The Econometrics Journal

Econometrics Journal (2017), volume **20**, pp. S61–S85.
doi: 10.1111/ectj.12078

Sparse estimation of huge networks with a block-wise structure

FRANCESCO MOSCONE[†], ELISA TOSETTI[†] AND VERONICA VINCIOTTI[‡]

[†]*Brunel Business School, Eastern Gateway Building, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK.*
E-mail: francesco.moscone@brunel.ac.uk, elisa.tosetti@brunel.ac.uk

[‡]*Department of Mathematics, John Crank Building, Brunel University London, Uxbridge, Middlesex UB8 3PH, UK.*
E-mail: veronica.vinciotti@brunel.ac.uk

First version received: January 2016; final version accepted: November 2016

Summary Networks with a very large number of nodes appear in many application areas and pose challenges for traditional Gaussian graphical modelling approaches. In this paper, we focus on the estimation of a Gaussian graphical model when the dependence between variables has a block-wise structure. We propose a penalized likelihood estimation of the inverse covariance matrix, also called Graphical LASSO, applied to block averages of observations, and we derive its asymptotic properties. Monte Carlo experiments, comparing the properties of our estimator with those of the conventional Graphical LASSO, show that the proposed approach works well in the presence of block-wise dependence structure and that it is also robust to possible model misspecification. We conclude the paper with an empirical study on economic growth and convergence of 1,088 European small regions in the years 1980 to 2012. While requiring *a priori* information on the block structure – e.g. given by the hierarchical structure of data – our approach can be adopted for estimation and prediction using very large panel data sets. Also, it is particularly useful when there is a problem of missing values and outliers or when the focus of the analysis is on out-of-sample prediction.

Keywords: *Block-wise dependence, Graphical LASSO, Graphical modelling, Panels, Spatial econometrics.*

1. INTRODUCTION

Estimation of large covariance matrices and their inverse has several applications in various areas, from economics and finance to health, biology, computer science and engineering. One important technique developed by the statistical and computer science literature is the graphical modelling approach, 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 and, in this case, the dependence structure is completely determined by the covariance matrix, or, equivalently, by its inverse, where the off-diagonal elements are proportional to partial correlations (Lauritzen, 1996). Specifically, variables i and j are conditionally independent given all other variables, if and only if the (i, j)th element of the inverse covariance matrix, referred to as the precision matrix, is zero. One result in the Gaussian graphical modelling literature is that there is a one-to-one correspondence between the joint Gaussian distribution of a vector

of random variables and its conditional Gaussian distribution. Under the latter, the distribution of a variable observed in a certain node, given values observed in all other nodes, depends only on the observations in its neighbourhood; see, e.g. Mardia (1988) and Meinshausen and Buhlmann (2006). Hence, the problem of estimating the (inverse) covariance matrix is equivalent to a neighbourhood selection problem. This observation has led to efficient node-wise LASSO approaches for sparse high-dimensional graphs; see, e.g. Meinshausen and Buhlmann (2006) and Peng et al. (2009). In contrast to these approaches, Friedman et al. (2008) have developed the Graphical LASSO (GLASSO) approach, where the inverse covariance matrix is directly estimated via penalized likelihood.

Conditional Gaussian models are known in the spatial econometrics literature as conditional autoregressive (CAR) models, representing data from a given spatial location as a function of data in neighbouring locations; see, e.g. Cressie (1993) and Anselin (2010). In a CAR model, the neighbourhood structure is represented by 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. It is interesting to observe that the problem of estimating the spatial weights matrix in a CAR model is equivalent to a neighbourhood selection problem in a graphical model; for more details, see Section 5. 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. 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 applications in the social sciences.

In this paper, we consider the case of networks with a very large number of nodes and we focus on the estimation of Gaussian graphical models when the dependency between variables has a block-wise structure. We assume that units can be split into a set of non-overlapping groups, or blocks, in such a way that the dependence between units only varies across blocks, instead of individual observations. Hence, rather than estimating the links between each pair of units in the sample, we propose to estimate the dependence (links) between groups of cross-sectional units. Our approach consists of applying the GLASSO methodology of Friedman et al. (2008) to block-level averages of observations rather than to single observations. When the size of the group is unity, our method collapses to the conventional GLASSO. A major advantage of this method is that its computational cost is greatly reduced and hence it can be adopted for estimation and prediction using very large, or huge, networks. Our approach is also particularly useful when there is a problem of missing values and outliers or when the focus of the analysis is out-of-sample prediction.

There exist several examples where it is reasonable to assume a block-wise dependence structure between units. In economics, preferences for consumer goods of individuals belonging to the same household may react similarly in response to consumption decisions of neighbouring households. Companies belonging to the same sector of economic activity and located within the same geographical area (e.g. the postcode, the region or the country) tend to behave similarly because they have similar characteristics or face similar opportunities and constraints. Thus, it is reasonable to assume that the way they interact with companies from other sectors and/or geographical areas is similar. A block-wise dependence structure is also a realistic assumption when the variable of interest displays an explicit hierarchical or group membership structure, namely, clustering of units in an organized fashion, such as students within classrooms, members of a household, General Practitioners in a clinic, etc. This is common, for example, when dealing with large, individual-level, microeconomic or health data sets. Other examples are in neuroscience, where the networks used to represent brain activity have a hierarchical structure,

with billions of neurons connected to each other through hub nodes, called voxels, and with connected voxels forming areas that are again connected with each other (Luo, 2015). In biology, regulatory networks are thought to have a hub-type structure, with groups of genes having a similar dependency structure and regulated by a small number of unobserved proteins (Hao et al., 2012). When the grouping is not fully known *a priori*, we could use methods that allow us to determine endogenously the optimal grouping of cross-sectional units, by employing techniques from the clustering literature; see, e.g. Lin and Ng (2012), Bonhomme and Manresa (2015) and Ando and Bai (2016).

Exploitation of *a priori* information on the group structure of variables is not new in the social interaction literature and in the statistical and graphical modelling literature. Empirical works from the social interaction literature typically assume that an individual reacts to the average of others in a predefined group; see Durlauf and Young (2001) and Blume et al. (2013) for a review. Such an assumption implies that the spatial weights matrix has a group-membership structure, where the weights are identical for all units belonging to the same group, while they are set to zero for the interaction between units belonging to different groups. Lee and Yu (2007) considered the identification and estimation of interaction effects in the context of a spatial autoregressive model where the spatial weights matrix (and the associated precision matrix) has such a block diagonal structure with equal entries. Note that this is a more restrictive assumption to that used in this paper, as it does not allow for dependences between groups. Nevertheless, this model has been widely adopted in several different areas of the social sciences, such as education (Calvó-Armengol et al., 2009), labour market outcomes (Bayer et al., 2008), crime (Sirakaya, 2006) and welfare participation (Bertrand et al., 2000). Similar models have been proposed by the statistical literature, where mixed effect models are commonly used to represent variables with a hierarchical or known group membership structure (Goldstein, 2011). When the random effects are assumed to be correlated, these models lead to a covariance matrix that has a block-wise structure of the same type that we use in this paper, with equal correlation within groups and equal correlation between any two elements of two specified groups (Laird and Ware, 1982). Maximum likelihood approaches are typically used for parameter estimation in these models. In the case of a large number of regressors, penalized approaches based on the L_1 penalty are used for estimation and variable selection (Schelldorfer et al., 2014). However, these methods typically require a small number of random effects (blocks).

A number of authors in the literature on graphical modelling have proposed sparse estimation of graphs with a block structure. These methods exploit *a priori* information on group membership of observations to propose fast, sparse estimation algorithms. Guo et al. (2011) consider a 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 maximizing 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). The inclusion of equality constraints, while reducing the number of parameters, often increases the computational complexity of the estimation procedures. For example, the general block structures considered by Wit and Abbruzzo (2015) and Vinciotti et al. (2016) imply a computational cost of the estimation

procedure that is higher compared to the approaches of Guo et al. (2011) and Mazumder and Hastie (2012), where the assumed block structure allows the large GLASSO problem to be split into many, smaller tractable problems.

In this paper, we use block structures with the intent to achieve computational efficiency, allowing us to infer networks of very large dimensions. Differently from Guo et al. (2011) and Mazumder and Hastie (2012), our approach does not need to impose block-diagonality of the precision matrix. However, we assume that units can be split into groups in such a way that the covariance (and associated precision matrix) only varies across blocks, rather than individual observations.

The rest of the paper is structured as follows. In Section 2, we describe the main features of our graphical model with block-wise dependence structure, while in Section 3 we propose our estimator based on GLASSO. In Section 4, we run Monte Carlo experiments to investigate the small-sample properties of the proposed estimator. In Section 5, we carry out an empirical study on the economic growth of a set of small regions in Europe. Finally, in Section 6, we provide some concluding remarks. The Appendix provides the proofs.

We use $|\lambda_1(\mathbf{A})| \geq |\lambda_2(\mathbf{A})| \geq \cdots \geq |\lambda_n(\mathbf{A})|$ to denote the eigenvalues of a matrix $\mathbf{A} \in \mathbb{M}^{n\times n}$, where $\mathbb{M}^{n\times n}$ is the space of $n \times n$ matrices. $\mathrm{Tr}(\mathbf{A})$ is the trace of $\mathbf{A} \in \mathbb{M}^{n\times n}$, while its Frobenius norm is $\|\mathbf{A}\|_F = (\sum_{i,j=1}^{m} a_{ij}^2)^{1/2}$. K is used for a fixed positive constant that does not depend on N; S^c is used to denote the complement of a set S.

2. BLOCK-WISE DEPENDENCE STRUCTURE IN HUGE NETWORKS

Let y_{it} be the observed data for the ith individual, $i = 1, 2, \ldots, N$, at time t, with $t = 1, 2, \ldots, T$, and assume that the N-dimensional vector $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an $N \times N$ symmetric and positive definite matrix, independent of t. For ease of exposition, we set $\boldsymbol{\mu} = \mathbf{0}$, although this assumption can be relaxed by setting $\boldsymbol{\mu}$ to a non-zero vector depending on a set of strictly or weakly exogenous regressors, including, for example, temporal lags of the dependent variable. Assume that the variables can be split into G non-overlapping groups, with $G \leq N$, such that the dependence between individuals belonging to different groups is the same for all individuals belonging to the same group. Suppose, for simplicity, that all groups are of the same size $M = N/G$, where M is an integer number. Under this assumption, $\boldsymbol{\Sigma}$ has the following block-wise structure:

$$\underset{N\times N}{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\sigma}_1 & \sigma_{12}\mathbf{1}_M & \ldots & \sigma_{1G}\mathbf{1}_M \\ \sigma_{21}\mathbf{1}_M & \boldsymbol{\sigma}_2 & \ldots & \sigma_{2G}\mathbf{1}_M \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{G1}\mathbf{1}_M & \sigma_{G2}\mathbf{1}_M & \ldots & \boldsymbol{\sigma}_G \end{pmatrix}, \tag{2.1}$$

where $\mathbf{1}_M$ is an $M \times M$ matrix of ones, and

$$\underset{M\times M}{\boldsymbol{\sigma}_g} = \begin{pmatrix} \delta_g & \sigma_{gg} & \ldots & \sigma_{gg} \\ \sigma_{gg} & \delta_g & \ldots & \sigma_{gg} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{gg} & \sigma_{gg} & \ldots & \delta_g \end{pmatrix}, \tag{2.2}$$

where σ_{gg} are intra-group covariances, while δ_g are group-specific variances, for $g = 1, 2, \ldots, G$. Let

$$\underset{G\times G}{\mathbf{\Sigma}_G} = \begin{pmatrix} \sigma_{11} & \sigma_{21} & \ldots & \sigma_{1G} \\ \sigma_{12} & \sigma_{22} & \ldots & \sigma_{2G} \\ \ldots & \ldots & \ldots & \ldots \\ \sigma_{G1} & \ldots & \ldots & \sigma_{GG} \end{pmatrix}, \qquad \underset{G\times G}{\mathbf{\Gamma}_G} = \begin{pmatrix} \gamma_1 & 0 & \ldots & 0 \\ 0 & \gamma_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \gamma_G \end{pmatrix}, \tag{2.3}$$

where $\gamma_g = \delta_g - \sigma_{gg} \geq 0$. Then, $\mathbf{\Sigma}$ can be written in compact form as

$$\mathbf{\Sigma} = (\mathbf{\Sigma}_G \otimes \mathbf{1}_M) + (\mathbf{\Gamma}_G \otimes \mathbf{I}_M), \tag{2.4}$$

where $\mathbf{\Sigma}_G$ is a $G \times G$ matrix assumed to be positive definite. If $\mathbf{\Sigma}$ has the above block-wise structure, then its inverse, namely the precision matrix, is also block-wise. To show this, rewrite

$$\begin{aligned} \mathbf{\Sigma} &= \Big(M\mathbf{\Sigma}_G \otimes \frac{1}{M}\mathbf{1}_M\Big) + \Big(\mathbf{\Gamma}_G \otimes \frac{1}{M}\mathbf{1}_M\Big) - \Big(\mathbf{\Gamma}_G \otimes \frac{1}{M}\mathbf{1}_M\Big) + (\mathbf{\Gamma}_G \otimes \mathbf{I}_M) \\ &= \Big((M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G) \otimes \frac{1}{M}\mathbf{1}_M\Big) + \Big(\mathbf{\Gamma}_G \otimes (\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M)\Big). \end{aligned} \tag{2.5}$$

Noting that $(1/M)\mathbf{1}_M$ and $(\mathbf{I}_M - (1/M)\mathbf{1}_M)$ are idempotent matrices such that their sum is the identity matrix, we can apply Lemma 2.1 (point (iv)) in Magnus (1982) to obtain

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = \Big((M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G)^{-1} \otimes \frac{1}{M}\mathbf{1}_M\Big) + \Big({\mathbf{\Gamma}_G}^{-1} \otimes (\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M)\Big). \tag{2.6}$$

Assuming that the matrix $(\mathbf{\Sigma}_G + (1/M)\mathbf{\Gamma}_G)^{-1}$ has generic elements ϕ_{gh}, the likelihood function has the simplified expression:

$$\begin{aligned} l(\theta) \approx &- \ln|M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G| - (M-1)\ln|\mathbf{\Gamma}_G| - \frac{1}{MT}\sum_{t=1}^{T}\sum_{g=1}^{G} \\ &\times\Big(\frac{1}{M}\sum_{h=1}^{G}\sum_{i\in g: j\in h} y_{it}y_{jt}\phi_{gh} + (M-1)\sum_{i\in g} y_{it}^2\gamma_g^{-1} - \sum_{i\neq j: i,j\in g} y_{it}y_{jt}\gamma_g^{-1}\Big). \end{aligned} \tag{2.7}$$

See Appendix A for a proof. Below, we propose a penalized maximum likelihood approach to estimate $\mathbf{\Sigma}$ and $\mathbf{\Theta}$, which exploits the block-wise dependence structure and is based on the GLASSO.

3. BLOCK-GLASSO APPROACH

To propose our estimator, consider the group averages

$$\bar{y}_{gt} = \frac{1}{M}\sum_{i\in g} y_{it}, \tag{3.1}$$

and note that, if $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is given by (2.5), then also $\bar{\mathbf{y}}_{G,t} = (\bar{y}_{1t}, \bar{y}_2, \ldots, \bar{y}_{Gt})' \sim N(\mathbf{0}, \mathbf{\Psi}_G)$, where $\mathbf{\Psi}_G$ is a $G \times G$, positive definite matrix with elements

$$\psi_{gh} = \frac{1}{M^2} \sum_{i \in g, j \in h} \sigma_{ij} = \sigma_{gh}, \qquad \text{for } g \neq h, \tag{3.2}$$

$$\psi_{gg} = \frac{1}{M^2} \sum_{i,j \in g} \sigma_{ij} = \sigma_{gg} + \frac{1}{M}\gamma_g, \tag{3.3}$$

or, in matrix form,

$$\mathbf{\Psi}_G = \mathbf{\Sigma}_G + \frac{1}{M}\mathbf{\Gamma}_G. \tag{3.4}$$

It follows that we can estimate $\mathbf{\Sigma}$ by applying the GLASSO to the vector of group means, $\bar{\mathbf{y}}_{G,t}$. More specifically, consider the following two-step procedure.

STEP 1. Estimate $\mathbf{\Phi}_G = \mathbf{\Psi}_G^{-1}$ by applying the GLASSO to $\bar{\mathbf{y}}_{G,t}$, $t = 1, 2, \ldots, T$. This allows us to obtain $\hat{\sigma}_{gh}$ for $g \neq h = 1, 2, \ldots, G$, and $\hat{\psi}_{gg}$, $g = 1, 2, \ldots, G$.

STEP 2. Estimate γ_g by exploiting identity (2.4) and (3.4). Noting that $E[(1/MT) \sum_{i \in g} \sum_{t=1}^{T} y_{it}^2] = \sigma_{gg} + \gamma_g$, while $E[(1/MT) \sum_{i \in g} \sum_{t=1}^{T} \bar{y}_{gt}^2] = \sigma_{gg} + (1/M)\gamma_g$, we can consider the following estimator for $\hat{\gamma}_g$:

$$\hat{\gamma}_g = \frac{M}{M-1}\Big(\frac{1}{MT} \sum_{i \in g} \sum_{t=1}^{T} y_{it}^2 - \hat{\psi}_{gg}\Big), \qquad g = 1, 2, \ldots, G. \tag{3.5}$$

Hence, use (2.6) to recover $\hat{\mathbf{\Theta}}$:

$$\hat{\mathbf{\Theta}} = \Big(\frac{1}{M}\widehat{\mathbf{\Phi}}_G \otimes \frac{1}{M}\mathbf{1}_M\Big) + \Big(\hat{\mathbf{\Gamma}}_G^{-1} \otimes (\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M)\Big). \tag{3.6}$$

In Step 1, the estimator that maximizes the penalized likelihood for $\bar{\mathbf{y}}_{G,t}$ is

$$\widehat{\mathbf{\Phi}}_G = \max_{\mathbf{\Phi}_G \succ 0} \{\ln |\mathbf{\Phi}_G| - Tr(\mathbf{S}_G \mathbf{\Phi}_G) - \rho_G \sum_{g,h=1, g \neq h}^{G} |\phi_{gh}|\}, \tag{3.7}$$

where the maximization is taken over symmetric positive definite matrices, $\mathbf{S}_G$ is the sample covariance matrix, and ρ_G is the tuning parameter controlling the degree of the sparsity in the estimated inverse covariance matrix.

The following theorems derive the asymptotic properties of estimator (3.6) when both N and T go to infinity.

THEOREM 3.1. (CONSISTENCY) *Let* $\mathbf{y}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$ *where* $\mathbf{\Sigma}$ *has the block structure in (2.5), with* $\mathbf{\Sigma}_G$ *given by (2.3) being a symmetric, positive definite matrix such that* $\lambda_1(\mathbf{\Sigma}_G) < K < \infty$. *Let* $\sum_{g,h=1,g \neq h}^{G} 1_{\{\phi_{gh} \neq 0\}} = s_G$, *where* ϕ_{gh} *are the elements of* $\mathbf{\Phi}_G$. *Let* $\widehat{\mathbf{\Theta}}$ *be an estimate of* $\mathbf{\Theta}$ *following Steps 1 and 2, where* $\rho_G = O(\sqrt{(\ln G/T)})$, *with* ρ_G *being the tuning parameter in (3.7). Then, we have*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F = O_p\Big(\frac{1}{M}\sqrt{\frac{(G + s_G)\ln G}{T}}\Big). \tag{3.8}$$

THEOREM 3.2. (SPARSISTENCY) *Suppose all conditions in Theorem 3.1 hold, and that* $\|\widehat{\mathbf{\Phi}}_G - \mathbf{\Phi}_G\|^2 = O(\eta_G)$ *where* η_G *is such that* $\rho_G = O(\sqrt{(\ln G/T) + \eta_G})$, *with* ρ_G *being the tuning parameter in (3.7). Let* $S = \{(i, j) : i \neq j, \theta_{ij} = 0\}$ *be the set of indices of all non-zero off-diagonal elements in* $\mathbf{\Theta}$. *Then, with probability tending to 1 we have* $\hat{\theta}_{ij} = 0$ *for all* $i, j \in S^c$.

Theorem 3.2 is a straightforward consequence of the sparsistency theorem of Lam and Fan (2009) applied to $\widehat{\mathbf{\Phi}}_G$; see also Rothman et al. (2008) and Guo et al. (2011).

Hence, for $\widehat{\mathbf{\Theta}}$ to be a good proxy of $\mathbf{\Theta}$, G needs to be small (or, equivalently, M large) and $\mathbf{\Phi}_G$ needs to be a sparse matrix, as measured by s_G. Note, however, that, from (2.6), the off-diagonal elements of $\mathbf{\Theta}$ are proportional to $1/M^2$. Hence, for fixed G, as M increases the (relative) effect of each individual neighbour on each unit would disappear and, in the limit, the precision matrix would become a diagonal matrix. A similar result has been obtained by Lee (2002) in the context of a spatial autoregressive (SAR) model where each spatial unit is influenced aggregately by a significant portion of other spatial units in the sample. Lee (2002) showed that if each spatial unit in the limit has infinitely many neighbours (which would happen in our case for G fixed and M increasing), then the ordinary least-squares (OLS) estimator for a SAR model would still be consistent and even asymptotically efficient. In Section 4, we investigate the properties of our estimator for different values of G relative to N.

A major advantage of our proposed estimation procedure is that it is considerably faster than the conventional GLASSO for estimating an $N \times N$ precision matrix. Using the algorithm proposed by Friedman et al. (2008), the computational cost associated with a coordinate descendent update would decrease from $O(N^2)$ to $O(G^2)$. This could decrease further to $O(G)$ using faster algorithms, such as QUIC (Hsieh et al., 2014). Another advantage of our approach is that using block averages rather than single observations greatly helps in the presence of missing values, a common problem in statistical analysis. Exploiting group membership information is also very useful for prediction purposes on a hold-out sample of units, for which the position in the (individual-level) network is usually unknown. It is important, however, to remark that our approach requires *a priori* information on the block structure. If this is not available, then one could exploit methods from the clustering literature that allow us to determine endogenously the optimal grouping of cross-sectional units, such as the k-means algorithm (Forgy, 1965) extended to allow for covariates in the model; see, in particular, Lin and Ng (2012) and Bonhomme and Manresa (2015), and also Ando and Bai (2016). Our approach also has potential application in the area of spatial econometrics. Given the equivalence between CAR models and the joint Gaussian distribution emphasized by many authors – see, among others, Mardia (1988) and Meinshausen and Buhlmann (2006) – this method provides a means for estimating spatial weights matrices in the context of very large panel data. Later in the paper, we offer a small empirical exercise using CAR models.

Finally, it is important to remark that our approach does not allow us to estimate consistently the precision matrix when this arises from one or more common, pervasive factors. Unobserved common factors occur in time series as a result of global shocks, namely unexpected events that may hit all statistical units, although with different intensities (Stock and Watson, 2010). These large-scale perturbations affect micro-level population units and are often responsible for observable co-movements of a large number of time series. We observe that our model is more parsimonious than the common factor specification and may be useful in situations where T is too short to allow for fully unrestricted common effects. However, in a large T setting, in the presence of unobserved common factors, our approach can be applied to de-factored residuals,

after estimating common factors using methods such as principal components (Bai, 2003) or the Common Correlated Effects methodology (Pesaran, 2006).

3.1. Case of blocks with unequal size

Suppose now we have blocks with unequal size, so that group g has size M_g, with $g = 1, 2, \ldots, G$. In this case, group averages in (3.1) are based on M_g observations. By applying recursively the theorem for block matrix inversion – see Bernstein (2005) – it is easy to see that in the case of blocks of unequal size, a block-wise structure for $\mathbf{\Sigma}$ still implies a block-wise $\mathbf{\Theta}$. In the case of blocks with unequal size, a convenient representation of $\mathbf{\Sigma}$ can be obtained using selection matrices. Let $M_{\max} = \max_{g=1,2,\ldots,G}\{M_g\}$ and consider

$$\mathbf{\Sigma}_{M_{\max}} = (\mathbf{\Sigma}_G \otimes \mathbf{1}_{M_{\max}}) + (\mathbf{\Gamma}_G \otimes \mathbf{I}_{M_{\max}}). \tag{3.9}$$

Then $\mathbf{\Sigma}$ can be extracted as follows:

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{\Sigma}_{M_{\max}}\mathbf{S}'. \tag{3.10}$$

Here, $\mathbf{S}$ is an $N \times GM_{\max}$ matrix of zeros and ones, selecting the correct number of rows and columns for each block in $\mathbf{\Sigma}_{M_{\max}}$, depending on the group size. Note that $\mathbf{SS}' = \mathbf{I}_{NT}$, and rewrite

$$\mathbf{\Sigma} = (\mathbf{S}\mathbf{\Sigma}_{M_{\max}}\mathbf{S}' + \mathbf{I}_{NT}) - \mathbf{I}_{NT} = \mathbf{S}(\mathbf{\Sigma}_{M_{\max}} + \mathbf{I}_{GM_{\max}})\mathbf{S}' - \mathbf{I}_{NT}, \tag{3.11}$$

where $\mathbf{I}_{GM_{\max}}$ is a $GM_{\max} \times GM_{\max}$ identity matrix. Using the matrix inversion lemma, we obtain

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = -\mathbf{S}\big((\mathbf{\Sigma}_{M_{\max}} + \mathbf{I}_{GM_{\max}})^{-1} - \mathbf{S}'\mathbf{S}\big)^{-1}\mathbf{S}' - \mathbf{I}_{NT}, \tag{3.12}$$

where

$$\begin{aligned}(\mathbf{\Sigma}_{M_{\max}} + \mathbf{I}_{GM_{\max}})^{-1} = &\Big((M_{\max}\mathbf{\Sigma}_G + \mathbf{\Gamma}_G + \mathbf{I}_G)^{-1} \otimes \frac{1}{M_{\max}}\mathbf{1}_{M_{\max}}\Big) \\ &+\Big((\mathbf{\Gamma}_G + \mathbf{I}_G)^{-1} \otimes (\mathbf{I}_{M_{\max}} - \frac{1}{M_{\max}}\mathbf{1}_{M_{\max}})\Big)\end{aligned} \tag{3.13}$$

and $\mathbf{S}'\mathbf{S}$ is a diagonal $GM_{\max}$-dimensional matrix of zeros and ones.[1] Steps 1 and 2 outlined above can still be carried to obtain $\widehat{\mathbf{\Phi}}_G$ and $\widehat{\mathbf{\Gamma}}_G$, where now TM_g observations are used to calculate $\hat{\gamma}_g$. The resulting $\widehat{\mathbf{\Sigma}}_G$ can then be plugged into (3.12) and (3.13). From (3.12) and (3.13), it can be seen that consistency and sparsistency of the resulting estimator continue to hold with rates that now depend on N, G and $M_{\max}$.

[1] The matrix inversion lemma states that (Bernstein, 2005)

$$(\mathbf{A} + \mathbf{BDC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}^{-1} + \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}.$$

3.2. Allowing for general intra-block correlation structure

The approach outlined in Section 3 can be extended to allow for a general, intra-block correlation matrix at the expense of reducing the computational efficiency. Suppose that

$$\sigma_{ij} = \sigma_{gg} + \pi_{ij}, \qquad \text{for all } i, j \in g = 1, 2, \ldots, G, \tag{3.14}$$

$$\sigma_{ij} = \sigma_{gh}, \qquad \text{for all } i \in g, j \in h, \quad \text{with } g \neq h = 1, 2, \ldots, G. \tag{3.15}$$

Under this framework, while the covariance between variables of different blocks is constant for all variables belonging to the same block, the intra-block covariance is allowed to vary across variables. In this case, the covariance matrix can be written as

$$\mathbf{\Sigma} = (\mathbf{\Sigma}_G \otimes \mathbf{1}_M) + \mathbf{\Pi},$$

where Π is a block-diagonal matrix.

We can show that, under the condition that $(1/M)\sum_{k\in g} \pi_{ik} \approx 0$ for all i, the Π matrix can be estimated by the covariance matrix of $y_{it} - \bar{y}_{gt}$ for each block. This results in a relatively easy implementation, whereby we first calculate $\bar{y}_{gt}$ and apply the block-GLASSO outlined in Section 3 to compute $\widehat{\mathbf{\Sigma}}_G$. Hence, we calculate the deviations of each value y_{it} from its corresponding group-level average, namely $y_{it} - \bar{y}_{gt}$, and we apply the conventional GLASSO to all $y_{it} - \bar{y}_{gt}$ for each block, separately. This approach requires that π_{ij}, namely the deviations of σ_{ij} from σ_{gg}, are not too large, so that the $\bar{y}_{gt}$ can be used to consistently estimate σ_{gg}. The computational complexity of this procedure rises to $O(G^2) + O(GM^2)$, as it is necessary to estimate G blocks of size M. In the rest of the paper, we refer to this approach as the flexible block-GLASSO.

4. MONTE CARLO EXPERIMENTS

In this section, we provide Monte Carlo evidence on the properties of the above estimation procedure. We consider the following data-generating process:

$$y_{it} = \alpha_i + \beta x_{it} + e_{it}, \qquad i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T. \tag{4.1}$$

Here

$$x_{it} = 0.4x_{i,t-1} + v_{it}, \qquad t = -19, -18, \ldots, -1, 0, 1, 2, \ldots, T \tag{4.2}$$

with $\alpha_i \sim IIDN(0, 0.5)$, $e_{it} \sim N(0, \Sigma)$ and $v_{it} \sim N(0, \Sigma_X)$. In generating x_{it}, we set $x_{i,-20} = 0$ and discard the first 20 observations to reduce the effect on estimates of initial values of x_{it}. To generate $\mathbf{\Sigma}$, we start from $\mathbf{\Theta}_G = \mathbf{\Sigma}_G^{-1}$ and assume that its elements, $\theta_{gh,G} \sim Bin(1, (3/G))$ for $g, h = 1, \ldots, G$. We obtain $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ by applying (2.6), where we assume $\gamma_G \sim U(0.2, 0.5)$. Letting $\mathbf{D}$ be the Choleski decomposition of $\mathbf{\Sigma}$, namely $\mathbf{\Sigma} = \mathbf{D}\mathbf{D}'$, we generate $\mathbf{e}_t = \mathbf{D}\boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{Nt})'$, with $\epsilon_{it} \sim IDN(0, 1)$. We generate $\mathbf{\Sigma}_X$ following the same procedure. As for β, in a first set of experiments we set $\beta = 0$, and apply our methodology to y_{it}, to test our procedure when there is no uncertainty regarding the mean of y_{it}. We then set $\beta = 1$ and apply our methodology to regression residuals after estimating β by OLS. As a robustness check, we carry an additional experiment where errors are non-normally distributed. In this case, when generating e_{it}, we set $\epsilon_{it} = (u_{it} - 1)/\sqrt{2}$, with $u_{it} \sim \chi^2_1$. Model (4.1)–(4.2) has strictly exogenous

regressors, an assumption that may not hold in practice. In a further set of experiments, we also consider a dynamic set-up, where we assume that y_{it} is generated by the first-order autoregressive model

$$y_{it} = \alpha_i + \lambda y_{i,t-1} + e_{it}, \qquad i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T, \tag{4.3}$$

where all elements are generated as above, and $\lambda = 0.4$.

Finally, we examine the performance of the more general flexible block-GLASSO approach outlined in Section 3.2 when Σ has a general intra-block correlation structure. Under this experiment, all parameters are the same as in (4.1) and (4.2), with $\beta = 1$ and

$$\sigma_{ij} = \sigma_{gg} + \pi_{ij}, \qquad \text{for all } i, j \in g = 1, 2, \ldots, G, \tag{4.4}$$

$$\sigma_{ij} = \sigma_{gh}, \qquad \text{for all } i \in g; j \in h, \text{ with } g \neq h = 1, 2, \ldots, G. \tag{4.5}$$

We generate each block in Π by assuming that its inverse has elements distributed as $Bin(1, (3/M))$.

In each experiment, we compute the block-GLASSO and the conventional GLASSO, for all pairs of N and T with $N = 50$ and 100 and $T = 10$, 50 and 200. As for the choice of G, we try $G = N/2$, and $N/5$. Each experiment is replicated $R = 250$ times. We also carry out another set of experiments with N much larger than T, and set $N = 500$, 1,000 and 2,000 and $T = 20$. In this set of experiments, given the computational difficulties and poor performance in computing conventional GLASSO for such large networks, we only provide results for the block-GLASSO. Under the dynamic set-up (4.3), we only run experiments for large T (i.e. $T = 50$ and 200) to avoid incurring bias of the OLS estimator for short panels.[2]

A number of statistics are used to assess the performance of our graph estimators. In terms of recovery of the network structure (provided by the non-zero coefficients in $\mathbf{\Theta}$), we consider the receiver operating characteristic (ROC) curve, which plots the true positive rate (percentage of non-zeros, i.e. links, correctly estimated as non-zero) versus the false positive rate (percentage of zeros incorrectly estimated as non-zeros), as the tuning parameter, ρ_G, varies. We summarize ROC curves by providing the maximum F1 score and the area under the curve (AUC), both averaged across the R replications. The F1 score is defined by $(2TP/(2TP + FN + FP))$, with TP, FP and FN being the true positive, the false positive and the false negatives (number of non-zeros incorrectly detected as zeros), respectively. In terms of estimation of the precision matrix, we report the average entropy loss (EL) and the average Frobenius loss (FL), defined by

$$EL = \mathrm{Tr}(\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}) - \ln|\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}| - N, \tag{4.6}$$

$$FL = \frac{\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}\|_F^2}{\|\mathbf{\Theta}\|_F^2}. \tag{4.7}$$

When computing EL and FL, we use the rotation information criterion (RIC) – see Lysen (2009) – to select the optimal regularization parameter (and associated optimal precision matrix). Only for selected combinations of N and T, we also provide graphs with the ROC curves. As for β, we

[2] When T is short, our approach can be used in combination with methods for estimating short dynamic panels, such as the generalized method of moments by Arellano and Bond (1991).

Table 1. Properties of block-GLASSO and conventional GLASSO in model (4.1)–(4.2), $\beta = 0$.

N	T	G	Block-GLASSO				Conventional GLASSO			
			F1	AUC	EL	FL	F1	AUC	EL	FL
50	200	25	0.929	0.881	2.894	0.015	0.869	0.551	15.063	0.491
50	200	10	0.923	0.906	0.800	0.003	0.638	0.285	19.694	0.472
50	50	25	0.828	0.818	6.099	0.056	0.719	0.509	27.918	0.679
50	50	10	0.817	0.786	1.562	0.010	0.670	0.457	27.650	0.678
50	10	25	0.665	0.400	13.167	0.571	0.578	0.172	43.232	0.829
50	10	10	0.707	0.640	3.668	0.063	0.548	0.147	65.296	0.827
100	200	50	0.948	0.894	6.458	0.015	0.863	0.529	35.085	0.538
100	200	20	0.944	0.912	1.970	0.003	0.668	0.303	45.417	0.531
100	50	50	0.819	0.772	12.855	0.053	0.689	0.415	61.453	0.717
100	50	20	0.801	0.812	3.821	0.010	0.597	0.281	84.888	0.710
100	10	50	0.620	0.207	26.570	0.601	0.523	0.079	86.485	0.827
100	10	20	0.675	0.475	8.299	0.064	0.498	0.071	135.156	0.838

Note: F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average Frobenius Loss in (4.7).

Table 2. Properties of block-GLASSO with large N in model (4.1)–(4.2), $\beta = 0$.

N	T	G	F1	AUC	EL	FL
500	20	50	0.657	0.421	13.757	0.011
500	20	100	0.656	0.248	33.660	0.029
500	20	250	0.616	0.092	94.290	0.168
1,000	20	50	0.649	0.402	12.306	0.005
1,000	20	100	0.631	0.232	30.079	0.011
1,000	20	250	0.613	0.094	90.020	0.040
2,000	20	50	0.641	0.388	11.429	0.003
2,000	20	100	0.624	0.228	28.523	0.010
2,000	20	250	0.609	0.090	87.742	0.009

Note: F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average Frobenius loss in (4.7).

report bias, root mean square error (RMSE), empirical size and power of the OLS estimator of β and the feasible generalized least-squares (GLS) estimator implemented using $\hat{\mathbf{\Theta}}$ as estimate of $\mathbf{\Theta}$. In computing the empirical size, we set the nominal size to 5%, while in calculating the power we assume as an alternative hypothesis $H_1 : \beta = 0.95$.

4.1. Results

The results are summarized in Tables 1–6 and Figures 1–2. The results from Table 1 show that, when data have block-wise dependence structure, our method greatly outperforms the conventional GLASSO for all combinations of N, T and G. In particular, the F1 score and AUC show that block-GLASSO has higher true positive rates and substantially lower false positive rates, while the EL and FL are always lower for block-GLASSO, indicating that the latter provides a better estimation of the precision matrix. However, it is interesting to note

Table 3. Properties of OLS and GLS estimators of β and of block-GLASSO applied to regression residuals in model (4.1)–(4.2), $\beta = 1$.

N	T	G	OLS				GLS				Block-GLASSO			
			Bias	RMSE	Size (%)	Power (%)	Bias	RMSE	Size (%)	Power (%)	F1	AUC	EL	FL
50	200	25	0.000	0.013	14.80	100.0	0.000	0.008	4.80	100.0	0.928	0.881	2.915	0.015
50	200	10	0.000	0.016	22.00	99.60	0.001	0.008	4.60	100.0	0.918	0.898	0.794	0.003
50	50	25	0.000	0.030	19.60	82.00	0.000	0.019	4.40	92.40	0.828	0.818	6.137	0.059
50	50	10	−0.003	0.036	23.60	76.80	0.002	0.016	4.40	96.00	0.811	0.829	1.654	0.012
50	10	25	0.005	0.056	13.60	47.20	0.000	0.048	10.40	43.60	0.667	0.401	14.332	0.924
50	10	10	−0.020	0.083	25.20	47.20	0.002	0.038	4.40	46.40	0.703	0.633	4.173	0.113
100	200	50	0.000	0.009	15.60	100.0	−0.001	0.005	4.90	100.0	0.948	0.895	6.520	0.015
100	200	20	0.000	0.011	23.60	100.0	0.000	0.006	4.50	100.0	0.937	0.912	1.990	0.002
100	50	50	−0.001	0.017	18.40	97.20	−0.001	0.012	6.00	99.20	0.818	0.774	12.809	0.060
100	50	20	0.002	0.026	24.80	91.20	0.000	0.011	5.40	100.0	0.798	0.806	3.925	0.012
100	10	50	0.002	0.044	16.80	57.20	−0.001	0.034	8.00	55.60	0.623	0.207	29.313	0.983
100	10	20	0.001	0.064	22.40	59.20	0.001	0.033	5.20	66.40	0.672	0.470	9.252	0.111

Note: In calculating the empirical size, the nominal size is set to 5%, while in calculating the power we assume as an alternative hypothesis H_1: $\beta = 0.95$. F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average Frobenius loss in (4.7).

Table 4. Properties of OLS and GLS estimators of λ and of block-GLASSO applied to regression residuals in model (4.3), $\lambda = 0.4$.

N	T	G	OLS				GLS				Block-GLASSO			
			Bias	RMSE	Size (%)	Power (%)	Bias	RMSE	Size (%)	Power (%)	F1	AUC	EL	FL
50	200	25	−0.006	0.017	33.20	98.00	−0.001	0.010	6.40	100.0	0.930	0.882	2.860	0.015
50	200	10	−0.008	0.022	31.10	92.00	0.001	0.011	5.60	100.0	0.912	0.896	0.803	0.003
50	50	25	−0.022	0.037	39.20	49.00	−0.003	0.021	10.00	82.30	0.828	0.818	6.042	0.057
50	50	10	−0.019	0.046	35.40	56.00	0.003	0.020	5.20	89.20	0.808	0.825	1.596	0.011
100	200	50	−0.004	0.011	27.10	100.0	0.000	0.007	5.00	100.0	0.949	0.895	6.519	0.015
100	200	20	−0.005	0.014	35.20	100.0	0.002	0.006	4.60	100.0	0.937	0.913	1.998	0.003
100	50	50	−0.020	0.027	49.50	69.0	0.000	0.014	5.80	98.10	0.809	0.768	12.727	0.058
100	50	20	−0.021	0.035	45.10	67.0	0.008	0.014	4.20	100.0	0.800	0.814	3.850	0.011

Note: In calculating the empirical size, the nominal size is set to 5%, while in calculating the power we assume as an alternative hypothesis H_1: $\beta = 0.95$. F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average FL in (4.7).

Table 5. Properties of block-GLASSO and conventional GLASSO in model (4.1)–(4.2): non-normal errors, $\beta = 0$.

N	T	G	Block-GLASSO				Conventional GLASSO			
			F1	AUC	EL	FL	F1	AUC	EL	FL
50	200	25	0.930	0.881	26.885	0.604	0.639	0.280	66.760	0.832
50	200	10	0.919	0.903	34.697	0.636	0.639	0.280	66.760	0.832
50	50	25	0.829	0.814	26.803	0.576	0.726	0.508	43.637	0.823
50	50	10	0.819	0.830	34.572	0.627	0.621	0.347	67.244	0.835
50	10	25	0.681	0.413	26.492	0.515	0.596	0.183	44.165	0.827
50	10	10	0.712	0.653	33.972	0.590	0.551	0.147	67.892	0.839
100	200	50	0.945	0.890	51.858	0.605	0.860	0.522	84.741	0.822
100	200	20	0.936	0.913	69.041	0.636	0.614	0.262	133.373	0.832
100	50	50	0.818	0.769	51.977	0.578	0.699	0.417	85.726	0.825
100	50	20	0.800	0.810	68.965	0.628	0.594	0.274	134.467	0.836
100	10	50	0.639	0.216	51.495	0.526	0.541	0.083	86.977	0.829
100	10	20	0.682	0.486	67.671	0.588	0.500	0.071	135.553	0.839

Note: In calculating the empirical size, the nominal size is set to 5%, while in calculating the power we assume as an alternative hypothesis H_1: $\beta = 0.95$. F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average FL in (4.7).

that when $T = 10$ and $G = N/2$ the block-GLASSO does not perform well relative to other cases, and its properties are much worse than the case $T = 10$ and $G = N/5$. More generally, Tables 1 and 2 show that for the same pair of N and T, the properties of block-GLASSO deteriorate as G rises, thus confirming our theoretical results that, holding N and T fixed, the estimation error is higher when G is large or, equivalently, M small. This result is also confirmed by Figure 1, showing the ROC curves for the block-GLASSO for varying N, T and G. As expected, the performance of the estimator improves as N increases (and hence M) for fixed T and G, and as T increases for fixed N and G, while it deteriorates as G rises, holding N and T constant.

Table 3 reports the small-sample properties of OLS and GLS estimators as well as of the block-GLASSO. As expected in the case of cross-sectionally correlated regression errors, the OLS estimator, while having a bias comparable to that of the GLS, has higher RMSE and is oversized for all combinations of N, T and G. Hence, ignoring the network leads to severe over-rejection of the null hypothesis. Looking at the GLS estimator, its empirical size is close to the nominal size of 5% in most cases, although some size distortions can be observed when $T = 10$ and $G = N/2$, namely, for short panels characterized by the presence of many, small groups. In fact, under this case the block-GLASSO does not perform well, having small F1 and AUC and large EL and FL, thus confirming our asymptotic results reported in Section 3. Similar results can be observed in Table 4 for the case where the dependent variable is generated by the first-order autoregressive model (4.3). Under non-normal errors (Table 5), the block-GLASSO still performs well in detecting the network, as confirmed by F1 and AUC values similar to those reported in Table 1, although its EL and FL are much higher than in the normal counterpart.

Table 6 shows results when the error covariance matrix displays general intra-block variation (see (4.4) and (4.5)). It is interesting to observe that the empirical size of the GLS estimator of β when ignoring the intra-block variation (block-GLASSO) is in some cases still close to the nominal value of 5%. The GLS estimator based on the more general procedure (flexible

Table 6. Properties of GLS estimators of β and of the flexible block-GLASSO applied to regression residuals in model (4.1)–(4.2) with general intrablock variation, $\beta = 1$.

N	T	G	GLS (flexible block-GLASSO)				GLS (block-GLASSO)				Flexible block-GLASSO			
			Bias	RMSE	Size (%)	Power (%)	Bias	RMSE	Size (%)	Power (%)	F1	AUC	EL	FL
50	200	25	0.002	0.015	10.05	98.49	0.002	0.015	10.55	98.50	0.902	0.885	2.451	0.060
50	200	10	0.000	0.016	4.55	95.48	0.000	0.017	5.52	95.50	0.907	0.902	2.382	0.081
50	50	25	0.000	0.033	10.05	66.83	0.000	0.033	11.05	67.35	0.772	0.766	4.072	0.126
50	50	10	−0.001	0.034	5.05	50.25	0.000	0.034	6.50	52.80	0.797	0.806	3.844	0.111
50	10	25	−0.008	0.080	7.65	21.43	−0.008	0.080	8.20	22.95	0.648	0.373	8.739	0.492
50	10	10	−0.015	0.090	5.00	16.58	−0.014	0.087	6.60	15.05	0.702	0.630	13.762	1.180
100	200	50	0.002	0.012	20.83	100.00	0.002	0.012	20.85	100.00	0.919	0.895	5.316	0.065
100	200	20	0.000	0.012	5.25	100.00	0.000	0.012	9.05	100.00	0.922	0.910	5.127	0.082
100	50	50	0.001	0.023	11.11	88.89	0.001	0.023	11.10	88.90	0.763	0.726	7.379	0.118
100	50	20	−0.001	0.022	5.35	72.97	−0.001	0.023	6.80	68.90	0.784	0.799	7.914	0.108
100	10	50	0.009	0.069	9.00	52.95	0.009	0.070	8.80	52.90	0.598	0.192	17.022	0.549
100	10	20	−0.005	0.061	5.05	20.60	−0.003	0.063	5.50	22.60	0.668	0.461	27.672	1.128

Note: In calculating the empirical size, the nominal size is set to 5%, while in calculating the power we assume as an alternative hypothesis H_1: $\beta = 0.95$. F1 is the F1 score, AUC is the area under the ROC, EL is the average EL in (4.6) and FL is the average FL in (4.7).

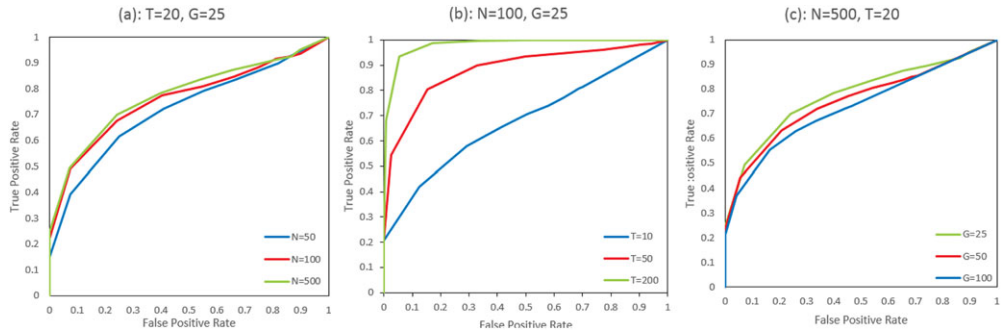

Figure 1. Block-GLASSO ROC curves: varying values of (a) N, (b) T and (c) G. [Colour figure can be viewed at wileyonlinelibrary.com]

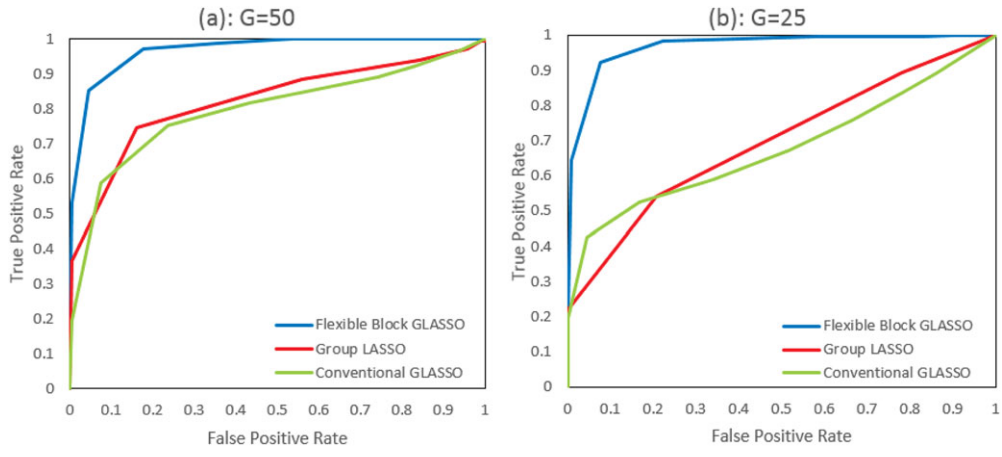

Figure 2. Flexible block-GLASSO, group LASSO and conventional GLASSO: within-block variation, $N = 100$, $T = 200$. [Colour figure can be viewed at wileyonlinelibrary.com]

block-GLASSO) shows a good performance only for smaller values of G, perhaps because under small G (and hence large M) the covariance of $\bar{y}_{gt}$ better approximates the part of the covariance that is block-wise. We also remark that the more flexible procedure is computationally much slower than the block-GLASSO. Figure 2 shows the ROC for the flexible block-GLASSO and the conventional GLASSO, as well as the group LASSO by Yuan and Lin (2006). The use of a group penalty in the group LASSO encourages the recovery of the block structure, although it does not impose it as in the block-GLASSO. Because the group LASSO has been developed in the context of regression analysis, we apply it to our model as a neighbourhood selection problem for each node of the network. It is interesting to see from Figure 2 that the group LASSO approach performs less well than the block-GLASSO, but slightly better than the conventional GLASSO, as the latter does not use any *a priori* information about the blocks.

5. AN EMPIRICAL EXAMPLE: SPATIAL SPILLOVERS IN REGIONAL GROWTH AND CONVERGENCE IN EUROPE

We use block-GLASSO for estimating a growth equation in per-capita gross value-added and for testing for economic convergence of European regions. The debate on whether there exists convergence in per-capita input and income across nations is still open, with results obtained that differ depending on the sample period and the regions included, as well as the estimation methods adopted. A number of authors have highlighted the importance of incorporating spatial effects when studying economic growth and regional convergence and have proposed the use of spatial econometric techniques; see, among others, Rey and Montouri (1999), Ertur and Koch (2007) and Cuaresma and Feldkircher (2013). Spatial dependence in regional economic growth is likely to arise from technology spillover across neighbouring regions and from factor mobility, as well as from the presence of spatial heterogeneity (Rey and Montouri, 1999). In the presence of spatial dependence in economic growth data, if ignored, estimates of the speed of income convergence across geographical regions will be biased.

We contribute to this literature by estimating a growth equation with spatial spillovers and we use the block-GLASSO procedure to estimate the spatial weights matrix. We use data on gross value-added per worker (GVA) for 1,088 NUTS3 observed over the period 1980–2012 in 14 European countries.[3] The NUTS classification is a hierarchical system for dividing up the economic territory of the European Union (EU) for the purpose of socio-economic analysis of the regions and design of EU regional policies. It subdivides the EU territory into regions at the three different levels, NUTS1, NUTS2 and NUTS3, moving from larger to smaller geographical units.

Standard neo-classical growth models state that countries will converge to the same level of per-capita income in the long run, independently of initial conditions, as long as there are diminishing returns to capital and labour and perfect diffusion of technology. Under this framework, poorer countries and regions grow faster than richer countries and a negative relationship between average growth rates and initial income levels is expected. Let $y_{i,t+k} = \ln(GVA_{i,t+k}/GVA_{it})$ be the growth in per-capita GVA (expressed in euros at 2005 prices) for the NUTS3 region i over a set of non-overlapping time intervals of length k. Our empirical model is the Gaussian CAR model for y_{it}

$$E(y_{i,t+k}|y_{j,t+k},\, j = 1, 2, \ldots, N,\, j \neq i) = \alpha + \beta \ln(GVA_{it}) + \sum_{j=1}^{N} w_{ij}(y_{j,t+k} - \alpha - \beta \ln(GVA_{jt})), \quad (5.1)$$

$$\mathrm{Var}(y_{i,t+k}|y_{j,t+k},\, j = 1, 2, \ldots, N,\, j \neq i) = \sigma_i^2, \quad (5.2)$$

where we set $k = 3$. Hence, a negative coefficient attached to the variable $\ln(GVA_{it})$ indicate that NUTS3 regions with a low initial level of income grow faster than regions with higher initial levels of income, supporting the hypothesis of absolute convergence. The use of non-overlapping time intervals is common practice in the cross-country growth literature, as this

[3] The countries included in the analysis are: Austria, Belgium, Germany, Denmark, Spain, Finland, France, Ireland, Italy, Netherlands, Norway, Portugal, Sweden and the United Kingdom.

Table 7. Descriptive statistics for NUTS3 regions.

	Average	Std dev.	Min	Max
Per-capita GVA (euros)	19,818.3	8,817.7	1,842.0	159,936.1
Growth in per-capita GVA (%)	5.005	7.611	−63.661	47.183

would decrease the influence of short-term shocks and business cycles on economic activity, while revealing long-run relationships. Compared to longer time intervals, the use of three-year non-overlapping intervals allows us to keep a sufficient number of observations to exploit the time dimension of panel data. Following existing studies on spatial interaction effects in regional economic growth models, the inclusion of the spatial lag of the dependent variable (growth rate) amongst the regressors in (5.1) aims at capturing the effect of inter-regional flows of labour, capital and technology on growth and convergence; see Rey and Montouri (1999), Ertur and Koch (2007) and Cuaresma and Feldkircher (2013).

In (5.1), w_{ij} is the (i, j)th element of an $N \times N$ matrix, $\mathbf{W}$, known as the spatial weights matrix, such that $w_{ii} = 0$. In spatial econometrics, $\mathbf{W}$ is often assumed to be known using *a priori* information (e.g. from economic theory) on how statistical units potentially interact. Spatial weights based on geographical or travel distance, or contiguity have been used for modelling spatial spillovers in the economic growth equation, although this has been pointed out as being unrealistic (Cuaresma and Feldkircher, 2013). In this application, we keep $\mathbf{W}$ as unknown and estimate it using our block-GLASSO approach. While the unit of analysis is the NUTS3 region, we take as groups larger geographical areas, given by 80 NUTS1 and then 211 NUTS2 European regions. Other grouping criteria may undoubtedly be suggested, for example by looking at the literature on club convergence – see, among others, Corrado et al. (2005) – or using methods for identifying communities in social networks from the graph modelling literature (Freeman, 1979).

It is interesting to observe that (5.1) and (5.2) for the conditional distribution imply the joint normal distribution (Besag, 1974).

$$\mathbf{y}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}), \tag{5.3}$$

where $\boldsymbol{\Sigma} = (\mathbf{I}_N - \mathbf{W})^{-1}\boldsymbol{\Lambda}$, with $\boldsymbol{\Lambda} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ and $\boldsymbol{\mu} = \alpha + \beta \ln(\mathbf{GVA}_t)$, provided that $(\mathbf{I}_N - \mathbf{W})$ is invertible and $(\mathbf{I}_N - \mathbf{W})^{-1}\boldsymbol{\Lambda}$ is symmetric and positive-definite. The reverse is also true: namely, if $\mathbf{y}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an $N \times N$ positive definite matrix, then also (5.1) and (5.2) hold, with

$$w_{ij} = -\frac{\theta_{ij}}{\theta_{ii}}, \tag{5.4}$$

$$\text{Var}(y_{it}|y_{jt},\, j = 1, 2, \ldots, n,\, j \neq i) = \theta_{ii}^{-1}; \tag{5.5}$$

see Mardia (1988) and Meinshausen and Buhlmann (2006). It follows that the problem of estimating w_{ij} in the CAR model (5.1)–(5.2) is equivalent to determining whether y_{it} and y_{jt} are conditionally independent, i.e. $\theta_{ij} = 0$. Hence, in this application, we estimate $\mathbf{W}$ via $\boldsymbol{\Theta}$ by imposing a block structure on $\boldsymbol{\Sigma}$ (and hence on $\boldsymbol{\Theta}$ and $\mathbf{W}$).

Table 7 offers some descriptive statistics on the variable under study, at the NUTS3 level. It is interesting to observe that the region with the highest level of per-capita GVA (159,936 euros) is the London area, while the region with the lowest per-capita GVA (1,842 euros) is North

Table 8. Regression results.

	OLS	GLS: NUTS1	GLS: NUTS2
$\ln(GVA_{it})$	-0.273^{*} (0.008)	-0.227^{*} (0.009)	-0.221^{*} (0.011)
Speed of convergence	0.106	0.086	0.083
Half-life	7.273	8.789	9.045
R^2	0.121	0.133	0.134
G	–	80	211
Percentage of links	–	36.22	17.23
Average path length	–	1.629	1.845
Graph centrality measures:			
Degree	–	0.126	0.065
Closeness	–	0.101	0.052
Betweenness	–	0.010	0.006

Note: NUTS3 regional dummies and time dummies have been included in all regressions. * denotes significant at the 5% level. Standard errors (given in parentheses) robust to unknown heteroscedasticity have been adopted.

Portugal, which is also the region with the highest growth in per-capita GVA (47.183%) over the three-year time interval.

Table 8 reports estimates of growth equations (5.1) and (5.2). The first column provides OLS estimates ignoring the spatial structure of data, while the second and third columns show GLS estimates where contemporaneous correlation is incorporated and estimated by block-GLASSO. The value of the coefficient of the initial per-capita GVA of NUTS 3 provinces is negative and significant, showing the presence of (absolute) convergence in all regressions. However, when adopting the GLS approach based on the block-GLASSO procedure, the coefficient is smaller, leading to lower speed of convergence towards the steady state, and a longer time necessary for the regional economies to cover half of the initial lag from their steady states, when compared to traditional OLS estimation. Goodness of fit for all regressions is low, ranging between 12% and 13%, indicating that some important factors have not been included in the models.

The lower panel of the table reports the percentage of links, the average path length and a set of centrality measures proposed by graph theory – see Borgatti and Everett (2006) and Freeman (1979) – that are widely used to characterize the compactness of graphs. The average path length is given by the average length of all the shortest paths from or to the vertices in the network, giving an indication of how dense the network is. The graph-level centrality measures are based on three node-level centrality indicators (i.e. degree, closeness and betweenness), which characterize different aspects of the relative importance of each node and are commonly used in the applied literature.[4] All graph-level measures vary between zero and one, and assume their highest value when the graph has a star or wheel shape. Looking at the percentage of links, it emerges that, as expected, the estimated networks are quite dense and connected when using either NUTS1 or NUTS2 as blocks. This is confirmed by the average path length, which is very low, being around 1.6–1.8. However, the graph centrality measures are close to zero, indicating that there is no single region dominating all other regions. This is also evident from Figure 3, which shows the adjacency graph resulting from the estimation of model (5.1)–(5.2) via block-GLASSO where NUTS1 regions are taken as blocks. We do not report the graph when

[4] Degree is the number of links for each unit, closeness is the inverse of the average length of the shortest paths to/from all the other vertices in the graph and betweenness is the number of times a node acts as a bridge between other nodes.

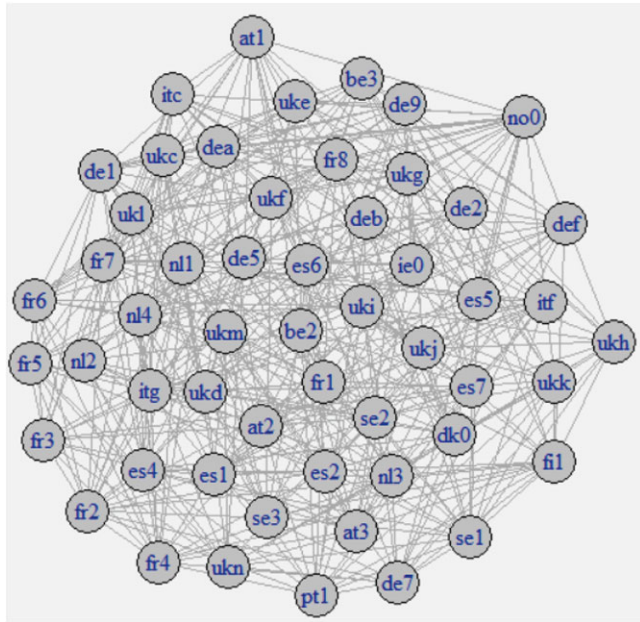

Figure 3. Adjacency graph of per-capita GVA growth: 1980–2012. [Colour figure can be viewed at wileyonlinelibrary.com]

using NUTS2 regions as blocks, because these are too many. It is interesting to observe that the most connected NUTS1 are also the regions with the highest per-capita GVA, namely Greater London, Norway and South Netherlands, while the areas with a smaller number of connections are Northern Ireland and northern areas of the United Kingdom, which are also geographically isolated from the other regions. Also, in most cases, regions from the same country are connected, thus supporting previous studies using geographical contiguity or geographical distance as a metric of distance.

6. CONCLUDING REMARKS

In the last few years, several methods have been proposed for reducing the dimensionality problem when estimating graphical models. These methods usually exploit *a priori* information on possible independence between groups of observations. In this paper, we focus on the estimation of a Gaussian graphical model with a large number of variables, where dependence between variables is block-wise because of, for example, a hierarchical or group membership structure. We propose an estimation strategy based on the GLASSO methodology applied to group averages of observations, and we derive the large-sample properties of the proposed estimator. Our Monte Carlo experiments show that our proposed estimator greatly outperforms the conventional GLASSO when data have block-wise dependence. These experiments also show that our procedure is robust to various deviations from block-wise dependence. For example, the method still delivers valid inference when there is some within-group variation, or under

non-normal errors. We have shown the usefulness of this procedure on an empirical study of economic convergence of European regions, showing that accounting for block-wise dependence helps us to better estimate convergence parameters. Although there are many examples in economics where the membership is given, in many others this is not true, making the assumption that the block structure is known *a priori* too restrictive. One interesting extension of this work would be to determine endogenously the inclusion of a unit in a group as well as the size and number of the groups, following the work by Lin and Ng (2012), Bonhomme and Manresa (2015) and Ando and Bai (2016). Future work should also consider a block-wise structure for the covariance matrix of a VAR model, within the setting proposed by Barigozzi and Brownlees (2016) and Abegaz and Wit (2013). Finally, while our approach does not allow us to estimate the covariance matrix arising from one or more common pervasive factors, it would be interesting to study the properties of an estimation procedure that first controls for common pervasive factors and then estimates the network structure using de-factored residuals.

ACKNOWLEDGEMENTS

F. Moscone and E. Tosetti acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) grant, Semantic Credit Risk Assessment of Business Ecosystems (SCRIBE).

REFERENCES

Abegaz, F. and E. Wit (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics 14*, 586–99.

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics 31*, 163–91.

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

Arellano, M. and S. R. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies 58*, 277–97.

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

Barigozzi, M. and C. Brownlees (2016). NETS: network estimation for time series. Working paper, Universität Pompeu Fabra (https://ssrn.com/abstract=2249909).

Bayer, P., S. Ross and G. Topa (2008). Place of work and place of residence: informal hiring networks and labor market outcomes. *Journal of Political Economy 116*, 1150–96.

Bernstein, D. S. (2005). *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton, NJ: Princeton University Press.

Bertrand, M., E. F. P. Luttermer and S. Mullainathan (2000). Network effects and welfare cultures. *Quarterly Journal of Economics 115*, 1019–55.

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.

Blume, L. E., W. A. Brock, S. N. Durlauf and R. Jayaraman (2013). Linear social interactions models. NBER Working Paper 19212, National Bureau of Economic Research.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica 83*, 1147–84.

Borgatti, S. P. and M. G. Everett (2006). A graph-theoretic perspective on centrality. *Social Networks 28*, 466–84.

Calvó-Armengol, A., E. Patacchini and Y. Zenou (2009). Peer effects and social networks in education. *Review of Economic Studies 76*, 1239–67.

Corrado, L., R. Martin and M. Weeks (2005). Identifying and interpreting regional convergence clusters across Europe. *Economic Journal 115*, C133–60.

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

Cuaresma, J. C. and M. Feldkircher (2013). Spatial filtering, model uncertainty and the speed of income convergence in Europe. *Journal of Applied Econometrics 28*, 720–41.

Durlauf, S. and H. Peyton Young (2001). The new social economics. In S. Durlauf and H. Peyton Young (Eds.), *Social Dynamics*, 1–14. Cambridge, MA: MIT Press.

Ertur, C. and W. Koch (2007). Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of Applied Econometrics 22*, 1033–62.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics 21*, 768–9.

Freeman, L. C. (1979). Centrality in social networks: conceptual clarification. *Social Networks 1*, 215–39.

Friedman, J., T. Hastie and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics 9*, 432–41.

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). New York, NY: Wiley.

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

Hao, D., C. Ren and C. Li (2012). Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC Systems Biology 6*, 34.

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*, 2911–47.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*, 963–74.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics 37*, 4254–78.

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

Lee, L. F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory 18*, 252–77.

Lee, L.-F. and J. Yu (2007). Near unit root in the spatial autoregressive model. Working paper, Ohio State University.

Lin, C. and S. Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods 1*, 42–55.

Luo, X. (2015). A hierarchical graphical model for big inverse covariance estimation with an application to fMRI. Working paper, Cornell University.

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

Magnus, J. R. (1982). Multivariate error components analysis of linear and nonlinear regression models by maximum likelihood. *Journal of Econometrics 19*, 239–85.

Mardia, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis 24*, 265–84.

Mazumder, R. and T. Hastie (2012). Exact covariance thresholding into connected components for large-scale graphical LASSO. *Journal of Machine Learning Research 13*, 1436–62.

Meinshausen, N. and P. Buhlmann (2006). High-dimensional graphs and variable selection with the LASSO. *Annals of Statistics 34*, 1436–62.

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*, 735–46.

Pesaran, M. H. (2006). Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica 74*, 967–1012.

Rey, S. and B. Montouri (1999). US regional income convergence: a spatial econometric perspective. *Regional Studies 33*, 143–56.

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

Schelldorfer, J., L. Meier and P. Bühlmann (2014). GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using L1-penalization. *Journal of Computational and Graphical Statistics 23*, 460–77.

Sirakaya, S. (2006). Recidivism and social interactions. *Journal of the American Statistical Association 101*, 863–75.

Stock, J. and M. Watson (2010). *Dynamic Factor Models*. Oxford: Oxford University Press.

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.

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

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Series B *68*, 49–67.

APPENDIX

Log-likelihood function for networks with block-wise dependence structure

Let S be the $N \times N$ sample covariance matrix based on a sample of size T from the random vector, y:

$$\mathbf{S} = \frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix} y_{1t}^2 & y_{1t}y_{2t} & \cdots & y_{1t}y_{Nt} \\ y_{2t}y_{1t} & y_{2t}^2 & \cdots & y_{2t}y_{Nt} \\ \cdots & \cdots & \cdots & \\ y_{Nt}y_{1t} & \cdots & \cdots & y_{Nt}^2 \end{pmatrix}.$$

To obtain the log-likelihood function, we first compute simplified expressions for $\ln|\mathbf{\Theta}|$ and $\mathrm{Tr}(\mathbf{S\Theta})$, with $\Theta = \Sigma^{-1}$ and Σ given by expression (2.4). Using results in Magnus (1982), we have

$$\ln|\mathbf{\Theta}| = \ln\left|\left((M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G)^{-1} \otimes \frac{1}{M}\mathbf{1}_M\right) + \left(\mathbf{\Gamma}_G{}^{-1} \otimes (\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M)\right)\right| \tag{A.1}$$

$$= -\ln|M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G| - (M-1)\ln|\mathbf{\Gamma}_G|. \tag{A.2}$$

Letting $\Psi_G = \Sigma_G + (1/M)\Gamma_G$ and ϕ_{gh} be the generic element of $\Phi_G = \Psi_G^{-1}$, we have

$$\mathrm{Tr}(\mathbf{S\Theta}) = \sum_{i,j=1}^{N} s_{ij}\theta_{ji}$$

$$= \sum_{i=1}^{N} s_{ii}\theta_{ii} + \sum_{i,j\in g, i\neq j}\sum_{g=1}^{G} s_{ij}\theta_{ji} + \sum_{g,h=1:g\neq h}^{G}\sum_{i\in g:j\in h} s_{ij}\theta_{ji}$$

$$= \sum_{g=1}^{G}\sum_{i\in g} s_{ii}\left(\frac{1}{M^2}\phi_{gg} + \frac{M-1}{M}\gamma_g^{-1}\right) + \sum_{g=1}^{G}\sum_{i,j\in g:i\neq j} s_{ij}\left(\frac{1}{M^2}\phi_{gg} - \frac{1}{M}\gamma_g^{-1}\right)$$

$$+\frac{1}{M^2}\sum_{g,h=1}^{G}\sum_{i\in g:j\in h:g\neq h} s_{ij}\phi_{hg}$$

$$= \frac{1}{M^2}\sum_{g,h=1}^{G}\sum_{i\in g,j\in h} s_{ij}\phi_{gh} + \frac{M-1}{M}\sum_{g=1}^{G}\sum_{i\in g} s_{ii}\gamma_g^{-1} - \frac{1}{M}\sum_{g=1}^{G}\sum_{i,j\in g:i\neq j} s_{ij}\gamma_g^{-1}.$$

Replacing the expressions for s_{ij}, we obtain

$$\mathrm{Tr}(\mathbf{S\Theta}) = \frac{1}{T}\frac{1}{M^2}\sum_{t=1}^{T}\sum_{g,h=1}^{G}\sum_{i\in g:j\in h} y_{it}y_{jt}\phi_{gh} + \frac{M-1}{M}\frac{1}{T}\sum_{t=1}^{T}\sum_{g=1}^{G}\sum_{i\in g} y_{it}^2\gamma_g^{-1}$$

$$-\frac{1}{M}\frac{1}{T}\sum_{t=1}^{T}\sum_{g=1}^{G}\sum_{i\neq j:i,j\in g} y_{it}y_{jt}\gamma_g^{-1}.$$

It follows that the likelihood function is

$$l(\theta) \approx -\ln|M\mathbf{\Sigma}_G + \mathbf{\Gamma}_G| - (M-1)\ln|\mathbf{\Gamma}_G|$$

$$-\frac{1}{MT}\sum_{t=1}^{T}\left(\frac{1}{M}\sum_{g,h=1}^{G}\sum_{i\in g:j\in h} y_{it}y_{jt}\phi_{gh} + (M-1)\sum_{g=1}^{G}\sum_{i\in g} y_{it}^2\gamma_g^{-1} - \sum_{g=1}^{G}\sum_{i\neq j:i,j\in g} y_{it}y_{jt}\gamma_g^{-1}\right).$$

Proof of Theorem 3.1: Note that, from (2.6), and using (3.4), we have

$$\mathbf{\Theta} = \left(\frac{1}{M}\mathbf{\Phi}_G \otimes \frac{1}{M}\mathbf{1}_M\right) + \left({\mathbf{\Gamma}_G}^{-1} \otimes \left(\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\right)\right).$$

Hence, it follows that

$$\widehat{\mathbf{\Theta}} - \mathbf{\Theta} = \left(\frac{1}{M}(\widehat{\mathbf{\Phi}}_G - \mathbf{\Phi}_G) \otimes \frac{1}{M}\mathbf{1}_M\right) + \left((\widehat{\mathbf{\Gamma}}_G^{-1} - \mathbf{\Gamma}_G^{-1}) \otimes (\mathbf{I}_M - \frac{1}{M}\mathbf{1}_M)\right).$$

Noting that, given two matrices A and B, $\|\mathbf{A}\otimes\mathbf{B}\|_F = \|\mathbf{A}\|_F\|\mathbf{B}\|_F$ – see, e.g. Bernstein (2005), p. 676 – and because $\|(1/M)\mathbf{1}_M\|_F = (1/M)\|\mathbf{1}_M\|_F = 1$ and $\|\mathbf{I}_M - (1/M)\mathbf{1}_M\|_F = \sqrt{M-1}$, we have

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F \leq \frac{1}{M}\|\widehat{\mathbf{\Phi}}_G - \mathbf{\Phi}_G\|_F + \sqrt{M-1}\|\widehat{\mathbf{\Gamma}}_G^{-1} - \mathbf{\Gamma}_G^{-1}\|_F.$$

By Theorem 1 in Rothman et al. (2008), we have

$$\|\widehat{\mathbf{\Phi}}_G - \mathbf{\Phi}_G\|_F = O_p\left(\sqrt{\frac{(G+s_G)\ln G}{T}}\right);$$

see also Theorem 1 in Lam and Fan (2009). Further, using the properties of moments of quadratic forms, it is easy to show that $\hat{\gamma}_g - \gamma_g = O_p(1/\sqrt{MT})$, so that

$$\|\widehat{\mathbf{\Gamma}}_G^{-1} - \mathbf{\Gamma}_G^{-1}\|_F = O_p\left(\sqrt{\frac{G}{MT}}\right). \tag{A.3}$$

It follows that

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F = O_p\left(\frac{1}{M}\sqrt{\frac{(G + s_G)\ln G}{T}}\right) + O_p\left(\sqrt{\frac{1}{MT}}\right)$$
$$= O_p\left(\frac{1}{M}\sqrt{\frac{(G + s_G)\ln G}{T}}\right).$$

□

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication files