NETS:
NETWORK ESTIMATION FOR TIME SERIES

Matteo Barigozzi[‡] Christian Brownlees[†]

March 9, 2016

Abstract

We model a large panel of time series as a VAR where the autoregressive matrices and the inverse covariance matrix of the system innovations are assumed to be sparse. The system has a network representation in terms of a directed graph representing predictive Granger relations and an undirected graph representing contemporaneous partial correlations. A LASSO algorithm called NETS is introduced to estimate the model. We apply the methodology to analyse a panel of volatility measures of ninety bluechips. The model captures an important fraction of the overall variability of the time series and improves out–of–sample forecasting.

Keywords: Networks, Multivariate Time Series, VAR, LASSO, Forecasting

JEL: C01, C32, C52

[‡] Department of Statistics, London School of Economics and Political Science,
e-mail: m.barigozzi@lse.ac.uk;
[†] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE,
e-mail: christian.brownlees@upf.edu.

The procedures presented in this paper are available in the package nets for R.

We would like to thank Vasco Carvalho, Frank Diebold, Tony Hall, Nikolaus Hautsch, Julien Idier, Clifford Lam, Marco Lippi, Gabor Lugosi, Tommaso Proietti, and Kimmo Soramaki for their helpful comments. Special thanks go to Giorgio Calzolari. All mistakes are ours.

1 Introduction

Over the last years, network analysis has become an active topic of research in time series econometrics, with numerous applications in macroeconomics and finance. Example of contributions in the literature include, *inter alia*, Billio, Getmansky, Lo, and Pellizzon (2012), Diebold and Yılmaz (2014, 2015), Hautsch, Schaumburg, and Schienle (2014a,b) and Härdle, Wang, and Yu (2016). In a nutshell, network analysis is concerned with representing the interconnections of a large panel as a graph: the vertices of the graph represent the variables in the panel, and the presence of an edge between two vertices denotes the presence of some appropriate measure of dependence between the two variables. From an economic perspective, the interest on networks has been boosted by the research of, *inter alia*, Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012), which shows that individual entities can have a non negligible effect on the aggregate economy when the system has a high degree of interconnectedness.

In this paper we propose network methodology for large panels of time series. We model the panel as a Vector Autoregression (VAR). We work under the assumption that the VAR is sparse, in the sense that the autoregressive matrices and the inverse covariance matrix of the system innovations are assumed to be sparse. Notice that the notion of sparsity used in this work is different from the one used in other papers such as Davis, Zang, and Zheng (2015), Kock and Callot (2015), and Medeiros and Mendes (2016) where sparsity assumptions are formulated for the autoregressive matrices only. Sparsity of the autoregressive matrices implies sparsity of the multivariate Granger causality structure of the system whereas sparsity of the inverse covariance matrix implies sparsity of the partial correlation structure (Dempster, 1972; Lauritzen, 1996).

Several network representations can be associated with a VAR system (Dahlhaus, 2000; Eichler, 2007; Diebold and Yılmaz, 2014). In this work we focus on two representations that are natural for the sparse VAR we work with. The first network representation consists of representing the system as a mixed graph containing both directed and undirected edges:

directed edges denote Granger causality linkages among time series while undirected edges represent contemporaneous partial correlation linkages. The second network representation we introduce is an undirected graph where edges denote long run partial correlation linkages among time–series. Long run partial correlation is a partial correlation measure constructed on the basis of the long run covariance matrix of the VAR. It synthesises simultaneously lead/lag and contemporaneous dependence among time series and can be thought of as a natural generalization for dependent data of the standard partial correlation model used in the statistics graphical literature.

In order to estimate large sparse VARs, we introduce a novel LASSO-based algorithm. The highlight of the procedure is that it simultaneously estimates the autoregressive matrices as well as the entries of the concentration matrix, avoiding to split up the estimation of the model parameters in two steps. The large sample properties of the proposed estimator are analysed and we establish conditions for consistent selection and estimation of the VAR parameters. The theory is derived in a high–dimensional setting, allowing the number of series in the system to increase with the sample size. Specifically, the number of series is allowed to be $O(T^{\zeta})$ for $\zeta > 0$ where T denotes the sample size of the panel.

The network methodology we introduce in this work has highlights in terms of interpretation and estimation. Understanding and synthesising the interdependence structure of a large multivariate system can be a daunting task. The network representation of the panel provides a more parsimonious synthesis of the data that can bring useful insights on their underlying structure. From an estimation perspective, carrying out inference on the VAR parameters can be challenging when the number of time series is large. The regularized estimation approach based on LASSO put forward in this work can lead to substantial gains in terms of estimation precision when the system is sparse and, ultimately, forecasting.

A natural application of network analysis techniques is the study of interdependence in panels of volatility measures. Detecting the interconnectedness structure of volatility panels is of interest to understand and monitor the risk transmission channels of these systems. See, for instance, the research of Diebold and Yılmaz (2014) on risk transmission in the 2007–2009 Great financial crisis or Engle, Gallo, and Velucchi (2012) in the 1997–1998 Asian financial crisis. We use the methodology derived in this work to analyse a panel of volatility measures for ninety US bluechips across different industry groups from January 2nd 2004 to December 31st 2015. An important feature of our application is that we study interconnectedness conditional on a market wide and sector specific volatility factors. We show that after conditioning on the factors the volatility panel has a sparse network structure capturing approximately 10% of the overall variability. The estimated networks connect the vast majority of the series in the panel and the interdependence is positive in the vast majority of cases. Results show that the financial sector is the most interconnected industry in this sample period. In particular, large financial institutions such as AIG, Bank of America and Citigroup are some of the most interconnected entities in the panel. An out–of–sample forecasting exercise is used to validate the methodology proposed in our work and shows that the sparse VAR model improves predictive ability over a number of benchmarks.

Our work relates to different strands of literature. First, it is related to the econometric literature on networks, which includes research by Billio *et al.* (2012), Diebold and Yılmaz (2014, 2015), Hautsch *et al.* (2014a,b), Härdle *et al.* (2016). This paper is also related to the literature on the estimation of sparse VARs, see Davis *et al.* (2015), Kock and Callot (2015), Medeiros and Mendes (2016), and, in a Bayesian setting, Ahelegbey, Billio, and Casarin (2015). Our contribution also relates to the statistical literature on large dimensional network estimation based on LASSO techniques. Contributions in this area include, *inter alia*, Meinshausen and Bühlmann (2006), Friedman, Hastie, and Tibshirani (2008), Peng, Wang, Zhou, and Zhu (2009).

The paper is structured as follows. Section 2 introduces the model, the network definitions and the estimation strategy. Section 3 derives the large sample properties of the estimator. Section 4 contains a simulation study that analyses the finite sample properties of the procedure. Section 5 contains the empirical application. Concluding remarks follow in Section 6.

2 Methodology

2.1 Model

We consider a zero-mean stationary n-dimensional multivariate time series $\mathbf{y}_t = (y_{1\,t}, \ldots, y_{n\,t})'$ generated by a p-th order VAR

$$\mathbf{y}_t = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{y}_{t-k} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim i.i.d.(\mathbf{0}, \mathbf{C}^{-1}), \tag{1}$$

where $\mathbf{A}_k$ and $\mathbf{C}$ are $n \times n$ matrices. Throughout the VAR is assumed to be stable and $\mathbf{C}$ to be positive definite. Notice that for convenience the distribution of the innovation terms is parametrized with the inverse covariance matrix $\mathbf{C}$, also known as concentration matrix, rather than the covariance. The (i,j)-th entries of the matrices $\mathbf{A}_k$ and $\mathbf{C}$ are denoted respectively as $a_{k\,ij}$ and c_{ij}.

In this work we focus on the analysis of sparse VAR systems, in the sense that the autoregressive matrices $\mathbf{A}_k$ and the concentration matrix $\mathbf{C}$ are assumed to be sparse matrices. More specific notions of sparsity are spelled out in Section 3, where precise assumptions are required by the estimation theory to establish the results of interest. In general, the sparsity assumption can be interpreted as a sparsity assumption on the lead/lag and contemporaneous dependence structure of the system.

The standard notion of dynamic interdependence used for time series is Granger causality. In this work we rely on a multivariate version of this concept. Formally, we say that $y_{j\,t}$ does not Granger cause $y_{i\,t}$ if adding $y_{j\,t}$ as predictor does not improve the mean square forecast error of $y_{i\,t+k}$ for any $k > 0$, that is

$$\mathsf{E}[(y_{i\,t+k} - \mathsf{E}(y_{i\,t+k}|\{y_{1\,t} \ldots y_{n\,t}\}))^2] = \mathsf{E}[(y_{i\,t+k} - \mathsf{E}(y_{i\,t+k}|\{y_{1\,t} \ldots y_{n\,t}\} \setminus y_{j\,t}))^2]. \tag{2}$$

It is immediate to see that the Granger causality structure of the model is encoded in the sparsity structure of the autoregressive matrices $\mathbf{A}_k$. We have indeed that if $a_{k\,ij} = 0$, for all k, then $y_{j\,t}$ does not Granger cause $y_{i\,t}$.

The classical measure of contemporaneous dependence used in the network literature is partial correlation. In this paper we consider partial correlation between two series conditional on the past realizations of the panel and contemporaneous realizations of the remaining series. This is encoded in the partial correlation between VAR innovations, which is defined as

$$\rho^{ij} = \mathsf{Cor}(\epsilon_{i\,t}, \epsilon_{j\,t} | \{\epsilon_{k\,t} : k \neq i, j\}). \tag{3}$$

It is well known that partial correlations are related to the entries c_{ij} of the concentration matrix $\mathbf{C}$ by means of the relation (Dempster, 1972)

$$\rho^{ij} = -\frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}. \tag{4}$$

Thus, the contemporaneous dependence sparsity structure is embedded in the sparsity structure of the concentration matrix $\mathbf{C}$. Indeed, if $c_{ij} = 0$, then series i and j are contemporaneously uncorrelated conditional on all other series in the system.

Networks are a useful tool to represent the interdependence structure of the time series in the panel $\mathbf{y}_t$. A network is defined as a graph $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. The set of vertices $\mathcal{V}$ is $\{1, ..., n\}$ where each element corresponds to a component of $\mathbf{y}_t$, while the set of edges $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$ such that the pair (i, j) is in $\mathcal{E}$ if and only if the components i and j are linked by an edge.

Several network definitions have been proposed for VAR models. In this work we introduce two network definitions that are natural for the sparse VAR model we focus on.

A natural representation of the sparse VAR model we work with in this paper is based on the union of two graphs: the first graph contains directed edges denoting Granger causality linkages among time series, while the second graph contains undirected edges representing contemporaneous partial correlation linkages. We label the two networks respectively as the Granger and contemporaneous networks. The Granger network is defined as a directed network $\mathcal{N}_G = (\mathcal{V}, \mathcal{E}_G)$ where the presence of an edge from i to j denotes that i Granger causes j in the sense of (2), that is

$$\mathcal{E}_G = \{(i,j) \in \mathcal{V} \times \mathcal{V} \,:\, a_{k\,ij} \neq 0, \text{ for at least one } k \in \{1, ..., p\}\} \,. \tag{5}$$

The contemporaneous network is defined as an undirected network $\mathcal{N}_C = (\mathcal{V}, \mathcal{E}_C)$ where an edge between i and j denotes that i is partially correlated to j, that is

$$\mathcal{E}_C = \left\{(i,j) \in \mathcal{V} \times \mathcal{V} \,:\, \rho^{ij} \neq 0\right\} . \tag{6}$$

An alternative way to represent the properties of the process consists of simultaneously summarising the lead/lag and contemporaneous information of the system by introducing a partial correlation measure based on the long run covariance matrix of the process. This idea is inspired by the HAC literature (see Newey and West, 1987; Andrews and Monahan, 1992; Den Haan and Levin, 1996). The long run covariance matrix of the process $\mathbf{y}_t$ can be defined as the covariance of the aggregated process:

$$\mathbf{\Sigma}_L = \lim_{M\to\infty} \frac{1}{M} \mathsf{Cov}\left(\sum_{t=1}^{M} \mathbf{y}_t, \sum_{t=1}^{M} \mathbf{y}_t\right) ,$$

assuming the limit exists. Equivalently, the long run covariance is defined in terms of the sum of all autocovariance functions of the process, that is the zero frequency spectral density matrix, which is given by

$$\mathbf{\Sigma}_L = \sum_{h=-\infty}^{+\infty} \mathsf{E}[\mathbf{y}_t \mathbf{y}'_{t-h}],$$

which shows how $\mathbf{\Sigma}_L$ synthesises the linear dependences of $\mathbf{y}_t$ at every lead and lag. Note that since the VAR is assumed to be stationary the sum above is well defined. As it is well known, in the case of a VAR model the long run covariance is given by

$$\boldsymbol{\Sigma}_L = \left(\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k \right)^{-1} \mathbf{C} \left(\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k' \right)^{-1} .$$

We propose a network definition based on the partial correlations constructed on the basis of the long run concentration matrix which is defined as

$$\mathbf{K}_L = \boldsymbol{\Sigma}_L^{-1} = \left(\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k \right)' \mathbf{C} \left(\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k \right) .$$

This is also known as the zero-frequency partial spectral coherence (Dahlhaus, 2000; Davis *et al.*, 2015). Notice that the expression of $\mathbf{K}_L$ is factorized in a sandwich form determined by the term $\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k$, which captures long run dynamic relations of the system, and the term $\mathbf{C}$, which accounts for the contemporaneous dependence of the system innovations. We can then express long run partial correlation coefficient for series i and j as a function of the entries $k_{L\,ij}$ of the long run concentration matrix $\mathbf{K}_L$

$$\rho_L^{ij} = \frac{-k_{L\,ij}}{\sqrt{k_{L\,ii} k_{L\,jj}}} .$$

The long run partial correlation network is then defined as a undirected network $\mathcal{N}_L = (\mathcal{V}, \mathcal{E}_L)$ where the set of edges $\mathcal{E}_L$ is defined as

$$\mathcal{E}_L = \left\{ (i,j) \in \mathcal{V} \times \mathcal{V} \, : \, \rho_L^{ij} \neq 0 \right\} . \tag{7}$$

A number of comments on the model and network definitions we propose are in order. First, this work assumes that the panel has a sparse dependence structure. In practice, this assumption can be quite restrictive. An important case in which the assumption of sparsity is violated, is when the components of the panel are a function of a set of common factors (Forni, Hallin, Lippi, and Reichlin, 2000; Stock and Watson, 2002a,b; Bai, 2003). It is straightforward to see that common factors induce a fully interconnected network structure among the variables in the panel.[1] In these cases the influence of the common factors ought to be filtered out before carrying out network analysis. Generally speaking, we view network analysis as a complement of factor analysis for the purpose of empirical applications (see for example the empirical results in De Mol, Giannone, and Reichlin, 2008 for a justification of this approach).

An important difference between the network modelling approached proposed here and other contributions in the literature is that we focus on representing the partial dependence structure of the panel. On the other hand, the contributions of, *inter alia*, Billio *et al.* (2012) and Diebold and Yılmaz (2014) propose network definitions that measure the overall degree of dependence between series. The advantage of the approach proposed here is that it is robust to spurious correlation effects among the variables in the system. Moreover, the network definitions we propose can be seen as natural extension for time series data of the popular partial correlation network models used in statistics.

2.2 Estimation

We are interested in detecting and estimating the non–zero entries of the autoregressive matrices $\mathbf{A}_k$ and the concentration matrix $\mathbf{C}$. A simple estimation approach for the sparse VAR would consists of using LASSO regression to estimate the autoregressive matrices $\mathbf{A}_k$ (as for example in Kock and Callot, 2015), and then using a LASSO procedure on the residuals to estimate the concentration matrix $\mathbf{C}$ (as for example in Friedman *et al.*, 2008; Peng *et al.*, 2009). The analysis of properties of the second step estimator is however challenging. Moreover, the rate of convergence of the estimator of the concentration matrix $\mathbf{C}$ would

[1] Consider an n–dimensional panel of time series y_{it} generated by a one factor model

$$y_{it} = \beta_i f_t + \epsilon_{it},$$

where f_t and ϵ_{it} are independent normals with zero mean and unit variance and $\epsilon_{i\,t}$ and $\epsilon_{j\,t}$ are independent for each $i \neq j$. Then the concentration matrix of the system is $\mathbf{K} = \mathbf{I}_n - \frac{1}{1+\boldsymbol{\beta}'\boldsymbol{\beta}}\boldsymbol{\beta}\boldsymbol{\beta}'$, where $\mathbf{I}_n$ is the identity matrix of size $n \times n$ and $\boldsymbol{\beta}$ is a $n \times 1$ vector of factor loadings β_i. If the vector of factor loading does not contain zero entries then $\mathbf{K}$ is not sparse.

depend on the rate of convergence of the estimator of the autoregressive matrices $\mathbf{A}_k$.[2] In this work we propose an estimation approach that avoids these hurdles by estimating both sets of parameters jointly.

For ease of notation, we re-parametrize the VAR as a function of: (i) the coefficients α_{ijk} contained in an n^2p-dimensional vector $\boldsymbol{\alpha}$ which correspond to the autoregressive coefficients $a_{k\,ij}$ in (1), (ii) the partial correlations ρ^{ij} contained in an $n(n-1)/2$-dimensional vector $\boldsymbol{\rho}$ and defined in (3), and (iii) the coefficients c_{ii} contained in an n-dimensional vector $\mathbf{c}$ which correspond to the diagonal of the concentration matrix $\mathbf{C}$. Then, in scalar notation the parameters of our model are given by the VAR equations

$$y_{i\,t} = \sum_{k=1}^{p} \sum_{j=1}^{n} \alpha_{ijk}\, y_{j\,t-k} + \epsilon_{i\,t}, \quad i = 1, \ldots, n, \tag{8}$$

and the contemporaneous equations (see Peng *et al.*, 2009)

$$\epsilon_{i\,t} = \sum_{\substack{h=1\\h\neq i}}^{n} \rho^{ih} \sqrt{\frac{c_{hh}}{c_{ii}}} \epsilon_{h\,t} + u_{i\,t}, \quad i = 1, \ldots, n, \tag{9}$$

where $u_{i\,t}$ is an error term uncorrelated with $\epsilon_{h\,t}$ for $i \neq h$.

In this section we define a novel LASSO based estimator for the parameters of (8) and (9). We call the estimation algorithm `nets` (Network Estimator for Time Series) and we describe it in detail in the next section. The main feature of the proposed procedure is that it estimates the autoregressive parameters, $\boldsymbol{\alpha}$, and partial correlations, $\boldsymbol{\rho}$, simultaneously, conditional on a pilot estimator of $\mathbf{c}$.

Consider the following regression representation of $y_{i\,t}$ as a function of the lags of all series as well as the contemporaneous realizations of all other series in the panel, that is

$$y_{i\,t} = \sum_{k=1}^{p} \sum_{j=1}^{n} \beta_{ijk}\, y_{j\,t-k} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\, y_{h\,t} + e_{i\,t}, \tag{10}$$

[2] This is shown in a previous working paper version of this manuscript.

where $e_{i\,t}$ is an error term. It is straightforward to see that (see Lemma 1) the β_{ijk} and γ_{ih} coefficients can be expressed as a function of the α_{ijk}, ρ^{ih} and c_{ii} parameters. In particular, (10) can be re-written as

$$y_{i\,t} = \sum_{k=1}^{p}\sum_{j=1}^{n} \underbrace{\left(\alpha_{ijk} - \sum_{\substack{l=1 \\ l\neq i}}^{n} \rho^{il} \sqrt{\frac{c_{ll}}{c_{ii}}} \alpha_{ljk} \right)}_{\beta_{ijk}} y_{j\,t-k} + \sum_{\substack{h=1 \\ h\neq i}}^{n} \underbrace{\rho^{ih} \sqrt{\frac{c_{hh}}{c_{ii}}}}_{\gamma_{ih}} y_{h\,t} + u_{i\,t}. \tag{11}$$

Notice that the lemma shows that the errors $e_{i\,t}$ and $u_{i\,t}$ are the same. We denote by $\boldsymbol{\theta}$ the vector of parameters of interest $(\boldsymbol{\alpha}', \boldsymbol{\rho}')'$ of dimension $m = n^2p + n(n-1)/2$. The regression representation in (11) suggests to associate the following quadratic loss function to the problem of determining $\boldsymbol{\theta}$, conditional on $\mathbf{c}$,

$$\ell(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{c}) = \sum_{i=1}^{n} \left(y_{i\,t} - \sum_{k=1}^{p}\sum_{j=1}^{n} \left(\alpha_{ijk} - \sum_{\substack{l=1 \\ l\neq i}}^{n} \rho^{il} \sqrt{\frac{c_{ll}}{c_{ii}}} \alpha_{ljk} \right) y_{j\,t-k} - \sum_{\substack{h=1 \\ h\neq i}}^{n} \rho^{ih} \sqrt{\frac{c_{hh}}{c_{ii}}} y_{h\,t} \right)^2 . \tag{12}$$

If a sample of T observations of the $\mathbf{y}_t$ process is available for $t = 1, \ldots, T$ then we propose to estimate the model parameters using a LASSO–type estimator

$$\widehat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \left[\frac{1}{T} \sum_{t=1}^{T} \ell(\boldsymbol{\theta}; \mathbf{y}_t, \widehat{\mathbf{c}}_T) + \lambda_T^{\mathbf{G}} \sum_{k=1}^{p}\sum_{i,j=1}^{n} \frac{|\alpha_{ijk}|}{|\widetilde{\alpha}_{T\,ijk}|} + \lambda_T^{\mathbf{C}} \sum_{\substack{l,h=1 \\ l>h}}^{n} \frac{|\rho^{lh}|}{|\widetilde{\rho}_T^{\,lh}|} \right], \tag{13}$$

where $\lambda_T^{\mathbf{G}} > 0$ and $\lambda_T^{\mathbf{C}} > 0$ are the LASSO tuning parameters and $\widetilde{\boldsymbol{\alpha}}_T$, $\widetilde{\boldsymbol{\rho}}_T$ and $\widehat{\mathbf{c}}_T$ are pre–estimators of the $\boldsymbol{\alpha}$, $\boldsymbol{\rho}$, and $\mathbf{c}$ coefficients, respectively. Due to the presence of autocorrelation in and across the components of $\mathbf{y}_t$ the regressors in (11) are likely to be dependent, therefore we adopt here an adaptive LASSO penalty as originally proposed by Zou (2006) and then studied by Kock (2012, 2016) for dependent data. If the sample size is sufficiently large, a natural pre–estimator of $\boldsymbol{\alpha}$ is the least squares estimator of the VAR autoregressive matrices while the pre–estimator of $\boldsymbol{\rho}$ is the partial correlation estimator obtained from the sample covariance of the VAR residuals. Otherwise, if the sample size is not sufficiently large a pre–estimator of $\boldsymbol{\alpha}$ could be obtained by estimating the autoregressive matrices via ridge regression while the pre–estimator of $\boldsymbol{\rho}$ could be obtained from a shrinkage estimator of the residual covariance (Ledoit and Wolf, 2004). Last, a possible choice for the pre–estimator of $\mathbf{c}$ is the reciprocal of the variance of each series (Peng *et al.*, 2009).

2.3 The `nets` algorithm

In this section we introduce the `nets` algorithm to solve the optimization problem of equation (13). Notice that the loss function in (12) is not the standard quadratic loss function of a linear regression model and the standard LASSO algorithms cannot be applied. However, it is still possible to design a coordinate descent algorithm that can be used to minimize the objective function of (13). The procedure we propose is a generalization of the `space` algorithm proposed by Peng *et al.* (2009) for the estimation of partial correlation networks, and its a variation of the shooting algorithm of Fu (1998) typically used for LASSO optimization.

Additional notation is required to describe the algorithm. We begin by introducing the matrix representation of the model of equation (11) obtained by stacking the time series in the panel. Let $\boldsymbol{\mathcal{Y}}$ denote a $nT \times 1$ vector defined as $(y_{1\,1}, ..., y_{1\,T},\ ...,\ y_{i\,1}, ..., y_{i\,T},\ ...,\ y_{n\,1}, ..., y_{n\,T})'$; let $\boldsymbol{\mathcal{X}}_G = (\boldsymbol{x}_{G\,111}, ..., \boldsymbol{x}_{G\,ijk}, ..., \boldsymbol{x}_{G\,nnp})$ be a $nT \times n^2p$ matrix with (i,j,k)-th column defined as

$$\boldsymbol{x}_{G\,ijk} = (\,0, ..., 0,\ \underbrace{y_{j\,-k}, ..., y_{j\,t-k}, ..., y_{j\,T-k}}_{i\text{–th block}},\ 0, ..., 0)',$$

and let $\boldsymbol{\mathcal{X}}_C = (\boldsymbol{x}_{C\,21}, \boldsymbol{x}_{C\,31}, \boldsymbol{x}_{C\,32}, ..., \boldsymbol{x}_{C\,ij}, ..., \boldsymbol{x}_{C\,n(n-1)})$ be a $nT \times n(n-1)/2$ matrix with (i,j)–th column defined as

$$\boldsymbol{x}_{C\,ij} = \left(\ 0, ..., 0,\ \underbrace{\sqrt{\frac{c_{jj}}{c_{ii}}}(y_{j\,1}, ..., y_{j\,t}, ..., y_{j\,T})}_{i\text{–th block}}\ 0, ..., 0,\ \underbrace{\sqrt{\frac{c_{ii}}{c_{jj}}}(y_{i\,1}, ..., y_{i\,t}, ..., y_{i\,T})}_{j\text{–th block}}\ 0, ..., 0\right)'.$$

Then, it is straightforward to check that model (11) can be represented as

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}_G\, \boldsymbol{\beta}(\boldsymbol{\alpha}, \boldsymbol{\rho}) + \boldsymbol{\mathcal{X}}_C\, \boldsymbol{\rho} + \boldsymbol{\mathcal{U}},$$

where $\boldsymbol{\mathcal{U}}$ is a $nT \times 1$ vector of residuals and $\boldsymbol{\beta}(\cdot, \cdot)$ denotes the function which maps the $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ parameter vectors onto the $\boldsymbol{\beta}$ parameter vector whose components are given in (10). Notice that the parameter $\boldsymbol{\beta}$ and the matrix $\boldsymbol{\mathcal{X}}_C$ depend implicitly on the parameter $\mathbf{c}$ and that in the estimation we set this to the pre-estimator $\widehat{\mathbf{c}}_T$. The dependence on $\mathbf{c}$ is suppressed in the notation for simplicity. In what follows it is convenient to introduce shorthand notation for the stacked vectors. Let $\boldsymbol{v}$ be a $nT \times 1$ stacked vector, then we use $v_{[it]}$ to refer to the t-th element of the i-th block of $\boldsymbol{v}$.

The `nets` algorithm is an iterative coordinate descent procedure for the minimization of the objective function of (13). Each iteration s of the algorithm updates one component of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\rho}')'$. The $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ parameter estimates at iteration s are denoted as $\widehat{\boldsymbol{\alpha}}^{(s)}$ and $\widehat{\boldsymbol{\rho}}^{(s)}$ respectively. We define the residual estimate at iteration s as

$$\widehat{\boldsymbol{\mathcal{U}}}^{(s)} = \boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}_G\, \boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}}^{(s)}, \widehat{\boldsymbol{\rho}}^{(s)}) - \boldsymbol{\mathcal{X}}_C\, \widehat{\boldsymbol{\rho}}^{(s)} \,.$$

The algorithm iterates until convergence, which is checked at the end of each full cycle of updates of $\boldsymbol{\theta}$. To describe the algorithm, it is useful to use two auxiliary $nT \times 1$ stacked vectors $\ddot{\boldsymbol{x}}$ and $\ddot{\boldsymbol{y}}$. The $\ddot{\boldsymbol{x}}$ vector denotes the regressors corresponding to the current coefficient being updated, while the $\ddot{\boldsymbol{y}}$ vector is the partial residual of the model with respect to all parameter besides the coefficient being currently updated (either α_{ijk} or ρ^{ij}). The α_{ijk} coefficient is updated as

$$\hat{\alpha}^{(s)}_{ijk} = \text{sign}\left(\ddot{\boldsymbol{y}}'\ddot{\boldsymbol{x}}\right) \left(\left| \frac{\ddot{\boldsymbol{y}}'\ddot{\boldsymbol{x}}}{\ddot{\boldsymbol{x}}'\ddot{\boldsymbol{x}}} \right| - \frac{\lambda^{\mathbf{G}}_T}{\widetilde{\alpha}_{ijk}} \frac{1}{\ddot{\boldsymbol{x}}'\ddot{\boldsymbol{x}}} \right)_+ ,$$

where $\ddot{\boldsymbol{x}}$ and $\ddot{\boldsymbol{y}}$ are defined as

$$\begin{aligned}
\ddot{x}_{[lt]} &= \begin{cases} y_{l\,t-k}, & \text{if } l = i \\ -\hat{\rho}^{il\,(s-1)}\sqrt{\frac{\widetilde{c}_{ll}}{\widetilde{c}_{ii}}}\, y_{j\,t-k} & \text{otherwise} \end{cases} \\
\ddot{y}_{[lt]} &= \mathcal{U}_{[lt]}^{(s-1)} + \hat{\alpha}_{ijk}^{(s-1)} \ddot{x}_{[lt]},
\end{aligned}$$

for each $l = 1, ..., n$ and $t = 1, ..., T$. The ρ^{ij} coefficient is updated as

$$\hat{\rho}^{ij\ (s)} = \operatorname{sign}\left(\ddot{\boldsymbol{y}}'\ddot{\boldsymbol{x}}\right)\left(\left|\frac{\ddot{\boldsymbol{y}}'\ddot{\boldsymbol{x}}}{\ddot{\boldsymbol{x}}'\ddot{\boldsymbol{x}}}\right| - \frac{\lambda_T^{\mathbf{C}}}{\widetilde{\rho}^{ij}}\frac{1}{\ddot{\boldsymbol{x}}'\ddot{\boldsymbol{x}}}\right)_+ ,$$

where $\ddot{\boldsymbol{x}}$ and $\ddot{\boldsymbol{y}}$ are defined as

$$\begin{aligned}
\ddot{x}_{[lt]} &= \sqrt{\frac{\widetilde{c}_{hh}}{\widetilde{c}_{ll}}}\left(y_{h\,t} - \sum_{j=1}^{n}\sum_{k=1}^{p} \alpha_{hjk}^{(s-1)} y_{j\,t-k}\right) \\
\ddot{y}_{[lt]} &= \mathcal{U}_{[lt]}^{(s-1)} + \hat{\rho}^{ij\,(s-1)} \ddot{x}_{[lt]},
\end{aligned}$$

for (l, h) equal (i, j) or (j, i) and $t = 1, ..., T$, and otherwise $\ddot{y}_{[lt]}$ and $\ddot{x}_{[lt]}$ are set to zero.

It is important to stress that the parameter vector $\boldsymbol{\theta}$ contains $n^2p + \frac{n(n-1)}{2}$ elements, whose optimization would require large amounts of memory to be stored when the panel is large. On the other hand, the coordinate wise minimization algorithm is appealing in this context in that it has mild storage requirements and can be applied in large dimensional applications.

As far as the estimation of $\mathbf{c}$ is concerned, we follow the two–step iterative procedure proposed in Peng *et al.* (2009): (i) Given an estimate of $\mathbf{c}$, we estimate the $\boldsymbol{\theta}$ parameter using `nets`. (ii) Given an estimate $\mathbf{c}$ and an estimate of $\boldsymbol{\theta}$ we update the estimate of $\mathbf{c}$. Notice, that c_{ii} is the reciprical of the residual variance of equation (11). These two steps are then iterated until convergence, which typically kicks in within very few iterations.

A number of hacks have been developed in the literature to optimize the estimation of LASSO models when the number of parameters is large. We point out here that the active shooting approach proposed in Peng *et al.* (2009) is particularly useful. Active shooting consists of carrying out the coordinate descent steps for the current non–zero parameters until convergence (the active set) and then for the zero parameters. When the LASSO solution is sparse, active shooting can lead to substantial advantages in terms of execution time.

3 Theory

In this section we show, under appropriate conditions, the estimation and selection consistency of the estimator defined in the previous section. We denote by $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0', \boldsymbol{\rho}_0')'$ and $\mathbf{c}_0$ the true value of the parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\rho}')'$ and $\mathbf{c}$. The proofs of the propositions of this section are provided in the Appendix.

The detailed assumptions are given in the Appendix and here we only review the main features of the model. First, we assume $\mathbf{y}_t$ to be generated by a stable VAR as in (1). Moreover, we require bounds on the higher order moments of $\mathbf{y}_t$, such that suitable Bernstein-type exponential inequalities for dependent processes apply (see e.g. Bosq, 1996; Doukhan and Neumann, 2007). Second, we assume positive definiteness of the spectral density matrix of $\mathbf{y}_t$ and of the precision matrix, $\mathbf{C}_0$, of the VAR innovations $\boldsymbol{\epsilon}_t$. This guarantees that the population Granger and contemporaneous network are both well defined. Given the loss function, $\ell(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{c})$ defined in (12), the unconstrained problem has a solution in population, i.e. the parameter vector $\boldsymbol{\theta}_0$ is identified.

PROPOSITION 1. *Under Assumption 1, the true value of the parameters is such that* $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathsf{E}[\ell(\boldsymbol{\theta}; \mathbf{y}_t, \mathbf{c}_0)]$.

The estimator defined in (13) can be equivalently formulated as

$$\widehat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \mathcal{L}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T), \tag{14}$$

where

$$\mathcal{L}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T) = \left[\frac{1}{T}\sum_{t=1}^{T} \ell(\boldsymbol{\theta}; \mathbf{y}_t, \widehat{\mathbf{c}}_T) + \lambda_T \sum_{i=1}^{m} w_i|\theta_i|\right], \tag{15}$$

where λ_T is the LASSO tuning parameter and w_i are the adaptive LASSO weights. The specification of the weight is $w_i = C_\bullet / |\widetilde{\theta}_{T\,i}|$ where $\widetilde{\theta}_{T\,i}$ denotes the pre–estimator of the θ_i coefficient and $C_\bullet$ denotes a known positive constant that is equal to C_α for the $\boldsymbol{\alpha}$ coefficients and C_ρ for the $\boldsymbol{\rho}$ coefficients. Put differently, in the theoretical analysis of the estimator we assume that the $\lambda_T^{\mathbf{G}} = \lambda_T\, C_\alpha$ and $\lambda_T^{\mathbf{C}} = \lambda_T\, C_\rho$. Thus, λ_T controls the overall degree of shrinkage of the parameters of the model.

In what follows we denote the sets of non–zero parameters as $\mathcal{A}_G = \{(i,j,k) : \alpha_{0\,ijk} \neq 0\}$, $\mathcal{A}_C = \{(i,j) : \rho_0^{ij} \neq 0\}$ and $\mathcal{A} = \mathcal{A}_G \cup \mathcal{A}_C$. The number of non–zero parameters in the model is $q_T = |\mathcal{A}|$. The set of zero parameters is then $\mathcal{A}^c$. Let also $\{s_T\}$ be a positive sequence of real numbers such that for any $i \in \mathcal{A}$ we have $|\theta_{0\,i}| \geq s_T$.

The solution of (14) when restricted to the set of parameters $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}_{\mathcal{A}^c} = \mathbf{0}$ is denoted as $\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}$. This is the estimator of the non–zero parameters obtained when we assume to know those that are zero. To obtain consistency we follow the same strategy as in Meinshausen and Bühlmann (2006), Peng *et al.* (2009), and Fan and Peng (2004). First, we prove consistency in the restricted problem.

PROPOSITION 2. (ESTIMATION CONSISTENCY). *Suppose that, as* $T \to \infty$, $q_T = o\left(\sqrt{\frac{T}{\log T}}\right)$, $\lambda_T \sqrt{\frac{T}{\log T}} \to \infty$, *and* $\sqrt{q_T}\lambda_T = o(1)$. *Then, under Assumptions 1 and 2, for any* $\eta > 0$, $\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}$ *exists with probability at least* $1 - O(T^{-\eta})$, *and there exists a constant* $\kappa_R > 0$ *such that*

$$\Pr\left(\left\| \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} - \boldsymbol{\theta}_{0\,\mathcal{A}} \right\| \leq \kappa_R \sqrt{q_T}\lambda_T \right) \geq 1 - O(T^{-\eta}).$$

Moreover, if the signal sequence s_T *is such that,* $\frac{s_T}{\sqrt{q_T}\lambda_T} \to \infty$, *then* $\Pr\big(\mathrm{sign}(\widehat{\theta}_{T\,i}^{\mathcal{A}}) = \mathrm{sign}(\theta_{0\,i})\big) \geq 1 - O(T^{-\eta})$, *for any* $i \in \mathcal{A}$.

Second, we show that the restricted estimator is also a solution of the unrestricted problem, i.e. when we do not know the zero coefficients (see Lemma 8 in the Appendix). Then, in the next Proposition, we prove consistency of edge selection and of the unrestricted estimator.

Proposition 3. (selection consistency and oracle property). *Suppose that the same conditions of Proposition 2 hold and that, as $T \to \infty$, $n = O(T^{\zeta})$ for some $\zeta > 0$ and $\sqrt{q_T}\sqrt{\frac{\log T}{T}} = o(\lambda_T)$. Then, under Assumptions 1 and 2, for any $\eta > 0$, $\widehat{\boldsymbol{\theta}}_T$ exists with probability at least $1 - O(T^{-\eta})$. Moreover,*

(a) $\Pr\big(\widehat{\theta}_{T\,i} = 0\big) \geq 1 - O(T^{-\eta})$*, for any* $i \in \mathcal{A}^c$*;*

(b) there exists a constant $\kappa_U > 0$ *such that*

$$\Pr\left(\left\|\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right\| \leq \kappa_U \sqrt{q_T}\lambda_T\right) \geq 1 - O(T^{-\eta}),$$

and $\Pr\big(\mathrm{sign}(\widehat{\theta}_{T\,i}) = \mathrm{sign}(\theta_{0\,i})\big) \geq 1 - O(T^{-\eta})$*, for any* $i \in \mathcal{A}$*.*

Since $\widehat{\boldsymbol{\theta}}_T = (\widehat{\boldsymbol{\alpha}}_T', \widehat{\boldsymbol{\rho}}_T')'$, as an immediate consequence of Proposition 3 we have consistency of all the estimated parameters of the model and we have selection consistency for the non–zero edges of the Granger and contemporaneous networks.

Corollary 1. *Define as* $\widehat{\alpha}_{T\,ijk}$ *the generic entry of* $\widehat{\boldsymbol{\alpha}}_T$ *and as* $\widehat{\rho}_T^{ij}$ *the generic entry of* $\widehat{\boldsymbol{\rho}}_T$*, and define the estimated edges' sets of the Granger and contemporaneous networks as*

$$\begin{aligned}
\widehat{\mathcal{E}}_{G\,T} &= \{(i,j) \in \mathcal{V} \times \mathcal{V} \,:\, \widehat{\alpha}_{T\,ijk} \neq 0, \ \textit{for at least one } k \in \{1,...,p\}\}, \\
\widehat{\mathcal{E}}_{C\,T} &= \left\{(i,j) \in \mathcal{V} \times \mathcal{V} \,:\, \widehat{\rho}_T^{ij} \neq 0\right\}.
\end{aligned}$$

Then, under the same Assumptions of Proposition 3, for T *sufficiently large and any* $\eta > 0$*, we have* $\Pr(\widehat{\mathcal{E}}_{G\,T} = \mathcal{E}_G) \geq 1 - O(T^{-\eta})$*, and* $\Pr(\widehat{\mathcal{E}}_{C\,T} = \mathcal{E}_C) \geq 1 - O(T^{-\eta})$*.*

Given the asymptotic conditions on the number of non–zero coefficients, q_T, the "worst case" scenario is when it is almost in the order of $\sqrt{\frac{T}{\log T}}$. In that case λ_T needs to be nearly in the order of $T^{-1/4}$. On the other hand, for the "best case" scenario, that is when $q_T = O(1)$ (for example, when the dimension n is fixed), then the order of λ_T can be nearly as small as $T^{-1/2}$ (within a factor of $\log T$). Consequently, the L_2-norm distance of the estimator from the true parameter is in the order of $\sqrt{\frac{\log T}{T}}$, with probability at least $1 - O(T^{-\eta})$.

4 Simulation Study

In this section we analyse the properties of our estimator using simulated data. The exercise consists of simulating a large sparse VAR process and then using the `nets` algorithm to estimate it.

FIGURE 1 ABOUT HERE

We simulate a 100 dimensional sparse VAR(1). Notice that the total number of parameters in this system is 15050. The sparse autoregressive matrix $\mathbf{A}_1$ and concentration matrix $\mathbf{C}$ are obtained from an Erdös–Renyi random graph model. The Erdös–Renyi random graph is a graph defined over a fixed set of vertices and a random set of edges, where the existence of an edge between vertices i and j is determined by a Bernoulli trial with probability p, independent of all other edges. The $\mathbf{A}_1$ matrix is constructed on the basis of a directed Erdös–Renyi model $\mathcal{G}_1$

$$[\mathbf{A}_1]_{ij} = \begin{cases} 0.275 & \text{if } (i,j) \in \mathcal{E}_1 \\ 0 & \text{otherwise} \end{cases},$$

where $\mathcal{E}_1$ is the set of (directed) edges of $\mathcal{G}_1$. The concentration matrix $\mathbf{C}$ is constructed on the basis of a undirected Erdös–Renyi model $\mathcal{G}_2$

$$[\mathbf{C}]_{ij} = \begin{cases} -\frac{1}{\sqrt{d_i d_j}} & i \neq j \text{ and } (i,j) \in \mathcal{E}_2 \\ 1.5 & i = j \\ 0 & \text{otherwise} \end{cases},$$

where $\mathcal{E}_2$ is the set of (undirected) edges of $\mathcal{G}_2$ and d_i denotes the degree of vertex i in this graph. Note that the simulation is designed in a way such that the sparsity structure of the Granger and contemporaneous networks of the VAR coincide with the one of the two random graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. Also, the specification guarantees that the VAR is stable and that the concentration matrix is positive definite. The edge probability p is set so that the expected number of links in each of the two networks is equal to the number of series, n, in the panel. We report in Figure 1 the plot of the Granger and contemporaneous networks associated with a randomly chosen realization of the model. Notice that despite the networks being sparse (in the sense that the expected number of links is $O(n)$), they are almost fully interconnected. We simulate samples of different sizes from the sparse VAR(1) we just described (T =250, 500, 750 and 1000) and then use the `nets` algorithm to estimate the model. For simplicity, the tuning parameters $\lambda_T^{\mathbf{G}}$ and $\lambda_T^{\mathbf{C}}$ are set equal to a common shrinkage tuning parameter λ_T. Our LASSO estimator requires pre–estimators of the $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ parameters to construct the LASSO penalty weights. The pre–estimator of $\boldsymbol{\alpha}$ is the least squares estimator of the VAR(1) autoregressive matrix, while the pre–estimator of $\boldsymbol{\rho}$ is the partial correlation estimator obtained from the sample covariance of the VAR(1) residuals. Last, we initialize $\mathbf{c}$ using the reciprocal of the sample variances of each series. The model is then estimated over a range of values of the common shrinkage tuning parameter λ_T.

FIGURE 2 AND TABLE 1 ABOUT HERE

The simulation is replicated 1000 times and the quality of the `nets` estimator is measured on the basis of the MSE and the ROC curve, which is the plot of the false discovery rate (FDR) of the estimator versus the true positive rate (TPR). We report in the left panel of Figure 2 the MSE of the `nets` estimator as a function of the tuning parameter λ_T for the sample size T equal to 500, 750 and 1000.[3] The picture displays the typical profile of shrinkage type estimators, that is the MSE is a convex function of the tuning parameter, and as the sample size increases the MSE of the estimator decreases. The right panel of Figure 2 reports the ROC curve associated with the `nets` estimator for the sample size T equal to 250, 500, 750 and 1000. Recall that the FDR is defined as the ratio of incorrectly detected non–zero parameters over the total number of zero parameters, while TPR is defined as the ratio of correctly detected non–zero parameters over the total number of non–zero parameters. Note that the penalization coefficient determines the FDR and TFR properties of the estimator:

[3] We omit from the picture for $T = 250$ because of scaling issues.

when λ_T is small (large), the proportion of type 1 errors is high (low) while the proportion of type 2 errors is low (high). The curves show that as the sample size T increases the performance of the `nets` estimator, as measured by the area underneath the ROC curve, increases steadily. In Table 1 we report detailed results on the MSE and TPR of the `nets` estimator when the FDR is controlled at 1%, 5% and 10%. For comparison purposes, the table also reports the MSE of the pre–estimator. The MSE of the `nets` estimator decreases steadily as the sample size get larger. When the sample size is 250 the efficiency gains with respect to the pre–estimator are substantial. As the sample size increases the pre–estimator becomes progressively more efficient relative to the LASSO estimator, however the efficiency gain of `nets` are still large. As far as the TPR is concerned, the table shows that when the TPR is controlled at 1%, 5% and 10% levels, the procedure has a fair amount of power even when the sample size T is 250, and that as the sample size increases power raises steadily. In particular, the power is roughly around 80% when the sample size is 750 and the FDR is controlled at the 1% level. Overall, the simulation results convey that the `nets` algorithm performs satisfactorily, and that the gains with respect to the traditional estimator can be large for sparse VAR systems.

5 Empirical Application

We use the methodology introduced in this work to study interconnectedness in a panel of volatility measures. The application is close in spirit to, among others, the research of Diebold and Yılmaz (2009, 2014, 2015) and Engle *et al.* (2012).

5.1 Data

TABLE 2 ABOUT HERE

We consider a panel of ninety U.S. bluechips across different industry sectors. The list of company names and industry groups is in Table 2. Our sample spans January 2nd 2004 to December 31st 2015, which corresponds to 3021 trading days. During this sample period most of the stocks in the list have been part of the S&P 100 index. Following Diebold and Yılmaz (2015) we measure volatility using the high–low range (Parkinson, 1980)

$$\tilde{\sigma}^2_{i\,t} = 0.361 \left(p^{\mathsf{high}}_{i\,t} - p^{\mathsf{low}}_{i\,t} \right) ,$$

where $p^{\mathsf{high}}_{i\,t}$ and $p^{\mathsf{low}}_{i\,t}$ denote respectively the max and the min log price of stock i on day t.[4]

We focus on analyzing volatility interconnectedness conditional on a market wide and sector specific volatility factors. There is a large literature documenting evidence of a factor structure in volatility (see, *inter alia*, Barigozzi, Brownlees, Gallo, and Veredas, 2014; Luciani and Veredas, 2015; Ghysels, 2014; Barigozzi and Hallin, 2015). As previously pointed out, it is straightforward to check that when common factors are present the dependence structure of the data is not sparse. To this extent, we study the interconnectedness of the residuals of the regression

$$\log \tilde{\sigma}^2_{i\,t} = \beta_0 + \beta_1 \log \tilde{\sigma}^2_{m\,t} + \beta_2 \log \tilde{\sigma}^2_{s\,t} + z_{i\,t}, \tag{16}$$

where $\tilde{\sigma}^2_{m\,t}$ and $\tilde{\sigma}^2_{s\,t}$ denote respectively a market wide and a sector specific volatility factors. The market and sectoral volatilities are measured using the high–low range estimator applied to the S&P 500 index and the SPDR sectoral indices of S&P 500.[5] The residuals are obtained after estimating the model by least squares. In what follows we refer to the volatility residual panel as the volatility panel for short.

TABLE 3 AND FIGURE 3 ABOUT HERE

[4] Several advanced estimators of volatility based on high frequency data have been proposed over the last years (Andersen, Bollerslev, Diebold, and Labys, 2003; Barndorff-Nielsen, Hansen, Lunde, and Shephard, 2008; Aït-Sahalia, Mykland, and Zhang, 2005). However, despite its simplicity a number of contributions have pointed out that the high–low range estimator performs satisfactorily relative to more sophisticated alternatives (Alizadeh, Brandt, and Diebold, 2002; Brownlees and Gallo, 2010).

[5] The SPDR sectoral indices of the S&P 500 we use are Consumer Discretionary (XLY), Consumer Staples (XLP), Energy (XLE), Financials (XLF), Health Care (XLV), Industrials (XLI), Materials (XLB), Technology (XLK) and Utilities (XLU).

Table 3 reports summary statistics on the variance, kurtosis, autocorrelation, average cross correlation and average cross autocorrelation of order one for the volatility residuals. Moreover, in Figure 3 we show the heatmaps of the sample autocorrelation matrix of order one and the sample correlation matrix. We note that after netting out the factors, the volatility residuals still exhibit autocorrelation. It is important to emphasize that while the raw volatility measure exhibit long range dependence, the volatility residuals exhibit a considerably weaker autocorrelation structure. Inspection of the average correlations and the heatmaps shows that contemporaneous and lagged cross correlation is still present in the volatility residuals. Interestingly, tickers in the same industry still exhibit a moderate degree of correlation even after conditioning on the sectoral factors.

5.2 In–Sample Estimation Results

We analyse the panel of volatility measures using the `nets` algorithm over the entire sample. The order of the VAR model p is set to one. The pre–estimator of the $\boldsymbol{\alpha}$ parameters is the least squares estimator of the VAR(1) autoregressive matrix, while the pre–estimator of the $\boldsymbol{\rho}$ parameters is the partial correlation estimator obtained from the sample covariance of the VAR(1) residuals. Last, we initialize $\mathbf{c}$ using the reciprocal of the sample variances of each series. The penalties $\lambda_T^{\mathbf{G}}$ and $\lambda_T^{\mathbf{C}}$ are determined by a cross–validation procedure. We split the entire sample in an estimation and a validation samples. The estimation sample corresponds to the first 75% of the entire sample and the validation sample to the last 25%. For given values of the tuning parameters, we first estimate the model in the estimation sample and then compute the residual sum of squares (RSS) in the validation sample. We perform these steps over a grid of $\lambda_T^{\mathbf{G}}$ and $\lambda_T^{\mathbf{C}}$ values and then choose the optimal tuning parameters as the ones that minimize the validation RSS. We then estimate the model over the entire sample using the optimal value of the tuning parameters.

TABLE 4 AND FIGURE 4 ABOUT HERE

We report the estimated Granger and contemporaneous networks in Figure 4. In the Granger network plot the diameter of each vertex is proportional to the out–degree (the number of non–zero spillovers effects toward others) while in the contemporaneous network the diameter is proportional to the degree. In both plots we use the vertex color to denote the different industry groups. We exclude from the graphs the vertices that do not have any connections, which is one ticker in the Granger network and seven tickers in the contemporaneous network.

Table 4 reports the number of linkages of the Granger and contemporaneous networks of the entire panel and individual sectors. The estimated Granger volatility network has a total of 251 edges (approximately 3% of the total edges), while the contemporaneous volatility network contains 294 edges (approximately 7% of the total edges). The estimated networks share some common features. For instance, the number of industry linkages of the two networks are highly correlated and the financial sector is in particular the sector that accounts for most linkages.

We compute an in–sample R^2 type goodness–of–fit criteria for each series in the panel to summarise the amount of variability explained by the sparse VAR, which is defined as the proportion of variance explained by the regression equation (11). Table 4 reports the average of the R^2 index over the entire panel as well as the individual sectors. The index has a strong positive correlation with the number of linkages in each sector and is on average around 22%. For comparison purposes, Table 4 also reports in–sample factor and sectoral R^2. The factor R^2 is defined as the R^2 obtained by regressing the volatility measure on the market factor and the sector R^2 is defined as the R^2 obtained by regressing the volatility measure on the market wide and sector factor minus the factor R^2. The market and sector factors account for most of the variability in the series, which is roughly 56%. A back of the envelop computation shows that the networks explains around an additional 11% of the overall variability, which roughly matches the amount of variability explained by sectoral factors.

TABLE 5 AND FIGURE 5 ABOUT HERE

In order to get better insights on the industry linkages in Table 5 we report the total number of links between industry groups. It is interesting to note that after conditioning on the sectoral factors there are still a moderate number of interconnections between firms within the same industry. The table also shows that firms in the financial sector in particular have a high degree of interconnectedness across industries. In Figure 5 we report the degree distribution of the estimated networks and the distribution of the non–zero $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ coefficients. As far as the degree distribution is concerned, the number of connections has a high degree of heterogeneity in the cross section. In particular, in the contemporaneous network the most interconnected tickers account for a large number of connections relative to the total. The histogram of the non–zero coefficients shows that the majority of the coefficients are positive and that positive coefficients are on average larger than the negative ones.

TABLE 6 ABOUT HERE

Last, we rank the firms in the panel on the basis of their influence in the Granger and contemporaneous networks. We measure the influence of series j in the Granger and contemporaneous networks using, respectively, the indices $\sum_{i \neq j}^{N} |\widehat{\alpha}_{ij1}|$ and $\sum_{i \neq j}^{N} |\widehat{\rho}_{ij}|$. We report the top ten most influential tickers of the Granger and contemporaneous networks according to this criteria in Table 6. The table shows clearly that large financials firms are highly influential. In particular, the results shows that the financial firms that have been heavily involved in the great financial crisis like Bank of America (BAC), AIG and Citigroup (C) are the stocks associated with the largest spillover effects in the Granger network.

Overall, the in–sample estimation results show that, after conditioning on market wide and sectoral factors, the sparse VAR captures an important proportion of overall variability, and that the financial industry in particular has the highest degree of interconnectedness.

5.3 Out–of–Sample Forecasting

We carry out a forecasting exercise to evaluate the out–of–sample performance of the methodology. The exercise is designed as follows. We split the sample in an in–sample period spanning January 2nd 2004 to December 31st 2013 and an out–of–sample period spanning January 2nd 2014 to December 31st 2015. We first estimate the sparse VAR in–sample using the same steps outlined in the previous section and we then evaluate the model in the out–of–sample period.

TABLE 7 ABOUT HERE

The prediction evaluation is divided into two parts. The first part focuses on the evaluation of the `nets` estimator of the autoregressive component by predicting one–step–ahead volatility residuals. The benchmark forecast for this exercise is the constant zero forecast. Notice that the constant zero forecast represents the optimal forecast in case the dependence in the panel is fully captured by the factor part of model (16) without exploiting the information in the residuals. The competing forecasts are the ones obtained from a VAR model estimated via `nets`, univariate AR(1) models estimated by least squares, and a VAR model estimated by ridge regression (with tuning parameter chosen by Generalized Cross Validation). Note that the volatility residuals are obtained from the estimation results of model (16) estimated over the entire sample.

We report the forecasting results in the top panel of Table 7. The first row of the table reports the MSE of the benchmark while the remaining rows report the out–of–sample R^2 of the competitors. The out–of–sample R^2 index is defined as one minus the ratio of the MSE of the competing models over the MSE of the benchmark. The performance indices are averaged over the entire panel and the industry sectors. The results show that the VAR forecasts obtained by the `nets` estimator systematically improve forecasting ability over the benchmark by roughly 8% on average and it is the best performing forecast method overall.

The second part focuses on the evaluation of the `nets` estimator of the contemporaneous component by predicting the contemporaneous volatility residuals conditional on the estimated autoregressive component. We construct the series of VAR residuals $\hat{\epsilon}_{it}$ of the autoregressive component estimated by `nets`, and the focus is on predicting each residuals series conditional on the remaining ones on the basis of the regression

$$\hat{\epsilon}_{i\,t} = \sum_{\substack{h=1 \\ h\neq i}}^{n} \gamma_{ij}\hat{\epsilon}_{h\,t} + u_{i\,t}, \quad i = 1, \ldots, n.$$

The benchmark forecast for this exercise is again the constant zero forecast, which is the optimal forecast in case the residuals do not have any cross–correlation. The competing forecasts are ones the ones obtained from the contemporaneous component of the VAR estimated by `nets`, the ones obtained from a linear regression estimated by least squares and a linear regression estimated by ridge regression (with tuning parameter chosen by Generalized Cross Validation). The linear regression and the ridge regression are estimated in the in–sample period using the in–sample one–step ahead forecast errors.

We report the forecasting results in the bottom panel of Table 7. The first row of the table reports the average MSE of the benchmark model while the remaining rows report the out–of–sample R^2 of the competitors. Results show that the `nets` forecasts systematically improve out–of–sample predictive ability across sectors and on average improve forecasting over the benchmark by 13%.

6 Conclusions

In this work we introduce network techniques for the analysis of lagre panels of time series. We model a panel as a VAR where the autoregressive matrices and the inverse covariance matrix of the system innovations are assumed to be sparse. The system has a natural network representation in terms of a directed graph representing predictive Granger relations and an undirected graph representing contemporaneous partial correlations. A LASSO estimation algorithm called `nets` is introduced to estimate simultaneously the autoregressive matrices and the inverse covariance matrix of the model. The large sample properties of the estimator are established in a high–dimensional setting. The methodology is used to analyse a panel of volatility measures of US bluechips between January 2004 and December 2015 conditional on market wide and sector specific volatility factors. The analysis shows that the series exhibit a hight degree of interconnectedness and that financial firms have the highest degree of interdependence. A forecasting exercise shows that the methodology introduced in this work allows to improve forecasting over a number of benchmarks.

References

Acemoglu, D., Carvalho, V., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, **80**, 1977–2016.

Ahelegbey, D. F., Billio, M., and Casarin, R. (2015). Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics*. available online.

Aït-Sahalia, Y., Mykland, P. A., and Zhang, L. (2005). How often to sample a continuous–time process in the presence of market microstructure noise. *The Review of Financial Studies*, **28**, 351–416.

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, **57**, 1047–1091.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, **60**, 953–966.

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

Barigozzi, M. and Hallin, M. (2015). Generalized dynamic factor models and volatilities: recovering the market volatility shocks. *The Econometrics Journal*. available online.

Barigozzi, M., Brownlees, C., Gallo, G. M., and Veredas, D. (2014). Disentangling systematic and idiosyncratic dynamics in panels of volatility measures. *Journal of Econometrics*, **182**, 364–384.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, **76**, 1481–1536.

Billio, M., Getmansky, M., Lo, A., and Pellizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, **104**, 535–559.

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction.* Springer, New York.

Brownlees, C. T. and Gallo, G. M. (2010). Comparison of volatility measures: A risk management perspective. *Journal of Financial Econometrics*, **8**, 29–56.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High–Dimensional Data: Methods, Theory and Applications*. Springer, New York.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157–172.

Davis, R. A., Zang, P., and Zheng, T. (2015). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*. available online.

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, **146**, 318–328.

Dedecker, J., Doukhan, P., Lang, G., León, J., Louhichi, S., and Prieur, C. (2007). *Weak Dependence: With Examples and Applications*. Springer, New York.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

Den Haan, W. J. and Levin, A. (1996). Inferences from parametrics and non–parametric covariance matrix estimation procedures. NBER Technical Working Paper 195.

Diebold, F. X. and Yılmaz, K. (2009). Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, **119**, 158–171.

Diebold, F. X. and Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, **182**, 119–134.

Diebold, F. X. and Yılmaz, K. (2015). *Financial and Macroeconomic Connectedness: A Network Approach to Measurement and Monitoring*. Oxford University Press.

Doukhan, P. and Neumann, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, **117**, 878–903.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, **137**, 334–353.

Engle, R. F., Gallo, G. M., and Velucchi, M. (2012). Volatility spillovers in East Asian financial markets: A MEM-based approach. *The Review of Economics and Statistics*, **94**, 222–223.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928–961.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Dynamic Factor Model: Identification and estimation. *The Review of Economics and Statistics*, **82**, 540–554.

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

Fu, W. J. (1998). Penalized regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Ghysels, E. (2014). Factor analysis with large panels of volatility proxies. CEPR Discussion Paper No. DP10034.

Härdle, W. K., Wang, W., and Yu, L. (2016). Tenet: Tail-event driven network risk. *Journal of Econometrics*, **forthcoming**.

Hautsch, N., Schaumburg, J., and Schienle, M. (2014a). Financial network systemic risk contributions. *Review of Finance*. available online.

Hautsch, N., Schaumburg, J., and Schienle, M. (2014b). Forecasting systemic impact in financial networks. *International Journal of Forecasting*, **30**, 781–794.

Kock, A. B. (2012). On the oracle property of the adaptive lasso in stationary and nonstationary autoregressions. CREATES Research Papers 2012-05.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. *Econometric Theory*, **32**, 243–259.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, **186**, 325–344.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large–dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Luciani, M. and Veredas, D. (2015). Estimating and forecasting large panels of volatilities with approximate dynamic factor models. *Journal of Forecasting*, **34**, 163–176.

Medeiros, M. C. and Mendes, E. F. (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, **191**, 255–271.

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

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, 703–708.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, **53**, 61–65.

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

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, **20**, 147–162.

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

Table 1: Simulation study

	FDR=1%		FDR=5%		FDR=10%		MSE
T	TPR	MSE	TPR	MSE	TPR	MSE	pre–estimator
250	0.49	0.32	0.55	0.31	0.60	0.34	5.33
500	0.75	0.10	0.84	0.17	0.88	0.25	1.45
750	0.79	0.06	0.89	0.11	0.92	0.15	0.81
1000	0.82	0.05	0.93	0.09	0.96	0.15	0.55

The table reports the results of the simulation exercise for different values of the sample size T. The table reports the True Positive Rate (TPR) and the MSE of the nets estimator when the False Discovery Rate (FDR) is controlled at the 1%, 5% and 10% levels. The last column of the table shows the MSE of the pre–estimator.

Table 2: U.S. Bluechips

Ticker	Company Name	Sector	Ticker	Company Name	Sector
AMZN	Amazon.com	Cons. Disc.	ABT	Abbott Laboratories	Health Care
CMCSA	Comcast	Cons. Disc.	AMGN	Amgen	Health Care
DIS	Walt Disney	Cons. Disc.	BAX	Baxter International	Health Care
F	Ford Motor	Cons. Disc.	BMY	Bristol-Myers Squibb	Health Care
FOXA	Twenty-First Century Fox	Cons. Disc.	GILD	Gilead Sciences	Health Care
HD	Home Depot	Cons. Disc.	JNJ	Johnson & Johnson	Health Care
LOW	Lowes	Cons. Disc.	LLY	Lilly (Eli) & Co.	Health Care
MCD	McDonalds	Cons. Disc.	MDT	Medtronic	Health Care
NKE	NIKE	Cons. Disc.	MRK	Merck & Co.	Health Care
SBUX	Starbucks	Cons. Disc.	PFE	Pfizer	Health Care
TGT	Target	Cons. Disc.	UNH	United Health	Health Care
TWX	Time Warner	Cons. Disc.	BA	Boeing Company	Industrials
CL	Colgate-Palmolive	Cons. Stap.	CAT	Caterpillar	Industrials
COST	Costco	Cons. Stap.	EMR	Emerson Electric	Industrials
CVS	CVS Caremark	Cons. Stap.	FDX	FedEx	Industrials
KO	The Coca Cola Company	Cons. Stap.	GD	General Dynamics	Industrials
MDLZ	Mondelez International	Cons. Stap.	GE	General Electric	Industrials
MO	Altria	Cons. Stap.	HON	Honeywell Intl	Industrials
PEP	PepsiCo	Cons. Stap.	LMT	Lockheed Martin	Industrials
PG	Procter & Gamble	Cons. Stap.	MMM	3M Company	Industrials
WMT	Wal-Mart Stores	Cons. Stap.	NSC	Norfolk Southern	Industrials
APA	Apache	Energy	RTN	Raytheon	Industrials
APC	Anadarko Petroleum	Energy	UNP	Union Pacific	Industrials
COP	ConocoPhillips	Energy	UPS	United Parcel Service	Industrials
CVX	Chevron	Energy	UTX	United Technologies	Industrials
DVN	Devon Energy	Energy	AAPL	Apple	Technology
HAL	Halliburton	Energy	ACN	Accenture plc	Technology
NOV	National Oilwell Varco	Energy	CSCO	Cisco Systems	Technology
OXY	Occidental Petroleum	Energy	EBAY	eBay	Technology
SLB	Schlumberger Ltd.	Energy	EMC	EMC	Technology
XOM	Exxon Mobil	Energy	HPQ	Hewlett-Packard	Technology
AIG	AIG	Financials	IBM	IBM	Technology
ALL	Allstate	Financials	INTC	Intel	Technology
AXP	American Express Co	Financials	MSFT	Microsoft	Technology
BAC	Bank of America	Financials	ORCL	Oracle	Technology
BK	Bank of New York	Financials	QCOM	QUALCOMM	Technology
C	Citigroup	Financials	TXN	Texas Instruments	Technology
COF	Capital One Financial	Financials	T	AT&T	Technology
GS	Goldman Sachs	Financials	VZ	Verizon	Technology
JPM	JPMorgan Chase	Financials	DD	Du Pont	Materials
MET	MetLife	Financials	DOW	Dow Chemical	Materials
MS	Morgan Stanley	Financials	FCX	Freeport-McMoran	Materials
SPG	Simon Property	Financials	MON	Monsanto	Materials
USB	U.S. Bancorp	Financials	AEP	American Electric Power	Utilities
WFC	Wells Fargo	Financials	EXC	Exelon	Utilities

The table reports the list of tickers, company names and industry sectors.

Table 3: Descriptive Stats

	Disc	Stap	Energy	Fin	Heal	Ind	Tech	Mat	Util	All
variance	0.064	0.049	0.042	0.064	0.059	0.050	0.059	0.064	0.041	0.056
kurtosis	4.401	5.434	4.981	4.929	4.993	4.676	4.467	4.413	5.740	4.807
ρ_1	0.260	0.231	0.206	0.309	0.254	0.222	0.257	0.320	0.193	0.253
ρ_5	0.170	0.137	0.144	0.241	0.156	0.133	0.163	0.237	0.120	0.168
ρ_{22}	0.142	0.104	0.123	0.183	0.118	0.109	0.120	0.202	0.075	0.132
$\rho_{0,\mathsf{others}}$	0.091	0.089	0.063	0.077	0.087	0.098	0.080	0.073	0.069	0.083
$\rho_{1,\mathsf{others}}$	0.045	0.048	0.024	0.028	0.046	0.045	0.042	0.034	0.034	0.039

The table reports average descriptive statistics over the industry sectors and the entire panel. The set of descriptive statistics considered contains the sample variance, kurtosis, autocorrelation of order 1, 5 and 22, the average contemporaneous correlation with all other tickers, and the average order 1 autocorrelation with all other tickers.

Table 4: Network Estimation Summary

	Disc	Stap	Ener	Fin	Heal	Ind	Tech	Mat	Util	All
Granger Links	47	15	19	41	34	39	35	16	5	251
Contemporaneous Links	33	20	29	71	30	46	51	11	3	294
`nets` R^2_{is}	22.5	18.6	18.9	32.2	18.9	24.4	20.0	19.5	11.0	22.3
factor R^2_{is}	45.3	37.2	42.7	56.6	33.6	51.3	39.1	45.9	36.8	44.4
sector R^2_{is}	7.3	9.6	25.3	16.1	9.0	5.8	8.4	10.6	25.8	11.6

The table reports summary estimation results over the industry sectors and the entire panel. The first row of the table reports the outer degree of the Granger network, the second row reports the degree of the contemporaneous network, the third row reports the (in–sample) average R^2_{is} of the `nets` regression. For comparison, the table reports in the fourth and fifth rows the average (in–sample) factor R^2_{is} and (in–sample) sector R^2_{is}, respectively.

Table 5: Sectoral Linkages

	Granger Component								
	Disc	Stap	Ener	Fin	Heal	Ind	Tech	Mat	Util
Disc	33.7	22.5	6.5	15.0	14.3	16.9	13.3	10.0	22.2
Stap	4.8	22.5	2.2	5.3	6.3	7.2	1.3	10.0	0.0
Ener	8.4	7.5	28.3	7.1	4.8	7.2	4.0	3.3	0.0
Fin	15.7	10.0	15.2	33.6	14.3	15.7	14.7	20.0	0.0
Heal	10.8	7.5	10.9	9.7	31.7	6.0	20.0	6.7	11.1
Ind	6.0	12.5	10.9	10.6	7.9	27.7	9.3	10.0	33.3
Tech	12.0	12.5	15.2	10.6	14.3	15.7	30.7	16.7	11.1
Mat	8.4	5.0	6.5	8.0	3.2	2.4	5.3	23.3	0.0
Util	0.0	0.0	4.3	0.0	3.2	1.2	1.3	0.0	22.2
	Contemporaneous Component								
	Disc	Stap	Ener	Fin	Heal	Ind	Tech	Mat	Util
Disc	5.8	18.2	14.5	13.9	13.5	18.8	14.7	14.9	10.3
Stap	13.0	6.1	8.5	6.7	13.5	9.2	11.3	10.3	13.8
Ener	8.2	6.8	8.5	13.0	5.7	4.4	8.0	10.3	6.9
Fin	15.0	10.1	24.8	6.3	15.1	18.8	18.5	16.1	13.8
Heal	12.6	17.6	9.4	13.0	5.7	14.8	16.8	11.5	17.2
Ind	20.8	14.2	8.5	19.3	17.7	6.1	18.9	17.2	13.8
Tech	16.9	18.2	16.2	19.7	20.8	19.7	5.9	12.6	10.3
Mat	6.3	6.1	7.7	6.3	5.2	6.6	4.6	4.6	6.9
Util	1.4	2.7	1.7	1.8	2.6	1.7	1.3	2.3	6.9

The table reports the fraction of Granger and contemporaneous linkages between the industrial sectors. The (i, j) entry of the Granger linkages table is defined as the total number of linkages for sector j to sector i standardized by the total number of linkages from sector j. The (i, j) entry of the contemporaneous linkages table is defined as the total number of linkages between sector i and j standardized by the total number of linkages of sector j.

Table 6: Rankings

	Granger		Contemporaneous	
Rank	Company	Sector	Company	Sector
1	BAC	Financials	UNP	Industrials
2	AIG	Financials	T	Technology
3	C	Financials	USB	Financials
4	ALL	Financials	GE	Industrials
5	MCD	Discretionary	TGT	Discretionary
6	HPQ	Technology	WFC	Financials
7	DOW	Material	MS	Financials
8	SPG	Financials	NSC	Industrials
9	GE	Industrials	F	Discretionary
10	CVS	Staples	NOV	Energy

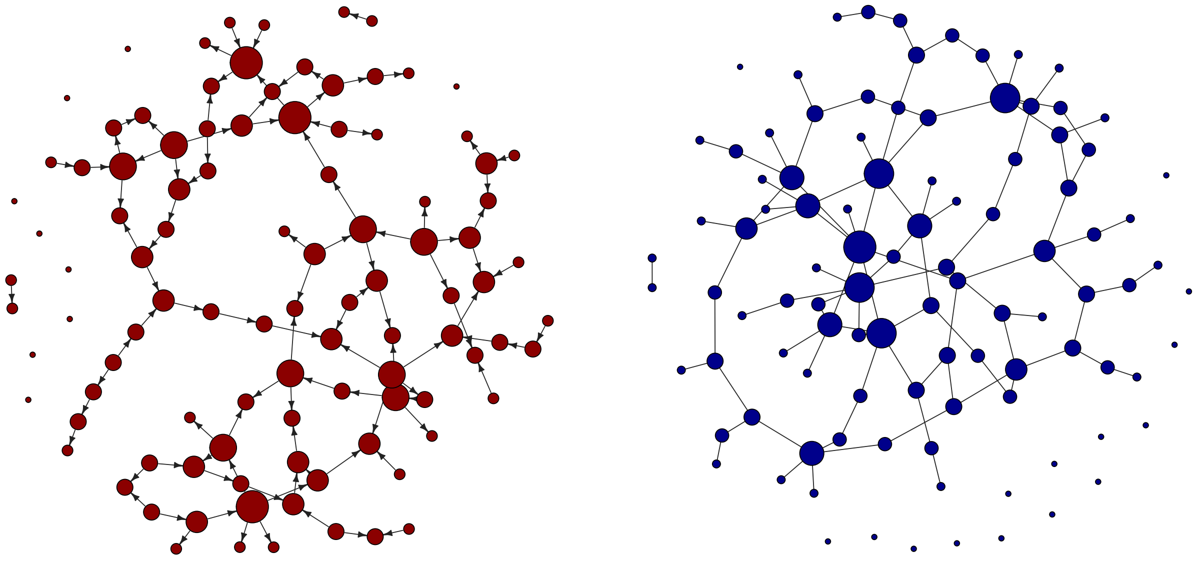
The table reports the top ten of the most interconnected series in the Granger and contemporaneous networks.

Table 7: Forecasting

	Granger Component									
	Disc	Stap	Energy	Fin	Heal	Ind	Tech	Mat	Util	All
benchmark MSE	5.67	3.96	4.52	3.90	5.15	4.32	5.86	8.16	3.14	4.91
`nets` R^2_{oos}	10.21	6.25	5.30	5.86	6.98	5.46	9.49	17.90	0.50	8.08
AR(1) R^2_{oos}	5.65	3.58	5.04	1.65	2.95	4.07	6.30	15.14	-3.49	5.06
ridge R^2_{oos}	5.47	1.79	2.74	-8.46	1.09	3.39	4.89	12.84	-8.20	2.57
	Contemporaneous Component									
	Disc	Stap	Energy	Fin	Heal	Ind	Tech	Mat	Util	All
benchmark MSE	5.12	3.74	4.30	3.66	4.82	4.11	5.35	6.78	3.11	4.54
`nets` R^2_{oos}	15.76	12.56	12.98	17.21	6.28	19.08	14.03	3.00	-1.17	13.20
reg R^2_{oos}	12.89	9.12	9.02	15.23	2.97	17.44	10.98	-1.94	-6.44	10.19
ridge R^2_{oos}	12.97	9.20	9.05	15.32	3.06	17.50	11.04	-1.87	-6.37	10.26

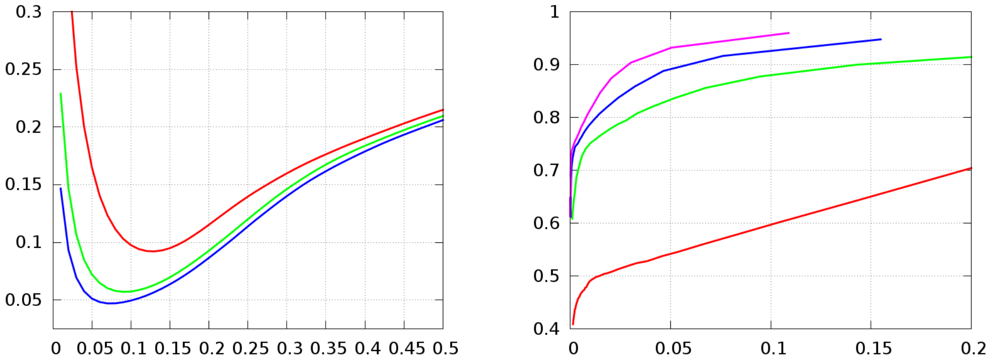
The table reports summary forecasting results over the industry sectors and the entire panel. The first panel reports forecasting exercise for the Granger component. The first row reports the MSE (times 100) of the benchmark. The second, third and fourth rows report the out–of–sample R^2 of, respectively, `nets`, an AR(1) estimated by least squares and a VAR estimated using ridge regression. The second panel reports the forecasting exercise of the contemporaneous component. The first row reports the MSE (times 100) of the benchmark. The second, third and fourth row report the out–of–sample R^2 of, respectively, `nets`, the linear regression estimator and the ridge estimator.

Figure 1: Simulated Granger and Contemporaneous Networks



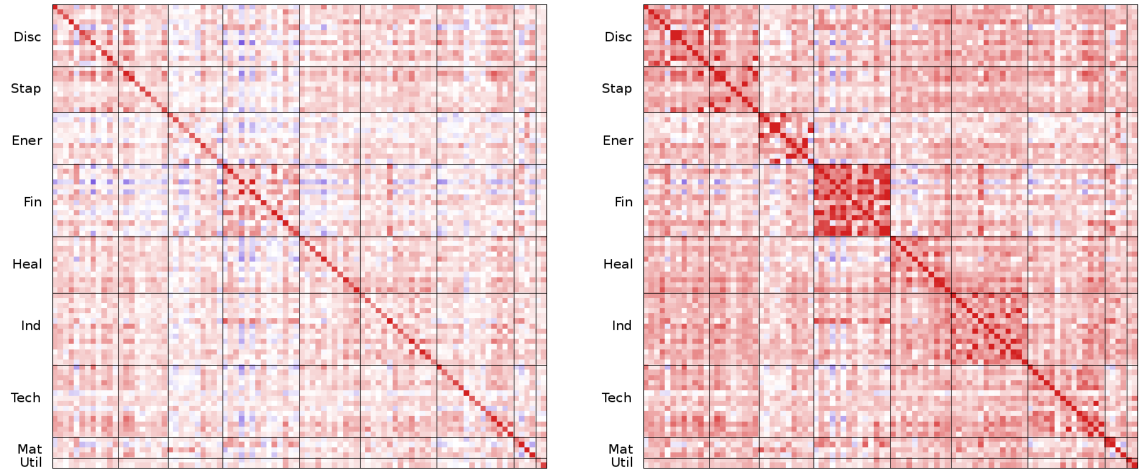
The figure displays realizations of the Erdös–Renyi random graph models used in the simulation exercise. The left picture displays a directed Erdös–Renyi graph used to generate the autoregressive matrix **A** while the right picture displays an undirected Erdös–Renyi random graph used to generate the contemporaneous concentration matrix **C**.

Figure 2: Simulation Study



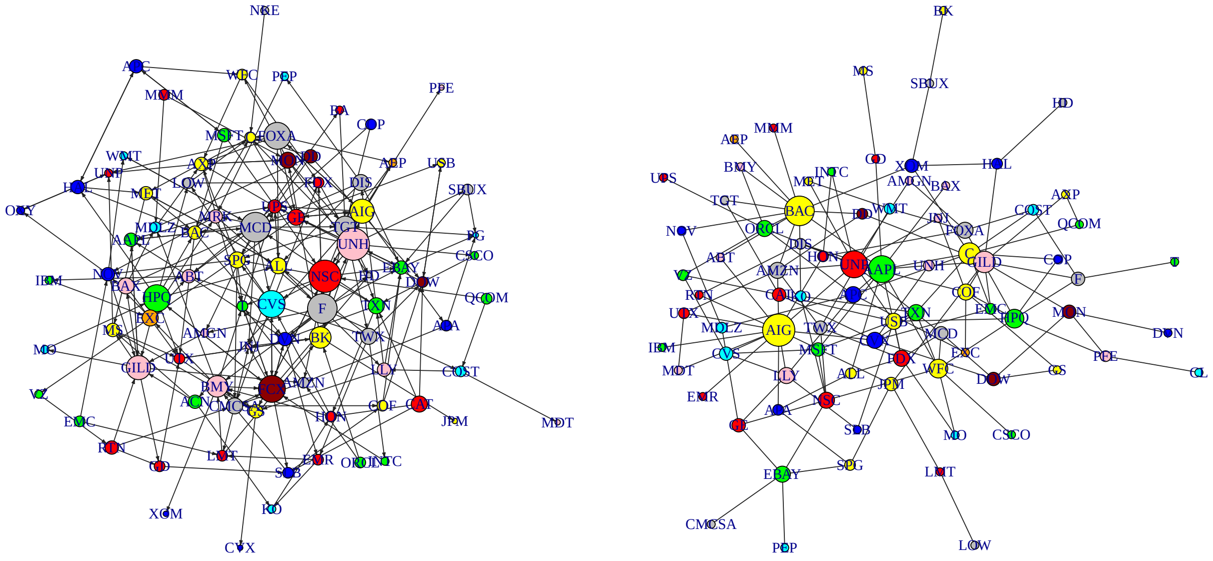
The left picture displays the MSE (multiplied by 100) of the nets estimator as a function of the tuning parameter λ_T for (from top to bottom) $T = 500, 750, 1000$. The right picture displays ROC curve of the nets estimator for (from bottom to top) $T = 250, 500, 750, 1000$.

Figure 3: Autocorrelation and Correlation Heatmaps



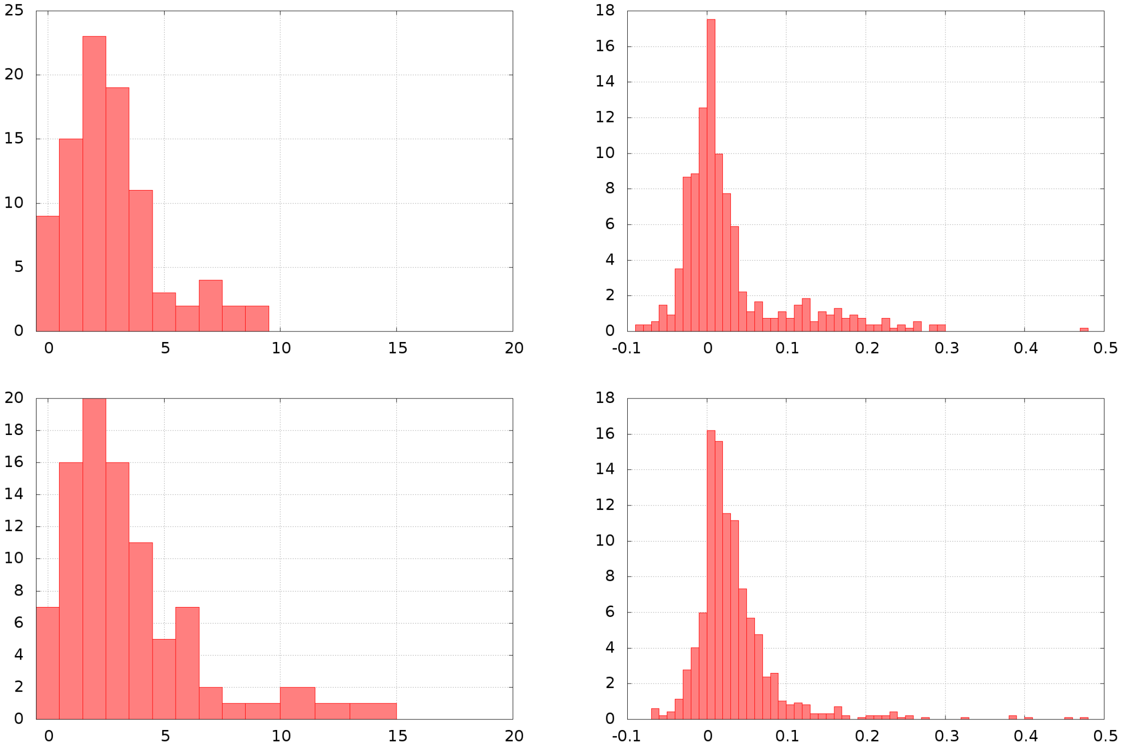
The figure displays the heatmap of the sample autocorrelation (left) and sample correlation matrices of the residuals of regression (16).

Figure 4: S&P 100 Volatility Granger and Contemporaneous Networks



The figure displays the estimated Granger and contemporaneous networks. The size of the vertices is proportional to their degree and the colour of the vertices depends on their industry sector.

Figure 5: Degree and Coefficient Distributions



The top left and bottom left pictures display the histograms of the degree distribution of the Granger and contemporaneous networks, respectively. The top right and bottom right pictures display the histograms of the estimated non–zero $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ coefficients, respectively.

A Technical appendix

A.1 Preliminary Definitions

Estimation is conditional on a given value of $\mathbf{c} = (c_{11} \dots c_{nn})'$. We define the dimension of the parameters' space as $m = n^2p + n(n-1)/2$. We collect the parameters of interest in (11) into the $m \times 1$ vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\rho}')'$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_{11} \dots \boldsymbol{\alpha}'_{1p} \dots \boldsymbol{\alpha}'_{n1} \dots \boldsymbol{\alpha}'_{np})'$ and $\boldsymbol{\alpha}'_{ik} = (\alpha_{i1k} \dots \alpha_{ink})$ is the i-th row of the VAR matrix $\mathbf{A}_k$ with $k = 1, \dots, p$. The $n(n-1)/2 \times 1$ vector $\boldsymbol{\rho}$ contains the stacked partial correlations of the VAR innovations. Similarly the parameters in (10) are collected into the $m \times 1$ vector $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{\rho}')'$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{11} \dots \boldsymbol{\beta}'_{1p} \dots \boldsymbol{\beta}'_{n1} \dots \boldsymbol{\beta}'_{np})'$ and $\boldsymbol{\beta}'_{ik} = (\beta_{i1k} \dots \beta_{ink})$ for $i = 1, \dots, n$ and $k = 1, \dots, p$. Define as $\boldsymbol{\theta}_0$, $\boldsymbol{\phi}_0$, and $\mathbf{c}_0$ the true values of the parameters.

With reference to the minimisation problem in (14)-(15), recall that the adaptive LASSO weights are defined as $w_i = C_\bullet / |\widetilde{\theta}_{T\,i}|$, with $\lambda_T^{\mathbf{G}} = C_\alpha \lambda_T$ and $\lambda_T^{\mathbf{C}} = C_\rho \lambda_T$. Hereafter, for simplicity and without loss of generality we assume that $C_\alpha = C_\rho = 1$. Moreover, we define the sample score and Hessian of the unconstrained problem as

$$\mathbf{S}_T(\boldsymbol{\theta}, \mathbf{c}) = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c}), \qquad \mathbf{H}_T(\boldsymbol{\theta}, \mathbf{c}) = \frac{1}{T} \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c}).$$

where $\nabla_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}$ and $\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$, and $\ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c})$ is the unconstrained loss function defined in (12). The population counterparts of the above are defined as

$$\mathbf{S}_0(\boldsymbol{\theta}, \mathbf{c}) = \mathsf{E}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c})], \qquad \mathbf{H}_0(\boldsymbol{\theta}, \mathbf{c}) = \mathsf{E}[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c})].$$

For a given symmetric matrix $\mathbf{A}$ we denote by $\mu_{\min}(\mathbf{A})$ and $\mu_{\max}(\mathbf{A})$ its smallest and largest eigenvalues respectively. For a generic matrix $\mathbf{B}$, the notation $\|\mathbf{B}\| = \sqrt{\mu_{\max}(\mathbf{B}\mathbf{B}')}$ is used for spectral norm. For a generic vector $\mathbf{b}$, the notation $\|\mathbf{b}\| = \sqrt{\sum_i b_i^2}$ indicates the Euclidean norm and $\|\mathbf{b}\|_\infty = \max_i |b_i|$.

In what follows we use the symbol K to denote a generic positive constant. The value of K needs not to be the same from line to line. When more than one distinct constant are present in the same equation we denote them by $K_0, K_1, K_2, \dots.$ The symbols $\kappa_0, \kappa_1, \kappa_2, \dots$ denote universal constants that are unique throughout the paper.

A.2 Assumptions

ASSUMPTION 1. *The n-dimensional random vector process $\mathbf{y}_t$ is purely non-deterministic, has zero-mean, and follows the* VAR*(p)* $\mathbf{y}_t = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{y}_{t-k} + \boldsymbol{\epsilon}_t$*, where* $\boldsymbol{\epsilon}_t \sim i.i.d.(\mathbf{0}, \mathbf{C}^{-1})$*. Moreover,*

(a) $\det(\mathbf{I} - \sum_{k=1}^{p} \mathbf{A}_k z^k) \neq 0$ *for any* $|z| \leq 1$*;*

(b) there exists a constant $c > 0$ *such that* $\mathsf{E}[|y_{it}|^k] \leq k! c^{k-2} \mathsf{E}[y_{it}^2] < \infty$*, for any* $i = 1, \dots, n$*,* $t = 1, \dots, T$*,* $k = 3, 4, \dots$*;*

(c) there exist couples of constants $\underline{M}_0, \overline{M}_0$ *and* $\underline{M}_1, \overline{M}_1$ *such that*

$$0 < \underline{M}_0 \leq \mu_{\min}\left(\boldsymbol{\Sigma}(\omega)\right) \leq \mu_{\max}\left(\boldsymbol{\Sigma}(\omega)\right) \leq \overline{M}_0 < \infty, \qquad 0 < \underline{M}_1 \leq \mu_{\min}(\mathbf{C}_0) \leq \mu_{\max}(\mathbf{C}_0) \leq \overline{M}_1 < \infty.$$

where $\boldsymbol{\Sigma}(\omega)$ *is the spectral density matrix of* $\mathbf{y}_t$*, defined for* $\omega \in [-\pi, \pi]$*.*

ASSUMPTION 2. *Define the pre-estimators* $\widehat{\mathbf{c}}_T$*,* $\widetilde{\boldsymbol{\theta}}_T = (\widetilde{\boldsymbol{\alpha}}_T, \widetilde{\boldsymbol{\rho}}_T)$*. Then, for T sufficiently large there exist constants* $C_1, C_2 > 0$ *such that, for any* $\eta > 0$*, with probability at least* $1 - O(T^{-\eta})$*, we have*

$$\max_{1 \leq i \leq n} |\widehat{c}_{T\,ii} - c_{0\,ii}| \leq C_1 \sqrt{\frac{\log T}{T}}, \qquad \max_{1 \leq i \leq m} |\widetilde{\theta}_{T\,i} - \theta_{0\,i}| \leq C_2 \sqrt{\frac{\log T}{T}}.$$

A.3 Lemmas and Conditions

LEMMA 1. *The parameters* $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$ *are related by means of the equations*

$$\beta_{ijk} = \alpha_{ijk} - \sum_{\substack{l=1 \\ l \neq i}}^{n} \rho^{il} \sqrt{\frac{c_{ll}}{c_{ii}}} \alpha_{ljk}, \qquad \gamma_{ij} = \rho^{ij} \sqrt{\frac{c_{jj}}{c_{ii}}}, \qquad i, j = 1, \dots, n, \quad k = 1, \dots, p.$$

Moreover, the error terms in (9) *and* (10) *coincide, that is,* $e_{i\,t} = u_{i\,t}$*.*

LEMMA 2. *For a given value of* $\mathbf{c}$*, define the* $n^2 \times n^2$ *matrix* $\mathbf{M}(\boldsymbol{\rho}; \mathbf{c}) = (\operatorname{diag} \mathbf{C})^{-1} \mathbf{C} \otimes \mathbf{I}_n$*. Then,*

$$\underbrace{\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\rho} \end{pmatrix}}_{\boldsymbol{\phi}} = \begin{pmatrix} \mathbf{M}(\boldsymbol{\rho}; \mathbf{c}) & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{M}(\boldsymbol{\rho}; \mathbf{c}) & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_{n(n-1)/2} \end{pmatrix} \underbrace{\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\rho} \end{pmatrix}}_{\boldsymbol{\theta}}.$$

LEMMA 3. *Consider the mapping* $\mathbf{g}_{\mathbf{c}} : \mathbb{R}^m \to \mathbb{R}^m$*, such that* $\mathbf{g}_{\mathbf{c}}(\boldsymbol{\theta}) = \boldsymbol{\phi}$*. Then, under Assumption 1, there exists a function* $\mathbf{h}_{\mathbf{c}_0} : \mathbb{R}^m \to \mathbb{R}^m$ *such that:* $\mathbf{h}_{\mathbf{c}_0}(\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)) = \boldsymbol{\theta}_0$*, that is* $\mathbf{g}_{\mathbf{c}_0}$ *is invertible in* $\boldsymbol{\theta}_0$*.*

Condition 1. *Under Assumption 1, there exist constants $\underline{L}, \overline{L}$ such that*

$$0 < \underline{L} \le \mu_{\min}(\mathbf{H}_0(\boldsymbol{\theta}_0, \mathbf{c}_0)) \le \mu_{\max}(\mathbf{H}_0(\boldsymbol{\theta}_0, \mathbf{c}_0)) \le \overline{L} < \infty.$$

Condition 2. *Under Assumption 2, for T sufficiently large there exist constants $C_3, C_4 > 0$ such that, for any $\eta > 0$, with probability at least $1 - O(T^{-\eta})$, we have*

$$\max_{1\le i\le m} \left|\mathbf{S}_{T\,i}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{S}_{T\,i}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\right| \le C_3\sqrt{\frac{\log T}{T}},$$
$$\max_{1\le i,j\le m} \left|\mathbf{H}_{T\,ij}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{H}_{T\,ij}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\right| \le C_4\sqrt{\frac{\log T}{T}}.$$

Lemma 4. *Under Assumptions 1 and 2 and the same conditions as in Proposition 2, for T sufficiently large there exist constants $\kappa_0, \kappa_1, \kappa_2, \kappa_3 > 0$ such that, for any $\eta > 0$ and any $\mathbf{u}$ in $\mathbb{R}^{q_T}$, with probability at least $1 - O(T^{-\eta})$, we have*

$$\begin{aligned}
&\text{(a)} \quad \left\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\right\| \le \kappa_0 \sqrt{q_T}\sqrt{\frac{\log T}{T}};\\
&\text{(b)} \quad \left|\mathbf{u}'\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\right| \le \kappa_1 \|\mathbf{u}\| \sqrt{q_T}\sqrt{\frac{\log T}{T}};\\
&\text{(c)} \quad \left\|\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\mathbf{u} - \mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u}\right\| \le \kappa_2 \|\mathbf{u}\|\, q_T\sqrt{\frac{\log T}{T}};\\
&\text{(d)} \quad \left|\mathbf{u}'\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\mathbf{u} - \mathbf{u}'\mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u}\right| \le \kappa_3 \|\mathbf{u}\|^2 q_T\sqrt{\frac{\log T}{T}}.
\end{aligned}$$

Lemma 5. *For any subset $\mathcal{S} \subseteq \mathcal{A} \cup \mathcal{A}^c$, we have $\widehat{\boldsymbol{\theta}}_T^{\mathcal{S}} = \operatorname{argmin}_{\boldsymbol{\theta}:\boldsymbol{\theta}_{\mathcal{S}}^c=\mathbf{0}} \mathcal{L}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T)$, where $\mathcal{L}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T)$ is defined in* (15)*, if and only if the i-th component of the sample score satisfies*

$$\begin{aligned}
S_{Ti}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{S}}, \widehat{\mathbf{c}}_T) &= -\frac{\lambda_T}{|\widetilde{\theta}_{Ti}|}\operatorname{sign}(\widehat{\theta}_{Ti}^{\mathcal{S}}), &&\text{if } \widehat{\theta}_{Ti}^{\mathcal{S}} \neq 0,\\
|S_{Ti}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{S}}, \widehat{\mathbf{c}}_T)| &\le \frac{\lambda_T}{|\widetilde{\theta}_{Ti}|}, &&\text{if } \widehat{\theta}_{Ti}^{\mathcal{S}} = 0.
\end{aligned}$$

If the solution is not unique then $|S_{Ti}(\overline{\boldsymbol{\theta}}_T^{\mathcal{S}}, \widehat{\mathbf{c}}_T)| \le \lambda_T(|\widetilde{\theta}_i|)^{-1}$ for some specific solution $\overline{\boldsymbol{\theta}}_T^{\mathcal{S}}$, then since $S_{Ti}(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T)$ is continuous in $\boldsymbol{\theta}$, then $\widehat{\theta}_i = 0$ for all solutions $\widehat{\boldsymbol{\theta}}$. Hence, if $\mathcal{S} = \mathcal{A} \cup \mathcal{A}^c$, we have the unconstrained optimisation and $\widehat{\boldsymbol{\theta}}_T^{\mathcal{S}} = \widehat{\boldsymbol{\theta}}_T$, while if $\mathcal{S} = \mathcal{A}$, we have the restricted optimisation and $\widehat{\boldsymbol{\theta}}_T^{\mathcal{S}} = \widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}$.

Lemma 6. *Under Assumptions 1 and 2 and the same conditions in Proposition 2, there exists a constant*

$\kappa_4 > 0$ *such that, for* T *sufficiently large and any* $\eta > 0$,

$$\Pr\Big(\exists\, \boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}^c_{\mathcal{A}}=\mathbf{0}}{\arg\min}\, \mathcal{L}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T) : \boldsymbol{\theta}^* \in D(\boldsymbol{\theta}_0)\Big) \geq 1 - O(T^{-\eta}),$$

where $D(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \kappa_4\sqrt{q_T}\lambda_T\}$.

Lemma 7. *Under Assumptions 1 and 2 and under the same conditions as in Proposition 2, there exists a constant* $\kappa_5 > 0$ *such that, for* T *sufficiently large and any* $\eta > 0$

$$\Pr\left(\|\mathbf{S}_T(\boldsymbol{\theta}, \widehat{\mathbf{c}}_T)\| > \sqrt{q_T}\frac{\lambda_T}{\min_{i\in\mathcal{A}}|\theta_{0i}|}\right) \geq 1 - O(T^{-\eta}),$$

for any $\boldsymbol{\theta} \in S(\boldsymbol{\theta}_0)$ *where* $S(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \kappa_5\sqrt{q_T}\lambda_T,\, \boldsymbol{\theta}_{\mathcal{A}^c} = \mathbf{0}\}$.

Lemma 8. *Under Assumptions 1 and 2 and under the same conditions as in Proposition 3, and if* $n = O(T^\zeta)$ *for some* $\zeta > 0$*, then for* T *sufficiently large and any* $\eta > 0$

$$\Pr\left(\max_{j\in\mathcal{A}^c}|S_{T\,j}(\widehat{\boldsymbol{\theta}}^{\mathcal{A}}_T, \widehat{\mathbf{c}}_T)| \leq \frac{\lambda_T}{\max_{j\in\mathcal{A}^c}|\widetilde{\theta}_{Tj}|}\right) \geq 1 - O(T^{-\eta}).$$

A.4 Proofs of Propositions

Proof of proposition 1. Notice that the loss related to (10) is given by

$$\ell(\boldsymbol{\phi}_0; \mathbf{y}_t, \mathbf{c}_0) = \sum_{i=1}^n \left(y_{i\,t} - \sum_{k=1}^p\sum_{j=1}^n \beta_{ijk}\, y_{j\,t-k} - \sum_{\substack{h=1\\h\neq i}}^n \rho^{ih}\sqrt{\frac{c_{0,hh}}{c_{0,ii}}}y_{h\,t}\right)^2. \tag{A-1}$$

Clearly $\boldsymbol{\phi}_0$ is a minimizer of (A-1) (using Assumption 1 for second order conditions):

$$\boldsymbol{\phi}_0 = \arg\min_{\boldsymbol{\phi}} \mathsf{E}[\ell(\boldsymbol{\phi}; \mathbf{y}_t, \mathbf{c}_0)]. \tag{A-2}$$

In order for $\boldsymbol{\theta}_0$ to be a minimum, we need to verify that first and second order conditions hold. The first order conditions are given by[6]

$$\mathsf{E}[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_0; \mathbf{y}_t, \mathbf{c}_0)] = \mathsf{E}[\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_0; \mathbf{y}_t, \mathbf{c}_0)\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)] = \mathsf{E}[\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_0; \mathbf{y}_t, \mathbf{c}_0)]\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0) = \mathbf{0}, \tag{A-3}$$

[6] Notice that we can exchange integral and differentiation operators as the loss function is such that $\ell \in \mathcal{C}^\infty(\mathbb{R}^m)$.

since $\mathsf{E}[\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_0;\mathbf{y}_t,\mathbf{c}_0)] = \mathbf{0}$ because of (A-2). The second order conditions are

$$\begin{aligned}\mathsf{E}[\nabla_{\boldsymbol{\theta\theta}}\ell(\boldsymbol{\theta}_0;\mathbf{y}_t,\mathbf{c}_0)] &= \mathsf{E}[\nabla_{\boldsymbol{\theta\phi}}\ell(\boldsymbol{\phi}_0;\mathbf{y}_t,\mathbf{c}_0)]\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0) + \mathsf{E}[\nabla_{\boldsymbol{\phi}}\ell(\boldsymbol{\phi}_0;\mathbf{y}_t,\mathbf{c}_0)]\nabla_{\boldsymbol{\theta\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)\\ &= \mathsf{E}[\nabla_{\boldsymbol{\phi\phi}}\ell(\boldsymbol{\phi}_0;\mathbf{y}_t,\mathbf{c}_0)]\left(\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)\right)^2, \qquad \text{(A-4)}\end{aligned}$$

which we used (A-3). Now, (A-4) is positive definite since the first term is positive definite because of (A-2) and the second term is positive definite because of Lemma 3. □

Proof of proposition 2. From Lemma 5, we have

$$\|\mathbf{S}_{T\,\mathcal{A}}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}})\|_\infty \le \lambda_T \max_{i\in\mathcal{A}} \frac{1}{|\widetilde{\theta}_{Ti}|}.$$

Moreover, for any $i\in\mathcal{A}$,

$$\frac{1}{|\widetilde{\theta}_{Ti}|} = \sqrt{\frac{1}{\widetilde{\theta}_{Ti}^2}} \le \frac{1}{|\theta_{0i}|} + \sqrt{\frac{2}{\theta_{0i}^3}|\widetilde{\theta}_{Ti}-\theta_{0i}| + o(|\widetilde{\theta}_{Ti}-\theta_{0i}|)}. \qquad \text{(A-5)}$$

Define $\theta_0^* = \min_{i\in\mathcal{A}}|\theta_{0i}|$ and notice that $\theta_0^*>0$ and define also $\nu_T = \sqrt{q_T}\lambda_T$, therefore $\nu_T\to 0$ as $T\to\infty$. Using Assumption 2 and (A-5), there exists a constant $K>0$ such that for T sufficiently large and for any $\eta>0$, we have with probability at least $1-O(T^{-\eta})$

$$\begin{aligned}\|\mathbf{S}_{T\,\mathcal{A}}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}})\| \le \sqrt{q_T}\;\|\mathbf{S}_{T\,\mathcal{A}}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}})\|_\infty &\le \nu_T \max_{i\in\mathcal{A}}\frac{1}{|\widetilde{\theta}_{Ti}|}\\ &\le \nu_T\left[\max_{i\in\mathcal{A}}\frac{1}{|\theta_{0i}|} + K\left(\frac{\log T}{T}\right)^{1/4}\right]\\ &\le \frac{\nu_T}{\theta_0^*} + \nu_T K\left(\frac{\log T}{T}\right)^{1/4}. \qquad \text{(A-6)}\end{aligned}$$

Notice that the last term on the rhs of (A-6) is $o(\nu_T)$, thus it can be neglected so that for T sufficiently large and for any $\eta>0$, we have

$$\Pr\left(\|\mathbf{S}_{T\,\mathcal{A}}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}})\| \le \frac{\nu_T}{\theta_0^*}\right) \ge 1 - O(T^{-\eta}). \qquad \text{(A-7)}$$

From Lemma 7 we also have that for T sufficiently large and for any $\eta>0$

$$\Pr\left(\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta})\| \le \frac{\nu_T}{\theta_0^*}\right) \ge 1 - O(T^{-\eta}). \qquad \text{(A-8)}$$

for any $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}_{\mathcal{A}^c}=\mathbf{0}$ and $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\kappa_5\nu_T$. Therefore, (A-8) implies that inside a disc of radius $\kappa_5\nu_T$ condition (A-7) is satisfied. In particular, (A-7) is a consequence of the Karush-Kuhn-Tucker condition in

Lemma 5 for $\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}$ to be a minimum. Moreover, by Lemma 6, such minimum always exists in a disc of radius $\kappa_4\nu_T$. Hence, if we define $\kappa_R = \min(\kappa_4, \kappa_5)$, for T sufficiently large and for any $\eta > 0$, we have

$$\Pr\left(\|\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} - \boldsymbol{\theta}_{0\,\mathcal{A}}\| \le \kappa_R\nu_T\right) \ge 1 - O(T^{-\eta}). \tag{A-9}$$

Finally, for any $i \in \mathcal{A}$ and for T sufficiently large, we have $|\theta_{0i}| > s_T > 2\kappa_R\nu_T$. Moreover, for any $i \in \mathcal{A}$, in general we have

$$\Pr\left(\operatorname{sign}(\widehat{\theta}_{T\,i}^{\mathcal{A}}) = \operatorname{sign}(\theta_{0\,i})\right) \ge \Pr\left(\|\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} - \boldsymbol{\theta}_{0\,\mathcal{A}}\| \le \kappa_R\nu_T, |\theta_{0i}| > 2\kappa_R\nu_T\right), \tag{A-10}$$

which by (A-9) implies sign consistency. This completes the proof. □

Proof of proposition 3. (a) By Proposition 2 and Lemma 7 the non-zero coefficients of $\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}$ satisfy the Karush-Kuhn-Tucker condition in Lemma 5. Moreover, by Lemma 8 for T sufficiently large and for any $\eta > 0$ also the zero coefficients satisfy the Karush-Kuhn-Tucker condition with probability at least $1 - O(T^{-\eta})$. Therefore, since with probability at least $1 - O(T^{-\eta})$ the restricted estimator $\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}}$ is also a solution of the unrestricted problem, we proved the existence of a solution of the unrestricted problem. On the other hand, by Lemma 8 and the Karush-Kuhn-Tucker condition in Lemma 5, with probability at least $1 - O(T^{-\eta})$, any solution of the unrestricted problem is a solution of the restricted problem. That is,

$$\Pr\left(\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) \ge 1 - O(T^{-\eta}). \tag{A-11}$$

As a consequence of (A-11), given the unrestricted estimator, $\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}$, for T sufficiently large, for any $\eta > 0$ and for all $j \in \mathcal{A}^c$ we have

$$\begin{aligned}\Pr\left(\widehat{\theta}_{T\,j} = 0\right) &= \Pr\left(\widehat{\theta}_{T\,j} = 0 \,\middle|\, \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) \Pr\left(\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) + \Pr\left(\widehat{\theta}_{T\,j} = 0 \,\middle|\, \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} \neq \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) \Pr\left(\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} \neq \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right)\\ &\ge \Pr\left(\widehat{\theta}_{T\,j} = 0 \,\middle|\, \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) \Pr\left(\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) = \Pr\left(\widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}^{\mathcal{A}} = \widehat{\boldsymbol{\theta}}_{T\,\mathcal{A}}\right) \ge 1 - O(T^{-\eta}).\end{aligned}$$

This proves part (a). Part (b) follows directly from Proposition 2 and (A-11). This completes the proof. □

B Proofs of Complementary Results

Proof of lemma 1. The VAR(p) model (1) has n equations given by

$$y_{i\,t} = \sum_{k=1}^{p}\sum_{j=1}^{n} \alpha_{ijk}\, y_{j\,t-k} + \epsilon_{i\,t}, \quad i = 1, \ldots, n, \tag{A-12}$$

where ϵ_{it} is the i-the element of the vector $\boldsymbol{\epsilon}_t$. Then, by substituting (A-12) in (10), we have, for any $i = 1, \ldots, n$,

$$\begin{aligned}
y_{i\,t} &= \sum_{k=1}^{p}\sum_{j=1}^{n} \beta_{ijk}\, y_{j\,t-k} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\, y_{h\,t} + e_{i\,t}\\
&= \sum_{k=1}^{p}\sum_{j=1}^{n} \beta_{ijk}\, y_{j\,t-k} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih} \left(\sum_{k=1}^{p}\sum_{j=1}^{n} \alpha_{hjk}\, y_{j\,t-k} + \epsilon_{h\,t}\right) + e_{i\,t}\\
&= \sum_{k=1}^{p}\sum_{j=1}^{n} \left(\beta_{ijk} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\alpha_{hjk}\right) y_{j\,t-k} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\, \epsilon_{h\,t} + e_{i\,t}.
\end{aligned} \tag{A-13}$$

By comparing the rhs of (A-13) with (A-12) we have

$$\alpha_{ijk} = \beta_{ijk} + \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\alpha_{hjk}, \quad i, j = 1, \ldots, n, \quad k = 1, \ldots, p, \tag{A-14}$$

$$\epsilon_{i\,t} = \sum_{\substack{h=1\\h\neq i}}^{n} \gamma_{ih}\, \epsilon_{h\,t} + e_{i\,t}, \quad i = 1, \ldots, n. \tag{A-15}$$

and therefore, from (9) we also have $e_{i\,t} = u_{i\,t}$. From (A-15) using Lemma 3 in Peng *et al.* (2009) we have

$$\gamma_{ih} = \rho^{ih}\sqrt{\frac{c_{hh}}{c_{ii}}}, \quad i, h = 1, \ldots, n, \tag{A-16}$$

and clearly when $i = h$, $\gamma_{ih} = \rho^{ih} = 1$. By substituting (A-16) into (A-14) we complete the proof. □

Proof of lemma 2. First define the $n \times n$ matrix $\mathbf{R} = \mathbf{I}_n - (\mathrm{diag}\ \mathbf{C})^{-1/2}\mathbf{C}\ (\mathrm{diag}\ \mathbf{C})^{-1/2}$. Form the definition of partial correlation (4), we see that $\mathbf{R}$ is a matrix with ρ^{ij} as generic (i, j) entry whenever $i \neq j$ and zero otherwise. Now from Lemma 1 we immediately have that, for any $k = 1, \ldots, p$

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{\beta}_{1k} \\ \vdots \\ \boldsymbol{\beta}_{nk} \end{pmatrix} &= \left\{\mathbf{I}_{n^2} - \left[(\mathrm{diag}\ \mathbf{C})^{-1/2}\mathbf{R}(\mathrm{diag}\ \mathbf{C})^{1/2} \otimes \mathbf{I}_n\right]\right\} \begin{pmatrix} \boldsymbol{\alpha}_{1k} \\ \vdots \\ \boldsymbol{\alpha}_{nk} \end{pmatrix}\\
&= \left\{\mathbf{I}_{n^2} - \left[\left(\mathbf{I}_n - (\mathrm{diag}\ \mathbf{C})^{-1}\mathbf{C}\right) \otimes \mathbf{I}_n\right]\right\} \begin{pmatrix} \boldsymbol{\alpha}_{1k} \\ \vdots \\ \boldsymbol{\alpha}_{nk} \end{pmatrix} = \mathbf{M}(\boldsymbol{\rho}; \mathbf{c}) \begin{pmatrix} \boldsymbol{\alpha}_{1k} \\ \vdots \\ \boldsymbol{\alpha}_{nk} \end{pmatrix}.
\end{aligned} \tag{A-17}$$

The statement of the lemma follow straightforwardly from (A-17). □

Proof of lemma 3. From Lemma 2 we have

$$\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)=\boldsymbol{\phi}_0=\begin{pmatrix}\mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0) & \dots & \mathbf{0} & \mathbf{0}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \dots & \mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0) & \mathbf{0}\\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_{n(n-1)/2}\end{pmatrix}\boldsymbol{\theta}_0.$$

Then consider the Jacobian $\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)$ which has (i,j)-th entry $\partial g_{\mathbf{c}_0,i}(\boldsymbol{\theta}_0)/\partial\theta_j=\partial\phi_i/\partial\theta_j$:

$$\nabla_{\boldsymbol{\theta}}\mathbf{g}_{\mathbf{c}_0}(\boldsymbol{\theta}_0)=\begin{pmatrix}\mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0) & \dots & \mathbf{0} & \nabla_{\boldsymbol{\rho}}\mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0)(\boldsymbol{\alpha}_{0,11}\dots\boldsymbol{\alpha}_{0,n1})'\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \dots & \mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0) & \nabla_{\boldsymbol{\rho}}\mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0)(\boldsymbol{\alpha}_{0,1p}\dots\boldsymbol{\alpha}_{0,np})'\\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I}_{n(n-1)/2}\end{pmatrix}. \tag{A-18}$$

Since $\mathbf{M}(\boldsymbol{\rho}_0;\mathbf{c}_0)$ is positive definite because of Assumption 1, the Jacobian in $\boldsymbol{\theta}_0$ is positive definite too and the mapping $\mathbf{g}_{\mathbf{c}_0}$ is invertible in $\boldsymbol{\theta}_0$ and this completes the proof. □

Proof of condition 1. The inequality on the lhs is proved in the proof of Proposition 1, while the inequality on the rhs is proved in condition B1 in the supplementary appendix of Peng *et al.* (2009). □

Proof of condition 2. This is an immediate consequence of consistency of the pre-estimator $\widehat{\mathbf{c}}_T$ given in Assumption 2 and the continuous mapping theorem. □

Proof of lemma 4. (a) We begin by noting that the sample averages of the partial derivatives of ℓ in $(\boldsymbol{\theta}_0',\boldsymbol{c}_0')'$ satisfy a Bernstein–type exponential inequality. The partial derivatives of ℓ are

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta}_0,\mathbf{y}_t,\mathbf{c}_0)}{\partial\alpha_{0\,ijk}} &= -2u_{i\,t}y_{j\,t-k}+\sum_{\substack{l=1\\ l\neq i}}^{n}2\rho^{il}\sqrt{\frac{c_{0\,ii}}{c_{0\,ll}}}u_{l\,t}y_{j\,t-k},\\ \frac{\partial\ell(\boldsymbol{\theta}_0,\mathbf{y}_t,\mathbf{c}_0)}{\partial\rho_0^{ij}} &= -2\sqrt{\frac{c_{0\,ii}}{c_{0\,jj}}}u_{i\,t}\epsilon_{j\,t}-2\sqrt{\frac{c_{0\,jj}}{c_{0\,ii}}}u_{j\,t}\epsilon_{i\,t}.\end{aligned}$$

We only show this for the partial derivatives with respect to the $\boldsymbol{\alpha}$ coefficients. The proof for the partial derivatives of the $\boldsymbol{\rho}$ coefficients follows analogous steps. In particular, we that show that the averages of the partial derivatives of the the $\boldsymbol{\alpha}$ coefficients satisfy an exponential inequality that does not depend on n. From (9) we have $\mathsf{Var}(u_{l\,t})\leq\mathsf{Var}(\epsilon_{l\,t})$ for any $l=1,\dots,n$, therefore, there exists a constant $K>0$ such that

$$\mathsf{Var}\left(\sum_{\substack{l=1\\ l\neq i}}^{n}\rho_0^{il}\sqrt{\frac{c_{0\,ii}}{c_{0\,ll}}}u_{l\,t}\right)\leq\mathsf{Var}\left(\sum_{\substack{l=1\\ l\neq i}}^{n}\rho_0^{il}\sqrt{\frac{c_{0\,ii}}{c_{0\,ll}}}\epsilon_{l\,t}\right)=\mathsf{Var}(\epsilon_{i\,t})\leq K, \tag{A-19}$$

where the last equality is given in (9). Define

$$A_{T\,ijk}=-\frac{2}{T}\sum_{t=1}^{T}u_{i\,t}y_{j\,t-k},\qquad B_{T\,ijk}=\frac{2}{T}\sum_{t=1}^{T}\left(\sum_{\substack{l=1\\ l\neq i}}^{n}\rho^{il}\sqrt{\frac{c_{0\,ii}}{c_{0\,ll}}}u_{l\,t}y_{j\,t-k}\right).$$

By Assumption 1, $y_{i\,t}$ is a strongly mixing process. Thus, $|T^{-1}\sum_{t=1}^{T}y_{i\,t}|$ satisfy the Bernstein-type exponential inequality in Theorem 1 by Doukhan and Neumann (2007) (see also Theorem 1.4 in Bosq, 1996), i.e. for any i there exists a constant $K_0>0$ such that, for any $\varepsilon>0$,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}y_{it}\right|>\varepsilon\right)\leq\exp\left\{-K_0T\varepsilon^2\right\}. \tag{A-20}$$

Since $u_{i\,t}$ is i.i.d. by construction, then as a consequence of (A-20) and Remark 2.2 in Dedecker, Doukhan, Lang, León, Louhichi, and Prieur (2007), there exists also a constant $K_1 > 0$ such that

$$\Pr\left(|A_{T\,ijk}| > \varepsilon\right) \leq \exp\left\{-K_1 T\varepsilon^2\right\}. \tag{A-21}$$

Moreover, by Assumption 1, for any $i = 1, \ldots, n$ we have $0 < c_{0\,ii} < \infty$ and therefore because of (A-19) each term in parenthesis in $B_{T\,ijk}$ has finite variance and zero mean, therefore, using arguments analogous to those used for (A-21), there exists also a constant $K_2 > 0$ such that

$$\Pr\left(|B_{T\,ijk}| > \varepsilon\right) \leq \exp\left\{-K_2 T\varepsilon^2\right\}.$$

Therefore, there exists a constant $K_3 > 0$ such that

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \ell(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{c}_0)}{\partial \alpha_{0\,ijk}}\right|^2 > \varepsilon^2\right) = \Pr\left(|A_{T\,ijk} + B_{T\,ijk}| > \varepsilon\right) \leq \Pr\left(|A_{T\,ijk}| + |B_{T\,ijk}| > \varepsilon\right) \leq 2\,\exp\left\{-K_3 T\varepsilon^2\right\}.$$

Note that as a consequence there exist a constant $K_4 > 0$ such that

$$\begin{aligned}
\Pr(\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\| > \varepsilon) &= \Pr(\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\|^2 > \varepsilon^2) = \Pr\left(\sum_{i=1}^{q_T}|S_{T\,\mathcal{A}\,i}|^2 > \varepsilon^2\right),\\
&\leq q_T \Pr\left(|S_{T\,\mathcal{A}\,i}|^2 > \frac{\varepsilon^2}{q_T}\right) = q_T \Pr\left(|S_{T\,\mathcal{A}\,i}| > \frac{\varepsilon}{\sqrt{q_T}}\right),\\
&\leq 2q_T \exp\left\{-K_4 T\frac{\varepsilon^2}{q_T}\right\}.
\end{aligned}$$

By setting the rhs of the last expression equal to $\delta = O(T^{-\eta})$ for $\eta > 0$ and solving with respect to ε we get that for T sufficiently large there exist a constant κ_0 such that

$$\varepsilon \leq \kappa_0\sqrt{q_T}\,\sqrt{\frac{\log T}{T}}. \tag{A-22}$$

Then, for T sufficiently large there exist a constant κ_0 such that

$$\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\| < \kappa_0\sqrt{q_T}\sqrt{\frac{\log T}{T}},$$

with at least probability $1 - O(T^{-\eta})$. Moreover, we have

$$\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\| \leq \|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\| + \|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\| \tag{A-23}$$

and for T sufficiently large the second term is $O(\sqrt{(q_T \log T)\,T^{-1}}\,) = o(1)$ by Condition 2. Part (a) follows by combining (A-22) and (A-23). Part (b) follows from (a) and the Cauchy-Schwarz inequality.

(c) We begin by noting that

$$\begin{aligned}
\|\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u} - \mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u}\|^2 &\leq 2\|\mathbf{u}\|^2\|\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\|^2,\\
&\leq 2\|\mathbf{u}\|^2\sum_{i=1}^{q_T}\sum_{j=1}^{q_T}\left[\mathbf{H}_{T\,\mathcal{A}\mathcal{A}\,ij}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{H}_{0\,\mathcal{A}\mathcal{A}\,ij}(\boldsymbol{\theta}_0, \mathbf{c}_0)\right]^2.
\end{aligned}$$

Next, we focus on showing that the differences

$$A_{T\,ij} = \mathbf{H}_{T\,\mathcal{A}\mathcal{A}\,ij}(\boldsymbol{\theta}_0, \mathbf{c}_0) - \mathbf{H}_{0\,\mathcal{A}\mathcal{A}\,ij}(\boldsymbol{\theta}_0, \mathbf{c}_0)$$

satisfy an appropriate Bernstein–type exponential inequality. We begin by noting that the first $n^2p \times n^2p$

diagonal block of the Hessian has entries

$$\frac{\partial^2 \ell(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{c}_0)}{\partial \alpha_{ij'k'} \partial \alpha_{ijk}} = 2 y_{j\, t-k} y_{j'\, t-k'} \left(1 + \sum_{\substack{l=1 \\ l \neq i}}^{n} \rho_0^{il} \sqrt{\frac{c_{0\, ll}}{c_{0\, ii}}} \left(\rho_0^{li} \sqrt{\frac{c_{0\, ii}}{c_{0\, ll}}} - 1 \right) \right)$$

for any $j, j' = 1 \dots n$ and any $k, k' = 1 \dots p$ and

$$\frac{\partial^2 \ell(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{c}_0)}{\partial \alpha_{i'j'k'} \partial \alpha_{ijk}} = 2 y_{j\, t-k} y_{j'\, t-k'} \left(\left(1 - \rho_0^{ii'} \sqrt{\frac{c_{0\, i'i'}}{c_{0\, ii}}} \right) + \sum_{\substack{l=1 \\ l \neq i}}^{n} \rho_0^{il} \sqrt{\frac{c_{0\, ll}}{c_{0\, ii}}} \left(\rho_0^{li} \sqrt{\frac{c_{0\, ii}}{c_{0\, ll}}} - 1 \right) \right),$$

for $i \neq i'$ and any $j, j' = 1 \dots n$ and any $k, k' = 1 \dots p$. The second $n(n-1)/2 \times n(n-1)/2$ diagonal block has entries

$$\frac{\partial^2 \ell(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{c}_0)}{\partial \rho^{ij'} \partial \rho^{ij}} = 2 \sqrt{\frac{c_{0\, jj} c_{0\, ii}}{c_{0\, ii} c_{0\, j'j'}}} \epsilon_{j'\, t} \epsilon_{j\, t},$$

for any $i, j, j' = 1 \dots n$ with $i \neq j$, $i \neq j'$ and $j \neq j'$. It is straightforward to check that the averages of the partial derivatives with respect to the $\boldsymbol{\rho}$ coefficients satisfy a Bernstein–type inequality. As far as the partial derivatives with respect to the $\boldsymbol{\alpha}$ coefficients we need to show that this term does not grow with n. Notice that by Assumption 1 and from (9), there exists a constant $K_1 > 0$ such that

$$\sum_{\substack{l=1 \\ l \neq i}}^{n} |\rho_0^{il}| \leq \sum_{\substack{l=1 \\ l \neq i}}^{n} (\rho_0^{il})^2 \leq \frac{\mathsf{Var}(\epsilon_{i\, t})}{\mu_{\min}(\mathbf{C}_0^{-1})} = \mathsf{Var}(\epsilon_{i\, t}) \mu_{\max}(\mathbf{C}_0) < K_1. \tag{A-24}$$

Thus, given (A-24), and since by Assumption 1, we have $0 < c_{0\, ii} < \infty$ for any $i = 1, \dots, n$, there exists a constant $K_2 > 0$ such that

$$\left| \sum_{\substack{l=1 \\ l \neq i}}^{n} \rho_0^{il} \sqrt{\frac{c_{0\, ll}}{c_{0\, ii}}} \left(\rho_0^{li} \sqrt{\frac{c_{0\, ii}}{c_{0\, ll}}} - 1 \right) \right| \leq \sum_{\substack{l=1 \\ l \neq i}}^{n} (\rho_0^{il})^2 + \sum_{\substack{l=1 \\ l \neq i}}^{n} |\rho_0^{il}| \sqrt{\frac{c_{0\, ll}}{c_{0\, ii}}} < K_2.$$

By the Cauchy–Schwartz inequality, we have that the mixed partial derivatives with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$ also not grow with n and satisfy a Bernstein–type concentration inequality. Thus, there exists a constant $K_3 > 0$ such that $|A_{1T\, ijk, i'j'k'}| \leq K_3 |y_{j\, t-k} y_{j'\, t-k'} - \mathsf{E}[y_{j\, t-k} y_{j'\, t-k'}]|$ for any (i, j, k) and (i', j', k'). Therefore, by Assumption 1 and the same arguments leading to (A-21) there exists a constant $K_4 > 0$ such that

$$\Pr\left(\sum_{i=1}^{q_T} \sum_{j=1}^{q_n} |A_{T\, ij}|^2 \geq \varepsilon^2 \right) \leq q_T^2 \Pr\left(|A_{T\, ij}| \geq \frac{\varepsilon}{q_T} \right) \leq 2 q_T^2 \exp\left\{ -K_4 T \frac{\varepsilon^2}{q_T^2} \right\}.$$

By setting the rhs of the last expression equal to $\delta = O(T^{-\eta'})$ for $\eta' > 0$ and solving with respect to ε we get that for T sufficiently large there exist a constant $\kappa_2 > 0$ such that

$$\varepsilon \leq \kappa_2\, q_T \sqrt{\frac{\log T}{T}}.$$

Finally, for $\eta > 0$ and T sufficiently large there exists a constant κ_2 such that

$$\| \mathbf{H}_{T\, \mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0) \mathbf{u} - \mathbf{H}_{0\, \mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0) \mathbf{u} \| < \kappa_2 \| \mathbf{u} \|^2 q_T \sqrt{\frac{\log T}{T}},$$

with at least probability $1 - O(T^{-\eta})$. Part (c) follows as part (a) by using Condition 2 and the conditions in the statements of Propositions 2 and 3. Part (d) follows from (c) and the Cauchy-Schwarz inequality. This completes the proof. □

Proof of lemma 5. See Lemma 2.1 in Bühlmann and van de Geer (2011). □

Proof of lemma 6. Define $\nu_T = \sqrt{q_T}\lambda_T$, therefore $\nu_T \to 0$ as $T \to \infty$. Consider a generic vector $\mathbf{u} \in \mathbb{R}^m$ such that $\mathbf{u}_{\mathcal{A}^c} = \mathbf{0}$ and $\|\mathbf{u}\| = C$. Define $L_T(\boldsymbol{\theta}, \mathbf{c}) = \frac{1}{T}\sum_{t=1}^T \ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c})$ and $\ell(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{c})$ is the unconstrained loss function defined in (12). The increment of the sample loss defined in (14)-(15) is

$$
\begin{aligned}
Q_T(\boldsymbol{\theta}_0 + \nu_T\mathbf{u}) &= \frac{1}{T}\Big(\mathcal{L}_T(\boldsymbol{\theta}_0 + \nu_t\mathbf{u}, \widehat{\mathbf{c}}_T) - \mathcal{L}_T(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\Big) = \\
&= \Big[L_T(\boldsymbol{\theta}_0 + \nu_t\mathbf{u}, \widehat{\mathbf{c}}_T) - L_T(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\Big] - \lambda_T \sum_{\substack{i=1\\ i\in\mathcal{A}}}^m \frac{|\theta_{0i}| - |\theta_{0i} + \nu_T u_i|}{|\widetilde{\theta}_{Ti}|} \\
&\geq \Big[L_T(\boldsymbol{\theta}_0 + \nu_t\mathbf{u}, \widehat{\mathbf{c}}_T) - L_T(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\Big] - \lambda_T\nu_T \sum_{\substack{i=1\\ i\in\mathcal{A}}}^m \frac{|u_i|}{|\widetilde{\theta}_{Ti}|}. \qquad \text{(A-25)}
\end{aligned}
$$

Start from the first term in (A-25). By Lemma 4, for T sufficiently large and for any $\eta > 0$, we have with probability at least $1 - O(T^{-\eta})$

$$
\begin{aligned}
L_T(\boldsymbol{\theta}_0 + \nu_t\mathbf{u}, \widehat{\mathbf{c}}_T) - L_T(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T) &= \nu_T\mathbf{u}_{\mathcal{A}}'\mathbf{S}_{T\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T) + \frac{1}{2}\nu_T^2\mathbf{u}_{\mathcal{A}}'\mathbf{H}_{T\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\mathbf{u}_{\mathcal{A}} \qquad \text{(A-26)}\\
= \nu_T\mathbf{u}_{\mathcal{A}}'\mathbf{S}_{T\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T) &+ \frac{1}{2}\nu_T^2\mathbf{u}_{\mathcal{A}}'\mathbf{H}_{0\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T)\mathbf{u}_{\mathcal{A}} + \frac{1}{2}\nu_T^2\mathbf{u}_{\mathcal{A}}'\Big(\mathbf{H}_{T\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T) - \mathbf{H}_{0\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\Big)\mathbf{u}_{\mathcal{A}} + o(\nu_T^2) \\
\geq -\kappa_1\|\mathbf{u}_{\mathcal{A}}\|\sqrt{q_T}\sqrt{\frac{\log T}{T}}\nu_T &- \kappa_3\|\mathbf{u}_{\mathcal{A}}\|^2 q_T\sqrt{\frac{\log T}{T}}\nu_T^2 + \frac{1}{2}\nu_T^2 u_{\mathcal{A}}'\mathbf{H}_{0\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u}_{\mathcal{A}}.
\end{aligned}
$$

By the conditions given in the statements of Propositions 2 and 3 and since $\|\mathbf{u}_{\mathcal{A}}\| = C$, for the first and second term on the rhs of (A-26) we have

$$
-\kappa_1 C\sqrt{q_T}\sqrt{\frac{\log T}{T}}\frac{\lambda_T}{\lambda_T}\nu_T = \nu_T^2 o(1) = o(\nu_T^2), \qquad \text{(A-27)}
$$

$$
-\kappa_3 C^2 q_T\sqrt{\frac{\log T}{T}}\nu_T^2 = \nu_T^2 o(1) = o(\nu_T^2), \qquad \text{(A-28)}
$$

and both terms can be neglected for T sufficiently large. Moreover, by Condition 1, we have

$$
\frac{1}{2}\nu_T^2 u_{\mathcal{A}}'\mathbf{H}_{0\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0, \mathbf{c}_0)\mathbf{u}_{\mathcal{A}} \geq \frac{1}{2}\nu_T^2 C^2\mu_{\min}(\mathbf{H}_{0\mathcal{A}\mathcal{A}}) \geq \frac{1}{2}\nu_T^2 C^2\underline{L} > 0. \qquad \text{(A-29)}
$$

Then, notice that, by Cauchy-Schwarz inequality,

$$
\Bigg(\sum_{\substack{i=1\\ i\in\mathcal{A}}}^m \frac{|u_i|}{|\widetilde{\theta}_{Ti}|}\Bigg)^2 \leq C^2 \sum_{\substack{i=1\\ i\in\mathcal{A}}}^m \frac{1}{\widetilde{\theta}_{Ti}^2}. \qquad \text{(A-30)}
$$

Moreover, for any $i \in \mathcal{A}$,

$$
\frac{1}{\widetilde{\theta}_{Ti}^2} = \frac{1}{\theta_{0i}^2} - \frac{2\theta_{0i}}{\theta_{0i}^4}(\widetilde{\theta}_{Ti} - \theta_{0i}) + o((\widetilde{\theta}_{Ti} - \theta_{0i})) \leq \frac{1}{\theta_{0i}^2} + \frac{2}{\theta_{0i}^3}|\widetilde{\theta}_{Ti} - \theta_{0i}| + o(|\widetilde{\theta}_{Ti} - \theta_{0i}|). \qquad \text{(A-31)}
$$

Define $\theta_{0\min}^2 = \min_{i\in\mathcal{A}}\theta_{0i}^2$ and notice that $|\theta_{0\min}| > 0$. Then, combining Assumption 2 and (A-31), there exists a constant $K > 0$ such that for T sufficiently large and for any $\eta > 0$, we have with probability at least

$1-O(T^{-\eta})$

$$C^2\sum_{\substack{i=1\\ i\in\mathcal{A}}}^{m}\frac{1}{\widetilde{\theta}_{Ti}^2}\leq\frac{C^2 q_T}{\theta_{0\,\min}^2}+C^2Kq_T\sqrt{\frac{\log T}{T}}.\tag{A-32}$$

Therefore, using (A-30) and (A-32), for the second term in (A-25), for T sufficiently large and for any $\eta>0$, with probability at least $1-O(T^{-\eta})$, we have

$$\begin{aligned}-\lambda_T\nu_T\sum_{\substack{i=1\\ i\in\mathcal{A}}}^{m}\frac{|u_i|}{|\widetilde{\theta}_{Ti}|}&\geq-\lambda_T\nu_TC\sqrt{q_T}\left[\frac{1}{|\theta_{0\,\min}|}+\sqrt{K}\left(\frac{\log T}{T}\right)^{1/4}\right]\\&\geq-\nu_T^2\frac{C}{|\theta_{0\,\min}|}+\nu_T^2\sqrt{K}\left(\frac{\log T}{T}\right)^{1/4},\end{aligned}\tag{A-33}$$

and notice that the last term is $o(\nu_T^2)$, thus it can be neglected for T sufficiently large. Then, by substituting (A-26) and (A-33) in (A-25), and using (A-27), (A-28), and (A-29), we have, for T sufficiently large and for any $\eta>0$,

$$\Pr\left(Q_T(\boldsymbol{\theta}_0+\nu_T\mathbf{u})\geq\frac{1}{2}\nu_T^2C^2\underline{L}-\frac{C}{|\theta_{0\,\min}|}\nu_T^2=\nu_T^2C\left(\frac{\underline{L}}{2}C-\frac{1}{|\theta_{0\,\min}|}\right)\right)\geq1-O(T^{-\eta}).$$

Thus, if we choose $C=2/(\underline{L}|\theta_{0\,\min}|)+\epsilon$, for any $\epsilon>0$, then for T sufficiently large and for any $\eta>0$

$$\Pr\left(\inf_{\substack{\mathbf{u}:\mathbf{u}_{\mathcal{A}^c}=\mathbf{0}\\ \|\mathbf{u}\|=C}}Q_T(\boldsymbol{\theta}_0+\nu_T\mathbf{u})>0\right)=\Pr\left(\inf_{\substack{\mathbf{u}:\mathbf{u}_{\mathcal{A}^c}=\mathbf{0}\\ \|\mathbf{u}\|=C}}\mathcal{L}_T(\boldsymbol{\theta}_0+\nu_T\mathbf{u},\widehat{\mathbf{c}}_T)>\mathcal{L}_T(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)\right)\geq1-O(T^{-\eta}),$$

which means that there exists a local minimum for the restricted problem within the disc $D(\boldsymbol{\theta}_0)=\{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\nu_TC\}$, with probability at least $1-O(T^{-\eta})$. By choosing $\kappa_4=C$, we complete the proof. □

Proof of lemma 7. Define $\nu_T=\sqrt{q_T}\lambda_T$, therefore $\nu_T\to0$ as $T\to\infty$. Then, any $\boldsymbol{\theta}\in S(\boldsymbol{\theta}_0)$ can be written as $\boldsymbol{\theta}=\boldsymbol{\theta}_0+\nu_T\mathbf{u}$, where $\mathbf{u}_{\mathcal{A}^c}=\mathbf{0}$, $\|\mathbf{u}\|\geq\kappa_5$, and $\|\mathbf{u}\|\leq C<\infty$. For any $\boldsymbol{\theta}\in S(\boldsymbol{\theta}_0)$, we can write

$$\begin{aligned}\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta},\widehat{\mathbf{c}}_T)&=\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)+\nu_T\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)\mathbf{u}\\&=\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)+\nu_T\Big(\mathbf{H}_{T\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)-\mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\mathbf{c}_0)\Big)\mathbf{u}+\nu_T\mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\mathbf{c}_0)\mathbf{u}+o(\nu_T).\end{aligned}$$

Thus, by Lemma 4, for T sufficiently large and for any $\eta>0$, we have, with probability at least $1-O(T^{-\eta})$,

$$\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta},\widehat{\mathbf{c}}_T)\|\geq-\kappa_0\sqrt{q_T}\sqrt{\frac{\log T}{T}}-\kappa_2\|\mathbf{u}\|\nu_Tq_T\sqrt{\frac{\log T}{T}}+\nu_T\|\mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\mathbf{c}_0)\mathbf{u}\|.$$

The first and second term on the rhs of the last expression are both $o(\nu_T)$. Then, using Condition 1, for T sufficiently large and for any $\eta>0$, with probability at least $1-O(T^{-\eta})$ we have

$$\|\mathbf{S}_{T\,\mathcal{A}}(\boldsymbol{\theta},\widehat{\mathbf{c}}_T)\|\geq\nu_T\|\mathbf{H}_{0\,\mathcal{A}\mathcal{A}}(\boldsymbol{\theta}_0,\mathbf{c}_0)\mathbf{u}\|\geq\nu_T\,\underline{L}\,\kappa_5.$$

Define $\theta_0^*=\min_{i\in\mathcal{A}}|\theta_{0i}|$ and notice that $\theta_0^*>0$. By choosing $\kappa_5=1/(\underline{L}\theta_0^*)+\epsilon$ for any $\epsilon>0$, we complete the proof. □

Proof of lemma 8. In the following define $\mathbf{v}_T=(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}-\boldsymbol{\theta}_0)$. For any $j\in\mathcal{A}^c$ we have

$$\begin{aligned}S_{T\,j}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}},\widehat{\mathbf{c}}_T)&=S_{T\,j}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)+H_{T\,j}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)\mathbf{v}_T+o(\|\mathbf{v}_T\|)\\&=\underbrace{S_{T\,j}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)}_{A_{1T\,j}}+\underbrace{H_{0\,j}(\boldsymbol{\theta}_0,\mathbf{c}_0)\mathbf{v}_T}_{A_{2T\,j}}+\underbrace{[H_{T\,j}(\boldsymbol{\theta}_0,\widehat{\mathbf{c}}_T)-H_{0\,j}(\boldsymbol{\theta}_0,\mathbf{c}_0)]\,\mathbf{v}_T}_{A_{3T\,j}}+o(\|\mathbf{v}_T\|).\end{aligned}\tag{A-34}$$

Start from the first term on the rhs of (A-34). By an argument analogous to the proof of Lemma 4 (a) we have that for T sufficiently large and for any $\eta' > 0$ there exists a constant $K_1 > 0$ such that

$$\Pr\left(|A_{1T\,j}| \leq K_1\sqrt{\frac{\log T}{T}}\right) \geq 1 - O(T^{-\eta'}). \tag{A-35}$$

For the second term on the rhs of (A-34) from Condition 1 we have

$$|A_{2T\,j}| \leq \|H_{0\,j}(\boldsymbol{\theta}_0, \mathbf{c}_0)\|\ \|\mathbf{v}_T\| \leq \mu_{\max}(\mathbf{H}_0(\boldsymbol{\theta}_0, \mathbf{c}_0))\ \|\mathbf{v}_T\| \leq \overline{L}\ \|\mathbf{v}_T\|.$$

Therefore, if we define $K_2 = \kappa_R \overline{L}$, by Proposition 2 we have that for T sufficiently large and for any $\eta' > 0$

$$\Pr\left(|A_{2T\,j}| \leq K_2\sqrt{q_T}\lambda_T\right) \geq 1 - O(T^{-\eta'}). \tag{A-36}$$

For the third term on the rhs of (A-34), we have

$$|A_{3T\,j}| \leq \|\ [H_{T\,j}(\boldsymbol{\theta}_0, \widehat{\mathbf{c}}_T) - H_{0\,j}(\boldsymbol{\theta}_0, \mathbf{c}_0)]\ \|\ \|\mathbf{v}_T\|.$$

Then, using an argument similar to the proof of Lemma 4 (c) and by Proposition 2 and by defining $K_3 = \kappa_2\kappa_R$ we have that for T sufficiently large and for any $\eta' > 0$

$$\Pr\left(|A_{3T\,j}| \leq K_3 q_T \lambda_T\sqrt{\frac{\log T}{T}}\right) \geq 1 - O(T^{-\eta'}). \tag{A-37}$$

Moreover, by Assumption 2 there exists a constant $K_4 > 0$ such that for T sufficiently large and for any $\eta' > 0$ we have

$$\Pr\left(\frac{1}{\max_{j\in\mathcal{A}^c}|\widetilde{\theta}_{Tj}|} > K_4\sqrt{\frac{T}{\log T}}\right) \geq 1 - O(T^{-\eta'}). \tag{A-38}$$

From (A-38) and (A-34) we have

$$\begin{aligned}\Pr\left(|S_{T\,j}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}, \widehat{\mathbf{c}}_T)| \leq \frac{\lambda_T}{\max_{j\in\mathcal{A}^c}|\widetilde{\theta}_{Tj}|}\right) &\geq \Pr\left(|A_{1T\,j}| + |A_{2T\,j}| + |A_{3T\,j}| \leq \frac{\lambda_T}{\max_{j\in\mathcal{A}^c}|\widetilde{\theta}_{Tj}|}\right)\\ &\geq \Pr\left(|A_{1T\,j}| + |A_{2T\,j}| + |A_{3T\,j}| \leq K_4\lambda_T\sqrt{\frac{T}{\log T}}\right).\end{aligned} \tag{A-39}$$

Notice that, as $T \to \infty$, we have

$$\lambda_T\sqrt{\frac{T}{\log T}} \to \infty, \quad \sqrt{q_T}\frac{\lambda_T}{T} \to 0, \quad q_T\frac{\lambda_T}{T}\sqrt{\frac{\log T}{T}} \to 0, \quad \sqrt{\frac{\log T}{T}} \to 0. \tag{A-40}$$

where the first three conditions are assumed in Proposition 3 while the last one is trivial.

Finally, consider the complementary of (A-39), then, by combining (A-35)-(A-37) with (A-40), we have that for T sufficiently large and for any $\eta' > 0$

$$\Pr\left(|A_{1T\,j}| + |A_{2T\,j}| + |A_{3T\,j}| \geq K_4\lambda_T\sqrt{\frac{T}{\log T}}\right) \leq \sum_{k=1}^{3}\Pr\left(|A_{kT\,j}| \geq K_4\lambda_T\sqrt{\frac{T}{\log T}}\right) = O(T^{-\eta'}),$$

which implies that

$$\Pr\left(|S_{T\,j}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}, \widehat{\mathbf{c}}_T)| \leq \frac{\lambda_T}{\max_{j\in\mathcal{A}^c}|\widetilde{\theta}_{Tj}|}\right) \geq 1 - O(T^{-\eta'}). \tag{A-41}$$

Given $n = O(T^{\zeta})$, define $\eta' = \eta + \zeta$, then for T sufficiently large and for any $\eta > 0$, from (A-41) we have

$$\Pr\left(\max_{j \in \mathcal{A}^c} |S_{T\,j}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}, \widehat{\mathbf{c}}_T)| \geq \frac{\lambda_T}{\max_{j \in \mathcal{A}^c} |\widetilde{\theta}_{Tj}|} \right) \leq n \Pr\left(|S_{T\,j}(\widehat{\boldsymbol{\theta}}_T^{\mathcal{A}}, \widehat{\mathbf{c}}_T)| \geq \frac{\lambda_T}{\max_{j \in \mathcal{A}^c} |\widetilde{\theta}_{Tj}|} \right) = O(T^{-\eta}).$$

By considering the complementary event we complete the proof. □