Journal of Machine Learning Research 16 (2015) 417-453 Submitted 9/13; Revised 5/14; Published 3/15

Network Granger Causality with Inherent Grouping Structure

Sumanta Basu SUMBOSE@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1092, USA

Ali Shojaie ASHOJAIE@U.WASHINGTON.EDU
Department of Biostatistics
University of Washington
Seattle, WA, USA

George Michailidis GMICHAIL@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1092, USA

Editor: Bin Yu

Abstract

The problem of estimating high-dimensional network models arises naturally in the analysis of many biological and socio-economic systems. In this work, we aim to learn a network structure from temporal panel data, employing the framework of Granger causal models under the assumptions of sparsity of its edges and inherent grouping structure among its nodes. To that end, we introduce a group lasso regression regularization framework, and also examine a thresholded variant to address the issue of group misspecification. Further, the norm consistency and variable selection consistency of the estimates are established, the latter under the novel concept of direction consistency. The performance of the proposed methodology is assessed through an extensive set of simulation studies and comparisons with existing techniques. The study is illustrated on two motivating examples coming from functional genomics and financial econometrics.

Keywords: Granger causality, high dimensional networks, panel vector autoregression model, group lasso, thresholding

1. Introduction

We consider the problem of learning a directed network of interactions among a number of entities from time course data. A natural framework to analyze this problem uses the notion of Granger causality (Granger, 1969). Originally proposed by C.W. Granger this notion provides a statistical framework for determining whether a time series X is useful in forecasting another one Y, through a series of statistical tests. It has found wide applicability in economics, including testing relationships between money and income (Sims, 1972), government spending and taxes on economic output (Blanchard and Perotti, 2002), stock price

©2015 Sumanta Basu, Ali Shojaie and George Michailidis.

and volume (Hiemstra and Jones, 1994), etc. More recently the Granger causal framework has found diverse applications in biological sciences including functional genomics, systems biology and neurosciences to understand the structure of gene regulation, protein-protein interactions and brain circuitry, respectively.

It should be noted that the concept of Granger causality is based on associations between time series, and only under very stringent conditions, true causal relationships can be inferred (Pearl, 2000). Nonetheless, this framework provides a powerful tool for understanding the interactions among random variables based on time course data.

Network Granger causality (NGC) extends the notion of Granger causality among two variables to a wider class of p variables. Such extensions involving multiple time series are handled through the analysis of vector autoregressive processes (VAR) (Lütkepohl, 2005). Specifically, for p stationary time series $X_1^t, \ldots, X_p^t$, with $X^t = (X_1^t, \ldots, X_p^t)'$, one considers the class of models

$$X^t = A^1 X^{t-1} + \ldots + A^d X^{t-d} + \epsilon^t, \tag{1}$$

where $A^1, A^2, \ldots, A^d$ are $p \times p$ real-valued matrices, d is the *unknown* order of the VAR model and the innovation process satisfies $\epsilon^t \sim N(0, \sigma^2 I)$. In this model, the time series $\{X_j^t\}$ is said to be Granger causal for the time series $\{X_i^t\}$ if $A_{i,j}^h \neq 0$ for some $h = 1, \ldots, d$. Equivalently we can say that there exists an edge $X_j^{t-h} \to X_i^t$ in the underlying network model comprising of $(d+1) \times p$ nodes (see Figure 1). We call $A^1, \ldots, A^d$ the adjacency matrices from lags $1, \ldots, d$. Note that the entries A_{ij}^h of the adjacency matrices are not binary indicators of presence/absence of edges between two nodes X_i^t and X_j^{t-h}. Rather, they represent the direction and strength of influence from one node to the other.

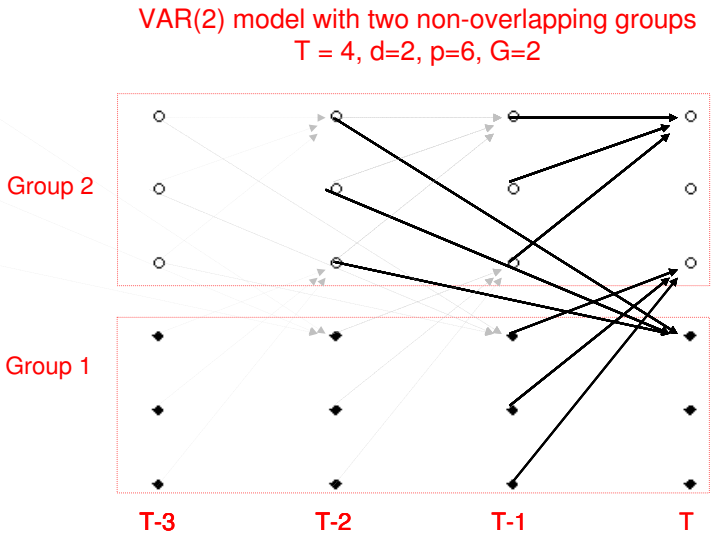


Figure 1: An example of a Network Granger causal model with two non-overlapping groups observed over T = 4 time points

The temporal structure induces a natural partial order among the nodes of this network, which in turn simplifies significantly the corresponding estimation problem (Shojaie and Michailidis, 2010a) of a directed acyclic graph. Nevertheless, one still has to deal with estimating a high-dimensional network (e.g., hundreds of genes) from a limited number of samples.

The traditional asymptotic framework of estimating VAR models requires observing a long, stationary realization $\{X^1, \ldots, X^T, T \to \infty, p, d \text{ fixed}\}$ of the p-dimensional time series. This is not appropriate in many biological applications for the following reasons. First, long stationary time series are rarely observed in these contexts. Second, the number of time series (p) being large compared to T, the task of consistent order (d) selection using standard criteria (e.g., AIC or BIC) becomes challenging. Similar issues arise in many econometric applications where empirical evidence suggests lack of stationarity over a long time horizon, although the multivariate time series exhibits locally stable distributional properties.

A more suitable framework comes from the study of panel data, where one observes several replicates of the time series, with possibly short T, across a panel of n subjects. In biological applications replicates are obtained from test subjects. In the analysis of macroeconomic variables, households or firms typically serve as replicates. After removing panel specific fixed effects one treats the replicates as independent samples, performs regression analysis under the assumption of common slope structure and studies the asymptotic properties under the regime $n \to \infty$. Recent works of Cao and Sun (2011) and Binder et al. (2005) analyze theoretical properties of short panel VARs in the low-dimensional setting ($n \to \infty$, T, p fixed).

The focus of this work is on estimating a *high-dimensional NGC model* in the panel data context (p, n large, T small to moderate). This work is motivated by two application domains, functional genomics and financial econometrics. In the first application (presented in Section 6) one is interested in reconstructing a gene regulatory network structure from time course data, a canonical problem in functional genomics (Michailidis, 2012). The second motivating example examines the composition of balance sheets of the $n = 50$ largest US banks by size, over $T = 9$ quarterly periods, which provides insight into their risk profile.

The nature of high-dimensionality in these two examples comes from both estimation of p^2 coefficients for each of the adjacency matrices $A^1, \ldots, A^d$, but also from the fact that the order of the time series d is often unknown. Thus, in practice, one must either "guess" the order of the time series (often times, it is assumed that the data is generated from a VAR(1) model, which can result in significant loss of information), or include all of the past time points, resulting in significant increase in the number of variables in cases where $d \ll T$. Thus, efficient estimation of the order of the time series becomes crucial.

Latent variable based dimension reduction techniques like principal component analysis or factor models are not very useful in this context since our goal is to reconstruct a network among the observed variables. To achieve dimension reduction we impose a group sparsity assumption on the structure of the adjacency matrices $A_1, \ldots, A_d$. In many applications, structural grouping information about the variables exists. For example, genes can be naturally grouped according to their function or chromosomal location, stocks according to their industry sectors, assets/liabilities according to their class, etc. This information can

be incorporated to the Granger causality framework through a group lasso penalty. If the group specification is correct it enables estimation of denser networks with limited sample sizes (Bach, 2008; Huang and Zhang, 2010; Lounici et al., 2011). However, the group lasso penalty can achieve model selection consistency only at a group level. In other words, if the groups are misspecified, this procedure can not perform within group variable selection (Huang et al., 2009), an important feature in many applications.

Over the past few years, several authors have adopted the framework of network Granger causality to analyze multivariate temporal data. For example, Fujita et al. (2007) and Lozano et al. (2009) employed NGC models coupled with penalized ℓ_1 regression methods to learn gene regulatory mechanisms from time course microarray data. Specifically, Lozano et al. (2009) proposed to group all the past observations, using a variant of group lasso penalty, in order to construct a relatively simple Granger network model. This penalty takes into account the average effect of the covariates over different time lags and connects Granger causality to this average effect being significant. However, it suffers from significant loss of information and makes the consistent estimation of the signs of the edges difficult (due to averaging). Shojaie and Michailidis (2010b) proposed a truncating lasso approach by introducing a truncation factor in the penalty term, which strongly penalizes the edges from a particular time lag, if it corresponds to a highly sparse adjacency matrix.

Despite recent use of NGC in applications involving high dimensional data, theoretical properties of the resulting estimators have not been fully investigated. For example, Lozano et al. (2009) and Shojaie and Michailidis (2010b) discuss asymptotic properties of the resulting estimators, but neither addresses in depth norm consistency properties, nor do they examine under what vector autoregressive structures the obtained results hold.

In this paper, we develop a general framework that accommodates different variants of group lasso penalties for NGC models. It allows for the simultaneous estimation of the order of the times series and the Granger causal effects; further, it allows for variable selection even when the groups are misspecified. In summary, the key contributions of this work are: (i) investigate in depth *sufficient conditions* that explicitly take into consideration the structure of the VAR(d) model to establish norm consistency, (ii) introduce the novel notion of *direction consistency*, which generalizes the concept of sign consistency and provides insight into the properties of group lasso estimates within a group, and (iii) use the latter notion to introduce an easy to compute thresholded variant of group lasso, that performs within group variable selection in addition to group sparsity pattern selection even when the group structure is misspecified.

All the obtained results are non-asymptotic in nature, and hence help provide insight into the properties of the estimates under different asymptotic regimes arising from varying growth rates of T, p, n, group sizes and the number of groups.

2. Model and Framework

Notation. Consider a VAR model

$$\underbrace{X^t}_{p\times 1} = \underbrace{A^1}_{p\times p} X^{t-1} + \ldots + A^d X^{t-d} + \epsilon^t, \quad \epsilon^t \sim N(0_{p\times 1}, \sigma^2 I_{p\times p}), \tag{2}$$

observed over T time points $t = 1, \ldots, T$, across n panels. The index set of the variables $\mathbb{N}_p = \{1, 2, \ldots, p\}$ can be partitioned into G non-overlapping groups $\mathcal{G}_g$, i.e., $\mathbb{N}_p = \cup_{g=1}^{G} \mathcal{G}_g$ and $\mathcal{G}_g \cap \mathcal{G}_{g'} = \phi$ if $g \neq g'$. Also $k_g = |\mathcal{G}_g|$ denotes the size of the g^{th} group with $k_{max} = \max_{1 \leq g \leq G} k_g$. In general, we use $\lambda_{\min}$ and $\lambda_{\max}$ to denote the minimum and maximum of a finite collection of numbers $\lambda_1, \ldots, \lambda_m$.

For any matrix A, we denote the i^{th} row by $A_{i:}$, j^{th} column by $A_{:j}$ and the collection of rows (columns) corresponding to the g^{th} group by $A_{[g]:}$ ($A_{:[g]}$). The transpose of a matrix A is denoted by A' and its Frobenius norm by $||A||_F$. For a symmetric/Hermitian matrix Σ, its maximum and minimum eigenvalues are denoted by $\Lambda_{\min}(\Sigma)$ and $\Lambda_{\max}(\Sigma)$, respectively. The symbol $A^{1:h}$ is used to denote the concatenated matrix $\left[A^1 : \cdots : A^h\right]$, for any $h > 0$. For any matrix or vector D, $\|D\|_0$ denotes the number of non-zero coordinates in D. For notational convenience, we reserve the symbol $\|.\|$ to denote the ℓ_2 norm of a vector and/or the spectral norm of a matrix. For a pre-defined set of non-overlapping groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$ on $\{1, \ldots, p\}$, the mixed norms of vectors $v \in \mathbb{R}^p$ are defined as $\|v\|_{2,1} = \sum_{g=1}^{G} \|v_{[g]}\|$ and $\|v\|_{2,\infty} = \max_{1 \leq g \leq G} \|v_{[g]}\|$. Also for any vector β, we use β_j to denote its j^{th} coordinate and $\beta_{[g]}$ to denote the coordinates corresponding to the g^{th} group. We also use $supp(v)$ to denote the support of v, i.e., $supp(v) = \{j \in \{1, \ldots, p\} | v_j \neq 0\}$.

Network Granger causal (NGC) estimates with group sparsity. Consider n replicates from the NGC model (2), and denote the $n \times p$ observation matrix at time t by $\mathcal{X}^t$. In econometric applications the data on p economic variables across n panels (firms, households etc.) can be observed over T time points. For time course microarray data one typically observes the expression levels of p genes across n subjects over T time points. After removing the panel specific fixed effects one assumes the common slope structure and independence across the panels. The data are high-dimensional if either T or p is large compared to n. In such a scenario, we assume the existence of an underlying group sparse structure, i.e., for every $i = 1, \ldots, p$, the support of the i^{th} row of $A^{1:T-1} = \left[A^1 : \cdots : A^{T-1}\right]$ in the model (2) can be covered by a small number of groups s_i, where $s_i \ll (T-1)G$. Note that the groups can be misspecified in the sense that the coordinates of a group covering the support need not be all non-zero. Hence, for a properly specified group structure we shall expect $s_i \ll \|A^{1:T}_{i:}\|_0$. On the contrary, with many misspecified groups, s_i can be of the same order, or even larger than $\|A^{1:T}_{i:}\|_0$.

Learning the network of Granger causal effects $\{(i,j) \in \{1, \ldots, p\} : A^t_{ij} \neq 0 \text{ for some } t\}$ is equivalent to recovering the correct sparsity pattern in $A^{1:(T-1)}$ and consistently estimating the non-zero effects A^t_{ij}. In the high-dimensional regression problems this is achieved by simultaneous regularization and selection operators like lasso and group lasso. The group Granger causal estimates of the adjacency matrices $A^1, \ldots, A^{T-1}$ are obtained by solving the following optimization problem

$$\hat{A}^{1:T-1} = \underset{A^1, \cdots, A^{T-1}}{\operatorname{argmin}} \frac{1}{2n} \left\| \mathcal{X}^T - \sum_{t=1}^{T-1} \mathcal{X}^{T-t} \left(A^t\right)' \right\|_F^2 + \lambda \sum_{t=1}^{T-1} \sum_{i=1}^{p} \sum_{g=1}^{G} w^t_{i,g} \|A^t_{i:[g]}\|, \qquad (3)$$

where $\mathcal{X}^t$ is the $n \times p$ observation matrix at time t, constructed by stacking n replicates from the model (2), w^t is a $p \times G$ matrix of suitably chosen weights and λ is a common

regularization parameter. The optimization problem can be separated into the following p penalized regression problems:

$$\hat{A}_{i:}^{1:T-1} = \underset{\theta^1,\cdots,\theta^{T-1}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\|\mathcal{X}_{:i}^T - \sum_{t=1}^{T-1}\mathcal{X}^{T-t}\theta^t\|^2 + \lambda\sum_{t=1}^{T-1}\sum_{g=1}^{G} w_{i,g}^t\|\theta_{[g]}^t\|, \quad i=1,\cdots,p. \tag{4}$$

The order d of the VAR model is estimated as $\hat{d} = \max_{1\le t\le T-1}\{t : \hat{A}^t \neq \mathbf{0}\}$.

Different choices of weights $w_{i:g}^t$ lead to different variants of NGC estimates. The regular NGC estimates correspond to the choices $w_{i,g}^t = 1$ or $\sqrt{k_g}$, while for adaptive group NGC estimates the weights are chosen as $w_{i,g}^t = \left\|\hat{A}_{i:[g]}^t\right\|^{-1}$, where $\hat{A}^t$ are obtained from a regular NGC estimation. For $\hat{A}_{i:[g]}^t = \mathbf{0}$, the weight $w_{i,g}^t$ is infinite, which is interpreted as discarding the variables in group g from the optimization problem.

Thresholded NGC estimates are calculated by a two-stage procedure. The first stage involves a regular NGC estimation procedure. The second stage uses a bi-level thresholding strategy on the estimates $\hat{A}^t$. First, the estimated groups with ℓ_2 norm less than a threshold ($\delta_{grp} = c\lambda$, $c > 0$) are set to zero. The second level of thresholding (within group) is applied if the *a priori* available grouping information is not entirely reliable. $\hat{A}_{ij}^t$within an estimated group $\hat{A}_{i:[g]}^t$ is thresholded to zero if $\left|\hat{A}_{ij}^t\right| / \left\|\hat{A}_{i:[g]}^t\right\|$ is less than a threshold $\delta_{misspec} \in (0,1)$. So, for every $t = 1,\ldots,T-1$, if $j \in \mathcal{G}_g$, the thresholded NGC estimates are

$$\tilde{A}_{ij}^t = \hat{A}_{ij}^t I\left\{\left|\hat{A}_{ij}^t\right| \ge \delta_{misspec}\left\|\hat{A}_{i:[g]}^t\right\|\right\} I\left\{\left\|\hat{A}_{i:[g]}^t\right\| \ge \delta_{grp}\right\}.$$

The tuning parameters λ_{grp} and $\delta_{misspec}$ are chosen via cross-validation. The rationale behind this thresholding strategy is discussed in Section 4.

3. Estimation Consistency of NGC estimates

In this section we establish the norm consistency of regular group NGC estimates. The regular NGC estimates in (3) are obtained by solving p separate group lasso programs with a *common* design matrix $\mathbf{X}_{n\times p(T-1)} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$. This design matrix has $\bar{p} = (T-1)p$ columns which can be partitioned into $\bar{G} = (T-1)G$ groups $\{\mathcal{G}_1,\ldots,\mathcal{G}_{\bar{G}}\}$. We denote the sample Gram matrix by $C = \mathcal{X}'\mathcal{X}/n$. For the i^{th} optimization problem, these $\bar{G} = (T-1)G$ groups are penalized by $\lambda_{(t-1)G+g} := \lambda\, w_{i,g}^t$, $1 \le t \le T-1$, $1 \le g \le G$, with the choice of weights $w_{i,g}^t$ described in Section 2. Following Lounici et al. (2011) one can establish a non-asymptotic upper bound on the ℓ_2 estimation error of the NGC estimates $\hat{A}^t$ under certain restricted eigenvalue (RE) assumptions. These assumptions are common in the literature of high-dimensional regression (Lounici et al., 2011; Bickel et al., 2009; van de Geer and Bühlmann, 2009) and are known to be sufficient to guarantee consistent estimation of the regression coefficients even when the design matrix is singular. Of main interest, however, is to investigate the validity of these assumptions in the context of NGC models. This issue is addressed in Proposition 3.2.

For $L > 0$, we say that a **Restricted Eigenvalue** (RE) assumption RE(s, L) is satisfied if there exists a positive number $\phi_{RE} = \phi_{RE}(s) > 0$ such that

$$\min_{\substack{J \subset \mathbb{N}_{\bar{G}}, |J| \le s \\ \Delta \in \mathbb{R}^{\bar{p}} \setminus \{\mathbf{0}\}}} \left\{ \frac{\|\mathbf{X}\Delta\|}{\sqrt{n}\|\Delta_{[J]}\|} : \sum_{g \in J^c} \lambda_g \|\Delta_{[g]}\| \le L \sum_{g \in J} \lambda_g \|\Delta_{[g]}\| \right\} \ge \phi_{RE}. \tag{5}$$

The following proposition provides a non-asymptotic upper bound on the ℓ_2-estimation error of the group NGC estimates under RE assumptions. The proof follows along the lines of Lounici et al. (2011) and is delegated to Appendix C.

Proposition 3.1 *Consider a regular NGC estimation problem* (4) *with* $s_{\max} = \max_{1 \le i \le p} s_i$ *and* $s = \sum_{i=1}^{p} s_i$. *Suppose* λ *in* (3) *is chosen large enough so that for some* $\alpha > 0$,

$$\lambda_g \ge \frac{2\sigma}{\sqrt{n}} \sqrt{\left\|C_{[g][g]}\right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log \bar{G}} \right) \quad \textit{for every } g \in \mathbb{N}_{\bar{G}}, \tag{6}$$

Also assume that the common design matrix $\mathbf{X} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$ *in the* p *regression problems* (4) *satisfy* $RE(2s_{\max}, 3)$. *Then, with probability at least* $1 - 2p\bar{G}^{1-\alpha}$,

$$\left\| \hat{A}^{1:T-1} - A^{1:T-1} \right\|_F \le \frac{4\sqrt{10}}{\phi_{RE}^2(2s_{\max})} \frac{\lambda_{\max}^2}{\lambda_{\min}} \sqrt{s}. \tag{7}$$

Remark. Consider a high-dimensional asymptotic regime where $\bar{G} \asymp n^B$ for some $B > 0$, $k_{\max}/k_{\min} = O(1)$, $s = O(n^{a_1})$ and $k_{\max} = O(n^{a_2})$ with $0 < a_1,\ a_2 < a_1 + a_2 < 1$ so that the total number of non-zero effects is $o(n)$. If $\{\|C_{[g][g]}\|,\ g \in \mathbb{N}_{\bar{G}}\}$ are bounded above (often accomplished by standardizing the data) and $\phi_{RE}^2(2s_{\max})$ is bounded away from zero (see Proposition 3.2 for more details), then the NGC estimates are norm consistent for any choice of $\alpha > 2 + a_2/B$.

Note that group lasso achieves faster convergence rate (in terms of estimation and prediction error) than lasso if the groups are appropriately specified. For example, if all the groups are of equal size k and $\lambda_g = \lambda$ for all g, then group lasso can achieve an ℓ_2 estimation error of order $O\left(\sqrt{s}(\sqrt{k} + \sqrt{\log \bar{G}})/\sqrt{n}\right)$. In contrast, lasso's error is known to be of the order $O\left(\sqrt{\|A^{1:d}\|_0 \log \bar{p}/n}\right)$, which establishes that group lasso has a lower error bound if $s \ll \|A^{1:d}\|_0$. On the other hand, lasso will have a lower error bound if $s \asymp \|A^{1:d}\|_0$, i.e., if the groups are highly misspecified.

Validity of RE assumption in Group NGC problems. In view of Theorem 3.1, it is important to understand how stringent the RE condition is in the context of NGC problems. It is also important to find a lower bound on the RE coefficient ϕ_{RE}, as it affects the convergence rate of the NGC estimates. For the panel-VAR setting, we can rigorously establish that the RE condition holds with overwhelming probability, as long as $n,\ p$ grow at the same rate required for ℓ_2-consistency.

The following proposition achieves this objective in two steps. Note that each row of the design matrix $\mathbf{X}$ (common across the p regressions) is independently distributed as $N(\mathbf{0}, \Sigma)$ where Σ is the variance-covariance matrix of the $(T-1)p$-dimensional random variable

$\left((X^1)', \ldots, (X^{T-1})'\right)'$. First, we exploit the spectral representation of the stationary VAR process to provide a lower bound on the minimum eigenvalue of Σ. In the next step, we establish a suitable deviation bound on $\mathbf{X} - \Sigma$ to prove that $\mathbf{X}$ satisfies RE condition with high probability for sufficiently large n.

Proposition 3.2 *(a) Suppose the VAR(d) model of* (2) *is stable, stationary. Let* Σ *be the variance-covariance matrix of the* $(T-1)p$*-dimensional random variable* $\left((X^1)', \ldots, (X^{T-1})'\right)'$. *Then the minimum eigenvalue of* Σ *satisfies*

$$\Lambda_{min}(\Sigma) \geq \sigma^2 \left[\max_{\theta \in [-\pi, \pi]} \|\mathcal{A}(e^{-i\theta})\|\right]^{-2} \geq \sigma^2 \left[1 + \sum_{t=1}^{d} \|A^t\|\right]^{-2} \geq \sigma^2 \left[1 + \frac{1}{2}(\mathbf{v}_{in} + \mathbf{v}_{out})\right]^{-2},$$

where $\mathcal{A}(z) := I - A^1 z - A^2 z^2 - \ldots - A^d z^d$ *is the reverse characteristic polynomial of the VAR(d) process, and* $\mathbf{v}_{in}$, $\mathbf{v}_{out}$ *are the maximum incoming and outgoing effects at a node, cumulated across different lags*

$$\mathbf{v}_{in} = \sum_{t=1}^{d} \max_{1 \leq i \leq p} \sum_{j=1}^{p} |A^t_{ij}|, \qquad \mathbf{v}_{out} = \sum_{t=1}^{d} \max_{1 \leq j \leq p} \sum_{i=1}^{p} |A^t_{ij}|.$$

(b) In addition, suppose the replicates from different panels are i.i.d. Then, for any $s > 0$, *there exist universal positive constants* c_i *such that if the sample size* n *satisfies*

$$n \quad > \frac{\Lambda^2_{\max}(\Sigma)}{\Lambda^2_{\min}(\Sigma)} (2 + L\lambda_{\max}/\lambda_{\min})^4 \, c_0 s(k_{\max} + c_1 \log(e\bar{G}/2s)),$$

then $\mathbf{X}$ *satisfies* $RE(s, L)$ *with* $\phi^2_{RE} \geq \Lambda_{\min}(\Sigma)/2$ *with probability at least* $1 - c_2 \exp(-c_3 n)$.

Remark. Proposition 3.2 has two interesting consequences. First, it provides a lower bound on the RE constant ϕ_{RE} which is independent of T. So if the high dimensionality in the Granger causal network arises only from the time domain and not the cross-section ($T \to \infty$, p, G fixed), the stationarity of the VAR process guarantees that the rate of convergence depends only on the true order (d), and not T. Second, this result shows that the NGC estimates are consistent even if the node capacities $\mathbf{v}_{in}$ and $\mathbf{v}_{out}$ grow with n, p at an appropriate rate.

4. Variable Selection Consistency of NGC estimates

In view of (4), to study the variable selection properties of NGC estimates it suffices to analyze the variable selection properties of p generic group lasso estimates with a common design matrix.

The problem of group sparsity selection has been thoroughly investigated in the literature (Wei and Huang, 2010; Lounici et al., 2011). The issue of selection and sign consistency within a group, however, is still unclear. Since group lasso does not impose sparsity within a group, all the group members are selected together (Huang et al., 2009) and it is not clear which ones are recovered with correct signs. This also leads to inconsistent variable selection if a group is misspecified, i.e., not all the members within a group have non-zero

effect. Several alternate penalized regression procedures have been proposed to overcome this shortcoming (Breheny and Huang, 2009; Huang et al., 2009). The main idea behind these procedures is to combine ℓ_2 and ℓ_1 norms in the penalty to encourage sparsity at both group and variable level. These estimators involve nonconvex optimization problems and are computationally expensive. Also their theoretical properties in a high dimensional regime are not well studied.

We take a different approach to deal with the issue of group misspecification. Although the group lasso penalty does not perform exact variable selection within groups, it performs regularization and shrinks the individual coefficients. We utilize this regularization to detect misspecification within a group. To this end, we formulate a generalized notion of sign consistency, henceforth referred as "direction consistency", that provides insight into the properties of group lasso estimates within a single group. Subsequently, these properties are used to develop a simple, easy to compute, thresholded variant of group lasso which, in addition to group selection, achieves variable selection and sign consistency within groups.

We consider a generic group lasso regression problem of the linear model $y = X\beta^0 + \epsilon$ with p variables partitioned into G non-overlapping groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$ of size k_g, $g = 1, \ldots, G$. Without loss of generality, we assume $\beta^0_{[g]} \neq \mathbf{0}$ for $g \in S = \{1, 2, \ldots, s\}$ and $\beta^0_{[g]} = \mathbf{0}$ for all $g \notin S$ and consider the following group lasso estimate of β^0:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\|, \tag{8}$$

$$\underbrace{\beta^0}_{p\times 1} = [\underbrace{\beta^0_{[1]}, \ldots, \beta^0_{[s]}}_{k_1+\ldots+k_s=q}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{p-q}] = [\beta^0_{(1)} : \beta^0_{(2)}], \tag{9}$$

$$\underbrace{\mathbf{X}}_{n\times p} = [\underbrace{\mathbf{X}_{(1)}}_{n\times q} : \underbrace{\mathbf{X}_{(2)}}_{n\times(p-q)}], \qquad C = \frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}. \tag{10}$$

Direction Consistency. For an m-dimensional vector $\tau \in \mathbb{R}^m \backslash \{\mathbf{0}\}$ define its direction vector $D(\tau) = \tau/\|\tau\|$, $D(\mathbf{0}) = \mathbf{0}$. In the context of a generic group lasso regression (10), for a group $g \in S$ of size k_g, $D(\beta^0_{[g]})$ indicates the direction of influence of $\beta^0_{[g]}$ at a group level in the sense that it reflects the relative importance of the influential members within the group. Note that for $k_g = 1$ the function $D(\cdot)$ simplifies to the usual $sgn(\cdot)$ function.

Definition. An estimate $\hat{\beta}$ of a generic group lasso problem (8) is ***direction consistent*** at a rate δ_n, if there exists a sequence of positive real numbers $\delta_n \to 0$ such that

$$\mathbb{P}\left(\|D(\hat{\beta}_{[g]}) - D(\beta^0_{[g]})\| < \delta_n,\ \forall g \in S,\ \hat{\beta}_{[g]} = \mathbf{0},\ \forall g \notin S\right) \to 1 \text{ as } n, p \to \infty. \tag{11}$$

Now suppose $\hat{\beta}$ is a direction consistent estimator. Consider the set $\tilde{S}^n_g := \{j \in \mathcal{G}_g : |{\beta^0}_j| \,/\, \|{\beta^0}_{[g]}\| > \delta_n\}$. $\tilde{S}^n_g$ can be viewed as a collection of influential group members within a group $\mathcal{G}_g$, which are "detectable" with a sample of size n. Then, it readily follows from the definition that

$$\mathbb{P}(sgn(\hat{\beta}_j) = sgn(\beta_j),\ \forall j \in \tilde{S}^n_g, \forall g \in \{1, \ldots, s\}) \to 1 \text{ as } n, p \to \infty. \tag{12}$$

The latter observation connects the precision of group lasso estimates to the accuracy of *a priori* available grouping information. In particular, if the pre-specified grouping structure is correct, i.e., all the members within a group have non-zero effects, then for a sufficiently large sample size we have $\tilde{S}_g^n = \mathcal{G}_g$ for all $g \in S$. Hence, if the group lasso estimate is direction consistent, it will correctly estimate the sign of all the variables in the support. On the other hand, in case of a misspecified *a priori* grouping structure (numerous zero coordinates in β_g for $g \in S$), group lasso will correctly estimate only the signs of the influential group members. This argument on zero vs. non-zero effects can be generalized to strong vs. weak effects, as well.

Example. We demonstrate the property of direction consistency using a small example. Consider a linear model with 8 predictors

$$y = 0.5x_1 - 3x_2 + 3x_3 + x_4 - 2x_5 + 3x_8 + e, \quad e \sim N(0,1).$$

The coefficient vector β^0 is partitioned into four groups of size 2, viz., $(0.5, -3)$, $(3, 1)$, $(-2, 0)$ and $(0, 3)$. The last two groups are misspecified. We generated $n = 25$ samples from this model and ran group lasso regression with the above group structure. Figure 2 shows the true coefficient vectors (solid) and their estimates (dashed) from five iterations of the above exercise. Note that even though the ℓ_2 errors between $\beta^0_{[g]}$ and $\hat{\beta}_{[g]}$ vary largely across the four groups, the distance between their projections on the unit circle, $\left\| D(\beta^0_{[g]}) - D(\hat{\beta}_{[g]}) \right\|$, are comparatively stable across groups. In fact, Theorem 4.1 shows that under certain irrepresentable conditions (IC) on the design matrix, it is possible to find a uniform (over all $g \in S$) upper bound δ_n on the ℓ_2 gap of these direction vectors. This motivates a natural thresholding strategy to correct for the misspecification in groups (cf. Proposition 4.2). Even though a group $\beta^0_{[g]}$ is misspecified (i.e., lies on a coordinate axis), direction consistency ensures, with high probability, that the corresponding coordinate in $D(\hat{\beta}_{[g]})$ will be smaller than a threshold δ_n which is common across all groups in the support.

Group Irrepresentable Conditions (IC). Next, we define the IC required for direction consistency of group lasso estimates. Irrepresentable conditions are common in the literature of high-dimensional regression problems (Zhao and Yu, 2006; van de Geer and Bühlmann, 2009) and are shown to be sufficient (and essentially necessary) for selection consistency of the lasso estimates. Further these conditions are known to be satisfied with high probability, if the population analogue of the Gram matrix belongs to the Toeplitz family (Zhao and Yu, 2006; Wainwright, 2009). In NGC estimation the population analogue of the Gram matrix $\Sigma = Var(\mathbf{X}^{1:(T-1)})$ is block Toeplitz, so the irrepresentable assumptions are natural candidates for studying selection consistency of the estimates. Consider the notations of (8) and (10). Define $K = diag\,(\lambda_1 \mathbf{I}_{k_1}, \lambda_2 \mathbf{I}_{k_2}, \ldots, \lambda_s \mathbf{I}_{k_s})$.

Uniform Irrepresentable Condition (IC) is satisfied if there exists $0 < \eta < 1$ such that for all $\tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \le g \le s} \|\tau_{[g]}\|_2 \le 1$,

$$\frac{1}{\lambda_g} \left\| \left[C_{21} (C_{11})^{-1} K \tau \right]_{[g]} \right\| < 1 - \eta, \ \forall g \notin S = \{1, \ldots, s\}. \tag{13}$$

Note that the definition reverts to the usual IC for lasso when all groups correspond are singletons.

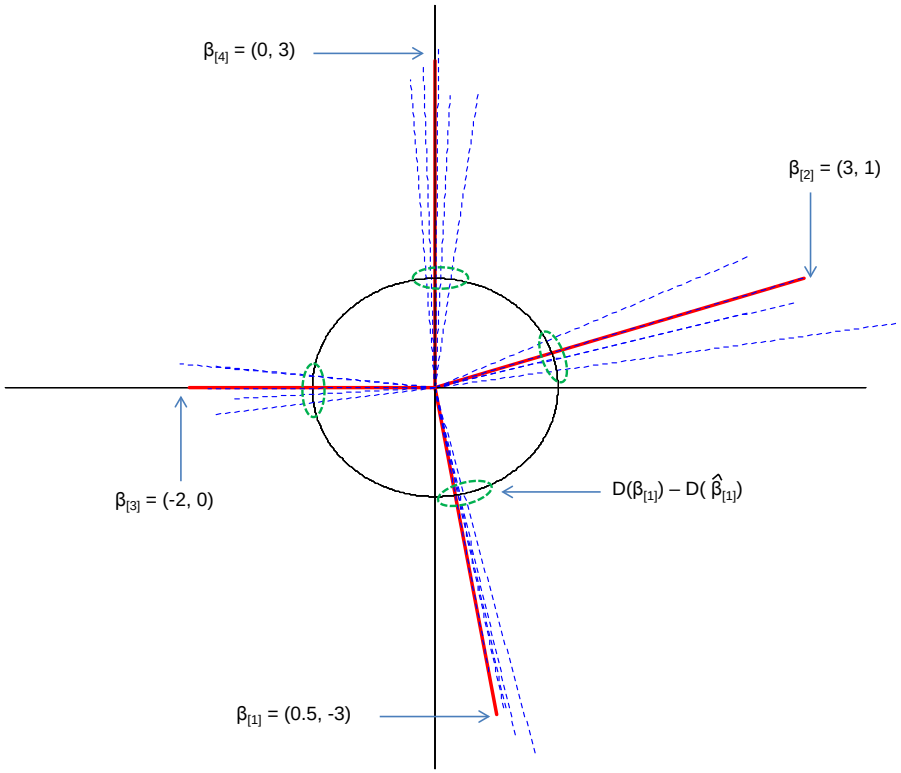


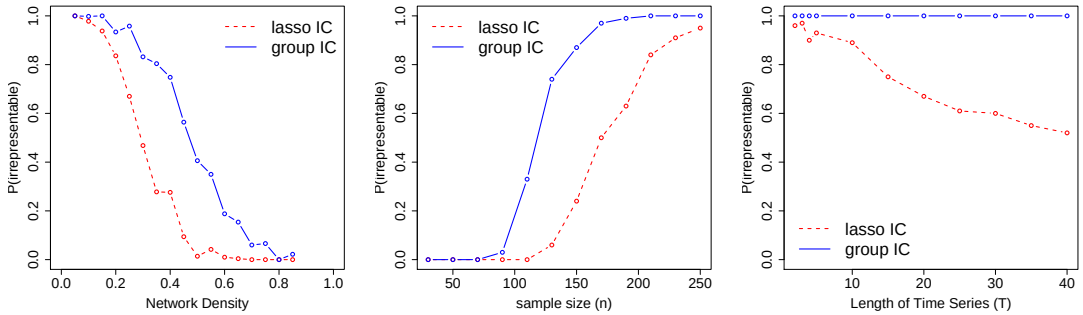
Figure 2: Example demonstrating direction consistency

The IC is more stringent than the RE condition and is rarely met if the underlying model is not sparse. It can be shown that a slightly weaker version of this condition is necessary for direction consistency. We refer the readers to Appendix D for further discussion on the different irrepresentable assumptions and their properties. Numerical evidence suggests that the group IC tends to be less stringent than the IC required for the selection consistency of lasso. We illustrate this using three small simulated examples.

Simulation 1. We constructed group sparse NGC models with $T = 5$, $p = 21$, $G = 7, k_g = 3$ and different levels of network densities, where the network edges were selected at random and scaled so that $\|A^1\| = 0.1$. For each of these models, we generated 100 samples of size $n = 150$ and calculated the proportions of times the two types of irrepresentable conditions were met. The results are displayed in Figure 3a.

Simulation 2. We selected a VAR(1) model from the above class and generated samples of size $n = 20, 50, \ldots, 250$. Figure 3b displays the proportions of times (based on 100 simulations) the two ICs were met.

Simulation 3. We generated $n = 200$ samples from the VAR(1) model of example 2 for $T = 2, 3, 4, 5, 10, \ldots, 40$. Figure 3c displays the proportions of times (based on 100 simulations) the two ICs were met.



(a) IC and Network Density (b) IC and Sample Size (c) IC and Time Series length

Figure 3: Comparison of lasso and group irrepresentable conditions in the context of group sparse NGC models. (a) group ICs tend to be met for dense networks where lasso IC fails to meet. (b) For the same network group IC is met with smaller sample size than required by lasso. (c) For longer time series group IC is satisfied more often than lasso IC.

Selection consistency for generic group lasso estimates. For simplicity, we discuss the selection consistency properties of a generic group lasso regression problem with a common tuning parameter across groups, i.e., $\lambda_g = \lambda$ for every $g \in \mathbb{N}_G$. Similar results can be obtained for more general choices of the tuning parameters.

Theorem 4.1 *Assume that the group uniform IC holds with $1-\eta$ for some $\eta > 0$. Then, for any choice of $\alpha > 0$,*

$$\lambda \geq \max_{g \notin S} \frac{1}{\eta} \frac{\sigma}{\sqrt{n}} \sqrt{\left\| (C_{22})_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right) \quad \text{and}$$

$$\delta_n \geq \max_{g \in S} \frac{1}{\left\| \beta^0_{[g]} \right\|} \left(\lambda \sqrt{s} \left\| (C_{11})^{-1} \right\| + \frac{\sigma}{\sqrt{n}} \sqrt{\left\| (C_{11})^{-1}_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right) \right),$$

with probability greater than $1 - 4G^{1-\alpha}$, there exists a solution $\hat{\beta}$ satisfying

1. *$\hat{\beta}_{[g]} = 0$ for all $g \notin S$,*
2. *$\left\| \hat{\beta}_{[g]} - \beta^0_{[g]} \right\| < \delta_n \left\| \beta^0_{[g]} \right\|$, and hence $\left\| D(\hat{\beta}_{[g]}) - D(\beta^0_{[g]}) \right\| < 2\delta_n$, for all $g \in S$. If $\delta_n < 1$, then $\hat{\beta}_{[g]} \neq 0$ for all $g \in S$.*

Remark. The tuning parameter λ can be chosen of the same order as required for ℓ_2 consistency to achieve selection consistency within groups in the sense of (12). Further, with the above choice of λ, δ_n can be chosen of the order of $O(\sqrt{s}(\sqrt{k_{max}} + \sqrt{\log G})/\sqrt{n})$. Thus, group lasso correctly identifies the group sparsity pattern and is direction consistent if $\sqrt{s}(\sqrt{k_{max}} + \sqrt{\log G})/\sqrt{n} \to 0$, the same scaling required for ℓ_2 consistency.

Thresholding in Group NGC estimators. As described in Section 2, regular group NGC estimates can be thresholded both at the group and coordinate levels. The first level of thresholding is motivated by the fact that lasso can select too many false positives [cf. van de Geer et al. (2011), Zhou (2010) and the references therein]. The second level of thresholding employs the direction consistency of regular group NGC estimates to perform within group variable selection with high probability. The following proposition demonstrates the benefit of these two types of thresholding. The second result is an immediate corollary of Theorem 4.1. Proof of the first result (thresholding at group level) requires some additional notations and is delegated to Appendix E.

Theorem 4.2 *Consider a generic group lasso regression problem* (8) *with common tuning parameter* $\lambda_g = \lambda$.
(i) Assume the RE(s, 3) condition of (5) *holds with a constant* ϕ_{RE} *and define* $\hat{\beta}_{[g]}^{thgrp} = \hat{\beta}_{[g]}\mathbf{1}_{\|\hat{\beta}_{[g]}\|>4\lambda}$. *If* $\hat{S} = \{g \in \mathbb{N}_G : \hat{\beta}_{[g]}^{thgrp} \neq \mathbf{0}\}$, *then* $|\hat{S}\backslash S| \leq \frac{s}{\phi_{RE}^2/12}$, *with probability at least* $1 - 2G^{1-\alpha}$.
(ii) Assume that uniform IC holds with $1 - \eta$ *for some* $\eta > 0$. *Choose* λ *and* δ_n *as in Theorem 4.1 and define*

$$\hat{\beta}_j^{thgrp} = \hat{\beta}_j \mathbf{1}\{|\hat{\beta}_j|/\|\hat{\beta}_{[g]}\| > 2\,\delta_n\} \text{ for all } j \in \mathcal{G}_g.$$

Then $sgn(\beta_j^0) = sgn(\hat{\beta}_j^{thgrp})\ \forall\, j \in \mathbb{N}_p$ *with probability at least* $1 - 4G^{1-\alpha}$, *if* $\min_{j \in supp(\beta^0)} |\beta_j^0| > 2\delta_n \|\beta_{[g]}^0\|$ *for all* $j \in \mathcal{G}_g$, *i.e., if the effect of every non-zero member in a group is "visible" relative to the total effect from the group.*

5. Performance Evaluation

We evaluate the performances of regular, adaptive and thresholded variants of the group NGC estimators through an extensive simulation study, and compare the results to those obtained from lasso estimates. The R package `grpreg` (Breheny and Huang, 2009) was used to obtain the group lasso estimates. The settings considered are:
(a) *Balanced groups of equal size*: i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 5$ time points, comprising of $p = 60, 120, 200$ nodes partitioned into groups of equal size in the range 3-5.
(b) *Unbalanced groups:* We retain the same setting as before, however the corresponding node set is partitioned into one larger group of size 10 and many groups of size 5.
(c) *Misspecified balanced groups:* i.i.d samples of size $n = 60, 110, 160$ are generated from lag-2 ($d = 2$) VAR models on $T = 10$ time points, comprising of $p = 60, 120$ nodes partitioned into groups of size 6. Further, for each group there is a 30% misspecification rate, namely that for every parent group of a downstream node, 30% of the group members do not exert any effect on it.

Using a 19 : 1 sample-splitting, the tuning parameter λ is chosen from an interval of the form $[C_1\lambda_e, C_2\lambda_e]$, $C_1, C_2 > 0$, where $\lambda_e = \sqrt{2 \log p/n}$ for lasso and $\sqrt{2 \log G/n}$ for group lasso. The thresholding parameters are selected as $\delta_{grp} = 0.7\lambda\sigma$ at the group level and $\delta_{misspec} = n^{-0.2}$ within groups. These parameters are chosen by conducting a 20-fold cross-validation on independent tuning data sets of same sizes, using intervals of the form

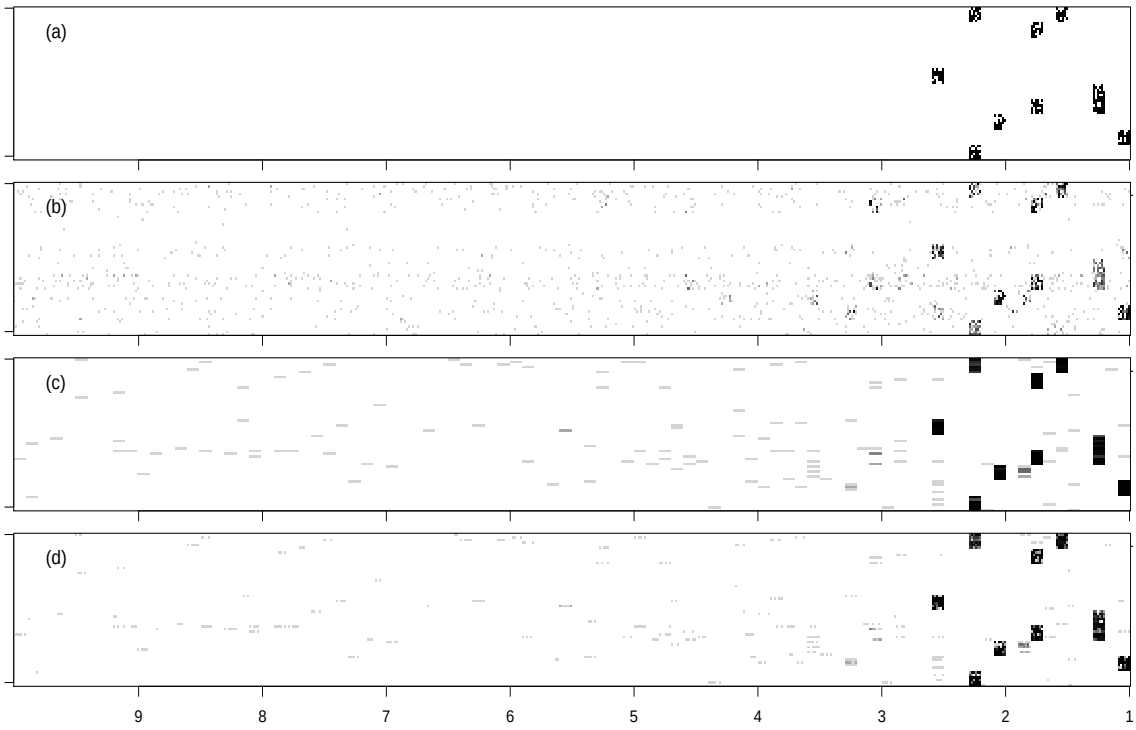


Figure 4: Estimated adjacency matrices of a misspecified NGC model with p = 60, T = 10, n = 60: (a) True, (b) Lasso, (c) Group Lasso, (d) Thresholded Group Lasso. The grayscale represents the proportion of times an edge was detected in 100 simulations.

$[C_3\lambda, C_4\lambda]$ for δ_{grp} and $\{n^{-\delta}, \delta \in [0,1]\}$ for $\delta_{misspec}$. Finally, within group thresholding is applied only when the group structure is misspecified.

The following performance metrics were used for comparison purposes: (i) $Precision = TP/(TP+FP)$, (ii) $Recall = TP/(TP+FN)$ and (iii) Matthew's Correlation coefficient (MCC) defined as

$$\frac{(TP \times TN) - (FP \times FN)}{((TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN))^{1/2}},$$

where TP, TN, FP and FN correspond to true positives, true negatives, false positives and false negatives in the estimated network, respectively. The average and standard deviations (over 100 replicates) of the performance metrics are presented for each setup.

The results for the balanced settings are given in Table 1. The Recall for $p = 60$ shows that even for a network with $60 \times (5-1) = 240$ nodes and $|E| = 351$ true edges, the group NGC estimators recover about 71% of the true edges with a sample size as low as $n = 60$, while lasso based NGC estimates recover only 31% of the true edges. The three group NGC estimates have comparable performances in all the cases. However thresholded lasso shows slightly higher precision than the other group NGC variants for smaller sample sizes (e.g., $n = 60, p = 200$). The results for $p = 60, n = 110$ also display that lower precision of

		$p = 60$, $\|E\| = 351$ Group Size=3			$p = 120$, $\|E\| = 1404$ Group Size=3			$p = 200$, $\|E\| = 3900$ Group Size=5		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	80(2)	75(2)	66(4)	69(1)	62(2)	52(2)	52(1)	47(1)	38(1)
	Grp	95(2)	91(4)	83(7)	91(3)	80(5)	68(7)	78(4)	72(3)	59(6)
	Thgrp	96(1)	92(3)	86(6)	93(3)	83(5)	70(7)	82(4)	76(3)	64(6)
	Agrp	96(2)	92(4)	83(7)	92(3)	82(5)	69(7)	81(3)	74(3)	60(6)
R	Lasso	71(2)	54(2)	31(2)	54(1)	40(1)	22(1)	38(1)	28(1)	15(1)
	Grp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(1)	70(2)	41(4)
	Thgrp	99(1)	93(3)	71(7)	91(2)	81(2)	48(8)	84(2)	69(2)	41(3)
	Agrp	99(1)	93(3)	71(7)	91(2)	81(2)	47(8)	84(1)	69(2)	40(4)
MCC	Lasso	75(2)	63(2)	45(3)	60(1)	49(1)	33(1)	43(1)	35(1)	23(1)
	Grp	97(1)	92(3)	76(5)	91(1)	80(2)	56(2)	81(2)	70(2)	48(2)
	Thgrp	98(1)	93(2)	78(5)	92(1)	81(2)	57(3)	83(2)	72(2)	50(3)
	Agrp	97(1)	92(3)	76(5)	91(1)	81(2)	56(3)	82(2)	71(2)	48(2)
ERR	Lasso	10.5	11.3	13.9	16.63	17.37	16.69	19.79	20	18.52
LAG	Grp	3.19	6.95	12.76	4.86	10.77	12.65	4.21	5.27	7.8
	Thgrp	2.83	5.87	10.01	3.98	9.03	11.19	3.06	3.91	5.68
	Agrp	3.13	6.89	12.59	4.63	10.37	12.34	3.58	4.87	7.59

Table 1: Performance of different regularization methods in estimating graphical Granger causality with **balanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

lasso is caused partially by its inability to estimate the order of the VAR model correctly, as measured by ERR LAG=Number of falsely connected edges from lags beyond the true order of the VAR model divided by the number of edges in the network ($|E|$). This finding is nicely illustrated in Figure 4 and Table 1. The group penalty encourages edges from the nodes of the same group to be picked up together. Since the nodes of the same group are also from the same time lag, the group variants have substantially lower ERR LAG. For example, average ERR LAG of lasso for $p = 200,\ n = 160$ is 19.79% while the average ERR LAGs for the group lasso variants are in the range $3.06\% - 4.21\%$.

The results for the unbalanced networks are given in Table 2. As in the balanced group setup, in almost all the simulation settings the group NGC variants outperform the lasso estimates with respect to all three performance metrics. However the performances of the different variants of group NGC are comparable and tend to have higher standard deviations than the lasso estimates. Also the average ERR LAGs for the group NGC variants are substantially lower than the average ERR LAG for lasso demonstrating the advantage of group penalty. Although the conclusions regarding the comparisons of lasso and group NGC estimates remain unchanged it is evident that the performances of all the estimators are affected by the presence of one large group, skewing the uniform nature of the network. For example the MCC measures of group NGC estimates in a balanced network with $p = 60$ and $|E| = 351$ vary around $97 - 98\%$ which lowers to $89\% - 90\%$ when the groups are unbalanced.

The results for misspecified groups are given in Table 3. Note that for higher sample size n, the MCC of lasso and regular group lasso are comparable. However, the thresholded version of group lasso achieves significantly higher MCC than the rest. This demonstrates the advantage of using the directional consistency of group lasso estimators to perform

		$p = 60$, $\|E\| = 450$ Groups=$1 \times 10, 11 \times 5$			$p = 120$, $\|E\| = 1575$ Groups=$1 \times 10, 23 \times 5$			$p = 200$, $\|E\| = 4150$ Groups=$1 \times 10, 39 \times 5$		
	n	160	110	60	160	110	60	160	110	60
P	Lasso	72(2)	69(3)	62(2)	51(1)	48(1)	41(1)	61(1)	53(1)	42(2)
	Grp	84(4)	79(6)	76(9)	55(5)	47(5)	40(6)	86(3)	77(5)	66(7)
	Thgrp	86(4)	82(7)	78(11)	60(6)	50(7)	40(5)	88(2)	79(6)	69(6)
	Agrp	85(3)	81(5)	77(9)	59(5)	51(5)	42(6)	88(2)	78(5)	67(6)
R	Lasso	45(2)	35(2)	22(2)	43(1)	34(1)	22(1)	23(1)	15(0)	7(0)
	Grp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
	Thgrp	95(2)	88(4)	62(8)	89(3)	77(4)	50(5)	73(3)	50(6)	21(5)
	Agrp	94(3)	87(5)	61(8)	88(2)	75(5)	48(6)	73(3)	49(6)	22(5)
MCC	Lasso	56(2)	48(2)	35(2)	46(1)	39(1)	29(1)	36(1)	28(1)	17(1)
	Grp	89(3)	82(4)	67(5)	68(3)	58(3)	42(3)	79(1)	61(3)	37(3)
	Thgrp	90(3)	84(4)	68(6)	72(4)	61(4)	43(2)	80(1)	62(3)	37(3)
	Agrp	89(3)	83(4)	67(6)	71(3)	60(3)	43(3)	79(1)	61(3)	37(3)
ERR	Lasso	10.59	10.74	11.76	18.3	18.72	18.76	11.54	10.93	9.29
LAG	Grp	7.04	9.85	13.04	12.53	14.71	13.06	4.8	6.41	6.85
	Thgrp	6.58	8.98	11.1	9.6	11.9	10.9	4.06	5.65	5.7
	Agrp	6.74	9.19	12.96	10.81	12.78	11.79	4.55	6.2	6.81

Table 2: Performance of different regularization methods in estimating graphical Granger causality with **unbalanced** group sizes and no misspecification; $d = 2$, $T = 5$, $SNR = 1.8$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

		$p = 60$, $\|E\| = 246$ Group Size=6			$p = 120$, $\|E\| = 968$ Group Size=6		
	n	160	110	60	160	110	60
P	Lasso	88(2)	85(3)	77(5)	59(1)	55(1)	49(2)
	Grp	65(2)	66(2)	66(3)	43(3)	44(4)	38(4)
	Thgrp	87(3)	88(3)	85(3)	56(6)	56(6)	51(7)
	Agrp	65(2)	66(2)	66(3)	45(2)	45(4)	39(4)
R	Lasso	80(3)	63(3)	37(2)	66(1)	54(1)	35(1)
	Grp	100(0)	98(2)	82(6)	87(2)	78(3)	59(4)
	Thgrp	100(0)	98(2)	79(6)	86(2)	79(3)	57(4)
	Agrp	100(0)	98(2)	82(6)	86(2)	78(3)	58(3)
MCC	Lasso	84(2)	73(2)	53(3)	62(1)	54(1)	41(1)
	Grp	81(1)	80(2)	74(4)	61(2)	58(3)	47(2)
	Thgrp	93(2)	93(2)	82(4)	69(4)	66(4)	53(3)
	Agrp	81(1)	80(2)	74(4)	62(2)	59(2)	47(2)
ERR	Lasso	12.63	17.05	22.41	45.09	49.68	53.4
LAG	Grp	9.43	8.78	15.12	18.22	18.43	29.26
	Thgrp	6.45	5.34	8.02	11.81	12.84	15.57
	Agrp	9.11	8.78	14.96	16.32	16.9	27.69

Table 3: Performance of different regularization methods in estimating graphical Granger causality with **misspecified** groups (30% misspecification); $d = 2$, $T = 10$, $SNR = 2$. Precision (P), Recall (R), MCC are given in percentages (numbers in parentheses give standard deviations). ERR LAG gives the error associated with incorrect estimation of VAR order.

	Lasso	Grp	Agrp	Thgrp
mean	0.649	0.456	0.457	0.456
stdev	0.340	0.252	0.251	0.252

Table 4: Mean and standard deviation of MSE for different NGC estimates

within group variable selection. We would like to mention here that a careful choice of the thresholding parameters δ_{grp} and $\delta_{misspec}$ via cross-validation improves the performance of thresholded group lasso; however, we do not pursue these methods here as they require grid search over many tuning parameters or an efficient estimator of the degree of freedom of group lasso.

In summary, the results clearly show that all variants of group lasso NGC outperform the lasso-based ones, whenever the grouping structure of the variables is known and correctly specified. Further, their performance depends on the composition of group sizes. On the other hand, if the a priori known group structure is moderately misspecified lasso estimates produce comparable results to regular and adaptive group NGC ones, while thresholded group estimates outperform all other methods, as expected.

6. Application

Example: T-cell activation. Estimation of gene regulatory networks from expression data is a fundamental problem in functional genomics (Friedman, 2004). Time course data coupled with NGC models are informationally rich enough for the task at hand. The data for this application come from Rangel et al. (2004), where expression patterns of genes involved in T-cell activation were studied with the goal of discovering regulatory mechanisms that govern them in response to external stimuli. Activated T-cells are involved in regulation of effector cells (e.g., B-cells) and play a central role in mediating immune response. The available data comprising of $n = 44$ samples of $p = 58$ genes, measure the cells response at 10 time points, $t = 0, 2, 4, 6, 8, 18, 24, 32, 48, 72$ hours after their stimulation with a T-cell receptor independent activation mechanism. We concentrate on data from the first 5 time points, that correspond to early response mechanisms in the cells.

Genes are often grouped based on their function and activity patterns into biological pathways. Thus, the knowledge of gene functions and their membership in biological pathways can be used as inherent grouping structures in the proposed group lasso estimates of NGC. Towards this, we used available biological knowledge to define groups of genes based on their biological function. Reliable information for biological functions were found from the literature for 38 genes, which were retained for further analysis. These 38 genes were grouped into 13 groups with the number of genes in different groups ranging from 1 to 5.

Figure 5 shows the estimated networks based on lasso and thresholded group lasso estimates, where for ease of representation the nodes of the network correspond to groups of genes. In this case, estimates from variants of group NGC estimator were all similar, and included a number of known regulatory mechanisms in T-cell activation, not present in the regular lasso estimate. For instance, Waterman et al. (1990) suggest that TCF plays a significant role in activation of T-cells, which may describe the dominant role of this group of genes in the activation mechanism. On the other hand, Kim et al. (2005) suggest that

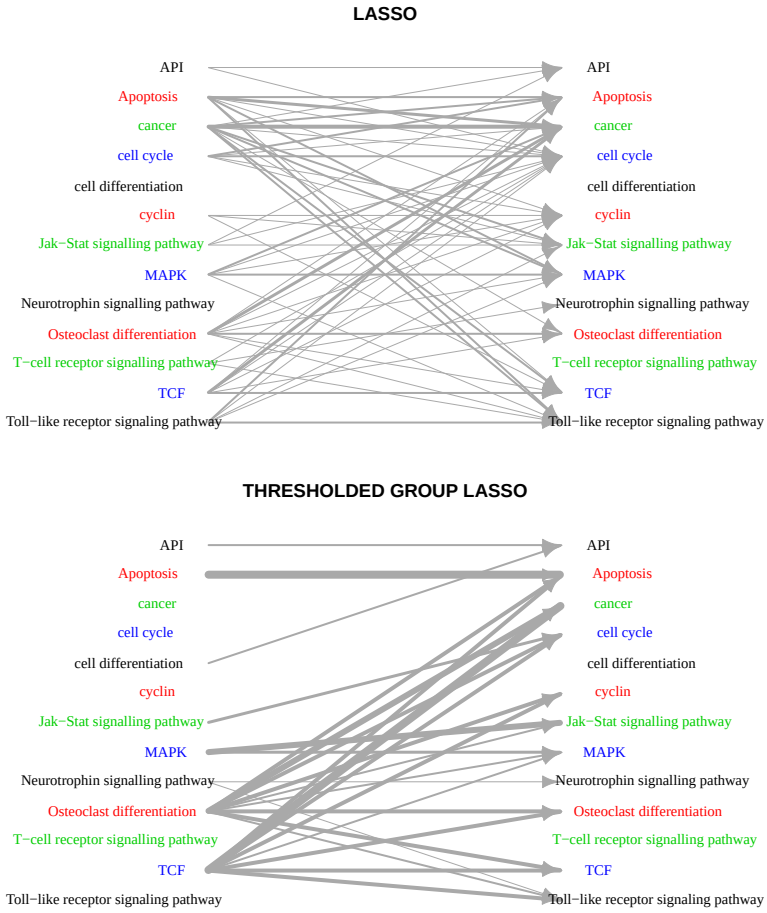


Figure 5: Estimated Gene Regulatory Networks of T-cell activation. Width of edges represent the number of effects between two groups, and the network represents the aggregated regulatory network over 3 time points.

activated T-cells exhibit high levels of osteoclast-associated receptor activity which may attribute the large number of associations between member of osteoclast differentiation and other groups. Finally, the estimated networks based on variants of group lasso estimator also offer improved estimation accuracy in terms of mean squared error (MSE) despite having having comparable complexities to their regular lasso counterpart (Table 4), which further confirms the findings of other numerical studies in that paper.

Example: Banking balance sheets application. In this application, we examine the structure of the balance sheets in terms of assets and liabilities of the $n = 50$ largest (in terms of total balance sheet size) US banking corporations. The data cover 9 quarters (September 2009-September 2011) and were directly obtained from the Federal Deposit Insurance Corporation (FDIC) database (available at `www.fdic.gov`). The $p = 21$ variables

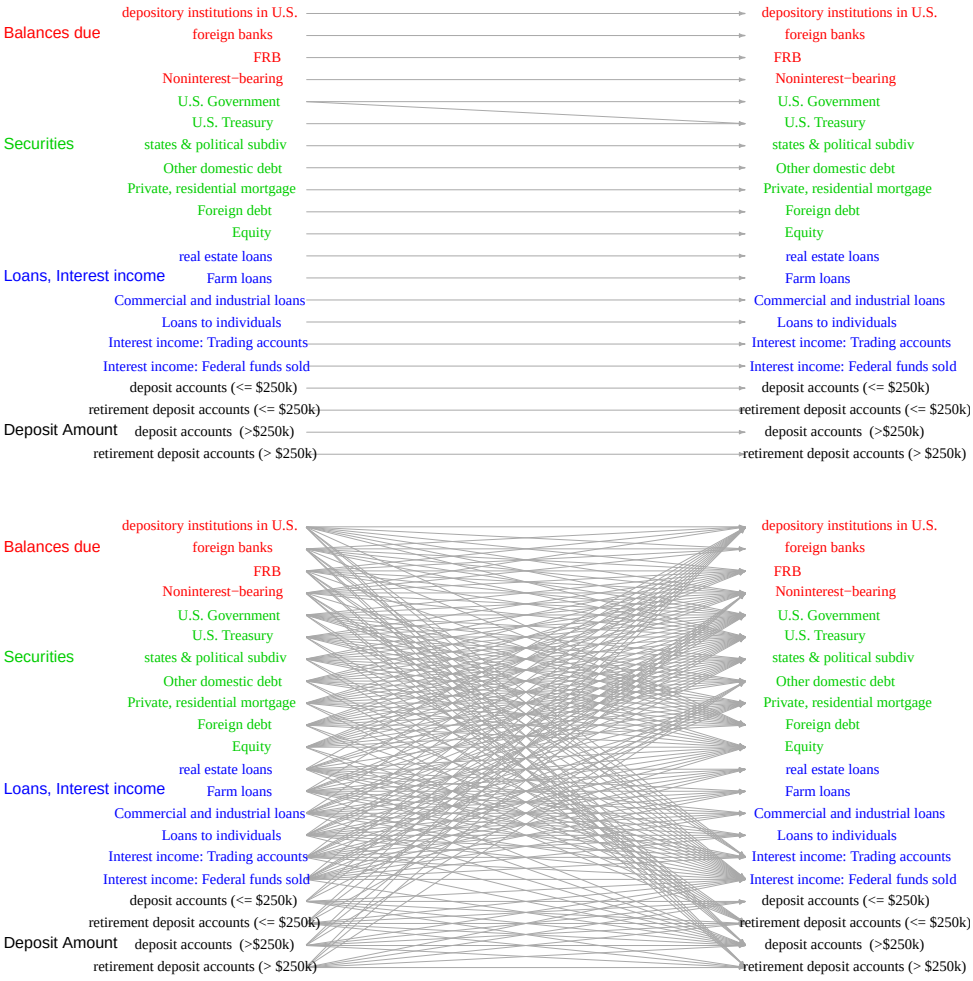


Figure 6: Estimated Networks of banking balance sheet variables using (a) lasso and (b) group lasso. The networks represent the aggregated network over 5 time points.

correspond to different assets (US and foreign government debt securities, equities, loans (commercial, mortgages), leases, etc.) and liabilities (domestic and foreign deposits from households and businesses, deposits from the Federal Reserve Board, deposits of other financial institutions, non-interest bearing liabilities, etc.) We have organized them into four categories: two for the assets (loans and securities) and two for the liabilities (Balances Due and Deposits, based on a $250K reporting FDIC threshold). Amongst the 50 banks examined, one discerns large integrated ones with significant retail, commercial and investment activities (e.g., Citibank, JP Morgan, Bank of America, Wells Fargo), banks primarily focused on investment business (e.g., Goldman Sachs, Morgan Stanley, American Express, E-Trade, Charles Schwab), regional banks (e.g., Banco Popular de Puerto Rico, Comerica Bank, Bank of the West).

Quarter	Lasso	Grp	Agrp	Thgrp
Dec 2010	1.59 (0.29)	0.36 (0.05)	0.36 (0.05)	0.37 (0.05)
Mar 2011	1.46 (0.30)	0.47 (0.23)	0.47 (0.23)	0.46 (0.22)
Jun 2011	1.33 (0.26)	0.36 (0.11)	0.36 (0.11)	0.35 (0.11)
Sep 2011	1.72 (0.32)	0.50 (0.18)	0.50 (0.18)	0.47 (0.16)

Table 5: Mean and standard deviation (in parentheses) of PMSE (MSE in case of Dec 2010) for prediction of banking balance sheet variables.

The raw data are reported in thousands of dollars. The few missing values were imputed using a nearest neighbor imputation method with $k = 5$, by clustering them according to their total assets in the most recent quarter in the data collection period (September 2011) and subsequently every missing observation for a particular bank was imputed by the median observation on its five nearest neighbors. The data were log-transformed to reduce non-stationarity issues. The data set was restructured as a panel with $p = 21$ variables and $n = 50$ replicates observed over $T = 9$ time points. Every column of replicates was scaled to have unit variance.

We applied the proposed variants of NGC estimates on the first $T = 6$ time points (Sep 2009 - Dec 2010) of the above panel data set. The parameters λ and δ_{grp} were chosen using a $19 : 1$ sample-splitting method and the misspecification threshold $\delta_{misspec}$ was set to zero as the grouping structure was reliable. We calculated the MSE of the fitted model in predicting the outcomes in the four quarters (December 2010 - September 2011). The Predicted MSE (MSE for Dec 2010) are listed in Table 5. The estimated network structures are shown in Figure 6.

It can be seen that the lasso estimates recover a very simple temporal structure amongst the variables; namely, that past values (in this case lag-1) influence present ones. Given the structure of the balance sheet of large banks, this is an anticipated result, since it can not be radically altered over a short time period due to business relationships and past commitments to customers of the bank. However, the (adaptive) group lasso estimates reveal a richer and more nuanced structure. Examining the fitted values of the adjacency matrices A^t, we notice that the dominant effects remain those discovered by the lasso estimates. However, fairly strong effects are also estimated within each group, but also between the groups of the assets (loans and securities) on the balance sheet. This suggests rebalancing of the balance sheet for risk management purposes between relatively low risk securities and potentially more risky loans. Given the period covered by the data (post financial crisis starting in September 2009) when credit risk management became of paramount importance, the analysis picks up interesting patterns. On the other hand, significant fewer associations are discovered between the liabilities side of the balance sheet. Finally, there exist relationships between deposits and securities such as US Treasuries and other domestic ones (primarily municipal bonds); the latter indicates that an effort on behalf of the banks to manage the credit risk of their balance sheets, namely allocating to low risk assets as opposed to more risky loans.

It is also worth noting that the group lasso model exhibits superior predictive performance over the lasso estimates, even 4 quarters into the future. Finally, in this case the

thresholded estimates did not provide any additional benefits over the regular and adaptive variants, given that the specification of the groups was based on accounting principles and hence correctly structured.

7. Discussion

In this paper, the problem of estimating Network Granger Causal (NGC) models with inherent grouping structure is studied when replicates are available. Norm, and both group level and within group variable selection consistency are established under fairly mild assumptions on the structure of the underlying time series. To achieve the second objective the novel concept of direction consistency is introduced.

The type of NGC models discussed in this study have wide applicability in different areas, including genomics and economics. However, in many contexts the availability of replicates at each time point is not feasible (e.g., in rate of returns for stocks or other macroeconomic variables), while grouping structure is still present (e.g., grouping of stocks according to industry sector). Hence, it is of interest to study the behavior of group lasso estimates in such a setting and address the technical challenges emanating from such a pure time series (dependent) data structure.

Acknowledgments

We thank the action editor and three anonymous reviewers for their helpful comments. The work of SB and GM was supported in part by DoD grant W81XWH-12-1-0130, and that of GM by NSF DMS-1106695 and NSA H98230-10-1-0203. The work of AS was partially supported by NSF grant DMS-1161565 and NIH grant 1R21GM101719-01A1.

Appendix A. Auxiliary Lemmas

Lemma A.1 (Characterization of the Group lasso estimate) *A vector $\hat{\beta} \in \mathbb{R}^p$ is a solution to the convex optimization problem*

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|Y - X\beta\|^2 + \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\| \tag{14}$$

if and only if $\hat{\beta}$ satisfies, for some $\tau \in \mathbb{R}^p$ with $\max_{1 \le g \le G} \left\|\tau_{[g]}\right\| \le 1$, $\frac{1}{n}\left[X'(Y - X\hat{\beta})\right]_{[g]} = \lambda_g \tau_{[g]} \, \forall g$. Further, $\tau_{[g]} = D\left(\hat{\beta}_{[g]}\right)$ whenever $\hat{\beta}_{[g]} \neq \mathbf{0}$.

Proof Follows directly from the KKT conditions for the optimization problem (14). ■

Lemma A.2 (Concentration bound for multivariate Gaussian) *Let $Z_{k\times 1} \sim N(0, \Sigma)$. Then, for any $t > 0$, the following inequalities hold:*

$$\mathbb{P}\left(\|\|Z\| - \mathbb{E}\|Z\|\| > t\right) \le 2 \exp\left(-\frac{2t^2}{\pi^2 \|\Sigma\|}\right), \qquad \mathbb{E}\|Z\| \le \sqrt{k}\sqrt{\|\Sigma\|}.$$

Proof The first inequality can be found in Ledoux and Talagrand (1991) (equation (3.2). To establish the second inequality note that,

$$\mathbb{E}\|Z\| \leq \sqrt{\mathbb{E}\|Z\|^2} = \sqrt{\mathbb{E}\left[\mathrm{tr}\left(ZZ'\right)\right]} = \sqrt{\mathrm{tr}\left(\Sigma\right)} \leq \sqrt{k}\sqrt{\|\Sigma\|}.$$

■

Lemma A.3 *Let $\beta,\ \hat{\beta} \in \mathbb{R}^m \backslash \{\mathbf{0}\}$. Let $\hat{u} = \hat{\beta} - \beta$ and $r = D(\hat{\beta}) - D(\beta)$. Then $\|r\| < 2\delta$ whenever $\|\hat{u}\| < \delta \|\beta\|$.*

Proof It follows from $\|\hat{u}\| < \delta \|\beta\|$ that

$$(1-\delta)\|\beta\| < \|\beta\| - \|\hat{u}\| \leq \|\hat{\beta}\| \leq \|\hat{u}\| + \|\beta\| < (1+\delta)\|\beta\|\,,$$

which implies that $\left|\|\beta\| - \|\hat{\beta}\|\right| < \delta\|\beta\|$. Now,

$$\|\hat{\beta}\|\,\|\beta\|\|r\| = \left\|\ \hat{\beta}\|\beta\| + (\hat{u}-\hat{\beta})\|\hat{\beta}\|\ \right\| \leq \left\|\hat{\beta}\left(\|\beta\| - \|\hat{\beta}\|\right) + \|\hat{\beta}\|\ \hat{u}\ \right\| < \|\hat{\beta}\|\,\|\beta\|(\delta+\delta),$$

since $\left|\|\beta\| - \|\hat{\beta}\|\right| < \delta\|\beta\|$ and $\|\hat{u}\| < \delta\,\|\beta\|$. ■

Lemma A.4 *Let $\mathcal{G}_1, \ldots, \mathcal{G}_G$ be any partition of $\{1, \ldots, p\}$ into G non-overlapping groups and $\lambda_1, \ldots, \lambda_G$ be positive real numbers. Define the cone sets $\mathcal{C}(J, L) = \{v \in \mathbb{R}^p : \sum_{g \notin J} \lambda_g \|v_{[g]}\| \leq L \sum_{g \in J} \lambda_g \|v_{[g]}\|\}$ for any subset of groups $J \subseteq \mathbb{N}_G$. Also define the set of group s-sparse vectors $\mathbb{D}(s) := \{v \in \mathbb{R}^p : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J$ for some $J \subseteq \mathbb{N}_G,\ |J| \leq s\}$. Then*

$$\bigcup_{J \subseteq \mathbb{N}_G, |J| \leq s} \mathcal{C}(J, L) \cap \mathbb{S}^{p-1} \subseteq (2 + L')cl\{conv\{\mathbb{D}(s)\}\}, \tag{15}$$

where $L' = L\lambda_{\max}/\lambda_{\min}$, $\mathbb{S}^{p-1} = \{v \in \mathbb{R}^p : \|v\| = 1\}$ is the ball of unit norm vectors in $\mathbb{R}^p$ and $cl\{.\}$, $conv\{.\}$ respectively denote the closure and convex hull of a set.

Proof Note that for any $J \subseteq \mathbb{N}_G$, $|J| \leq s$, and $v \in \mathcal{C}(J, L) \cap \mathbb{S}^{p-1}$, we have

$$\sum_{g \notin J} \|v_{[g]}\| \leq L \frac{\lambda_{\max}}{\lambda_{\min}} \sum_{g \in J} \|v_{[g]}\|,$$

which implies

$$\|v\|_{2,1} \leq (L'+1) \sum_{g \in J} \|v_{[g]}\| \leq (L'+1)\sqrt{s}\|v_{[J]}\| \leq (L'+1)\sqrt{s}.$$

Hence the union of the cone sets on the left hand side of (15) is a subset of $A := \{v \in \mathbb{R}^p : \|v\| \leq 1,\ \|v\|_{2,1} \leq (L'+1)\sqrt{s}\}$.

We will show that the set A is a subset of $B := (2 + L')cl\{conv\{\mathbb{D}(s)\}\}$, the closed convex hull on the right hand side of (15). Since both sets A and B are closed convex, it is enough to show that the support function of A is dominated by the support function of B.

The support function of A is given by $\phi_A(z) = \sup_{\theta \in A}\langle \theta, z\rangle$. For any $z \in \mathbb{R}^p$, let $S \subseteq \{1, \ldots, G\}$ be a subset of top s groups in terms of the ℓ_2 norm of $z_{[g]}$. Thus, $\|z_{[S^c]}\|_{2,\infty} \leq \|z_{[g]}\|$ for all $g \in S$. This implies $\|z_{[S^c]}\|_{2,\infty} \leq (1/s)\|z_{[S]}\|_{2,1} \leq (1/\sqrt{s})\|z_{[S]}\|$. So, we have

$$\begin{aligned} \phi_A(z) = \sup_{\theta \in A}\langle \theta, z\rangle &\leq \sup_{\|\theta_{[S]}\| \leq 1} \langle \theta_{[S]}, z_{[S]}\rangle + \sup_{\|\theta_{[S^c]}\|_{2,1} \leq \sqrt{s}(L'+1)} \langle \theta_{[S^c]}, z_{[S^c]}\rangle \qquad (16) \\ &\leq \|z_{[S]}\| + (L'+1)\sqrt{s}\|z_{[S^c]}\|_{2,\infty} \leq (L'+2)\|z_{[S]}\|. \qquad (17) \end{aligned}$$

On the other hand, support function of $B := (L'+2)cl\{conv\{\mathbb{D}(s)\}\}$ is given by

$$\phi_B(z) = \sup_{\theta \in B}\langle \theta, z\rangle = (L'+2) \max_{|U|=s,\ U \subseteq \mathbb{N}_G} \sup_{\|\theta_{[U]}\| \leq 1} \langle \theta_{[U]}, z_{[U]}\rangle = (L'+2)\|z_{[S]}\|.$$

This concludes the proof. ■

Lemma A.5 *Consider a matrix $X_{n\times p}$ with rows independently distributed as $N(0, \Sigma)$, $\Lambda_{\min}(\Sigma) > 0$. Let $\mathcal{G}_1, \ldots, \mathcal{G}_G$ be any partition of $\{1, \ldots, p\}$ into G non-overlapping groups of size $k_1, \ldots, k_g$, respectively. Let $C = X'X/n$ denote the sample Gram matrix and $\mathbb{D}(s)$ denote the set of group s-sparse vectors defined in Lemma A.4. Then, for any integer $s \geq 1$ and any $\eta > 0$, we have*

$$\begin{aligned} &\mathbb{P}\left[\sup_{v \in cl\{conv\{\mathbb{D}(s)\}\}} |v'(C - \Sigma)v| > 6\eta\|\Sigma\|\right] \\ \leq\ & c_0 \exp\left[-n\min\{\eta, \eta^2\} + c_1 s(k_{\max} + c_2 \log(eG/2s))\right] \qquad (18) \end{aligned}$$

for some universal positive constants c_i.

Proof We consider a fixed vector $v \in \mathbb{R}^p$ with $\|v\| \leq 1$, the support of which can be covered by a set J of at most s groups, i.e., $supp(v) \subseteq \mathcal{G}_J$, $J \subseteq \mathbb{N}_G$, $|J| \leq s$. Define $Y = Xv$. Then each coordinate of Y is independently distributed as $N(0, \sigma_y^2)$, where $\sigma_y^2 = v'\Sigma v \leq \|\Sigma\|$.

Then, for any $\eta > 0$, Hanson-Wright inequality of Rudelson and Vershynin (2013) ensures

$$\mathbb{P}\left[\left|v'(C - \Sigma)v\right| > \eta\|\Sigma\|\right] \leq \mathbb{P}\left[\frac{1}{n}\left|Y'Y - \mathbb{E}Y'Y\right| > \eta\sigma_y^2\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\}\right].$$

Next, we extend this deviation bound on all vectors v in the sparse set

$$\mathbb{D}(2s) = \{v \in \mathbb{R}^p : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J \text{ for some } J \subseteq \mathbb{N}_G,\ |J| \leq 2s\}. \qquad (19)$$

For a given $J \subseteq \mathbb{N}_G$, $|J| = 2s$, we define $\mathbb{D}_J = \{v \in \mathbb{R}^p : \|v\| \leq 1, supp(v) \subseteq \mathcal{G}_J\}$ and note that $\mathbb{D}(2s) = \cup_{|J|=2s}\mathbb{D}_J$. For an $\epsilon > 0$ to be specified later, we construct an ϵ-net $\mathcal{A}$ of $\mathbb{D}_J$. Since $\sum_{g \in J} k_g \leq 2s\, k_{\max}$, it is possible to construct such a net $\mathcal{A}$ with cardinality at most $(1 + 2/\epsilon)^{2s\, k_{\max}}$ (Vershynin, 2009).

We want a tail inequality for $M := \sup_{v \in \mathbb{D}_J} |v'\Delta v|$, where $\Delta = C - \Sigma$. Since $\mathcal{A}$ is an ϵ-cover of $\mathbb{D}_J$, for any $v \in \mathbb{D}_J$, there exists $v_0 \in \mathcal{A}$ such that $w = v - v_0$ satisfies $\|w\| \leq \epsilon$. Then

$$|v'\Delta v| = |(w + v_0)'\Delta(w + v_0)| \leq |w'\Delta w| + |v_0'\Delta v_0| + 2|v_0'\Delta w|.$$

Taking supremum over all $v \in \mathbb{D}_J$, and noting that $w/\epsilon \in \mathbb{D}_J$, we obtain

$$M \leq \epsilon^2 M + \max_{v_0 \in \mathcal{A}} |v_0'\Delta v_0| + \sup_{u,v \in \mathbb{D}_J} 2\epsilon |u'\Delta v|. \tag{20}$$

To upper bound the third term, note that $(u + v)/2 \in \mathbb{D}_J$, and

$$2|u'\Delta v| \leq |(u + v)'\Delta(u + v)| + |u'\Delta u| + |v'\Delta v|.$$

Hence

$$\sup_{u,v \in \mathbb{D}_J} 2\epsilon |u'\Delta v| \leq 4\epsilon M + \epsilon M + \epsilon M = 6\epsilon M.$$

From equation (20), we now have

$$M \leq (1 - 6\epsilon - \epsilon^2)^{-1} \max_{v_0 \in \mathcal{A}} |v_0'\Delta v_0|.$$

Choosing $\epsilon > 0$ small enough so that $(1 - 6\epsilon - \epsilon^2) > 1/2$, we obtain

$$\begin{aligned} \mathbb{P}\left[\sup_{v \in \mathbb{D}_J} |v'\Delta v| > 2\eta\|\Sigma\|\right] &\leq \mathbb{P}\left[\max_{v_0 \in \mathcal{A}} |v_0'\Delta v_0| > \eta\|\Sigma\|\right] \\ &\leq 2\,(1 + 2/\epsilon)^{2s\,k_{\max}} \exp[-cn\min\{\eta, \eta^2\}]. \end{aligned}$$

Taking supremum over $\begin{pmatrix} G \\ 2s \end{pmatrix} \leq (eG/2s)^{2s}$ choices of J, we get

$$\mathbb{P}\left[\sup_{v \in \mathbb{D}(2s)} |v'\Delta v| > 2\eta\|\Sigma\|\right] \leq 2\exp\left[-cn\min\{\eta, \eta^2\} + 2s\log\left(\frac{eG}{2s}\right) + 2s\,k_{\max}\log\left(1 + \frac{2}{\epsilon}\right)\right].$$

In order to extend this deviation inequality to $cl\{conv\{\mathbb{D}(s)\}\}$, we note that any v in the convex hull of $\mathbb{D}(s)$ can be expressed as $v = \sum_{i=1}^m \alpha_i v_i$, where $v_1, \ldots, v_m$ are in $\mathbb{D}(s)$ and $0 \leq \alpha_i \leq 1$, $\sum \alpha_i = 1$. Then

$$|v'\Delta v| \leq \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j |v_i'\Delta v_j|.$$

Also, for every $i,\ j$, $(v_i + v_j)/2 \in \mathbb{D}(2s)$, and

$$|v_i'\Delta v_j| \leq \frac{1}{2}\left[|(v_i + v_j)'\Delta(v_i + v_j)| + |v_i'\Delta v_i| + |v_j'\Delta v_j|\right].$$

Hence

$$\sup_{v \in conv\{\mathbb{D}(s)\}} |v'\Delta v| \leq \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j\ \frac{1}{2}[4 + 1 + 1] \sup_{v \in \mathbb{D}(2s)} |v'\Delta v|.$$

Together with the continuity of quadratic forms, this implies

$$\sup_{v \in cl\{conv\{\mathbb{D}(s)\}\}} |v'\Delta v| \leq 3 \sup_{v \in \mathbb{D}(2s)} |v'\Delta v|.$$

The result then readily follows from the above deviation inequality. ■

Appendix B. Proof of Main Results

Proof [Proof of Proposition 3.2] (a) Note that Σ is a $p(T-1) \times p(T-1)$ block Toeplitz matrix with $(i,j)^{th}$ block $(\Sigma_{ij})_{1 \leq i,j \leq (T-1)} := \Gamma(i-j)$, where $\Gamma(\ell)_{p \times p}$ is the autocovariance function of lag ℓ for the zero-mean VAR(d) process (2), defined as $\Gamma(\ell) = \mathbb{E}[\mathbf{X}^t(\mathbf{X}^{\mathbf{t}-\ell})']$.

We consider the cross spectral density of the VAR(d) process (2)

$$f(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\ell\theta}, \quad \theta \in [-\pi, \pi]. \tag{21}$$

From standard results of spectral theory we know that $\Gamma(\ell) = \int_{-\pi}^{\pi} e^{i\ell\theta} f(\theta)\, d\theta$, for every ℓ.

We want to find a lower bound on the minimum eigenvalue of Σ, i.e., $\inf_{\|x\|=1} x'\Sigma x$. Consider an arbitrary $p(T-1)$-variate unit norm vector x, formed by stacking the p-tuples $x^1, \ldots, x^{T-1}$.

For every $\theta \in [-\pi, \pi]$, define $G(\theta) = \sum_{t=1}^{T-1} x^t\, e^{-it\theta}$ and note that

$$\begin{aligned}
\int_{-\pi}^{\pi} G^*(\theta) G(\theta)\, d\theta &= \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)'(x^\tau) \int_{-\pi}^{\pi} e^{i(t-\tau)\theta}\, d\theta \\
&= \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)'(x^\tau)\ (2\pi\, \mathbf{1}_{\{t=\tau\}}) = 2\pi \sum_{t=1}^{T-1} (x^t)'(x^t) = 2\pi\ \|x\|^2 = 2\pi.
\end{aligned}$$

Also let $\mu(\theta)$ be the minimum eigenvalue of the Hermitian matrix $f(\theta)$. Following Parter (1961) we have the result

$$\begin{aligned}
x'\Sigma x &= \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)'\Gamma(t-\tau)x^\tau = \sum_{t=1}^{T-1} \sum_{\tau=1}^{T-1} (x^t)' \left(\int_{-\pi}^{\pi} e^{i(t-\tau)\theta} f(\theta) d\theta \right) x^\tau \\
&= \int_{-\pi}^{\pi} \left(\sum_{t=1}^{T-1} (x^t)' e^{it\theta} \right) f(\theta) \left(\sum_{\tau=1}^{T-1} x^\tau e^{-i\tau\theta} \right) d\theta = \int_{-\pi}^{\pi} G^*(\theta)\, f(\theta)\, G(\theta)\, d\theta \\
&\geq \int_{-\pi}^{\pi} \mu(\theta)\, (G^*(\theta)G(\theta))\ d\theta \geq \left(\min_{\theta \in (-\pi,\pi)} \mu(\theta) \right) \int_{-\pi}^{\pi} G^*(\theta) G(\theta)\, d\theta = 2\pi \min_{\theta \in (-\pi,\pi)} \mu(\theta).
\end{aligned}$$

So $\Lambda_{min}(\Sigma) \geq 2\pi \min_{\theta \in (-\pi,\pi)} \mu(\theta)$. Since $\mathcal{A}(z) = I - A^1 z - A^2 z^2 - \ldots - A^d z^d$ is the (matrix-valued) characteristic polynomial of the VAR(d) model (2), we have the following representation of the spectral density (see Priestley, 1981, eqn 9.4.23):

$$f(\theta) = \frac{1}{2\pi} \sigma^2 (\mathcal{A}(e^{-i\theta}))^{-1} (\mathcal{A}^*(e^{-i\theta}))^{-1}.$$

Thus, $2\pi\mu(\theta) = 2\pi\Lambda_{min}(f(\theta)) = 2\pi/\Lambda_{max}(f(\theta)^{-1}) \geq \sigma^2/\left\|\mathcal{A}(e^{-i\theta})\right\|^2$. But $\left\|\mathcal{A}(e^{-i\theta})\right\| \leq 1 + \sum_{t=1}^{d}\left\|A^t\right\|$ for every $\theta \in [-\pi, \pi]$. The result then follows at once from the standard matrix norm inequality (see e.g., Golub and Van Loan, 1996, Cor 2.3.2)

$$\|A^t\|_2 \leq \sqrt{\|A^t\|_1\|A^t\|_\infty} \leq \frac{\|A^t\|_1 + \|A^t\|_\infty}{2} \quad t = 1, \ldots, d,$$

where

$$\|A^t\|_1 = \max_{1\leq i\leq p}\sum_{j=1}^{p}|A^t_{ij}|, \quad \|A^t\|_\infty = \max_{1\leq j\leq p}\sum_{i=1}^{p}|A^t_{ij}|.$$

(b) The first part of the proposition ensures that $\Lambda_{min}(\Sigma) \geq \sigma^2\left[1 + \frac{1}{2}(\mathbf{v}_{in} + \mathbf{v}_{out})\right]^{-2}$. If the replicates available from different panels are i.i.d, each row of the design matrix is independently and identically distributed according to a $N(\mathbf{0}, \Sigma)$ distribution.

To show that RE(s, L) of (5) holds with high probability for sufficiently large n, it is enough to show that

$$\min_{\substack{v\in\mathcal{C}(J,L)\setminus\{0\}\\ J\subset\mathbb{N}_{\bar{G}},\ |J|\leq s}} \frac{1}{n}\frac{\|\mathbf{X}v\|^2}{\|v\|^2} \geq \phi^2_{RE} \tag{22}$$

holds with high probability, where the cone sets $\mathcal{C}(J, L)$ are defined as

$$\mathcal{C}(J, L) := \{v \in \mathbb{R}^{\bar{p}} : \sum_{g\notin J}\lambda_g\|v_{[g]}\| \leq L\sum_{g\in J}\lambda_g\|v_{[g]}\|\} \tag{23}$$

for all $J \subset \mathbb{N}_{\bar{G}}$ with $|J| \leq s$. Denote the ball of unit norm vectors in $\mathbb{R}^{\bar{p}}$ by $\mathbb{S}^{\bar{p}-1}$. By scale invariance of $\|\mathbf{X}v\|^2/n\|v\|^2$, it is enough to show that with high probability

$$\min_{\substack{v\in\mathbb{S}^{\bar{p}-1}\cap\mathcal{C}(J,L)\\ J\subset\mathbb{N}_{\bar{G}},\ |J|\leq s}} v'Cv \geq \phi^2_{RE}, \tag{24}$$

where $C = \mathbf{X}'\mathbf{X}/n$ is the sample Gram matrix.

By part (a), we already know that $v'\Sigma v \geq \Lambda_{\min}(\Sigma) > 0$ for all $v \in \mathbb{S}^{\bar{p}-1}$. So we only need to show that $|v'(C - \Sigma)v| \leq \Lambda_{\min}(\Sigma)/2$ with high probability, uniformly on the set

$$\bigcup_{J\subseteq\mathbb{N}_{\bar{G}},|J|\leq s}\mathcal{C}(J, L)\cap\mathbb{S}^{\bar{p}-1}. \tag{25}$$

The proof relies on two key parts. In the first part, we use an extremal representation to show that the above union of the cone sets sits within the closed convex hull of a suitably defined set of group s-sparse vectors. In particular, it follows from Lemma A.4 that

$$\bigcup_{J\subseteq\mathbb{N}_{\bar{G}},\ |J|\leq s}\mathcal{C}(J, L)\cap\mathbb{S}^{\bar{p}-1} \subseteq (L' + 2)cl\{conv\{\mathbb{D}(s)\}\}, \tag{26}$$

where $\mathbb{D}(s) = \{v \in \mathbb{R}^{\bar{p}} : \|v\| \leq 1,\ supp(v) \subseteq \mathcal{G}_J \text{ for some } J \subseteq \mathbb{N}_{\bar{G}},\ |J| \leq s\}$, $L' = L\lambda_{\max}/\lambda_{\min}$ and $cl\{.\}$, $conv\{.\}$ respectively denote the closure and convex hull of a set.

The next part of the proof is an upper bound on the tail probability of $v'(C-\Sigma)v$, uniformly over all $v \in cl\{conv\{\mathbb{D}(s)\}\}$, presented in Lemma A.5. In particular, setting $\eta = \Lambda_{\min}(\Sigma)/12\|\Sigma\|(2+L')^2$ in the above lemma yields

$$\mathbb{P}\left[\sup_{v\in(2+L')cl\{conv\{\mathbb{D}(s)\}\}} |v'(C-\Sigma)v| > \Lambda_{\min}(\Sigma)/2\right] \leq c_0 \exp[-c_1\, n] \tag{27}$$

for the proposed choice of n. Together with the lower bound on $\Lambda_{\min}(\Sigma)$ established in part (a), this concludes the proof. ■

Proof [Proof of Theorem 4.1] Consider any solution $\hat{\beta}_R \in \mathbb{R}^q$ of the restricted regression

$$\underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \frac{1}{2n}\left\|\mathbf{Y} - X_{(1)}\beta\right\|_2^2 + \lambda\sum_{g=1}^{s}\left\|\beta_{[g]}\right\|_2 \tag{28}$$

and set $\hat{\beta} = \left[\hat{\beta}_R' : \mathbf{0}_{1\times(p-q)}\right]'$. We show that such an augmented vector $\hat{\beta}$ satisfies the statements of Theorem 4.1 with high probability.

Let $\hat{u} = \hat{\beta}_{(1)} - \beta^0_{(1)} = \hat{\beta}_R - \beta^0_{(1)}$. In view of lemmas A.1 and A.3, it suffices to show that the following events happen with probability at least $1 - 4\,G^{1-\alpha}$:

$$\left\|\hat{u}_{[g]}\right\| < \delta_n \left\|\beta^0_{[g]}\right\|, \text{ for all } g \in S, \tag{29}$$

$$\frac{1}{n}\left\|\left[X'\left(\epsilon - X_{(1)}\hat{u}\right)\right]_{[g]}\right\| \leq \lambda, \text{ for all } g \notin S. \tag{30}$$

Note that, in view of Lemma A.1, $\hat{u} = (C_{11})^{-1}\left(\frac{1}{\sqrt{n}}Z_{(1)} - \lambda\tau\right)$ for some $\tau \in \mathbb{R}^q$ with $\left\|\tau_{[g]}\right\| \leq 1$ for all $g \in S$, and $Z = \frac{1}{\sqrt{n}}X'\epsilon = \left[Z'_{(1)} : Z'_{(2)}\right]'$. Thus, for any $g \in S$,

$$\begin{aligned}\mathbb{P}\left(\left\|\hat{u}_{[g]}\right\| > \delta_n\left\|\beta^0_{[g]}\right\|\right) &\leq \mathbb{P}\left(\left\|\left[(C_{11})^{-1}\left(\frac{1}{\sqrt{n}}Z_{(1)} - \lambda\tau\right)\right]_{[g]}\right\| > \delta_n\left\|\beta^0_{[g]}\right\|\right)\\ &\leq \mathbb{P}\left(\left\|\left[(C_{11})^{-1}Z_{(1)}\right]_{[g]}\right\| > \sqrt{n}\left[\delta_n\left\|\beta^0{}_{[g]}\right\| - \lambda\left\|\left[(C_{11})^{-1}\tau\right]_{[g]}\right\|\right]\right).\end{aligned}$$

Note that $V = (C_{11})^{-1}Z_{(1)} \sim N(\mathbf{0}, \sigma^2(C_{11})^{-1})$. So $V_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{11}^{[g][g]})$, where $\Sigma^{[g][g]} := (\Sigma^{-1})_{[g][g]}$. Also, by the second statement of lemma A.2 we have $\mathbb{E}\left\|V_{[g]}\right\| \leq \sigma\sqrt{k_g}\sqrt{\left\|C_{11}^{[g][g]}\right\|}$. Therefore $\mathbb{P}\left(\left\|\hat{u}_{[g]}\right\| > \delta_n\left\|\beta^0_{[g]}\right\|\right)$ is bounded above by

$$\begin{aligned}&\mathbb{P}\left(\left|\left\|V_{[g]}\right\| - \mathbb{E}\left\|V_{[g]}\right\|\right| > \sqrt{n}\left[\delta_n\left\|\beta^0{}_{[g]}\right\| - \lambda\left\|(C_{11})^{-1}\right\|\sqrt{s}\right] - \sigma\sqrt{k_g\left\|C_{11}^{[g][g]}\right\|}\right)\\ \leq\, &2\exp\left[-\frac{2}{\pi^2\sigma^2\|C_{11}^{[g][g]}\|}\left(\sqrt{n}\delta_n\|\beta^0_{[g]}\| - \sqrt{n}\lambda\|C_{11}^{-1}\|\sqrt{s} - \sigma\sqrt{k_g\|C_{11}^{[g][g]}\|}\right)^2\right].\end{aligned}$$

For the proposed choice of δ_n, this expression is bounded above by $2\,G^{-\alpha}$.
Next, for any $g \notin S$, we get

$$\mathbb{P}\left(\frac{1}{n}\left\|\left[X'\left(\epsilon - X_{(1)}\hat{u}\right)\right]_{[g]}\right\| > \lambda\right) \leq \mathbb{P}\left(\left\|\left[Z_{(2)} - C_{21}C_{11}^{-1}Z_{(1)}\right]_{[g]}\right\| > \sqrt{n}\lambda\left(1 - \left\|\left[C_{21}C_{11}^{-1}\tau\right]_{[g]}\right\|\right)\right).$$

Defining $W = Z_{(2)} - C_{21}C_{11}^{-1}Z_{(1)} \sim N(\mathbf{0}, \sigma^2(C_{22} - C_{21}C_{11}^{-1}C_{12}))$, the uniform irrepresentable condition implies that the above probability is bounded above by $\mathbb{P}\left(\left\|W_{[g]}\right\| > \sqrt{n}\lambda\eta\right)$.

It can then be seen that $W_{[g]} \sim N(\mathbf{0}, \sigma^2\bar{C}_{[g][g]})$, where $\bar{C} = C_{22} - C_{21}C_{11}^{-1}C_{12}$ denotes the Schur complement of C_{22}. As before, lemma A.2 establishes that

$$\begin{aligned}\mathbb{P}\left(\left\|W_{[g]}\right\| > \sqrt{n}\lambda\eta\right) &\leq \mathbb{P}\left(\left|\left\|W_{[g]}\right\| - \mathbb{E}\left\|W_{[g]}\right\|\right| > \sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right) \\ &\leq 2\exp\left[-\frac{2}{\pi^2\|\sigma^2\bar{C}_{[g][g]}\|}\left(\sqrt{n}\lambda\eta - \sigma\sqrt{k_g\|\bar{C}_{[g][g]}\|}\right)^2\right],\end{aligned}$$

and the last probability is bounded above by $2G^{-\alpha}$ for the proposed choice of λ.
The results in the proposition follow by considering the union bound on the two sets of the probability statements made across all $g \in \mathbb{N}_G$. ■

Appendix C. Proof of results on ℓ_2-consistency

We first note that each of the p optimization problems in (4) is essentially a generic group lasso regression on n independent samples from a linear model $Y = X\beta^0 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \sum_{g=1}^{\bar{G}} \lambda_g\|\beta_{[g]}\|, \tag{31}$$

where $\mathbf{Y}_{n\times 1} = \mathcal{X}_i^T$, $\mathbf{X}_{n\times\bar{p}} = [\mathcal{X}^1 : \cdots : \mathcal{X}^{T-1}]$, $\beta^0_{\bar{p}\times 1} = vec(A_{i:}^{1:(T-1)})$, $\{1, \ldots, \bar{p}\} = \cup_{g=1}^{\bar{G}}\mathcal{G}_g$, $\bar{p} = (T-1)p$, $\bar{G} = (T-1)G$ and $\lambda_g = \lambda w_{i,g}^t$. In Proposition C.1, we first establish the upper bounds on estimation error in the context of a generic group lasso penalized regression problem. The results for regular group NGC then readily follows by applying the above Proposition on the p separate regressions.

Recall the Restricted Eigenvalue assumption required for the derivation of ℓ_2 estimation and prediction error. Following van de Geer and Bühlmann (2009), we introduce a slightly weaker notion called **Group Compatibility** (GC). For a constant $L > 0$ we say that GC(S, L) condition holds, if there exists a constant
$\phi_{compatible} = \phi_{compatible}(S, L) > 0$ such that

$$\min_{\Delta\in\mathbb{R}^p\setminus\{\mathbf{0}\}}\left\{\frac{\left(\sum_{g\in S}\lambda_g^2\right)^{1/2}\|X\Delta\|}{\sqrt{n}\sum_{g\in S}\lambda_g\|\Delta_{[g]}\|} : \sum_{g\notin S}\lambda_g\|\Delta_{[g]}\| \leq L\sum_{g\in S}\lambda_g\|\Delta_{[g]}\|\right\} \geq \phi_{compatible}. \tag{32}$$

The fact that GC(S, L) holds whenever RE(s, L) is satisfied (and $\phi_{RE} \leq \phi_{compatible}$) follows at once from Cauchy Schwarz inequality. We shall derive upper bounds on the prediction and $\ell_{2,1}$ estimation error of group lasso estimates involving the compatibility constant. This notion will also be used later to connect the irrepresentable conditions to the consistency results of group lasso estimators.

Proposition C.1 *Suppose the GC condition* (32) *holds with* $L = 3$. *Choose* $\alpha > 0$ *and denote* $\lambda_{min} = \min_{1 \leq g \leq G} \lambda_g$. *If*

$$\lambda_g \geq \frac{2\sigma}{\sqrt{n}} \sqrt{\left\| C_{[g][g]} \right\|} \left(\sqrt{k_g} + \frac{\pi}{\sqrt{2}} \sqrt{\alpha \log G} \right)$$

for every $g \in \mathbb{N}_G$, *then, the following statements hold with probability at least* $1 - 2G^{1-\alpha}$,

$$\frac{1}{n} \left\| X \left(\hat{\beta} - \beta^0 \right) \right\|^2 \leq \frac{16}{\phi^2_{compatible}} \sum_{g=1}^{s} \lambda_g^2, \tag{33}$$

$$\|\hat{\beta} - \beta^0\|_{2,1} \leq \frac{16}{\phi^2_{compatible}} \frac{\sum_{g=1}^{s} \lambda_g^2}{\lambda_{min}}. \tag{34}$$

If, in addition, RE(2s, 3) holds, then, with the same probability we get

$$\|\hat{\beta} - \beta^0\| \leq \frac{4\sqrt{10}}{\phi^2_{RE}(2s)} \frac{\sum_{g=1}^{s} \lambda_g^2}{\lambda_{min} \sqrt{s}}. \tag{35}$$

Proof [Proof of Proposition (C.1)] Since $\hat{\beta}$ is a solution of the optimization problem (31), for all $\beta \in \mathbb{R}^p$, we have

$$\frac{1}{n} \|Y - X\hat{\beta}\|^2 + 2 \sum_{g=1}^{G} \lambda_g \|\hat{\beta}_{[g]}\| \leq \frac{1}{n} \|Y - X\beta\|^2 + 2 \sum_{g=1}^{G} \lambda_g \|\beta_{[g]}\|.$$

Plugging in $Y = X\beta^0 + \epsilon$, and simplifying the resulting equation, we get

$$\frac{1}{n} \|X(\hat{\beta} - \beta^0)\|^2 \leq \frac{1}{n} \|X(\beta - \beta^0)\|^2 + \frac{2}{n} \sum_{g=1}^{G} \left\| (X'\epsilon)_{[g]} \right\| \left\| (\hat{\beta} - \beta)_{[g]} \right\|$$
$$+ 2 \sum_{g=1}^{G} \lambda_g \left(\|\beta_{[g]}\| - \|\hat{\beta}_{[g]}\| \right).$$

Fix $g \in \mathbb{N}_G$ and consider the event $\mathcal{A}_g = \left\{ \epsilon \in \mathbb{R}^n : \frac{2}{n} \left\| (X'\epsilon)_{[g]} \right\| \leq \lambda_g \right\}$. Note that $Z = \frac{1}{\sqrt{n}} X'\epsilon \sim N(\mathbf{0}, \sigma^2 C)$. So $Z_{[g]} \sim N(\mathbf{0}, \sigma^2 C_{[g][g]})$. Then,

$$\begin{aligned} \mathbb{P}\left(\mathcal{A}_g^c\right) &= \mathbb{P}\left(\left\| Z_{[g]} \right\| > \frac{1}{2} \lambda_g \sqrt{n} \right) \\ &\leq \mathbb{P}\left(\left| \left\| Z_{[g]} \right\| - \mathbb{E} \left\| Z_{[g]} \right\| \right| > \frac{\lambda_g \sqrt{n}}{2} - \sigma \sqrt{k_g} \sqrt{\left\| C_{[g][g]} \right\|} \right), \end{aligned}$$

where the last inequality follows from the second statement of Lemma A.2. Now, let $x_g = \frac{\lambda_g\sqrt{n}}{2} - \sigma\sqrt{k_g}\sqrt{\left\|C_{[g][g]}\right\|}$. Then, for $x_g > 0$, if

$$2\exp\left(-\frac{2\,x_g^2}{\pi^2\sigma^2\left\|C_{[g][g]}\right\|}\right) \leq 2\,G^{-\alpha}\,,$$

we get

$$\mathbb{P}\left(\mathcal{A}_g^c\right) \leq 2G^{-\alpha}.$$

But this happens if,

$$\sqrt{2}x_g \geq \sqrt{\alpha\,\log\,G}\pi\sigma\sqrt{\left\|C_{[g][g]}\right\|},$$

which is ensured by the proposed choice of λ_g.

Next, define $\mathcal{A} := \cap_{g=1}^{G}\mathcal{A}_g$. Then, $\mathbb{P}\left(\mathcal{A}\right) \geq 1 - 2G^{1-\alpha}$, and on the event $\mathcal{A}$, we have, for all $\beta \in \mathbb{R}^p$,

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|^2 + \sum_{g=1}^{G}\lambda_g\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| \leq \frac{1}{n}\|X(\beta - \beta^0)\|^2 + 2\sum_{g=1}^{G}\lambda_g\left(\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| + \left\|\beta_{[g]}\right\| - \left\|\hat{\beta}_{[g]}\right\|\right).$$

Note that $\left(\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| + \left\|\beta_{[g]}\right\| - \left\|\hat{\beta}_{[g]}\right\|\right)$ vanishes if $g \notin S$ and is bounded above by $\min\{2\left\|\beta_{[g]}\right\|, 2\left(\left\|\beta_{[g]} - \hat{\beta}_{[g]}\right\|\right)\}$ if $g \in S$.

This leads to the following sparsity oracle inequality, for all $\beta \in \mathbb{R}^p$,

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|^2 + \sum_{g=1}^{G}\lambda_g\left\|\hat{\beta}_{[g]} - \beta_{[g]}\right\| \leq \frac{1}{n}\|X(\beta - \beta^0)\|^2 + 4\sum_{g\in S}\lambda_g \min\left\{\left\|\beta_{[g]}\right\|, \left\|\beta_{[g]} - \hat{\beta}_{[g]}\right\|\right\}. \tag{36}$$

The sparsity oracle inequality (36) with $\beta = \beta^0$, and $\Delta := \hat{\beta} - \beta^0$ leads to the following two useful bounds on the prediction and $\ell_{2,1}$-estimation errors:

$$\frac{1}{n}\left\|X\Delta\right\|^2 \leq 4\sum_{g\in S}\lambda_g\left\|\Delta_{[g]}\right\|, \tag{37}$$

$$\sum_{g\notin S}\lambda_g\left\|\Delta_{[g]}\right\| \leq 3\sum_{g\in S}\lambda_g\left\|\Delta_{[g]}\right\|. \tag{38}$$

Now, assume the group compatibility condition 32 holds. Then,

$$\frac{1}{n}\left\|X\Delta\right\|^2 \leq 4\sum_{g\in S}\lambda_g\left\|\Delta_{[g]}\right\| \leq \sqrt{\sum_{g\in S}\lambda_g^2}\,\frac{\|X\Delta\|}{\sqrt{n}}\,\frac{4}{\phi_{compatible}}, \tag{39}$$

which implies the first inequality of proposition C.1. The second inequality follows from

$$\begin{aligned}\lambda_{min}\left\|\hat{\beta}-\beta\right\|_{2,1} &\leq \sum_{g=1}^{G}\lambda_g\left\|\Delta_{[g]}\right\| \leq 4\sum_{g\in S}\lambda_g\left\|\Delta_{[g]}\right\| \\ &\leq 4\sqrt{\sum_{g\in S}\lambda_g^2}\,\frac{\|X\Delta\|}{\sqrt{n}}\,\frac{1}{\phi_{compatible}} \leq \frac{16}{\phi^2_{compatible}}\sum_{g\in S}\lambda_g^2,\end{aligned}$$

where the last step uses (39).

The proof of the last inequality of proposition C.1, i.e., the upper bound on ℓ_2 estimation error under $RE(2s)$, is the same as in Theorem 3.1 in Lounici et al. (2011) and is omitted. ■

Proof [Proof of Proposition 3.1] Applying the ℓ_2-estimation error of (35) on the i^{th} group lasso regression problem of regular group NGC, we have

$$\|\hat{A}_{i:}^{1:T-1}-A_{i:}^{1:T-1}\| \leq \frac{4\sqrt{10}}{\phi^2_{RE}(2s_i)}\frac{\sum_{g=1}^{s_i}\lambda_g^2}{\lambda_{\min}\sqrt{s_i}} \leq \frac{4\sqrt{10}}{\phi^2_{RE}(2s_{\max})}\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{s_i}$$

with probability at least $1-2\bar{G}^{1-\alpha}$. Combining the bounds for all $i=1,\ldots,p$ and noting that $s=\sum_{i=1}^{p}s_i$, we have the required result. ■

Appendix D. Irrepresentable assumptions and consistency

In this section, we discuss two results involving the compatibility and irrepresentable conditions for group lasso. We first show that a stronger version of the uniform irrepresentable assumption implies the group compatibility (32), and hence, consistency in $\ell_{2,1}$ norm. Next we argue that a weaker version of the irrepresentable assumption is indeed necessary for the direction consistency of the group lasso estimates. These results generalize analogous properties of lasso (van de Geer and Bühlmann, 2009; Zhao and Yu, 2006) to the group penalization framework. The proofs are given under a special choice of tuning parameter $\lambda_g=\lambda\sqrt{k_g}$. Similar results can be derived for the general choice of λ_g, although their presentation is more involved.

Proposition D.1 *Assume uniform irrepresentable condition* (13) *holds with* $\eta\in(0,1)$*, and* $\Lambda_{min}(C_{11})>0$*. Then group compatibility(S, L)* (32) *condition holds whenever* $L<\frac{1}{1-\eta}$*.*

Proof First note that with the above choice of λ_g the Group Compatibility (S,L) condition simplifies to

$$\phi_{compatible}:=\min_{\Delta\in\mathbb{R}^p\setminus\{\mathbf{0}\}}\left\{\frac{\sqrt{q}\|X\Delta\|}{\sqrt{n}\sum_{g\in S}\sqrt{k_g}\|\Delta_{[g]}\|}:\sum_{g\notin S}\sqrt{k_g}\|\Delta_{[g]}\|\leq L\sum_{g\in S}\sqrt{k_g}\|\Delta_{[g]}\|\right\}>0. \tag{40}$$

Also, the uniform irrepresentable condition guarantees that there exists $0 < \eta < 1$ such that $\forall \tau \in \mathbb{R}^q$ with $\|\tau\|_{2,\infty} = \max_{1 \le g \le s} \|\tau_{[g]}\|_2 \le 1$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21} \left(C_{11} \right)^{-1} K^0 \tau \right]_{[g]} \right\|_2 < 1 - \eta \ \forall g \notin S.$$

Here $K^0 = K/\lambda$ is a $q \times q$ block diagonal matrix with diagonal blocks $\sqrt{k_1}\, \mathbf{I}_{k_1 \times k_1}, \ldots, \sqrt{k_s}\, \mathbf{I}_{k_s \times k_s}$. Define

$$\Delta^0 := \underset{\Delta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathbf{X}\Delta\|_2^2 : \quad \sum_{g \in S} \sqrt{k_g} \|\Delta_{[g]}\|_2 = 1, \ \sum_{g \notin S} \sqrt{k_g} \|\Delta_{[g]}\|_2 \le L \right\}. \tag{41}$$

Note that $\frac{1}{n}\|\mathbf{X}\Delta^0\|_2^2 = \phi^2_{compatible}/q$, and introduce two Lagrange multipliers λ and λ' corresponding to the equality and inequality constraints for solving the optimization problem in (41). Also, partition $\Delta^0 = \left[\Delta^0_{(1)} : \Delta^0_{(2)} \right]$ and $\mathbf{X} = \left[\mathbf{X}_{(1)} : \mathbf{X}_{(2)} \right]$ into signal and nonsignal parts as in (10). The first q linear equations of the KKT conditions imply that there exists $\tau^0 \in \mathbb{R}^q$ such that

$$C_{11}\Delta^0_{(1)} + C_{12}\Delta^0_{(2)} \quad = \quad \lambda K^0 \tau^0 \tag{42}$$

and, for every $g \in S$,

$$\begin{aligned} \tau^0_{[g]} \quad &= \quad D(\Delta^0_{[g]}) \text{ if } \Delta^0_{[g]} \neq \mathbf{0}, \\ \|\tau^0_{[g]}\|_2 \quad &\le \quad 1 \text{ if } \Delta^0_{[g]} = \mathbf{0}. \end{aligned}$$

It readily follows that $\left(\tau^0\right)^T K^0 \Delta^0_{(1)} = \sum_{g \in S} \sqrt{k_g} \|\Delta^0_{[g]}\|_2 = 1$.

Multiplying both sides of (42) by $\left(\Delta^0_{(1)}\right)^T$ we get

$$\left(\Delta^0_{(1)}\right)^T C_{11} \Delta^0_{(1)} + \left(\Delta^0_{(1)}\right)^T C_{12} \Delta^0_{(2)} = \lambda. \tag{43}$$

Also, (42) implies

$$\Delta^0_{(1)} + \left(C_{11}\right)^{-1} C_{12} \Delta^0_{(2)} = \lambda \left(C_{11}\right)^{-1} K^0 \tau^0. \tag{44}$$

Multiplying both sides of the equation by $\left(K^0\tau^0\right)^T = \left(\tau^0\right)^T K^0$ we obtain

$$1 = -\left(\tau^0\right)^T K^0 \left(C_{11}\right)^{-1} C_{12} \Delta^0_{(2)} + \lambda \left(K^0\tau^0\right)^T \left(C_{11}\right)^{-1} \left(K^0\tau^0\right). \tag{45}$$

Note that the absolute value of the first term,

$$\left| \sum_{g \notin S} \left(\Delta^0_{[g]}\right)^T \left[C_{21} (C_{11})^{-1} K^0 \tau^0 \right]_{[g]} \right|, \tag{46}$$

is bounded above by

$$(1-\eta) \left(\sum_{g \notin S} \sqrt{k_g} \|\Delta^0_{[g]}\|_2 \right) \le (1-\eta) L \tag{47}$$

by virtue of the uniform irrepresentable condition and the Cauchy-Schwartz inequality. Assuming the minimum eigenvalue of C_{11}, i.e., $\Lambda_{min}(C_{11})$, is positive and considering $\|K^0\tau^0\|_2 \leq \sqrt{q}$, the second term is at most $\lambda\, q/\Lambda_{min}(C_{11})$. So (45) implies

$$1 \leq (1-\eta)L + \frac{\lambda q}{\Lambda_{min}(C_{11})}. \tag{48}$$

In particular, $\lambda \geq \Lambda_{min}(C_{11})\,(1-(1-\eta)L)\,/q$ is positive whenever $L < 1/(1-\eta)$.
Next, multiply both sides of (44) by $(\Delta^0_{(2)})^T C_{21}$ to get

$$\left(\Delta^0_{(2)}\right)^T C_{21}\Delta^0_{(1)} + \left(\Delta^0_{(2)}\right)^T C_{21}\,(C_{11})^{-1}\,C_{(12)}\Delta^0_{(2)} = \lambda\left(\Delta^0_{(2)}\right)^T C_{21}\,(C_{11})^{-1}\,K^0\tau^0. \tag{49}$$

Using the upper bound in (47), the right hand side is at least $-\lambda(1-\eta)L$.
Also a simple consequence of the block inversion formula of the non-negative definite matrix C guarantees that the matrix $C_{22} - C_{21}\,(C_{11})^{-1}\,C_{12}$ is non-negative definite. Hence,

$$\left(\Delta^0_{(2)}\right)^T\left[C_{22} - C_{21}\,(C_{11})^{-1}\,C_{12}\right]\Delta^0_{(2)} \geq 0$$

$$\text{and} \quad \left(\Delta^0_{(2)}\right)^T C_{22}\Delta^0_{(2)} \geq \left(\Delta^0_{(2)}\right)^T C_{21}\,(C_{11})^{-1}\,C_{12}\Delta^0_{(2)}.$$

Putting all the pieces together we get

$$\begin{aligned}
\phi^2_{compatible}/q &= \frac{1}{n}\|\mathbf{X}\Delta^0\|_2^2 \\
&= {\Delta^0_{(1)}}^T C_{11}\Delta^0_{(1)} + 2{\Delta^0_{(2)}}^T C_{21}\Delta^0_{(1)} + {\Delta^0_{(2)}}^T C_{22}\Delta^0_{(2)} \\
&= \lambda + {\Delta^0_{(2)}}^T C_{21}\Delta^0_{(1)} + {\Delta^0_{(2)}}^T C_{22}\Delta^0_{(2)}\ \text{, by (43)} \\
&\geq \lambda - \lambda(1-\eta)L\ \text{, by (49)} \\
&= \lambda(1-(1-\eta)L).
\end{aligned}$$

Plugging in the lower bound for λ we obtain the result; namely,

$$\phi^2_{compatible} = \Lambda_{min}(C_{11})\,(1-(1-\eta)L)^2 > 0$$

for any $L < \frac{1}{1-\eta}$. ■

In this section we investigate the necessity of irrepresentable assumptions for direction consistency of group lasso estimates. To this end we first introduce the notion of weak irrepresentability.

For a q-dimensional vector τ define the stacked direction vector

$$\underbrace{\tilde{D}(\tau)}_{q\times 1} = [\underbrace{D(\tau_{[1]})'}_{k_1\times 1}, \ldots, \underbrace{D(\tau_{[s]})'}_{k_s\times 1}]'.$$

Weak Irrepresentable Condition is satisfied if

$$\frac{1}{\lambda_g}\left\|\left[C_{21}(C_{11})^{-1}K\tilde{D}(\beta^0_{(1)})\right]_{[g]}\right\| \leq 1,\ \forall g \notin S = \{1,\ldots,s\}. \tag{50}$$

We argue the necessity of weak irrepresentable condition for group sparsity selection and direction consistency under two regularity conditions on the design matrix, as n, $p \to \infty$:
(A1) The minimum eigenvalue of the signal part of the Gram matrix, viz. $\Lambda_{min}(C_{11})$, is bounded away from zero.
(A2) The matrices C_{21} and C_{22} are bounded above in spectral norm.

As in the last proposition, we set $\lambda_g = \lambda\sqrt{k_g}$ and $K^0 = K/\lambda$. Suppose that the weak irrepresentable condition does not hold, i.e., for some $g \notin S$ and $\xi > 0$, we have,

$$\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\beta^0_{(1)}) \right]_{[g]} \right\| > 1 + \xi$$

for infinitely many n. Also suppose that there exists a sequence of positive reals $\delta_n \to 0$ such that the event

$$E_n := \{ \|D(\hat{\beta}_{[g]}) - D(\beta_{[g]})\|_2 < \delta_n, \ \forall g \in S, \text{ and } \hat{\beta}_{[g]} = \mathbf{0} \, \forall \, g \notin S \}$$

satisfies $\mathbb{P}(E_n) \to 1$ as p, $n \to \infty$.

Note that for large enough n so that $\delta_n < \min_g \|D(\beta_{[g]})\|$, we have $\hat{\beta}_{[g]} \neq \mathbf{0}$, $\forall \, g \in S$ on the event E_n.

Then, as in the proof of Theorem 4.1, we have, on the event E_n,

$$\hat{\mathbf{u}} = (C_{11})^{-1} \left[\frac{1}{\sqrt{n}} Z_{(1)} - \lambda K^0 \tilde{D}(\hat{\beta}_{(1)}) \right] \tag{51}$$

$$\text{and} \quad \frac{1}{n} \left\| \left[{\mathbf{X}_{(2)}}^T (\epsilon - \mathbf{X}_{(1)} \hat{u}) \right]_{[g]} \right\| \leq \lambda \sqrt{k_g}, \ \forall g \notin S. \tag{52}$$

Substituting the value of $\hat{u}$ from (51) in (52), we have, on the event E_n,

$$\frac{1}{\sqrt{n}} \left\| \left[\mathbf{Z}_{(2)} - C_{21}(C_{11})^{-1} \mathbf{Z}_{(1)} + \lambda\sqrt{n} C_{21}(C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| \leq \lambda\sqrt{k_g},$$

which implies that

$$\begin{aligned} &\left\| \left[Z_{(2)} - C_{21} \left(C_{11}\right)^{-1} Z_{(1)} \right]_{[g]} \right\| \\ &\qquad \geq \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\hat{\beta}_{(1)}) \right]_{[g]} \right\| - 1 \right]. \end{aligned} \tag{53}$$

Now note that for large enough n, if $\|C_{21}\|$ is bounded above, direction consistency guarantees that the expression on the right is larger than

$$\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \left[\frac{1}{\sqrt{k_g}} \left\| \left[C_{21}(C_{11})^{-1} K^0 \tilde{D}(\beta_{(1)}) \right]_{[g]} \right\| - 1 \right],$$

which in turn is larger than $\frac{1}{2} \lambda \sqrt{n} \sqrt{k_g} \, \xi$, in view of the weak irrepresentable condition.

This contradicts $\mathbb{P}(E_n) \to 1$, since the left-hand side of (53) corresponds to the norm of a centered Gaussian random variable with bounded variance structure $\left[C_{22} - C_{21} C_{11}^{-1} C_{12} \right]_{[g][g]}$ while $\lambda \sqrt{n} \sqrt{k_g}$ diverges with $\sqrt{\log G}$.

Appendix E. Thresholding Group Lasso Estimates.

Proof [Proof of Theorem 4.2] We use the notations developed in the proof of Proposition C.1. First note that, (ii) follows directly from Theorem 4.1. For (i), since the falsely selected groups are present after the initial thresholding, we get $\|\hat{\beta}_{[g]}\| > 4\lambda$ for every such group. Next, we obtain an upper bound for the number of such groups. Specifically, denoting $\Delta = \hat{\beta} - \beta^0$, we get

$$\left|\hat{S}\backslash S\right| \leq \frac{\|\hat{\beta}_{S^c}\|_{2,1}}{4\lambda} = \frac{\sum_{g\notin S}\|\Delta_{[g]}\|}{4\lambda}. \tag{54}$$

Next, note that from the sparsity oracle inequality (37), the following holds on the event $\mathcal{A}$,

$$\sum_{g\notin S}\|\Delta_{[g]}\| \leq 3\sum_{g\in S}\|\Delta_{[g]}\|.$$

It readily follows that

$$4\sum_{g\notin S}\|\Delta_{[g]}\| \leq 3\|\Delta\|_{2,1} \leq \frac{48}{\phi^2}\, s\lambda,$$

where the last inequality follows from the $\ell_{2,1}$-error bound of (34). Using this inequality together with (54) gives the result. ■

References

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008. ISSN 1532-4435.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

M. Binder, C. Hsiao, and M.H. Pesaran. Estimation and inference in short panel vector autoregressions with unit roots and cointegration. *Econometric Theory*, 21:795–837, 2005. ISSN 1469-4360. doi: 10.1017/S0266466605050413.

O. Blanchard and R. Perotti. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *The Quarterly Journal of Economics*, 117 (4):1329–1368, 2002.

P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Stat. Interface*, 2(3):369–380, 2009. ISSN 1938-7989.

B. Cao and Y. Sun. Asymptotic distributions of impulse response functions in short panel vector autoregressions. *Journal of Econometrics*, 163(2):127 – 143, 2011. ISSN 0304-4076.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science's STKE*, 303(5659):799, 2004.

A. Fujita, J. Sato, H. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. Sogayar, and C. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1(1):39, 2007. ISSN 1752-0509.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996. ISBN 0-8018-5413-X; 0-8018-5414-8.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *Journal of Finance*, pages 1639–1664, 1994.

J. Huang and T. Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004, 2010. ISSN 0090-5364.

J. Huang, S. Ma, H. Xie, and C-H. Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009. ISSN 0006-3444.

K. Kim, J.H. Kim, J. Lee, H.M. Jin, S.H. Lee, D.E. Fisher, H. Kook, K.K. Kim, Y. Choi, and N. Kim. Nuclear factor of activated t cells c1 induces osteoclast-associated receptor gene expression during tumor necrosis factor-related activation-induced cytokine-mediated osteoclastogenesis. *Journal of Biological Chemistry*, 280(42):35209–35216, 2005.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. Isoperimetry and processes.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110, 2009.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

G. Michailidis. Statistical challenges in biological networks. *Journal of Computational and Graphical Statistics*, 21(4):840–855, 2012. doi: 10.1080/10618600.2012.738614.

S. V. Parter. Extreme eigenvalues of Toeplitz forms and applications to elliptic difference equations. *Trans. Amer. Math. Soc.*, 99:153–192, 1961. ISSN 0002-9947.

J. Pearl. *Causality: Models, Reasoning, and Inference*, volume 47. Cambridge, 2000.

M. B. Priestley. *Spectral Analysis and Time Series. Vol. 2*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1981. ISBN 0-12-564902-9. Multivariate series, prediction and control, Probability and Mathematical Statistics.

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

M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:no. 82, 1–9, 2013. ISSN 1083-589X. doi: 10.1214/ECP.v18-2865.

A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010a.

A. Shojaie and G. Michailidis. Discovering graphical granger causality using a truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010b.

C.A. Sims. Money, income, and causality. *The American Economic Review*, 62(4):540–552, 1972.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. ISSN 1935-7524.

S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.*, 5: 688–749, 2011. ISSN 1935-7524.

R. Vershynin. *Lectures in Geometric Functional Analysis*. available at http://www-personal.umich.edu/ romanv/papers/GFA-book/GFA-book.pdf, 2009.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183 –2202, May 2009. ISSN 0018-9448.

M. L. Waterman, K. A. Jones, et al. Purification of tcf-1 alpha, a t-cell-specific transcription factor that activates the t-cell receptor c alpha gene enhancer in a context-dependent manner. *The New Biologist*, 2(7):621, 1990.

F. Wei and J. Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369–1384, 2010. ISSN 1350-7265.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, December 2006. ISSN 1532-4435.

S. Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation. *Arxiv preprint arXiv:1002.1583*, 2010.