhlmann (2006). This method estimates the precision matrix in a column-by-column fashion using the SQRT-Lasso by Belloni, Chernozhukov, and Wang (2012), which asymptotically does not depend on any unknown parameters.

5 Graphical models for limited dependent variables

5.1 The Ising graphical model

Assume now that $\mathbf{y}_t$ is a N-dimensional discrete random variable, where each y_{it} can only take two values, 0 or 1. These values might for example represent health outcomes such as death, or 30-days hospital readmission, or the outcome of political election. One important model used for estimating the graph associated with a binary random variable is the Ising model. Historically, the Ising model was introduced in statistical physics for describing magnetic interactions and the phase transitions in complex systems. Recently, it has been used for modeling several phenomena in social networks, including analysis of political elections (ADD REF), technology diffusion (ADD REF), and statistical genomics (ADD REF: Cheng et al 2014, A Sparse Ising Model with Covariates, Biometrics). Due to both its importance and difficulty, the problem of structure learning for discrete graphical models has recently attracted considerable attention. Under the Ising model, the log-likelihood function associated to the vector of binary random variable, $\mathbf{y}_t$, is

$$l\left(\mathbf{y}_t, \boldsymbol{\Theta}\right) = \sum_{i=1}^{N} \theta_{ii} y_{it} + \sum_{i,j=1}^{N} \theta_{ij} y_{it} y_{jt} - \Psi(\boldsymbol{\Theta}), \tag{11}$$

where $\Psi(\boldsymbol{\Theta})$ is a normalisation constant also known as log-partition function, and θ_{ij} is a canonical parameter measuring the coupling strength between the random variables y_{it} and y_{jt}. By convention, $\theta_{ii} = 0$. It is possible to show that the vertices corresponding to y_{it} and y_{jt} are unconnected in the graph if and only if $\theta_{ij} = 0$. The sparse maximum likelihood problem in this case is to maximize (11) with an added L_1-norm penalty on terms θ_{ij}. However, a complication with this approach is that direct estimation of a regularized likelihood involves calculation of the partition function in the likelihood, which is computationally intractable in most cases, except for some simple graph structures like the tree-structured graph. Approximations of the log-partition function have been suggested by Wainwright and Jordan (2006), and Banerjee, Ghaoui, and d'Aspremont (2008), leading to approximate sparse maximum likelihood estimate, and by Hoefling and Tibshirani (2009) and Ravikumar, Wainwright, and Lafferty (2010). In particular, Ravikumar, Wainwright, and Lafferty (2010) propose a method based on carrying N nodewise L_1-regularized logistic regressions, which performs well for any bounded-degree graph having a number of observations, T, growing logarithmically in the number of nodes, N. Loh and Wainwright (2013) propose a method that is consistent under similar asymptotic rates, and requires performing node-wise L_1-regularized linear regressions for neighborhood selection. This approach however, is only valid for certain, simple graph structures.

5.2 Graphical discrete choice models

(need to better summarise)

An approach alternative to the Ising model assumes a discrete choice model for $\mathbf{y}_t$, such as the Probit or Logit specification. In the case of the Probit model, using the latent response model, assume that y_{it} is generated by tresholding the latent variable y^*_{it}:

$$\begin{aligned} y^*_{it} &= \beta' \mathbf{x}_{it} + e_{it}, & (12) \\ y_{it} &= 1 \text{ if } y^*_{it} \geq 0, \text{ 0 otherwise}, & (13) \end{aligned}$$

where e_{ir} are Gaussian random errors, and assume that $E\left(\mathbf{e}_t \mathbf{e}'_t\right) = \mathbf{\Sigma}$. The log-likehood of the observed data is:

$$l(\mathbf{y}, \vartheta) = \ln \int f_{\mathbf{y},\mathbf{y}^*,\mathbf{u}}\left(\mathbf{y}, \mathbf{y}^*, \mathbf{u}|\vartheta\right) d\mathbf{y}^* d\mathbf{u}. \quad (14)$$

Because of the high dimensional integral present in (14), it is difficult to maximize $l(\vartheta)$, as well as its penalised version, directly. To deal with the numerical difficulties, methods approximating the likelihood by Gauss-Hermite quadrature or Monte Carlo integration and then maximizing it by either Newton-Raphson or Expectation-Maximisation algorithms have been proposed (Breslow and Clayton (1993); Schilling and Bock (2005)). However, these methods can only be applied in the presence of a limited number of variables because the number of evaluation points in Gauss-Hermite quadrature increases exponentially with the number of variables. One alternative, widely used approach for estimating models for correlated binary variables combines Monte Carlo integration with various Expectation-Maximisation (EM) algorithms, leading to the so-called Monte Carlo EM algorithm (Ashford and Sowden (1970), Chib and Greenberg (1998), Gueorguieva and Agresti (2001) and McCulloch (1997)).

QUI INTRODURREI IL MODELLO MIXED MODEL E L'IDEA DI GROUP-SPECIFIC CONSTRAINS NELLA MATRICE DI COVARIANZA. Focusing on a mixed Probit model with independent random effects, McCulloch (1994) proposes Monte Carlo versions of the EM algorithm for ML estimation based on the Gibbs sampling. This approach has been extended in various directions. McCulloch (1997) considers a Metropolis-Hastings algorithm at each E-step for ML estimation of generalised linear mixed models. Chan and Kuk (1997) propose an EM algorithm where the E-step is made feasible by Gibbs sampling, for ML estimation of a mixed Probit model with correlated random effects. The authors also suggest to approximate standard error of estimates by inverting a Monte Carlo estimate of the information matrix. The proposed approach however is computationally very intensive, as it requires sampling from a multivariate Truncated Normal distribution. To deal with this problem, Tan, Tian, and Fang (2007) propose a non-iterative importance sampling approach to evaluate the first and the second order moments of a truncated multivariate normal distribution associated with the Monte Carlo EM algorithm. Although this method is faster than direct estimation of the moments, it is still computationally very demanding for large scale problems. An alternative, direct sampling-based EM algorithm is advanced by An and Bentler (2012). The authors propose to draw random samples from the prior, Gaussian distribution of random effects, which is computationally much easier than from the corresponding unknown posterior distribution, although at the expenses of a higher Monte Carlo error. Guo, Levina, Michailidis, and Zhu (2015) propose an EM estimation algorithm for graphical model for ordinal variables based on a conditional independence assumption that simplifies computation considerably (see also Behrouzi, Johannes, and Wit (2016)).

6 Potential applications in health and health economics

(to be done)

There are potentially many real world applications associated with network structure learning. Intuitively, such research can be used to analyze underlying relationships embedded under complex networks. The variants of graphical lasso can also be applied to problems where the network is evolving with the time, such as political election and gene network.

- biology applications.
- Health economics applications: Estimating the number of rivals in studying competition between health care providers.

Up to today, estimating graphical models in an efficient way is still an open research problem.

References

An, X. and P. M. Bentler (2012). Efficient direct sampling mcem algorithm for latent variable models with binary responses. *Compuutational Statistics and Data Analysis 56*, 231–244.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science 89*, 3–25.

Ashford, J. R. and R. R. Sowden (1970). Multivariate probit analysis. *Biometrics 26*, 535–546.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–171.

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.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research 9*, 485–516.

Behrouzi, P., F. Johannes, and E. Wit (2016). Detecting genetic epistatic interactions via a latent gaussian copula graphical model. *Biostatistics 0*, 1–24.

Belloni, A., V. Chernozhukov, and L. Wang (2012). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 791–806.

Besag, A. A. J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B 36*, 192–236.

Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. *Annals of Statistics 36*, 199227.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of American Statistical Association 88*, 9–25.

Chan, J. S. K. and A. Y. C. Kuk (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics 53*, 86–97.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika*, 347–361.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

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

Fan, J., Y. Feng, and Y. Wu (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics 3*, 521–541.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*, 1348–1360.

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

Friedman, J., T. Hastie, H. Hoefling, and R. Tibshirani (2007). Pathwise coordinate optimization. *Annals of Applied Statistics 2*.

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

Friedman, J., T. Hastie, and R. Tibshirani (2010). A note on the group lasso and a sparse group lasso. Technical Report, Stanford University.

Gueorguieva, R. V. and A. Agresti (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association 96*, 1102–1112.

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

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics 24*, 183–204.

Hoefling, H. and R. Tibshirani (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research 10*, 883–906.

Hsieh, C., M. A. Sustik, I. S. Dhillon, and P. Ravikumar (2014). QUIC: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research 15*, 2911–2947.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals Of Statistics 37*, 4254–4278.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Oxford Statistical Science Series.

Lee, D., A. Rushworth, and S. K. Sahu (2014). A bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics 70*, 419–429.

Li, J. F. F. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association 96*, 1348–1360.

Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 1–14.

Liu, H. and L. Wang (2012). Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. Tech. rep., Department of Operations Research and Financial Engineering, Princeton University.

Loh, P. and M. J. Wainwright (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics 41*, 30223049.

Mardia, K. V. (1988). Multi-dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis 24*, 265–284.

Mazumder, R. and T. Hastie (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research 13*, 1436–1462.

McCulloch, C. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*, 330–335.

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

Meinshausen, N. and P. Buhlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*, 735–746.

Ravikumar, P., M. Wainwright, and J. Lafferty (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics 38*, 1287–1319.

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

Schilling, S. and R. D. Bock (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika 70*, 533–555.

Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*, 223–232.

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Sympossium on Mathematical and Statistical Probability*, Volume I, pp. 197–206.

Stern, H. S. and N. Cressie (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine 19*, 2377–2397.

Sun, T. and C. H. Zhang (2013). Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research 14*, 3385–3418.

Tan, M., G. Tian, and H. Fang (2007). An efficient mcem algorithm for fitting generalized linear mixed models for correlated binary data. *J. Stat. Comput. Simul*, 929–943.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Vinciotti, V., L. Augugliaro, A. Abbruzzo, and E. Wit (2016). Model selection for factorial gaussian graphical models with an application to dynamic regulatory networks. *Statistical Applications in Genetics and Molecular Biology 15*, 193–212.

Wainwright, M. and M. Jordan (2006). Log-determinant relaxation for approximate inference in discrete markov random fields. *IEEE Transactions on Signal Processing*.

Whittle, P. (1954). On stationary processes on the plane. *Biometrika 41*, 434–449.

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

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

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*, 19–35.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.