Precision Least Squares: Estimation and Inference in High-Dimensions

Luca Margaritella*
Rosnel Sessinou†

March 14, 2023

Abstract

The least squares estimator can be cast to depend only on the precision matrix. We show that a consistent estimator of the latter can be directly used to obtain a consistent estimator of the former even in high-dimensional regression problems where the number of covariates can be larger than the sample size. We call this the *precision least squares* estimator. We show that it is asymptotically Gaussian and delivers uniformly valid inference irrespective of the sparsity within the data generating process. Since bias can still hinder the estimates when using consistent but regularized precision matrix estimators, we show how to construct a nearly unbiased least squares estimator. We illustrate the relevance of regularized precision matrix on both simulated and real data. Contrary to the systemic risk literature based on multivariate autoregressive models for stock returns, and more in line with the theory of financial market fragility, we find evidence that returns connectedness of 88 global banks drastically decreases during crisis periods.

Keywords: Precision Least Squares, High-Dimensional Inference, Systemic Risk.
JEL codes: C55, C12, G19.

*Department of Economics, Lund University Email: luca.margaritella@nek.lu.se.
†Department of Decision Sciences, HEC Montréal. Email: rosnel.sessinou@hec.ca (corr. author)

We thank Sébastien Laurent, Karim Abadir, Sessi Tokpavi, Sullivan Hué, Stephan Smeekes, Joakim Westerlund, Simon Reese, Emmanuel Flachaire, Christophe Hurlin, Eléna-Ivona Dumitrescu, Anna Simoni, René Garcia, Anne Mackay, David Aridia, Keven Blateu, Benoît Perron and the participants of the 2022 Africa Meeting of the Econometric Society, the participant of the Economics and Statistics Séminars at HEC Montréal, Sherbrooke University, Lund University School of Economics and Management. Rosnel Sessinou acknowledges research support by Fin-ML, the French National Research Agency Grants ANR-17-EURE-0020 and ANR-21-CE26-0007-01 (project MLEforRisk) and by the Excellence Initiative of Aix-Marseille University- A*MIDEX.

1 Introduction

We introduce the *Precision Least Squares* (PrLS) estimator by making use of the fact that by *only* knowing the inverse covariance, i.e., the *precision matrix*, of a simple linear combination of the data, the whole least squares solution can be easily obtained. This is a well-known fact already in low-dimensional settings, and especially in finance, although it is often overlooked in other fields. Here, we extend its use to the high-dimensional setting, namely those frameworks where the number of covariates p is large and potentially larger than the sample size T. The conveniences of passing through the precision matrix to obtain the least squares solution are several. First and foremost, it is sufficient to substitute in the PrLS estimator *any* (nearly) unbiased and consistent estimator of the precision matrix in order to obtain a consistent estimator of the regression parameters. Also, the precision matrix, as opposed to the covariance, carries important information with regard to the conditional correlations among the covariates.[1]

When in high-dimensions, it still holds true that a simple plug-in of a (nearly) unbiased and consistent estimate of the precision matrix in the PrLS is all that is needed to obtain the least squares solution. However, and specifically whenever the concentration ratio p/T is bounded away from zero, sample covariance matrices are singular with probability one and thus the estimation of *large* precision matrices becomes more challenging.[2] In fact, it is nowadays well known how the estimation of high-dimensional linear regression models with standard techniques such as ordinary least squares leads to overfitting, high-variance and numerically unstable estimates (cf. the *curse* of dimensionality). Therefore, while in low dimensions a simple plug-in for the PrLS is e.g., the inverse sample covariance, finding an immediate plug-in in high dimensions is less straightforward. In fact, this requires

[1]The latter fact is something interesting in its own right. It can be used, among others, to build undirected graphs under additional Gaussianity assumption.

[2]Besides the computational burden (or impossibility when $p > T$) of inverting large matrices, it is well known how small errors in the covariance gets amplified by the inversion to obtain the precision matrix.

resorting to regularization techniques or "shrinkage-inducing estimators", as commonly done in the literature (see a.o., Pourahmadi, 2011; Fan et al., 2016b; Lam, 2020; Ledoit and Wolf, 2019). Regularization acts by reducing the excess degrees of freedom (overfit) of the precision matrix estimator.

Sparsity is the underlying assumption that justifies regularization in linear regression models. Stepwise regression, best subset selection, and ℓ_q-norm penalization are all commonly used techniques which induce regularization or even perform variable selection when setting some regression coefficient exactly equal to zero (e.g., the ℓ_1-norm regularizers such as lasso). By reducing p to something manageable via variable selection, usual low-dimensional techniques such as a simple least squares become viable options again. However, the sparsity assumption can be a strong statement regarding the data generating process (DGP). All the above mentioned techniques need to assume the vector of regression coefficients to have some of its entries exactly equal to zero (cf. *exact* sparsity) or close enough to it (cf. *approximate/weak* sparsity).[3] The recent literature has argued how sparsity, at least in fields such as economics and finance, might be just an "illusion" (Giannone et al., 2021) and that the reality is actually more complex, or *dense*. However, should be noted that the assumption of sparsity over the precision matrix entries is generally speaking a more reasonable assumption than assuming it directly on the coefficient vector. In fact, a regularized or shrunk estimate of the precision matrix might or might not imply sparsity on the estimated parameter vector (but at least avoids assuming it upfront on the DGP, see: Archakov and Hansen 2022; Bradic et al. 2022). It is easy to build examples where even though the precision matrix is sparse, the linear regression vector of coefficients is fully dense (Bradic et al., 2022). Likewise, it is easy to build examples where even though the covariance is completely dense, the precision matrix is quite sparse.[4] The reason for

[3]Approximate sparsity is discussed in Belloni et al. (2014) where the exact sparse model is assumed to be a (good) approximation to the true DGP. Weak sparsity is discussed in Adamek et al. (2022), where many non-zero but small coefficients are allowed.

[4]E.g.: A covariance matrix with a toeplitz structure.

this latter fact is that even though two covariates might be marginally correlated, they can be conditionally uncorrelated given the other variables in the information set. Then, as PrLS depends only on the precision matrix, it inherits the advantage that assuming sparsity on (a function of) the precision matrix entries, avoids assuming it a priori directly on the coefficient vector.

However, it is also known how the direct use of these regularization techniques on the precision matrix, still leads to biased parameter estimates (Janková and van de Geer, 2018). In fact, the idea behind all these procedures is to purposely introduce bias (towards zero or towards a structured matrix such as the identity matrix) in order to help reduce the variance of the sample covariance/sample precision matrix. The consequence, though, is that a direct application of any of such methods would also introduce bias in the PrLS estimator. Recent literature that focuses on high-dimensional inference has proposed several solutions to prevent regularization bias. In fact, such misspecification of the selected model makes also statistical inference after model selection unreliable (see Leeb and Pötscher, 2005). In particular, several works have introduced changes to various regularized methods to improve the estimation accuracy of the parameter vector (a.o., Belloni et al., 2014, 2015; Javanmard and Montanari, 2014; Nickl et al., 2013; Van de Geer et al., 2014; Zhang and Zhang, 2014). These approaches always correctly de-bias the estimates for valid high-dimensional post-selection inference. However, there are no guarantees of validity unless various sparsity structures are directly assumed on the parameter vector. We show here that, in a similar fashion to some existing debiasing methods for the precision matrix (Janková and van de Geer, 2018) and for the regression parameters (Javanmard and Montanari, 2014; Bühlmann et al., 2013), we can formulate a consistent PrLS estimator which delivers uniformly valid inference.

On the topic of inference, another nontrivial task when in a high-dimensional settings is how to set up hypothesis testing. First, we show under the physical dependence (Zhang et al., 2017) of the underlying stochastic process that the PrLS estimator has an asymptotically normal distribution and hence provides a simple way to use standard inferential techniques. Second, by means of subseries (cf. "batch-mean method", Zhang et al., 2017), we are able to estimate the covariance matrix of the PrLS estimator and obtain standard errors which are robust to moderate misspecification of the precision matrix. In addition, the asymptotic distribution of the PrLS estimator is found to be normal irrespective of the precision matrix estimator used. This provides a versatile tool for large-scale inference in linear regression models. Our results are validated through a Monte Carlo exercise.

We also illustrate the effectiveness of our approach on real data. Using the PrLS estimator, we obtain networks of predictive connectedness among daily asset returns of 88 global banks between 2005-2020. We find evidence that such connections drastically decrease during crisis periods. Network density (modularity) is then proposed as an empirical measure of crisis proximity. Our empirical results reconcile the literature on measuring stock returns systemic risk using the vector autoregressive (VAR) model with theories of financial market fragility and reputation (Lagunoff and Schreft, 2001; Ordoñez, 2013).

The remainder of the paper is organized as follows: Section 2 defines an oracle estimator for the high-dimensional coefficient vector and shows that the precision least squares estimator is asymptotically equivalent to the oracle. Section 3 shows how a finite sample bias could emerge if a poor estimate of the precision matrix is employed within the PrLS. As a consequence, two high-dimensional debiasing strategies are presented. In Section 4 Monte Carlo simulations are reported, showing how the PrLS outperforms state-of-the-art methods in terms of bias, size distortion, and power in finite samples. In Section 5 our empirical application is presented. Finally, Section 6 concludes.

The Online Appendix is organized as follows: Appendix A1 gives more details about our oracle approximation theory. Appendix A3 collects all the proofs. Appendix A4 reports some details on the regression-based modified Cholesky decomposition which is the precision matrix estimator we used throughout for illustration purposes. Finally, in Appendix A5 data details and additional results about the empirical application are presented.

A few words on notation. For any vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$, let the ℓ_s-norm $||\boldsymbol{v}||_s = \left(\sum_{j=1}^p |v_j|^s\right)^{1/s}$, $s \geq 1$. Also, $||\boldsymbol{v}||_\infty = \max_j |v_j|$. Similarly, for a matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{p \times q}$, the matrix ∞ norm $\|\boldsymbol{A}\|_\infty = \max_{1 \leq i \leq q} \sum_{j=1}^p |a_{ij}|$. For any two real numbers, $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$. Throughout, we write $A \gtrsim B$ if there exists an absolute constant c, independent of the model parameters, such that $A \geq cB$ and similarly for $A \lesssim B$. We use $C, C_1, c, c_1, \ldots$ to denote positive constants and $\xrightarrow{p}$ and $\xrightarrow{d}$ to denote convergence in probability and distribution, respectively. The superscripts p and d are omitted when the distinction is not relevant to the statement given. Throughout we use $\boldsymbol{I}$ for the identity matrix and $\mathbb{1}$ for vectors of ones and we omit to write their dimension as subscripts when there is no possible ambiguity.

2 Precision Least Squares

2.1 Estimation

Let y_t, $x_{1,t}, \ldots, x_{p,t}$ be a set of covariance-stationary time series of interest for a sample size T and dimension p, which is large, potentially larger than (and growing with) T. The linear regression model is

$$y_t = \sum_{j=1}^p \beta_j x_{j,t} + \epsilon_t = \boldsymbol{x}_t' \boldsymbol{\beta} + \epsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{x}_t$ is a $p \times 1$ vector of observables, $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients object of the estimation and ϵ_t is a scalar realization of a stationary stochastic process with mean zero, variance σ^2 and at least four finite moments. For notational simplicity we assume the variables have zero mean; if not, they can be demeaned prior to the analysis, or equivalently intercepts are added.[5]

Assumption 1 (Parameter space)**.** Consider the distribution of the data being indexed by the parameter $\boldsymbol{\vartheta} := (\boldsymbol{\beta}, \boldsymbol{\Sigma}_x, \sigma^2)$, where $\boldsymbol{\Sigma}_x := \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')$. Then, we employ the following set

$$\boldsymbol{\Psi} := \left\{ \boldsymbol{\vartheta} : c_0^{-1} \leq \Lambda_{\min}(\boldsymbol{\Sigma}_x) \leq \Lambda_{\max}(\boldsymbol{\Sigma}_x) \leq c_0,\ 0 < \sigma < c_1,\ \|\boldsymbol{\beta}\|_q \leq c_2 \right\},$$

where $c_0 > 1$, c_1, c_2 are positive finite constants, $\Lambda_{\min}(\boldsymbol{\Sigma}_x)$ and $\Lambda_{\max}(\boldsymbol{\Sigma}_x)$ are the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}_x$ and $q \in [1, \infty]$. We assume $\boldsymbol{\Psi}$ is a compact set.

The linear regression model (1) can be rewritten as a linear combination of the transformed variable $\boldsymbol{f}_t := (y_t - x_{1,t}, \ldots, y_t - x_{p,t}, y_t)'$ such that

$$\boldsymbol{f}_t' \boldsymbol{w}_0 = \epsilon_t, \tag{2}$$

where $\boldsymbol{w}_0 := (\boldsymbol{\beta}', \beta_0)'$ is a $p+1$-dimensional vector, $\beta_0 = 1 - \sum_{j=1}^{p} \beta_j$ and for which it holds $\sum_{j=0}^{p} w_{0,j} = 1$.

Assumption 2 (Modified Parameter space)**.** Given the transformations in $\boldsymbol{f}_t$, the modified parameter is $\tilde{\boldsymbol{\vartheta}} := (\boldsymbol{w}_0, \boldsymbol{\Sigma}, \sigma^2)$, where $\boldsymbol{\Sigma}_f := \mathbb{E}(\boldsymbol{f}_t \boldsymbol{f}_t')$ is the $(p+1) \times (p+1)$ full rank population covariance matrix of $\boldsymbol{f}_t$. Then, for the same constants as in Assumption 1 the parameter space is now

$$\tilde{\boldsymbol{\Psi}} := \left\{ \tilde{\boldsymbol{\vartheta}} : c_0^{-1} \leq \Lambda_{\min}(\boldsymbol{\Sigma}_f) \leq \Lambda_{\max}(\boldsymbol{\Sigma}_f) \leq c_0,\ 0 < \sigma < c_1,\ \|\boldsymbol{w}_0\|_q \leq c_2 \right\}.$$

We assume $\tilde{\boldsymbol{\Psi}}$ is a compact set.

[5] One can also allow the vector $\boldsymbol{x}_t$ to contain the lags of y_t as well as lags of other relevant covariates. In fact, the linear regression framework encompasses both autoregressive and seemingly unrelated regression (SUR) models.

Assumption 1 and 2 are standard and draw a parallel between the usual regression framework where the parameter space is compact to the transformed framework mapped by $\boldsymbol{f}_t$. The compactness of the parameter space in turn implies the boundedness of $\boldsymbol{w}_0$.

Then, the standard least squares problem for (1) can be reformulated as

$$\boldsymbol{w}_0 = \underset{\substack{\sum_{j=0}^{p} \bar{w}_j = 1, \\ \bar{\boldsymbol{w}} \in \mathbb{R}^{p+1}}}{\arg\min}\ \bar{\boldsymbol{w}}' \boldsymbol{\Sigma}_f \bar{\boldsymbol{w}}. \tag{3}$$

The analytical solution of (3) is

$$\boldsymbol{w}_0 = \boldsymbol{\Theta} \mathbb{1} \left(\mathbb{1}' \boldsymbol{\Theta} \mathbb{1} \right)^{-1}, \tag{4}$$

where $\mathbb{1}$ is a $(p+1)$ vector of ones and $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ is the (population) precision matrix of $\boldsymbol{f}_t$. The rewriting in (1)-(4) is well known, especially in the global minimum variance portfolio literature.[6] Note that if $\boldsymbol{\Theta}$ is known, a simple plug-in within (4) and extraction of the first p elements of $\boldsymbol{w}_0$ automatically gives $\boldsymbol{\beta}$. As in practice $\boldsymbol{\Theta}$ is unknown, it needs to be estimated.

Let $\boldsymbol{J}$ be a $(p+1) \times p$ matrix such that $\boldsymbol{J} = [\boldsymbol{I}_p, -\mathbb{1}_p]'$ with $\boldsymbol{e}_j$ being the unit vector with the one appearing in position j.[7] Let also $\hat{\boldsymbol{\Theta}}$ be an unbiased precision matrix estimator. By plugging $\hat{\boldsymbol{\Theta}}$ in (4), we get

$$\hat{\boldsymbol{w}} := (\mathbb{1}' \hat{\boldsymbol{\Theta}} \mathbb{1})^{-1} \hat{\boldsymbol{\Theta}} \mathbb{1} = \boldsymbol{J} \hat{\boldsymbol{\beta}} + \boldsymbol{e}_{p+1}, \tag{5}$$

and by extracting its first p elements we obtain $\hat{\boldsymbol{\beta}}$ which we refer to as the "Precision Least

[6] Since $\beta_0 + \sum_{j=1}^{p} \beta_j = 1$ and $\sum_{j=0}^{p} \beta_j y_t = \sum_{j=1}^{p} \beta_j x_{j,t} + \epsilon_t$ we deduce Eq. (2) from $\beta_0 y_t + \sum_{j=1}^{p} \beta_j (y_t - x_{j,t}) = \epsilon_t$. The least-square estimator of β follows from the minimization of the variance of ϵ_t under the constraint $\sum_{j=0}^{p} \beta_j = 1$ as in Eq. (3). Then, we obtain Eq. (4) by setting the first-order derivative of the objective function to zero as usual. See also Kempf and Memmel (2006) for more details.

[7] E.g., let $p = 2$, $\boldsymbol{J} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}$. Then, $\boldsymbol{J}\boldsymbol{\beta} + \boldsymbol{e}_{p+1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ 1 - (\beta_1 + \beta_2) \end{bmatrix}$.

Squares" (PrLS) estimator.

Remark 1. Let $\boldsymbol{Q}$ be a square matrix of order $p+1$ such that $\boldsymbol{Q}_{p+1,i} = 1$, for all i and $\boldsymbol{Q}_{i,i} = -1$ for $i > 1$. Let also $\boldsymbol{z}_t := (\boldsymbol{x}_t', y_t)'$. One has $\boldsymbol{f}_t = \boldsymbol{Q}'\boldsymbol{z}_t$ and denote $\boldsymbol{\Omega} = \text{Var}(\boldsymbol{z}_t)$ the joint covariance of $\boldsymbol{z}_t$. One also has that $\boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{Q}^{-1} = \boldsymbol{Q}$ and the following properties hold: (i) $\boldsymbol{\Sigma} = \boldsymbol{Q}'\boldsymbol{\Omega}\boldsymbol{Q}$; (ii) $\boldsymbol{\Theta} = \boldsymbol{Q}\boldsymbol{\Omega}^{-1}\boldsymbol{Q}'$. Properties (i) and (ii) state that there is a one-to-one mapping between the covariance and the precision matrix of both the vector $\boldsymbol{z}_t$ and the transformed vector $\boldsymbol{f}_t$. This is a convenient fact in light of the earlier discussion on sparsity, given that the matrix $\boldsymbol{Q}$ might not preserve the structure of $\boldsymbol{\Omega}$ or its inverse. For instance, one can pass from a sparse structure for $\boldsymbol{\Omega}$ or $\boldsymbol{\Omega}^{-1}$ to a dense structure for $\boldsymbol{\Sigma}$ or $\boldsymbol{\Theta}$. Therefore, in order to estimate (4), one only needs to estimate $\boldsymbol{\Omega}$, or its inverse.

The following theorem is our main result. It gives the asymptotic normality of a single coordinate of the PrLS estimator. We state it here first, and then discuss the steps leading to the result. As this follows from some propaedeutic Lemmas (Lemma A1.1, A1.2), we just mention them here but discuss them in full in the Online Appendix.

Theorem 1. Given the results in Lemma A1.1, one has for all $j = 1, \dots, p$,

$$\sqrt{T}(\hat{\beta}_j - \beta_j) \xrightarrow{d} N(0, \boldsymbol{V}_{j,j}) \quad \text{as} \quad T \to \infty, \tag{6}$$

where $\boldsymbol{V} = \sum_{l=-\infty}^{\infty} \mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}_{t+l}')$ with $\boldsymbol{u}_t = \boldsymbol{\Sigma}_x^{-1} \boldsymbol{x}_t \epsilon_t$.

We summarize here the oracle framework (also presented in the Online Appendix) that leads to our Theorem 1. Let us denote $\check{\boldsymbol{\Sigma}}$ an oracle estimator of $\boldsymbol{\Sigma}$ and $\check{\boldsymbol{\Theta}} = \check{\boldsymbol{\Sigma}}^{-1}$ its inverse. By "oracle", we refer to an *infeasible* estimator which is *almost* (see the estimation error term $\boldsymbol{\Delta}$ below) as good as an ideal estimator which relies on perfect information supplied by an oracle, but which can be unavailable in practice.[8]We define an oracle PrLS estimator

[8]See for instance Examples 1 and 2 fo an example of oracle estimator. See also Example 3 and the Monte Carlo simulation in Section 4 were we use the OLS estimator with an artificially augmented sample size to mimic an oracle estimator.

such as $\boldsymbol{w}^{\text{Oracle}} = \check{\boldsymbol{\Sigma}}^{-1}\check{\boldsymbol{v}}$, for $\check{\boldsymbol{v}} = \mathbb{1}(\mathbb{1}'\check{\boldsymbol{\Theta}}\mathbb{1})^{-1}$. Applying the matrix inversion lemma to $\hat{\boldsymbol{\Theta}}^{-1} = \check{\boldsymbol{\Sigma}} + \boldsymbol{\Delta}$, we obtain the following decomposition:

$$\hat{\boldsymbol{\Theta}} = (\boldsymbol{I} - \boldsymbol{L})\check{\boldsymbol{\Theta}}, \tag{7}$$

where $\boldsymbol{L} = \check{\boldsymbol{\Theta}}(\check{\boldsymbol{\Sigma}} + \boldsymbol{\Delta})^{-1}\check{\boldsymbol{\Sigma}}\boldsymbol{\Delta}$ and $\boldsymbol{\Delta} = \hat{\boldsymbol{\Theta}}^{-1} - \check{\boldsymbol{\Sigma}}$. The key to this oracle framework is that $\boldsymbol{\Delta}$ gives the mistake between the estimated inverse precision matrix and its oracle version. Hence, by plugging (7) into $\hat{\boldsymbol{w}}$ one gets:

$$\hat{\boldsymbol{w}} = (\boldsymbol{I} - \boldsymbol{L})\check{\boldsymbol{\Theta}}\mathbb{1}\left(\mathbb{1}'\check{\boldsymbol{\Theta}}\mathbb{1} - \mathbb{1}'\boldsymbol{L}\check{\boldsymbol{\Theta}}\mathbb{1}\right)^{-1}. \tag{8}$$

It follows from (8) that $\hat{\boldsymbol{w}}$ is an asymptotically unbiased estimator of $\boldsymbol{w}_0$ if $\boldsymbol{\Delta} \to 0$. In fact, if $\boldsymbol{\Delta} \to 0$, we immediately get $\hat{\boldsymbol{w}} \to \check{\boldsymbol{\Theta}}\mathbb{1}(\mathbb{1}'\check{\boldsymbol{\Theta}}\mathbb{1})^{-1} = \boldsymbol{w}^{\text{Oracle}}$. In the low-dimensional setting, this condition can easily be satisfied if $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Sigma}}^{-1}$, where $\hat{\boldsymbol{\Sigma}}^{-1}$ is the empirical counterpart of $\boldsymbol{\Theta}$. This is true since $\hat{\boldsymbol{\Sigma}}$ is an (asymptotically) unbiased and consistent estimator of $\boldsymbol{\Sigma}$ in low dimensions, implying $\mathbb{E}(\boldsymbol{\Delta}) = 0$ in (7). In high-dimensional settings, even though complications arise for obtaining an estimator of $\boldsymbol{\Theta}$, it remains the case that if $\boldsymbol{\Delta} \to 0$, Eq. (8) paired with an unbiased and consistent plug-in estimator of $\boldsymbol{\Theta}$ is all that is needed to recover the correct distribution of the PrLS estimator. For the rest of this section, we assume that $\boldsymbol{\Delta} \to 0$ and relax this assumption in Section 3.

Thus, if $\hat{\boldsymbol{\Theta}}$ is unbiased and consistent then $\hat{\boldsymbol{\Theta}}^{-1} \to \check{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{w}}$ behaves asymptotically like $\boldsymbol{w}^{\text{Oracle}}$. This is convenient since the desirable properties of an oracle estimator transfer in an asymptotic sense to any plug-in estimator $\hat{\boldsymbol{w}}$, provided that unbiasedness and consistency of the precision matrix estimate are satisfied. Let $\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t' y_t$. To emphasize the importance of the convergence of $\hat{\boldsymbol{\Sigma}}_x$, the empirical counterpart of $\boldsymbol{\Sigma}_x := \mathbb{E}[(\boldsymbol{x}_t\boldsymbol{x}_t')]$, we employ the following least-squares type oracle estimator: $\boldsymbol{w}^{\text{Oracle}} := \left(\boldsymbol{\beta}^{\text{Oracle}\prime}, 1 - \sum_{j=1}^{p}\beta_j^{\text{Oracle}}\right)' =$

$\boldsymbol{J}\boldsymbol{\beta}^{\text{Oracle}} + \boldsymbol{e}_{p+1}$, where

$$\boldsymbol{\beta}^{\text{Oracle}} = \boldsymbol{\Theta}_x \hat{\boldsymbol{\mu}}. \tag{9}$$

Substituting y_t in (9), we obtain the following rewriting

$$\boldsymbol{\beta}^{\text{Oracle}} = \boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{u}, \tag{10}$$

for $\boldsymbol{A} := \boldsymbol{\Theta}_x \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{I}$, $\boldsymbol{u}_t := \boldsymbol{\Theta}_x \boldsymbol{x}_t \epsilon_t$, $\boldsymbol{\Theta}_x = \boldsymbol{\Sigma}_x^{-1}$ and $\boldsymbol{u} := T^{-1} \sum_{t=1}^{T} \boldsymbol{u}_t$. We show in Lemma A1.1 that $\|\boldsymbol{A}\|_\infty$ converges in probability to zero. As a consequence, the second term on the right-hand side of (10) is controlled. Furthermore, in Lemma A1.2, we employ a high-dimensional Gaussian approximation for time series processes as in Zhang et al. (2017) on $\boldsymbol{u}_t$ and obtain a coordinate-wise central limit theorem (CLT) for $\boldsymbol{u}$. As a consequence of these results, in Theorem 1 we are able to state the asymptotic normality of the PrLS estimator, which follows immediately from the normality of the oracle. Note that Lemma A1.2, which is adapted from Theorems 3.2 and 3.3 of Zhang et al. (2017), guarantees uniform convergence of $\sqrt{T}\|\hat{\boldsymbol{\beta}}\|_\infty$ to the Gaussian distribution assuming that $\hat{\boldsymbol{\Theta}}$ is unbiased. As a consequence, the distribution of the maximum of the test statistics associated with the joint null hypothesis $H_0 : \beta_i = 0$ for a large set of parameters, after some rescaling and centering and under some regularity conditions, can be approximated with a Gumbel distribution (James et al., 2007).[9]

Remark 2. Direct replacing of $\boldsymbol{\Theta}_x$ in (9) with some sample estimate $\hat{\boldsymbol{\Theta}}_x$ would make the oracle estimator *feasible* as it would only depend on sample quantities. However, this way as $\hat{\boldsymbol{\Theta}}_x = \boldsymbol{\Theta}_x + \text{error}_1$ and $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \text{error}_2$, the random errors of the two estimated components would compound in the product and lead to an erratic behavior in high-dimensions (Bradic

[9] It is worth mentioning that Bradic et al. (2022) also analyse the properties of (9), restricting their attention to a single coefficient, and assuming i.i.d. data. In this paper, we study the properties of the whole vector (9), under the physical dependence framework of Zhang et al. (2017).

et al., 2022). Note in fact that $\hat{\boldsymbol{\mu}}$ is a likely non-sparse, $p \times 1$ vector, where p is assumed here to be large. Therefore, the product $\hat{\boldsymbol{\Theta}}_x \hat{\boldsymbol{\mu}}$ would be highly unstable as it depends on three sources of error: $\text{error}_1, \text{error}_2$ and their product.

In the following two examples, we illustrate the above theoretical properties of the PrLS estimator using Monte Carlo simulations.

Example 1. (*Sparse regression*)

We consider the following data generating process (DGP) as in Eq. (1). We set $T = 100$, while the number of variables is set to $p = 1003$. We let the $p \times 1$ sparse coefficient vector to be $\boldsymbol{\beta} = (2, 2, 2, 0, \ldots, 0)'$ and draw $\boldsymbol{x}_t$ from a constant correlation model with covariance $\boldsymbol{\Sigma}_x = \boldsymbol{d}\boldsymbol{d}' + \boldsymbol{I}_p$ for $\boldsymbol{d} = (3, \ldots, 3)'$ and $\mathbb{E}(\epsilon_t^2) = 1$. Also, we let $\boldsymbol{z}_t = (\boldsymbol{x}_t', y_t)'$ be i.i.d $N(0, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x \boldsymbol{\beta} \\ \boldsymbol{\beta}' \boldsymbol{\Sigma}_x & v_y \end{pmatrix}$, where v_y denotes the variance of y_t. To estimate $\boldsymbol{\beta}$ via PrLS and to test its elementwise significance, we only need to estimate $\text{Var}(\boldsymbol{z}_t) = \boldsymbol{\Omega}$ (see also Remark 1). As clear from its composition, $\boldsymbol{\Omega}$ is a block constant covariance matrix, and $\boldsymbol{\beta}$ is also block constant. The best estimator of an entry of the covariance matrix $\boldsymbol{\Omega}$ (or $\boldsymbol{\Sigma}_x$) in this setting is the average of the empirical counterpart of the constant block to which the entry belongs.[10] Table 1 reports both the estimated coefficient and the rejection frequencies associated with the test of significance for the first ten coefficients in $\boldsymbol{\beta}$.[11] We can see that the estimator bias is negligible and that significance tests based on the PrLS estimator have power close to one under the alternative hypotheses and size equal to zero under the null.

Insert Table 1 and Figure 1 approximately here

[10]It is immediate to see that $\boldsymbol{\Sigma}_x$ has constant covariances and variances and thus $\boldsymbol{\Omega}$ has constant blocks. Take $p = 4$ for instance, then $\boldsymbol{\Sigma}_x$ has 10s on the main diagonal and 9s everywhere else. The product with $\boldsymbol{\beta}$ gives the first three entries equal to 56 and the last equal to 54. To estimate the elements of $\boldsymbol{\Omega}$ is sufficient to take the cross-sectional average of their empirical counterpart.

[11]In Table 1 $p = 1003$ and $T = 100$. The variance of $\boldsymbol{\beta}^{PrLS}$ is estimated using a consistent estimator of $\boldsymbol{\Sigma}_x$. $Var(\hat{\boldsymbol{\beta}}^{PrLS}) = T\sigma^2 \boldsymbol{\Theta}_x$, where $\sigma^2 = \mathbb{V}ar(\epsilon_t)$. In the third column, the average of the random variable β_j^{PrLS} is reported for $j = 1, \ldots, 10$. In the last column, the frequencies of significance tests at 5% are computed over the 1000 replications.

Example 2. (*Dense regression*)

Consider the same setting as in Example 1 with the following modifications: $\boldsymbol{\beta} = (2, \ldots, 2)$, i.e., $\boldsymbol{\beta}$ is now dense. In Figure 1, we report the simulated density of the first estimated coefficient and compare it to a normal density centered around 2 with variance equal to the variance of the PrLS estimator. Unreported boxplots of the distribution of the average bias of the PrLS estimator over the 1003 estimated parameters show how the bias collapses to zero. Likewise, the rejection frequencies of the test $H_0^j: \ \beta_j = 0$ computed over 1000 replications is equal to one for all $j = 1, \ldots, 1003$.

In the above examples, a good estimation and inferential procedure are obtained by accounting for the fact that $\boldsymbol{\Omega}$ is a block constant matrix. Even though this assumption is restrictive, a.o., Archakov and Hansen (2022) show that the block constant covariance matrix assumption is realistic for several financial time series. In fact, Archakov and Hansen (2022, 2021) introduced a new representation for estimating such a matrix and showed that a plug-in approach based on the traditional OLS formula will lead to a consistent estimator of $\boldsymbol{\beta}$. This estimator is equivalent to the PrLS estimator whenever $\boldsymbol{\Omega}$ is estimated using their estimator.

In the examples above, the overall type I error of the individual tests is lower than the significance level α, while the power of the test converges to 1 given a consistent estimator of $\boldsymbol{\Omega}$. In Section 2.2 we shade more light on this property of the PrLS. In fact, we show that unbiased estimation and uniformly valid inference on $\boldsymbol{\beta}$ can be obtained using the PrLS estimator, regardless of the degree of sparsity of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}_x$, contrary to what has been previously assumed in the literature (see among others: Javanmard and Montanari, 2014; Bühlmann et al., 2013; Belloni et al., 2014; Van de Geer et al., 2014) but provided a consistent estimator of $\boldsymbol{\Omega}$ or $\boldsymbol{\Sigma}$ or their inverses.

2.2 Statistical Inference

We have seen that if $\hat{\boldsymbol{\Theta}}$ is unbiased, i.e., if $\boldsymbol{\Delta} \to 0$, we immediately get $\hat{\boldsymbol{w}} \to \check{\boldsymbol{\Theta}}\mathbb{1}(\mathbb{1}'\check{\boldsymbol{\Theta}}\mathbb{1})^{-1} = \boldsymbol{w}^{\text{Oracle}} = \boldsymbol{J}\boldsymbol{\beta}^{\text{Oracle}} + \boldsymbol{e}_{p+1}$, even in high dimensions. Thus, we find that the asymptotic property of $\boldsymbol{\beta}^{\text{Oracle}}$ is that of $\hat{\boldsymbol{\beta}}$. Recall our oracle estimator from (10) and its deviation from the true coefficient vector $\boldsymbol{\beta}$

$$\sqrt{T}(\boldsymbol{\beta}^{\text{Oracle}} - \boldsymbol{\beta}) = \sqrt{T}\boldsymbol{A}\boldsymbol{\beta} + \frac{1}{\sqrt{T}}\boldsymbol{u}. \tag{11}$$

As before in (10), the first term on the right-hand side of (11) represents the bias. We showed in Lemma A1.1 how this bias is negligible, i.e., with high probability $\|\boldsymbol{A}\|_\infty = o_p(1)$. The same tail bound as in (A1.1) can be applied on $\sqrt{T}\boldsymbol{A}\boldsymbol{\beta} = o_p(1)$ modulo a proportionality constant of the order $T^{-1/2}$.[12] The second term is the noise, which by Lemma A1.2, is shown to be approximately Gaussian and its covariance matrix is given by $\boldsymbol{V}$. Let $\widehat{\boldsymbol{V}}$ be a consistent estimator of $\boldsymbol{V}$. For instance, later in Section 3.2 we introduce the "batch mean estimator" of $\boldsymbol{V}$, see also Eq. (20). Then, for $j \in [1:p]$, the test statistics for a given null hypothesis $H_{0,j}$ for each element of the oracle vector of coefficients is defined as

$$T_j := \frac{\sqrt{T}\beta_j^{\text{Oracle}}}{\sqrt{\widehat{\boldsymbol{V}}_{jj}}}. \tag{12}$$

Thus, we reject the null $H_{0,j}$ if $|T_j| > th$ where th is a given quantile of the Gaussian distribution. Denote $\boldsymbol{R}(th) := \sum_{j=1}^{p} \boldsymbol{I}(|T_j| \geq th)$ the number of false discoveries obtained after this thresholding process. Theorem 2, shows how as the sample size and the dimension increase, control over the familywise error rate ($FWER$) remains guaranteed.

Theorem 2. Let $R = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the set of truly relevant covariates within $\boldsymbol{x}_t$. Let $\hat{R}$ be the set of estimated relevant features using PrLS. Then, there exist $a > 0$

[12] The factors T^{α_1+1} in the tail bound in (A1.1) become $T^{\alpha_1+1/2}$ while the rest is unchanged.

such that if $th \geq T^a$ then $\mathbb{P}\left(\boldsymbol{R}(th) \geq 1\right) = 0$ as $T, p \to 0$ and the rate of covergence is the minimum between that of $\|\boldsymbol{A}\|_\infty$ and $\max_j |\hat{\boldsymbol{V}}_{j,j} - \boldsymbol{V}_{j,j}|$.

The familywise error control property of the PrLS estimator does not mean perfect support recovery after p-value thresholding on a finite sample (see Figure 3a). It implies that the probability of getting perfect support recovery with the PrLS should be almost insensitive after p-value adjustment to control for FWER. We illustrate this in Figure 2.

Insert Figure 2 approximately here

Figure 2 compares the discrimination power between an OLS-based test of significance for the coefficients and the one based on the PrLS. Figure 3a shows that in finite samples identifying the relevant coefficients by simple thresholding of the p-values would yield an accumulation of type I error for both estimators. The two estimators exhibit similar power except for extreme values of the significance level where the PrLS outperforms. This already suggests that the probability of getting perfect support recovery with the PrLS will likely be insensitive to a Bonferroni type of correction. This is confirmed by Figure 3b. Indeed, unlike the OLS estimator, the aforementioned power of the PrLS is (almost) not impacted by the Bonferroni correction. The power is close to 1 for very small significant values, i.e., $< 1\%$. It quickly reaches 1 above this threshold. This power is almost identical to Holm (FWER error control) and Benjamini-Hochberg (FDR control) procedures which are known to be uniformly more powerful than the Bonferroni correction. The OLS is outperformed by the PrLS as its power is drastically reduced after p-value adjustment. We can observe that this is not the case for the OLS. Such improvement of the PrLS in terms of efficiency over the OLS comes from the fact that, unlike the latter, the former is adaptive to the linear dependence structure of the data and also holds in high-dimension (see Table 1).

Thus, estimating –and testing the significance of– the entries of $\boldsymbol{\beta}$ in high-dimension is only akin to a good structural and unbiased estimation of large covariance or precision matrices. That is, depending on the field of application, one can always justify the choice of a given covariance or precision matrix estimator based on the literature. For instance, in finance, approximate factor models, covariance thresholding estimators and lasso-based graphical models are also state-of-the-art estimators commonly used for asset pricing and portfolio optimization (see e.g., Gagliardini et al., 2016; Chang et al., 2018; Koike, 2020). One may therefore want to rely on these estimators for estimation with PrLS. However, such estimators can suffer from regularization bias and behave differently from an oracle. In Section 3, we illustrate the adverse consequences of regularization bias on the estimated PrLS and advocate two solutions to overcome them. We also show that regularized estimators can be used for oracle approximation.

3 Estimation and inference after regularization

3.1 Estimation with regularized precision matrices

In this section, we relax the unbiasedness assumption in Section 2 assuming that a regularized estimator $\tilde{\boldsymbol{\Theta}}$ is used as an approximation of the oracle precision matrix $\check{\boldsymbol{\Theta}}$. Plugging $\tilde{\boldsymbol{\Theta}}$ in (4) we obtain what we call the "Naive" PrLS. $\tilde{\boldsymbol{\Theta}}$ can belong to one of the three main families of regularized precision matrix estimators. The first family contains non-parametric methods. These apply regularization to the eigenvalues of the sample covariance matrix (Lam et al., 2016; Ledoit and Wolf, 2019, and references therein). The second family instead contains parametric methods. These provide parsimonious models for large precision or covariance matrix estimation (Pourahmadi, 2011; Lam and Fan, 2009, and references therein). The third family contains semi-parametric methods. These first extract the common factors in the data and then apply one of the former two approaches to the idiosyncratic components (Fan et al., 2018; Caner et al., 2022, and references therein). A reconstruction formula is then used to get the regularized precision matrix.

Let us call $\tilde{\boldsymbol{\beta}}$, a sub-vector of $\tilde{\boldsymbol{w}} := \tilde{\boldsymbol{\Theta}}\mathbb{1}/\mathbb{1}\tilde{\boldsymbol{\Theta}}\mathbb{1} = \boldsymbol{J}\tilde{\boldsymbol{\beta}} + \boldsymbol{e}_{p+1}$ where $\tilde{\boldsymbol{\Theta}}$ is a regularized precision matrix. Let us mention that the consistency and the rate of convergence of $\tilde{\boldsymbol{w}}$ have been extensively studied in the portfolio optimization literature, for various regularization methods as well as their empirical risk minimization property (Ledoit and Wolf, 2003, 2004, 2017; Fan et al., 2015, 2016a; Huang et al., 2006; Cai et al., 2017; Callot et al., 2019). In this section we show the negative effect of regularization bias and propose two bias correction methods. To fix ideas, let us consider the following example.

Example 3. We consider the following regularized precision matrix estimators: (i) adaptive lasso based modified Cholesky decomposition[13] (Cholesky); (ii) the linear shinkage estimamtor of $\boldsymbol{\Sigma}$ (Ledoit and Wolf, 2004); (iii) the nonlinear shrinkage estimator of $\boldsymbol{\Sigma}$ (Ledoit and Wolf, 2017); (iv) the covariance thresholding estimator (POET) (Fan et al., 2018). In the case (ii)-(iv) the regularized covariance is then inverted to obtain the precision matrix. For each of these estimators we compare the maximum excess bias ($\times 100$) with respect to an infeasible estimator of the PrLS. To obtain such infeasible estimator we employ an OLS estimator where the sample size has been raised to 500 observations against the $T = 250$ used in all other settings. For the number of variables we consider two settings: $p \in \{100, 250\}$. We consider two level of sparsity $s_0 = \{10, 60\}$, i.e. the number of non-zero coefficients in $\boldsymbol{\beta}$. In either cases we set the first s_0 elements of β_j to 2 while the remaining elements in the vector are set to zero. We simulate $x_{j,t}$ and ϵ_t from a student T distribution with 5 degrees of freedom and let $\text{Cov}(x_{j,t}, x_{i,t}) = 0.8^{|i-j|}$ for $i, j = 1, \ldots, p$. The results are reported in Table 2 under the column "Toeplitz". The bias of the resulting PrLS estimators (Naive) is shown to increase with the level of sparsity. The non-linear shrinkage estimator is on average less biased than the linear-shrinkage estimator, suggesting that the PrLS estimator is sensitive to the quality of the plugged-in precision matrix. The POET estimator is the overall most biased estimator in this setting. The Chosleky

[13] See Online Appendix A4 for more details.

estimator is instead up to 33 times less biased compared to the other estimators. Its bias is the closest to 0. However, the bias increase as p increases when s_0 is large. Since $\boldsymbol{\Theta}_x$ is sparse in this setting, we test the robustness of this results assuming that $\boldsymbol{\Theta}_x$ is dense. To do that, we add a single factor simulated from a standard normal distribution to the $x_{j,t}$'s. The results are qualitatively similar and reported under the column "Factor augmented" of the same table.

Insert Table 2 approximately here

Table 2 shows that a naive plug-in of a regularized estimator into the PrLS in (4) can lead to a poor estimator of $\boldsymbol{\beta}$. We find the Cholesky estimator to give a good approximation whenever $\boldsymbol{\Theta}_x$ is sparse or when the dimension is not too large. As clear, the unbiasedness of the PrLS based on any of the estimators mentioned above depends crucially on the level of sparsity in $\boldsymbol{\beta}$. To reduce this regularization bias of the estimated PrLS for a given precision matrix estimator, we rely on the debiasing formula given below in (13). We find that it can reduce up to 40% the bias of a Naive estimator when the latter is moderately biased (Table 2). Hereafter, we provide a theoretical justification of this debiasing procedure. Although for simplicity and illustration purposes, only the Cholesky estimator will be used in the subsequent examples or applications, the theoretical results below hold for any moderately biased estimator.

Denote $\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ the estimation error for the regularized precision matrix $\tilde{\boldsymbol{\Theta}}$. Since $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma} = \boldsymbol{I} + \tilde{\boldsymbol{\Delta}}\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma} = \boldsymbol{I} + \boldsymbol{\Sigma}\tilde{\boldsymbol{\Delta}}$, it follows from simple algebraic manipulations that $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}} = \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Delta}} + \boldsymbol{\Delta}\Sigma\tilde{\boldsymbol{\Delta}}$ and $2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta} = \tilde{\boldsymbol{\Delta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Delta}}$. Therefore, if $\tilde{\boldsymbol{\Theta}}$ is moderately biased, i.e., $\tilde{\boldsymbol{\Delta}} \approx 0$ for a large sample size, then $2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}}$ is a bias-corrected estimator of $\tilde{\boldsymbol{\Theta}}$. In other words, as the sample size increases, the quadratic form $\tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}}$ will converge to $\tilde{\boldsymbol{\Theta}}$ only if the precision matrix is perfectly estimated. Otherwise, the bias will surge. A simple subtraction of the bias from the regularized estimate of the precision matrix leads asymptotically to an unbiased version of $\tilde{\boldsymbol{\Theta}}$. As $2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}}$ depends on the population covariance matrix $\boldsymbol{\Sigma}$, we shall consider the plug-in estimator

$$\hat{\boldsymbol{\Theta}} = 2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\Theta}}, \tag{13}$$

where $\hat{\boldsymbol{\Sigma}}$ is the empirical counterpart of $\boldsymbol{\Sigma}$. Theorem 3 establishes that $\hat{\boldsymbol{\Theta}}$ is asymptotically a consistent and unbiased estimator of $\boldsymbol{\Theta}$ if $\tilde{\boldsymbol{\Theta}}$ is a moderately biased estimator of $\boldsymbol{\Theta}$. Let us stress here that the theorem below is not specific to any particular estimator of the precision matrix. Hence a specific rate is not given. However, a general error bound is given for *any* regularized precision matrix estimator.

Theorem 3. Denote $\tilde{\boldsymbol{\Delta}} \equiv \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ as the bias of $\tilde{\boldsymbol{\Theta}}$. Then, the following decomposition and bound are in order:

$$\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} = \tilde{\boldsymbol{\Delta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Delta}} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \tag{14a}$$

$$||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty \lesssim o_p\left(\delta_{T,p}^2\right)||\boldsymbol{\Sigma}||_\infty + o_p\left(a_{T,p}\right)\left(o_p\left(\delta_{T,p}\right) + ||\boldsymbol{\Theta}||_\infty\right)^2, \tag{14b}$$

where $\delta_{T,p}$ and $a_{T,p}$ are the rate of convergence of $\tilde{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Sigma}}$ with respect to an infinite norm $||.||_\infty$.

Theorem 3 implies that the main requirement for Eq. (13) to be valid is that $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Theta}}$ should be consistent estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$, respectively. It is clear that $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty \to 0$ as long as $\delta_{T,p} \to 0$ and $a_{T,p} \to 0$. However, Eq. (14a), given the term $\tilde{\boldsymbol{\Delta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Delta}}$, shows that some finite sample biases can emerge. Yet, it is easy to show that bias reduction indeed occurs, i.e., $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty < o_p\left(\delta_{T,p}\right)$ if $a_{T,p} < \frac{o_p\left(\delta_{T,p}^2\right) - o_p\left(\delta_{T,p}\right)||\boldsymbol{\Sigma}||_\infty}{\left(o_p\left(\delta_{T,p}\right) - ||\boldsymbol{\Theta}||_\infty\right)^2}$ and $\delta_{T,p} < 1/||\boldsymbol{\Sigma}||_\infty$. Furthermore, one has $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty \leq o_p\left(\delta_{T,p}^2\right)$ if $a_{T,p} \leq \frac{o_p\left(\delta_{T,p}^2\right)(1 - ||\boldsymbol{\Sigma}||_\infty)}{\left(o_p\left(\delta_{T,p}\right) - ||\boldsymbol{\Theta}||_\infty\right)^2}$ assuming wlog that $0 < ||\boldsymbol{\Sigma}||_\infty < 1$. Consequently, bias reduction is possible (using (13)) for the regularized. For instance, any estimator that achieves the optimal rate in the class of sparse precision matrix of Cai and Zhou (2012) leads to $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty \leq o_p\left(\delta_{T,p}^2\right)$. See for instance their ACLIME estimator.

Similarly, the estimators in e.g., Fan et al. (2008); Lam and Fan (2009); Fan et al. (2008); Cai and Zhou (2012); Cai et al. (2016) and the Cholesky one (see also Online Appendix A4) also imply that $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_\infty \leq o_p\left(\delta^2_{T,p}\right)$. Their exact rates of convergence depend on their regularity conditions.

Remark 3. The bound in Theorem 3 can in principle be used as a tool to rank different regularization estimators of the precision matrix in terms of their moderate biasedness, as the bound in (14b) depends on the specific rates of convergence of $\tilde{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\Sigma}}$. In this sense, one could think of plugging rates of different estimators of the precision matrix and rank them from the least to the most biased. However, as typically different estimators rely on substantially different sets of assumptions, whose impact on the bias is not necessarily clear, immediate comparisons via plugging rates of different estimators could be hard, if not misleading. Still, this result could be of independent interest and be used to test the effect of certain assumptions of a single estimator of the precision matrix on its moderate bias. In practice, one would also need a test for goodness of fit in order to rank different precision matrix estimators. Yet, the development of such a test is out the scope of this paper and it is left for future research.

Remark 4. Eq. (13) has been derived by Janková and van de Geer (2018); Koike (2020) to build confidence region around precision matrices of multivariate Gaussian data using the graphical lasso estimator or the nodewise regression. Kashlak (2021) conjectures that it can be used to correct the bias of other regularized estimators too. We show that Eq. (13) can indeed be used to correct the bias of any moderately biased precision matrix estimator. Note that we have derived this equation without making any specific distributional assumption and we used it only in the context of linear models.

The following proposition shows how the proposed bias correction method helps in safeguarding against the regularization effect in the estimation of $\boldsymbol{w}_0$.

Proposition 1. Denote $a = \dfrac{\mathbb{1}\boldsymbol{\Theta}\mathbb{1}}{\mathbb{1}\hat{\boldsymbol{\Theta}}\mathbb{1}}$ and recall that $k^{-1} = \dfrac{\mathbb{1}\tilde{\boldsymbol{\Theta}}\mathbb{1}}{\mathbb{1}\hat{\boldsymbol{\Theta}}\mathbb{1}}$, $\hat{\boldsymbol{w}} = \dfrac{\hat{\boldsymbol{\Theta}}\mathbb{1}}{\mathbb{1}\hat{\boldsymbol{\Theta}}\mathbb{1}}$. Then, one has:

$$\hat{\boldsymbol{w}} = 2k^{-1}\tilde{\boldsymbol{w}} - k^{-1}\tilde{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{w}}, \tag{15}$$

$$\boldsymbol{w}_0 - \hat{\boldsymbol{w}} = (1-a)\boldsymbol{w}_0 + \boldsymbol{\Delta}\boldsymbol{\Sigma}\left[k^{-1}\tilde{\boldsymbol{w}} - a\boldsymbol{w}_0\right] + k^{-1}\tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{w}}. \tag{16}$$

Eq. (15) shows how debiasing the precision matrix is equivalent to debiasing the naive PrLS estimator. Eq. (16) is the bias expansion of the resulting PrLS estimator. The quantities a and k^{-1} are expected to converge to one if the underlying parameters are consistently estimated. In these cases, the bias is expected to vanish under regularity conditions of $\tilde{\boldsymbol{\Theta}}$. See Example 3 for illustrations.

Note that, given the naive estimator $\tilde{\boldsymbol{w}}$, one can also construct an unbiased estimator $\hat{\boldsymbol{\beta}}^D$ defined below in Eq. (17). Recall that $\tilde{\boldsymbol{w}} = (\tilde{\boldsymbol{\beta}}', 1 - \tilde{\boldsymbol{\beta}}\mathbb{1})'$ and let

$$\hat{\boldsymbol{\beta}}^D := \tilde{\boldsymbol{\beta}} + T^{-1}\boldsymbol{K}\sum_{t=1}^{T}\boldsymbol{x}_t'\left(y_t - \boldsymbol{x}_t'\tilde{\boldsymbol{\beta}}\right). \tag{17}$$

Proposition 2.

$$\hat{\boldsymbol{\beta}}^D - \boldsymbol{\beta} = (\boldsymbol{K}\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{I})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + T^{-1}\boldsymbol{K}\sum_{t=1}^{T}\boldsymbol{x}_t'\epsilon_t, \tag{18}$$

$$||\hat{\boldsymbol{\beta}}^D - \boldsymbol{\beta}||_\infty \leq ||\boldsymbol{K}\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{I}||_\infty \; ||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||_\infty + o_p(1), \tag{19}$$

for $c > 0$ with $\hat{\boldsymbol{\Sigma}}_x$ the empirical covariance of $\boldsymbol{x}_t$.

Proposition 2 shows that $\hat{\boldsymbol{\beta}}^D$ is very similar to our estimator $\hat{\boldsymbol{w}}$ as one can see by comparing Eq. (15) and (18). In fact, one can see from (19) that the validity of the bias correction only depends on the accuracy of $\tilde{\boldsymbol{\beta}}$ or that of $\boldsymbol{K}\hat{\boldsymbol{\Sigma}}_x$, while the $o_p(1)$ term comes from the exogeneity condition. A key role is played by the matrix $\boldsymbol{K}$, whose function is to "decorrelate" the columns of $\boldsymbol{x}_t$. The idea is to assume that $\boldsymbol{K} \approx \boldsymbol{\Omega}_x$, the precision matrix of $\boldsymbol{x}_t$ (Van de Geer et al., 2014).[14] The interaction between the rate of convergence of $\boldsymbol{K}\hat{\boldsymbol{\Sigma}}_x$ to $\boldsymbol{I}$ and that of $\tilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$ makes this bias correction appealing. In other words, as long as the rate of convergence of one of the two elements in the product is fast enough, consistency of $\hat{\boldsymbol{\beta}}^D$ is established, and $\hat{\boldsymbol{\beta}}^D$ behaves asymptotically like $\hat{\boldsymbol{\beta}}$: the procedure compensates for the bias introduced by the estimation of $\boldsymbol{\Theta}$. Such a correction has been extensively used in the literature to construct unbiased estimators for the lasso and ridge estimators (Bühlmann et al., 2013; Javanmard and Montanari, 2014; Adamek et al., 2022, and references therein). In this paper, we extend this to the naive PrLS estimator.

3.2 Variance estimation

Having obtained a nearly unbiased estimator of $\boldsymbol{\beta}$, a consistent estimator of the variance – that we denoted $\boldsymbol{V}$– remains necessary for inference. In the different illustrations above, as we simulated i.i.d data, we used the oracle least-square variance formula, which is also that of the PrLS.[15] However, in practice, estimation of $\boldsymbol{V}$ is not trivial when using a regularized estimator of the precision matrix or when the error term ϵ_t is heteroskedastic or exhibits autocorrelation. In that case, the i.i.d. formula is no longer valid. The traditional solution would be to use kernel or bootstrap-based estimators. These, though, involve the selection of a bandwidth or are usually computationally intensive. Here, we propose a heteroskedastic and autocorrelation robust variance estimator based on the principle of "batch-mean" or "subseries method". This type of variance estimator is computationally cheap, kernel-free, has good finite sample properties and it has been extensively studied in the literature Carlstein et al. (1986); Flegal et al. (2010); Zhang et al. (2017). Assume that the sample is split into Π folds and denote I_π as the index set of the observations in the fold $\pi = 1, \ldots, \Pi$.

[14] $\boldsymbol{K}$ can also be estimated using some a priori knowledge about the linear dependencies among the $\boldsymbol{x}_t$. E.g., $\boldsymbol{x}_t$ is generated from a factor model or a sparse graphical model. Therefore $\boldsymbol{K}$ can be a consistent regularized estimator of the precision matrix of $\boldsymbol{x}_t$ which can be readily subtracted from $\tilde{\boldsymbol{\Theta}}$ for instance.

[15] See Footnote 11.

A consistent estimator of $\boldsymbol{V} := \text{Var}(\hat{\boldsymbol{\beta}})$ is given by

$$\hat{\boldsymbol{V}} = \frac{1}{\Pi - 1} \sum_{\pi=1}^{\Pi} (\hat{\boldsymbol{\beta}}^{\pi} - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{\pi} - \bar{\boldsymbol{\beta}})', \tag{20}$$

where $\bar{\boldsymbol{\beta}} = \frac{1}{\Pi} \sum_{\pi=1}^{\Pi} \hat{\boldsymbol{\beta}}^{\pi}$ and $\hat{\boldsymbol{\beta}}^{\pi}$ is the estimated coefficient ($\hat{\boldsymbol{\beta}}$ or $\hat{\boldsymbol{\beta}}^{D}$) on the fold I_{π}. The consistency of $\hat{\boldsymbol{V}}$ is guaranteed by application of Zhang et al. (2017) Theorem 5.1 to the process $T^{-1/2}\boldsymbol{u}_t$, where $\boldsymbol{u}_t = \boldsymbol{\Theta}_x \boldsymbol{x}_t \epsilon_t$. However, as a direct application of (20) would involve re-estimation of $\tilde{\boldsymbol{\Theta}}$ Π times, we propose equivalent variance estimators specific to $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{D}$ that avoid re-estimation of $\tilde{\boldsymbol{\Theta}}$. The following propositions state the consistency of the batch mean procedure in the case of the PrLS estimator $\hat{\boldsymbol{w}}$ and the bias-corrected estimator $\boldsymbol{\beta}^{D}$, respectively.

Proposition 3. Let

$$\hat{\boldsymbol{\Theta}}^{\pi} := 2\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}^{-1}\widehat{\text{Var}}(\boldsymbol{f}_t)\tilde{\boldsymbol{\Theta}}, \quad \forall t \in I_{\pi}. \tag{21}$$

Then, $\widehat{\text{Var}}(\hat{\boldsymbol{w}}^{\pi}) - \boldsymbol{V} = o_p(1)$ with $\widehat{\text{Var}}(\hat{\boldsymbol{w}}^{\pi})$ being the empirical variance of $\hat{\boldsymbol{w}}^{\pi}$.

Proposition 4. Let

$$\hat{\boldsymbol{\beta}}^{D,\pi} = \tilde{\boldsymbol{\beta}} + (|\boldsymbol{I}_{\pi}| - 1)^{-1} \boldsymbol{K} \sum_{t \in I_{\pi}} \boldsymbol{x}_t{}' \left(y_t - \boldsymbol{x}_t \tilde{\boldsymbol{\beta}} \right), \quad \forall \pi = 1, \ldots, \Pi, \tag{22}$$

where $|I_{\pi}|$ denotes the carnality of I_{π}. Then, $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{D,\pi}) - \boldsymbol{V} = o_p(1)$ with $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{D,\pi})$ being the empirical variance of $\hat{\boldsymbol{\beta}}^{D,\pi}$.

According to Flegal et al. (2010), a fixed-width rule $\Pi = [z]$, where $[z]$ denotes the integer part of z, can be used to choose the number of folds Π. Their analysis showed that z can be set to $2/3$, $1/2$ or $1/3$ in practice.[16] Note that testing the significance of a parameter $G(\boldsymbol{\beta})$, where $G(.)$ is a specific measurable function, can be done readily. In

[16] It is well known that the parameter Π makes a trade-off between size and power especially when the time series serial dependence is high.

fact, under some regularity condition (Carlstein et al., 1986), $G(\hat{\boldsymbol{\beta}})$ and $G(\hat{\boldsymbol{\beta}}^D)$ converge to $G(\boldsymbol{\beta})$ in probability. By replacing $G(\hat{\boldsymbol{\beta}}^\pi)$ with $G(\boldsymbol{\beta})$ or $G(\hat{\boldsymbol{\beta}}^{D,\pi})$ in the corresponding formula, statistical inference can be conducted using Gaussian critical values. Note also that the batch mean-variance estimator performs asymptotically similarly to the bootstrap but it is computationally cheaper and is also robust to fat tails and skewed innovations ϵ_t (Carlstein et al., 1986; Flegal et al., 2010; Zhang et al., 2017).[17]

4 Monte-Carlo

We give here an overview of the Monte-Carlo setting and its main results. Section A2 in the Online Appendix contains all figures and additional details. Consider the linear model (1). We simulate normal i.i.d data such that $\beta_j = 2$, for $j = 1, \ldots, 50$, $\boldsymbol{\Sigma}_x = (0.6)^{|i-j|}$, $p = 500$, $T = 250$, and $\mathbb{E}(\epsilon_t^2) = 0.1$. The number of replications is 500. Firstly, we compare the naive PrLS $\hat{\boldsymbol{\beta}}^{Naive}$ with its two proposed debiased versions i.e., $\hat{\boldsymbol{\beta}}^{PrLS}$, namely the PrLS with debiasing as in Eq. (13) and $\hat{\boldsymbol{\beta}}^D$ as presented in Eq. (17). Refitted adaptive lasso-based modified Cholesky decomposition (presented in Online Appendix A4) is used as a plug-in estimator of the precision matrix to get $\hat{\boldsymbol{\beta}}^{Naive}$, $\hat{\boldsymbol{\beta}}^{PrLS}$ and $\hat{\boldsymbol{\beta}}^D$. Furthermore, we add the comparisons with: the debiased lasso in Van de Geer et al. (2014) and Adamek et al. (2022), the debiased ridge in Bühlmann et al. (2013). The variance of $\hat{\boldsymbol{\beta}}^{PrLS}$ and $\hat{\boldsymbol{\beta}}^D$ is obtained via the subseries method presented in Section 3.2. The number of subseries values is fixed to $T^{1/3}$ across the simulations. We also include as a benchmark the infeasible OLS estimator to mimic the behavior of an oracle estimator. This estimator is obtained by augmenting the simulated sample size of the data such that $T = 750$. Below, we report the main findings.

[17]It is worth mentioning that Using the batch mean estimator (20) of $\boldsymbol{V}$, joint hypothesis testing is possible whenever the size of $\boldsymbol{b}$ is lower than the number of batches Π. In fact, given the result of Lemma A1.1, the low-dimensional joint normality of any sub-vector of $\hat{\boldsymbol{\beta}}$ follows from a simple application of the continuous mapping theorem.

- (Fig A.1) $\boldsymbol{\beta}^{PrLS}$, $\boldsymbol{\beta}^{D}$, and $\boldsymbol{\beta}^{Naive}$ behave similarly in terms of bias. This bias is close to 0 and it is similar to that of the infeasible OLS. Instead, Van de Geer et al. (2014), Bühlmann et al. (2013) and Adamek et al. (2022) debiased estimators exhibit substantial bias.

- (Fig A.3a, A.3b) $\boldsymbol{\beta}^{PrLS}$, $\boldsymbol{\beta}^{D}$, and $\boldsymbol{\beta}^{Naive}$ are shown to be uniformly unbiased over the set of all coefficients. Conversely, the other estimators display some degree of bias and even show erratic behaviors of some coefficients.

- (Fig A.5a) For the rejection frequencies of the tests of individual significance of the irrelevant coefficients, the test based on the infeasible OLS and Van de Geer et al. (2014) estimators have size around the nominal 5% level. The tests based on the other estimators have size lower than 5%. This is expected, as a PrLS-based test discriminates strongly between relevant and irrelevant variables.

- (Fig A.5b) for the rejection frequency of the tests of individual significance of the relevant coefficient, a part from the test based on the Adamek et al. (2022) estimator, all the other tests have power close to one.

Overall, our Monte Carlo simulation results illustrate the good performance of the PrLS estimator. It compares favorably in terms of bias, size, and power to traditional state-of-the-art estimators. Although the reported results are based on i.i.d. normal data, similar results are obtained for ARMA-GARCH processes with skewed student innovations for the covariates and error terms. These results are omitted for the sake of space and are available upon request.

5 Empirical Application

We examine the yearly financial connectedness of $n = 88$ banks from 28 countries around the world using daily stock log-returns, spanning from 2005 to 2020. The list of these banks is reported in Table A5.1 in the Online Appendix. Stock prices $P_{i,t}$ are downloaded from Datastream in local currency and their log-returns $y_{i,t}$ for $i = 1, \dots, n$ are computed as $y_{i,t} = \log P_{i,t} - \log P_{i,t-1}$. Our analysis relies on networks based on conditional Granger causality tests (Granger, 1969) as well as the concept of modularity (Clauset et al., 2004). We briefly introduce both concepts hereafter and we give the full mathematical details in Section A5.

Granger causality captures predictability given a particular information set. Conditional on an information set containing the past returns of all institutions, if the past values of the return of institution i' improve the prediction of the return of institution i at time t, then i' is Granger causal for i. For each institution $i = 1, \dots n$, we estimate a VAR(5) model for each year ($T \approx 250$).[18] Since $\boldsymbol{\beta}_i \in \mathbb{R}^{440}$, estimation and inference via OLS is infeasible, and a natural solution is to use our PrLS estimator. The refitted adaptive lasso-based modified Cholesky decomposition is used as a consistent estimator of the precision matrices.[19] The number of subseries values is also fixed to $T^{1/3}$. Furthermore, to prevent the accumulation of type I errors due to multiple hypothesis testing, we apply a Benjamini and Yekutieli (2001) false-discovery control procedure at each estimation step. Thus, for each year from 2005 to 2020, we obtain an adjacency matrix $\boldsymbol{A}$ containing all the information about the connectedness among the institutions, such that $A_{i,i'} = 1$ if at least one of the parameters $\beta_{i,6}, \dots, \beta_{i,10}$ has its associated adjusted p-value lower than 20% and zero otherwise.

[18]For an overview of lag selection procedures, see Hecq et al. (2021). Following their approach, we find that the optimal lag is lower than 5 in the different estimations, but we overspecified the number of optimal lags to avoid potential lag truncation issues. We use 5 to account for the potential dynamic in a week.

[19]See Online Appendix A4 for its presentation and Remark A4.1 for the justification of this choice in this context.

The concept of *modularity* comes from graph theory. Its interpretation can be formalized as follows. If the expected fraction of within-community edges for a randomized network is not different from that observed in the estimated network, then the modularity will be zero. Therefore, nonzero values represent deviations from randomness, where a large value indicates significant community structure in a network. We employ the hierarchical agglomeration algorithm of Clauset et al. (2004) to detect communities. The algorithm uses a greedy optimization strategy that starts with each vertex as a unique member of a community of one and repeatedly joins the two communities whose amalgamation produces the largest increase in modularity.

5.1 Results

Insert Figure 4 approximately here

Figure 4 displays the Granger causality networks obtained with the precision least squares estimator between 2006 and 2010. We focus on this period to observe the behavior of the connections before, during, and after the 2008 financial crisis. The patterns observed are very insightful. By looking at the causal networks, we observe how financial connectedness substantially decreases when approaching the financial crisis and peaks at its lowest in the crisis year, to only slightly recover in 2009 and fully return to dense in 2010. Prediction in the presence of a structural break as in 2008 (subprime crisis) is very difficult and is reflected by the lack of connections in stock returns during 2008. Interestingly, 2008 does not look empty, meaning that a subset of series can still have predictive power over an unprecedented crisis. Figure 5(a) reports the dynamic of the number of communities (membership) and connections obtained using the greedy optimization algorithm of Clauset et al. (2004) based on modularity. We can deduce that financial predictive connectedness is useful for characterizing financial crises but fails to capture the relative importance of the crisis. Membership, on the other hand, can capture the magnitude of the crisis. In fact, it peaks with large shocks, such as in 2008 (subprime), 2016 (downside macroeconomic concerns), and 2020 (COVID-19), but it is less sensitive to small shocks, such as in 2011 (European Debt) and 2018 (Turkish currency and debt). In Figure 5(b), we show that modularity summarizes the information contained in both the number of connections and the number of communities.

Our analysis is therefore relevant for at least three reasons. First, we show that in practice, the PrLS estimator is reliable and capable of handling high-dimensional systems and producing sensible results. Second, the variation in the number of financial connections among stock returns, as measured in our predictive Granger causal sense, seems to serve as an indicator of a financial crisis when the overall number of connections decreases substantially. Third, in parallel to financial connectedness, modularity also serves as an (early warning) indicator of financial crises.

Insert Figures 5 approximately here

6 Conclusion

In this paper we introduce the Precision Least Squares estimator (PrLS). PrLS spans a class of least square estimators for both low and high-dimensional models. It is based on a simple plug-in approach that only requires estimating an unbiased and consistent precision matrix of a linear transformation of the data. Under covariance stationarity of the data the PrLS is shown to be asymptotically Gaussian. Furthermore, using the subseries method the covariance matrix of PrLS can be estimated with robustness to moderate misspecifications of the plugged-in precision matrix. The power of tests of variables significance based on the PrLS estimates are also shown to be almost insensitive to p-value adjustments in finite samples. In fact, PrLS guarantees asymptotic control over the directional familywise error rate.

We also show that in high-dimension, any regularized estimator of the precision matrix —with a rate of convergence at least as fast as that of the sample covariance matrix — can be used to get a near-oracle PrLS estimator after a simple regularization bias correction. We also show that for some state-of-the-art regularized methods, the sparsity assumption on the regression coefficient is ineluctable. We introduce a subseries method for regularized precision matrix-based estimators which avoids refitting and makes the proposed inferential procedure as fast as a kernel-based estimator. We illustrate the benefits of using our estimation and inferential procedure on simulated data. This shows how PrLS compares favorably to states-of-the-arts methodologies.

Empirically, we employ PrLS to investigate financial connectedness by estimating high-dimensional Granger causal networks of 88 banks worldwide, using daily asset returns from 2005 until 2020. Focusing on each specific year, we find evidence that predictive connections among bank asset returns had a significant decrease in the year of a financial or an economic crisis due to the unpredictability of such an economic or financial collapse. The network density increases at financial or economic crisis proximity. Our empirical result reconciles the empirical literature on measuring systemic risk using the VAR model with theories of financial market fragility and reputation.

Overall, our precision least squares estimator provides a versatile and almost "off the shelf" tool for large-scale inference in linear regression models with any moderately biased precision matrix estimator.

References

R. Adamek, S. Smeekes, and I. Wilms. Lasso inference for high-dimensional time series. *Journal of Econometrics*, 2022.

I. Archakov and P. R. Hansen. A new parametrization of correlation matrices. *Econometrica*, 89 (4):1699–1715, 2021.

I. Archakov and P. R. Hansen. A canonical representation of block matrices with applications to covariance and correlation matrices. *Review of Economics and Statistics*, pages 1–39, 2022.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94, 2015.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

J. Bradic, J. Fan, and Y. Zhu. Testability of high-dimensional linear models with nonsparse structures. *Annals of Statistics*, 50(2):615, 2022.

P. Bühlmann et al. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, 44(2):455, 2016.

T. T. Cai and H. H. Zhou. Minimax estimation of large covariance matrices under ℓ_1-norm. *Statistica Sinica*, pages 1319–1349, 2012.

T. T. Cai, Z. Guo, et al. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, 45(2):615–646, 2017.

L. Callot, M. Caner, A. Ö. Önder, and E. Ulaşan. A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, pages 1–12, 2019.

M. Caner, M. Medeiros, and G. F. Vasconcelos. Sharpe ratio analysis in high dimensions: Residual-based nodewise regression in factor models. *Journal of Econometrics*, 2022.

E. Carlstein et al. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, 14(3):1171–1179, 1986.

J. Chang, Y. Qiu, Q. Yao, and T. Zou. Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57–82, 2018.

A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

J. Fan, Y. Liao, and X. Shi. Risks of large portfolios. *Journal of Econometrics*, 186(2):367–387, 2015.

J. Fan, A. Furger, and D. Xiu. Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics*, 34(4):489–503, 2016a.

J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *Econometrics Journal*, 19(1):C1–C32, 2016b.

J. Fan, H. Liu, and W. Wang. Large covariance estimation through elliptical factor models. *Annals of Statistics*, 46(4):1383, 2018.

J. M. Flegal, G. L. Jones, et al. Batch means and spectral variance estimators in markov chain monte carlo. *Annals of Statistics*, 38(2):1034–1070, 2010.

P. Gagliardini, E. Ossola, and O. Scaillet. Time-varying risk premium in large cross-sectional equity data sets. *Econometrica*, 84(3):985–1046, 2016.

D. Giannone, M. Lenza, and G. E. Primiceri. Economic predictions with big data: The illusion of sparsity. *Econometrica*, 89(5):2409–2437, 2021.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

A. Hecq, L. Margaritella, and S. Smeekes. Granger Causality Testing in High-Dimensional VARs: A Post-Double-Selection Procedure. *Journal of Financial Econometrics*, 2021.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

B. James, K. James, and Y. Qi. Limit distribution of the sum and maximum from multivariate gaussian sequences. *Journal of Multivariate Analysis*, 98(3):517–532, 2007.

J. Janková and S. van de Geer. Inference in high-dimensional graphical models. In *Handbook of Graphical Models*, pages 325–350. CRC Press, 2018.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

A. B. Kashlak. Non-asymptotic error controlled sparse high dimensional precision matrix estimation. *Journal of Multivariate Analysis*, 181:104690, 2021.

A. Kempf and C. Memmel. Estimating the global minimum variance portfolio. *Schmalenbach Business Review*, 58(4):332–348, 2006.

Y. Koike. De-biased graphical lasso for high-frequency data. *Entropy*, 22(4):456, 2020.

R. Lagunoff and S. L. Schreft. A model of financial fragility. *Journal of Economic Theory*, 99 (1-2):220–264, 2001.

C. Lam. High-dimensional covariance matrix estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(2):e1485, 2020.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254, 2009.

C. Lam et al. Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Annals of Statistics*, 44(3):928–953, 2016.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119, 2004.

O. Ledoit and M. Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

O. Ledoit and M. Wolf. The power of (non-) linear shrinking: a review and guide to covariance matrix estimation. *Journal of Financial Econometrics*, 2019.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, pages 21–59, 2005.

R. Nickl, S. Van De Geer, et al. Confidence sets in sparse regression. *Annals of Statistics*, 41(6): 2852–2876, 2013.

G. L. Ordoñez. Fragility of reputation and clustering of risk-taking. *Theoretical Economics*, 8(3): 653–700, 2013.

M. Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.

D. Zhang, W. B. Wu, et al. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

7 Figures and Tables

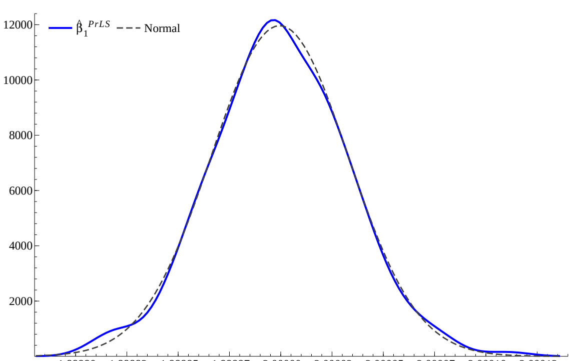
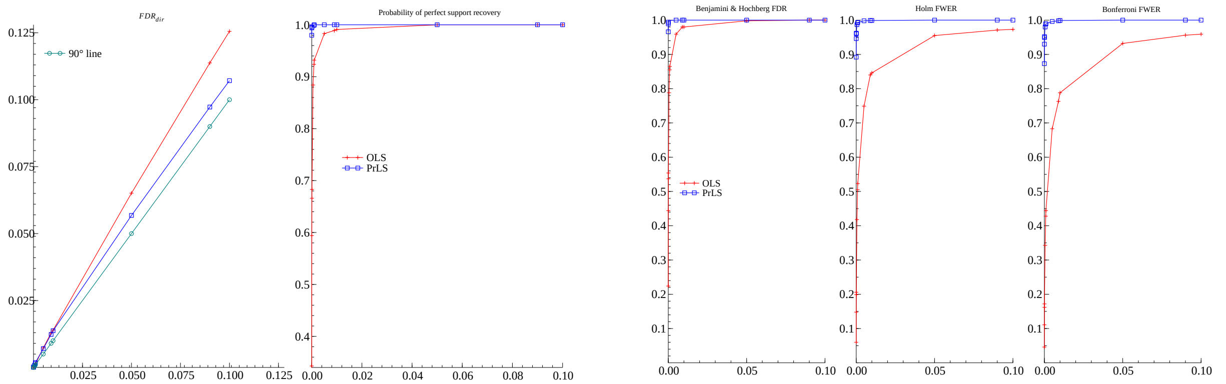


Figure 1: Simulated vs theoretical density of $\hat{\boldsymbol{\beta}}^{\text{PrLS}}$ when $\boldsymbol{\beta}$ is dense

	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}^{PrLS}$	RF
β_1	2	1.9489	0.915
β_2	2	1.9489	0.915
β_3	2	1.9489	0.915
$\beta_4 - \beta_{1003}$	0	0	0

Table 1: Average estimates of $\hat{\boldsymbol{\beta}}^{\text{PrLS}}$ and the rejection frequencies (RF) of significance tests at 5% when $\boldsymbol{\beta}$ is sparse



(a) FDR_{dir} and Probability of perfect support recovery without p-value adjustment.

(b) Probability of perfect support recovery after p-value adjustment.

Figure 2: Insensitive power of the PrLS after p-value adjustment

Notes: We simulate $\boldsymbol{x}_t \in \mathbb{R}^{50}$, for $t = 1, \ldots, 100$, from a single factor model. The factor follows a standard normal distribution. The factor loadings are drawn from a normal distribution and set fixed across the 1000 simulations. The error process is i.i.d zero-mean normal with covariance $|0.8|^{i-j}$. We randomly select 20 coefficients and set them to 2 while the other coefficients are set to 0. $\boldsymbol{\Theta}$ is estimated using a refitted adaptive lasso-based modified Cholesky decomposition estimator and obtained an initial estimator $\tilde{\boldsymbol{\Theta}}$ (see Online Appendix A4 for details). $\hat{\boldsymbol{\Theta}}$ is obtained after (13). The variance of the PrLS estimator is given by the formula in footnote 11. $FDR_{dir} := \mathbb{E}(FDP_{dir}) \geq FWER_{dir}$ where $FDP_{dir} := \left|\left\{j \in \widehat{R} : \widehat{\text{sign}}_j \neq \text{sign}\,(\beta_j)\right\}\right| / \max(|\widehat{R}|, 1)$. The significant level is on the x-axis.

Table 2: Maximum excess bias ×100 with respect to an infeasible estimator

		Toeplitz				Factor augmented			
					$s_0 = 10$				
p	Type	Cholesky	Linear Shrinkage	Non-linear Shrinkage	POET	Cholesky	Linear Shrinkage	Non-linear Shrinkage	POET
100	Naive	3.10	132.87	128.88	205.30	3.61	108.87	71.71	201.00
	PrLS	3.04	91.55	89.00	301.48	3.60	73.49	41.73	240.56
250	Naive	3.80	178.94	172.78	201.43	4.15	138.33	175.36	200.46
	PrLS	3.70	154.08	148.71	256.45	4.12	115.33	150.82	236.52
					$s_0 = 60$				
100	Naive	2.02	108.68	67.31	230.97	1.88	161.16	44.21	201.03
	PrLS	1.99	73.09	45.38	430.32	1.88	130.74	29.09	257.80
250	Naive	8.63	170.03	170.18	221.84	27.28	174.36	167.66	201.07
	PrLS	7.09	144.54	147.41	276.87	14.52	155.32	144.52	253.83

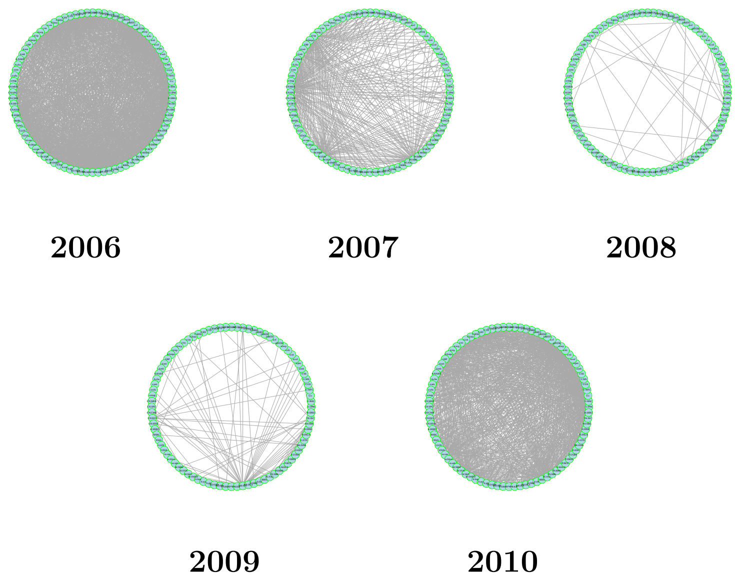


Figure 4: Granger-causality networks via PrLS

(a) Connections and Membership

(b) Financial Connectedness

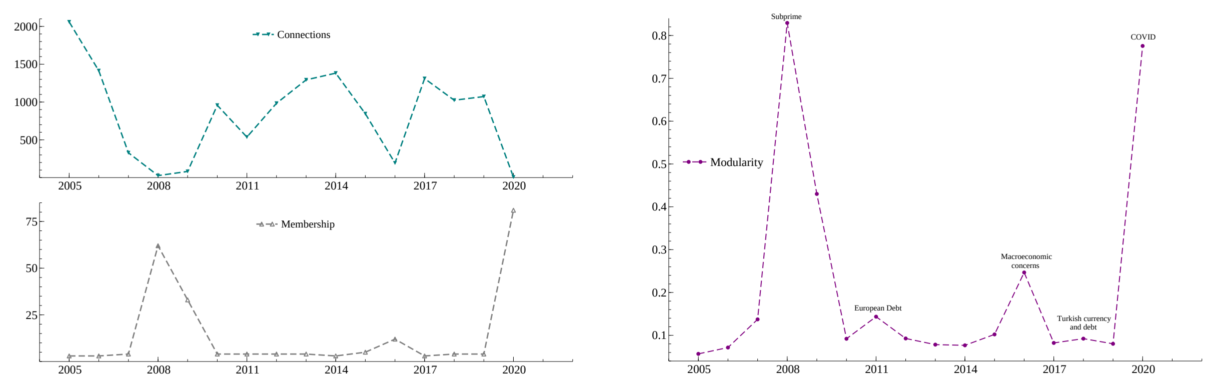


Figure 5: Connections, Membership, modularity

Online Appendix to: "Precision Least Squares: Estimation and Inference in High-Dimensions"

Luca Margaritella*
Rosnel Sessinou†

March 14, 2023

Abstract

This online appendix supplement provides (i) the details of a framework for oracle approximation used in Section 2 of the main paper, including a concentration inequality for dependent processes; (ii) the Monte-Carlo details; (iii) the proofs of (a.) the concentration inequality, (b.) all theorems and propositions of the main paper; (iv) details of the regression-based modified Cholesky decomposition, which is the main precision matrix estimator used in the paper for illustrations; (v) details on the methodology and the sample of banks used in the empirical application of Section 5 of the main paper, along with additional figures and a discussion.

*Department of Economics, Lund University Email: luca.margaritella@nek.lu.se.
†Department of Decision Sciences, HEC Montréal. Email: rosnel.sessinou@hec.ca

A1 A framework for oracle approximation

Section 2 shows that the consistency of the precision least square estimator, given an unbiased and consistent estimator $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$, depends on the rate of convergence of $\boldsymbol{\Sigma}_x$. In this section, we give more insights on the required rate of convergence of $\boldsymbol{\Sigma}_x$ and discuss the asymptotic normality of the PrLS. First, we give the following lemma as the starting point of our discussion.

Lemma A1.1. Given the decomposition of the oracle least squares provided in Eq. (10), the term $\boldsymbol{A}$ converges to zero in probability i.e., $||\boldsymbol{A}||_\infty = ||\boldsymbol{\Theta}_x \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{I}||_\infty = o_p(1)$.

Lemma A1.1 follows directly from the deviation bound given below in Proposition 1. Its proof is given in Section A3. The statement of Lemma A1.1 has a one to one correspondence with showing $\left\| \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x \right\|_\infty = o_p(1)$. Let $\boldsymbol{\chi}_t := \boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')$ and call $\boldsymbol{\Delta}_{x,t} := \hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x \equiv T^{-1} \sum_{t=1}^T \boldsymbol{\chi}_t$ the $p \times p$ matrix of differences between the population covariance for $\boldsymbol{x}_t$ and its empirical counterpart. It follows that the key ingredient to prove the claim of Lemma A1.1 is obtaining a deviation bound for the $p(p+1)/2$ process $\bar{\boldsymbol{\chi}}_T := \text{vech}(\boldsymbol{\Delta}_{x,t})$ where vech stacks on top of one another the columns of the lower triangle part (including the diagonal) of $\boldsymbol{\Delta}_{x,t}$. Therefore, we shall bound the probability for the maximum element of $\bar{\boldsymbol{\chi}}_T$ being large.[1] This means that the coordinates of $\bar{\boldsymbol{\chi}}_T$ shall uniformly concentrates around zero i.e., for a strictly positive sequence κ_T, $\mathbb{P}\left(\|\bar{\boldsymbol{\chi}}_T\|_\infty \geq \kappa_T\right) \leq Q_{T,p,*}$, where $Q_{T,p,*}$ is a quantity which depends on the sample size T, the number of variables p and other terms (indicated with a $*$) related to the type of dependence assumed over the stochastic process. We rely on the physical dependence framework as introduced by Wei Biao Wu in a sequence of papers (see e.g., Wu, 2005). As this is needed for the proofs, we formally introduce it in Section A3. In broad terms, as the Wold decomposition applies to a vast

[1] To make matters clear we illustrate this with a toy-example. Let $p = 2$ such that $\boldsymbol{\Delta}_{x,t} = \begin{bmatrix} x_{1,t}x_{1,t} - \mathbb{E}(x_{1,t}x_{1,t}) & \\ x_{2,t}x_{1,t} - \mathbb{E}(x_{2,t}x_{1,t}) & x_{2,t}x_{2,t} - \mathbb{E}(x_{2,t}x_{2,t}) \end{bmatrix}$; $\bar{\boldsymbol{\chi}}_T = T^{-1}\sum_{t=1}^T \begin{bmatrix} (x_{1,t}x_{1,t} - \mathbb{E}(x_{1,t}x_{1,t})) \\ (x_{2,t}x_{1,t} - \mathbb{E}(x_{2,t}x_{1,t})) \\ (x_{2,t}x_{2,t} - \mathbb{E}(x_{2,t}x_{2,t})) \end{bmatrix}$.

variety of stationary processes, many time series processes can be casted in an MA(∞) form as $\boldsymbol{x}_t = \sum_{\ell=0}^{\infty} \psi_\ell e_{t-\ell}$, for $\psi_0 = 1, \sum_{\ell=0}^{\infty} \psi_j^2 < \infty$ and e_t a white noise. As such, the white noise elements e_t forming this linear process can be viewed as independent input of a physical system where all the dependencies among the outputs $\boldsymbol{x}_t$ result from the underlying data generating mechanism $G(\cdot)$ such that $\boldsymbol{x}_t = (\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}, \dots, \boldsymbol{x}_{p,t})^\top = G(\mathcal{F}^t)$ for the infinite-past filtration $\mathcal{F}^t = (\dots, e_{t-1}, e_t)$ and likewise $x_{j,t} = g_j(\mathcal{F}^t)$, $t \in \mathbb{Z}$, where $g_j(\cdot)$, $1 \leq j \leq p$, is the j-th coordinate projection of $G(\cdot)$. The measure induced by the functional dependence considers the dependence of $x_{j,t}$ on e_k for some $k < t$, when e_k is replaced by an i.i.d. copy e_k while "freezing" the rest of the innovations. The deviation bound in the next proposition gives the probability rate of Lemma A1.1.

Proposition 1. Let $\boldsymbol{x}_{j,\cdot} = \{x_{j,t}\}_{t\in\mathbb{Z}}$ and $\boldsymbol{x}_{\cdot} = \{x_{j,t}\}_{t\in\mathbb{Z}, j=1,\dots,p}$. Assume finite $\mathcal{L}^\infty$ dependence adjusted norm $\||\boldsymbol{x}_{\cdot}|_\infty\|_{2,\alpha} < \infty$ for $q > 2, \alpha \geq 0$ and finite $q = 2$ uniform dependence adjusted norm $\Psi_{q,\alpha} = \max_{1\leq j\leq p} \|\boldsymbol{x}_{j,\cdot}\|_{2,\alpha} < \infty$ for $\alpha \geq 0$.[2] Let $\ell = \ell(p) = 1 \vee \log p$ then[3] for some $\alpha \geq 0$, there exists a strictly positive sequence κ_T and a constant α_1 such that the following (polynomial) Nagaev-type deviation bound holds:

$$\mathbb{P}\left(\|\bar{\boldsymbol{\chi}}_T\|_\infty \geq \kappa_T\right) \lesssim \frac{C_{q,\alpha} T^{\alpha_1+1} \ell^{q/2} \||\boldsymbol{x}_{\cdot}|_\infty\|_{q,\alpha}^q}{\kappa_T^q} + C_{q,\alpha} \exp\left(-\frac{C_{q,\alpha}\kappa_T^2}{T^{\alpha_1+1}\Psi_{2,\alpha}^2}\right), \tag{A1.1}$$

where $C_{q,\alpha}$ is a constant that depends only on q, α.

Both a "weaker" and a "stronger" dependence can be accounted for in the bound in (A1.1). To elaborate, the strength of the dependence is dictated by how slowly the cumulative functional dependence measures decays (see Section A3); therefore, for smaller $\alpha > 0$ the higher the dependence. In the "weaker" dependence case i.e., if $\alpha > 1/2 - 1/q$,

[2] See Section A3, Remark A3.1 for the formal definition of the dependence adjusted norms.

[3] The condition $\ell = \ell(p) = 1 \vee \log p$ is just to remark that the bound also holds for the univariate case. However, as we are in high dimensions, henceforth this should interpreted as $\ell = \log p$.

then for $\alpha_1 = 1, \ell = \log p$, we have

$$\|\bar{\boldsymbol{\chi}}_T\|_\infty = O_{\mathbb{P}}\left(T^{(\alpha_1+1)/q}\sqrt{\log p}\| \||\boldsymbol{x}_\cdot||_\infty\|_{q,\alpha} + T^{(\alpha_1+1)/2}\Psi_{2,\alpha}\right),$$

and for $\kappa_T \gtrsim T^{(\alpha_1+1)/q}\log p^{3/2}\| \||\boldsymbol{x}_\cdot||_\infty\|_{q,\alpha} + \sqrt{pT^{(\alpha_1+1)}}\Psi_{2,\alpha}$ then $\|\bar{\boldsymbol{\chi}}_T\|_\infty = o_p(1)$ for $p \to \infty$. For the "stronger" dependence case: $0 < \alpha < 1/2 - 1/q$, then the same expression for the rates can be obtained, now for $\alpha_1 = q/2 - \alpha q, \ell = \log p$. Note how as (A1.1) is a Nagaev-type of inequality, two bounds for the tail probability are displayed, namely a polynomial tail bound and a sub-Gaussian type one. For large κ_T, the polynomial one dominates, while for small κ_T the sub-Gaussian type does.

In the proof in Section A3 we follow closely Zhang et al. (2017). However, we give partly a different proof strategy, where instead of using a Rosnethal-Burkholder type bound on moments of Banach-spaced martingales, we obtain the same result with a simple, albeit more lengthy, chain of inequalities. Given Lemma A1.1, the claim $\|\boldsymbol{A\beta}\|_\infty = o_p(1)$ follows from assumption of compact parameter space for $\boldsymbol{\beta}$.

Remark A1.1. Note that we stated the deviation bound (A1.1) assuming finiteness of both the $\mathcal{L}^\infty$ dependence adjusted norm and the uniform dependence adjusted norm for the process $\boldsymbol{x}_t$ instead of $\bar{\boldsymbol{\chi}}_T = T^{-1}\sum_{t=1}^T \text{vech}(\boldsymbol{\Delta}_{x,L,t})$. The reason is that the former can be shown to imply the latters. The functional dependence measure for the process $\bar{\boldsymbol{\chi}}_T$ can be shown to be upper bounded by twice the functional dependence measure of $\boldsymbol{x}_t$. As such, similarly, both the $\mathcal{L}^\infty$ dependence adjusted norm and the uniform dependence adjusted norm of $\bar{\boldsymbol{\chi}}_T$ are upper bounded by twice the corresponding norms of $\boldsymbol{x}_t$. We show this in Section A3, Lemma A3.1. However, one needs to account the averaging factor T^{-1} in the bound but this follows by simple substitution of variable from Theorem 6.1 of Zhang et al. (2017). In fact, to account for this the sequence κ_T is inflated by a $\sqrt{T}$ in the first term if compared to Theorem 6.2 of Zhang et al. (2017). This is also justified by majorating the right hand side inequality of Lemma 8 in Chernozhukov et al. (2015).

Remark A1.2. Let us mention that in a more recent paper Zhang and Wu (2021) derived a similar Nagaev-type tail bound for (locally) stationary processes, directly on the deviation of the autocovariance from its true counterpart. This could analogously be used here too, without changing any conclusion. While the rates are similar –albeit in their case depending on some bandwidth– their Proposition 3.3 requires at least five finite moments while here we only need three.

Remark A1.3. The bound in (A1.1) is polynomial and can be strengthened to exponential by means of stronger assumptions on the moments for $\boldsymbol{x}_t$. Note that to get (A1.1) only a bit more than two finite moments are needed for $\boldsymbol{x}_t$ to exist.[4] Strengthening the moment requirement can lead to faster rates. For instance, by assuming finiteness of the sub-exponential (Orlicz) dependence adjusted norm i.e., $\|||\boldsymbol{x}_{\cdot}||_\infty\|_{\psi_\ell,\alpha} := \sup_{q\geq 2} q^{-\ell}\|||\boldsymbol{x}_{\cdot}||_\infty\|_{q,\alpha} < \infty$ one can obtain exponentially fast rates (see Section C.2 Zhang et al., 2017, for the univariate case). For even stronger assumption on $\boldsymbol{x}_t$ being a random matrix directly drawn from the $\boldsymbol{\Sigma}-$Gaussian ensemble (i.e., i.i.d. Gaussian rows) one can obtain similar polynomial bounds of the order $\log p/T$ directly on $\boldsymbol{A}$; likewise, assuming the rows of $\boldsymbol{x}_t$ to be zero mean i.i.d. sub-Gaussian, then exponentially fast Chernoff-like bounds for $\boldsymbol{A}$ can be derived. We show these in Section A3.

The next Lemma shows how the term $\boldsymbol{u}$ in (10), namely the term that gives the asymptotic distribution of the oracle least square and as such of the precision least squares as well, is approximately Gaussian.

Lemma A1.2. Consider the p-dimensional zero-mean stationary process under physical dependence $\boldsymbol{u}_t = \boldsymbol{\Theta}_x \boldsymbol{x}_t \epsilon_t$ and the p-dimensional sample mean vector $\boldsymbol{u} = T^{-1}\sum_{t=1}^{T} \boldsymbol{u}_t$.

[4]This is implied by $q > 2$ in the adjusted norm.

Then, for $T, p \to \infty$, the following Gaussian approximation applies as

$$\sup_{\kappa_3 \geq 0} \left| \mathbb{P}\left(\sqrt{T}\|\boldsymbol{u}\|_\infty \geq \kappa_3\right) - \mathbb{P}\left(\|\boldsymbol{\eta}\|_\infty \geq \kappa_3\right) \right| \to 0, \tag{A1.2}$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{V})$ for $\boldsymbol{V} = \sum_{l=-\infty}^{\infty} \mathbb{E}(\boldsymbol{u}_t \boldsymbol{u}'_{t+l})$ being the long-run covariance matrix.

Proof of Lemma A1.2 follows directly from the main result in Zhang et al. (2017); hence, we refer to this work for a full treatment of the Gaussian approximation.

Remark A1.4. While the details of the Gaussian approximation are given in Theorem 3.2 of Zhang et al. (2017) a couple of remarks are in order. First, like for the deviation bound in (A1.1) this Gaussian approximation only requires finite polynomial moments. In fact, letting $\boldsymbol{\Gamma}^u_{q,\alpha} = \left(\sum_{j=1}^p \|\boldsymbol{u}_t\|^q_{q,\alpha}\right)^{1/q}$ be the overall dependence adjusted norm and $\boldsymbol{\Xi}^u_{q,\alpha} = \boldsymbol{\Gamma}^u_{q,\alpha} \wedge \left(\left\|||\boldsymbol{u}_t||_{q,\alpha}(\log p)^{3/2}\right\|\right)$, the requirement for the Gaussian approximation to hold in both the weaker dependence case where $\alpha > 1/2 - 1/q$ and for the stronger dependence case $0 < \alpha < 1/2 - 1/q$ is that $\boldsymbol{\Xi}^u_{q,\alpha} < \infty$ for $q \geq 4$.[5] Furthermore, for stronger assumptions on the finiteness of the sub-exponential (Orlicz) dependence adjusted norm, the Gaussian approximation is valid for ultra high-dimension p. Second, (A1.2) is stated in the nonnormalized form i.e., both $\boldsymbol{u}$ and $\boldsymbol{\eta}$ are not pre-multiplied by the reciprocal of the square root of the diagonal matrix of $\boldsymbol{V}$. This implicitly assumes boundedness from above of the maximal long-run variance. Finally, an error bound for the Gaussian approximation is provided in Theorem 7.4 of Zhang et al. (2017).

A2 Monte-Carlo - Details

Consider the linear model (1). We simulate normal i.i.d data such that $\beta_j = 2$, for $j = 1, \ldots, 50$, $\boldsymbol{\Sigma}_x = (0.6)^{|i-j|}$, $p = 500$, $T = 250$, and $\mathbb{E}(\epsilon_t^2) = 0.1$. The number of repli-

[5] Other technical conditions on the rate of shrinking of both the uniform and overall dependence adjusted norms are reported in (Assumption 3.1 of Zhang et al., 2017).

cations is 500. Firstly, we compare the naive PrLS $\hat{\boldsymbol{\beta}}^{Naive}$ with its two proposed debiased versions i.e., $\hat{\boldsymbol{\beta}}^{PrLS}$, namely the PrLS with debiasing as in Eq. (13) and $\hat{\boldsymbol{\beta}}^{D}$ as presented in Eq. (17). Refitted adaptive lasso-based modified Cholesky decomposition (presented in Online Appendix A4) is used as a plug-in estimator of the precision matrix in all these. Furthermore we add the comparisons with: the debiased lasso in Van de Geer et al. (2014) (Lasso) and Adamek et al. (2022) (Lasso2), the debiased ridge in Bühlmann et al. (2013) (Ridge). The variance of $\hat{\boldsymbol{\beta}}^{PrLS}$ and $\hat{\boldsymbol{\beta}}^{D}$ is obtained via the subseries method presented in Section 3.2. The number of subseries values is fixed to $T^{1/3}$ across the simulations. We also include as a benchmark the infeasible OLS estimator to mimic the behavior of an oracle estimator. This estimator is obtained by augmenting the simulated sample size of the data such that $T = 750$.

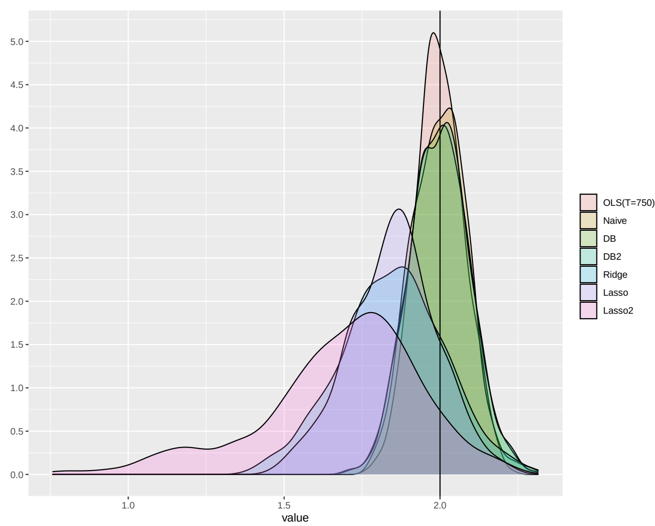


Figure A.1: Comparison of the densities of the estimated $\boldsymbol{\beta}_j$ of different estimators

Notes: `Naive`, `DB` and `DB2` correspond to $\boldsymbol{\beta}^{Naive}$, $\boldsymbol{\beta}^{PrLS}$ and $\boldsymbol{\beta}^{D}$, respectively. `Lasso` refers to Van de Geer et al. (2014), ridge refers to Bühlmann et al. (2013) debiased version and `Lasso2` refer to the Adamek et al. (2022) debiased lasso for time series. OLS($T = 750$) refers to the OLS estimator computed on $T = 750$ observations from which the first $T = 250$ are drawn to compare it to the other estimators. $p = 500$.

Figure A.1 displays the densities of an estimated regression coefficient using the different methods under comparison. One can observe how $\boldsymbol{\beta}^{PrLS}$, $\boldsymbol{\beta}^{D}$, and $\boldsymbol{\beta}^{Naive}$ behave similarly in terms of bias. This bias is close to 0 and it is similar to that of the infeasible OLS obtained by artificially augmenting the data. This is not the case for the Van de Geer et al. (2014), Bühlmann et al. (2013) and Adamek et al. (2022) debiased estimators. These estimators exhibit substantial bias and the Adamek et al. (2022) one is also left skewed.

(a) Comparison of the boxplot of the bias of the relevant coefficients of different estimators

(b) Comparison of the boxplot of the bias of the irrelevant coefficients of different estimators

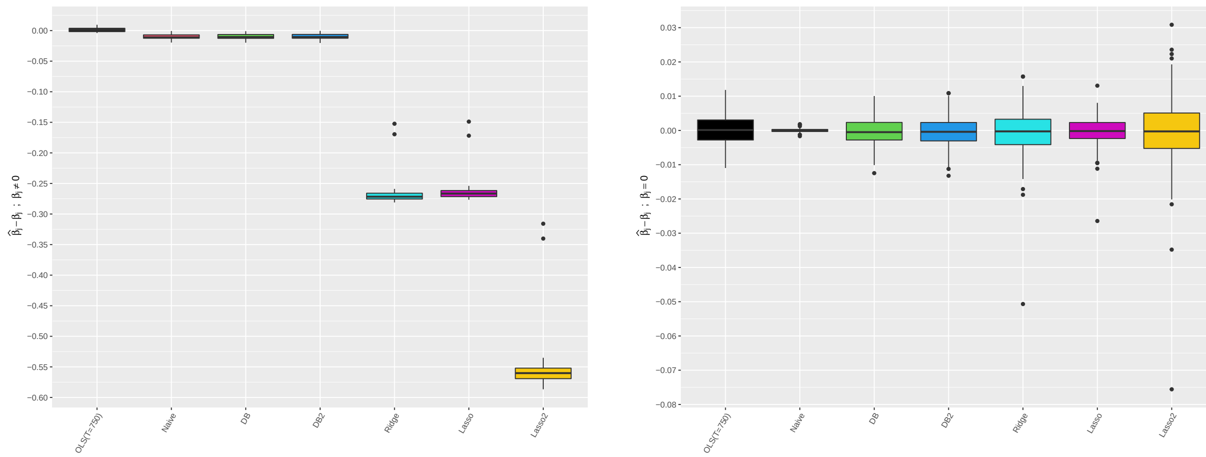


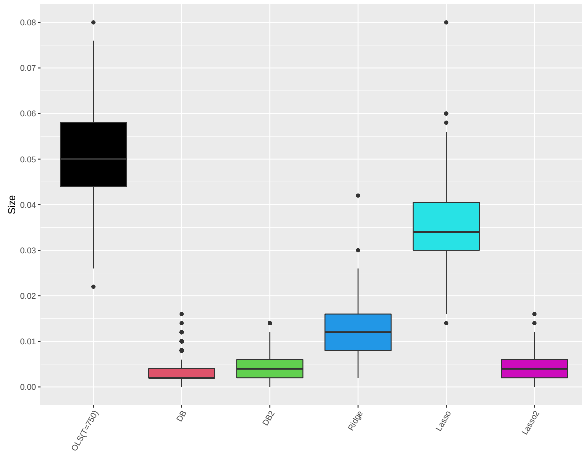
Figure A.2: Comparison of the biases

Notes: Bias=$E(\hat{\beta}_j - \beta_j)$. See Figure A.1 Notes for details about the data generating process.

Figure A.3a and Figure A.3b display the boxplots for the bias of the estimated coefficients obtained using the different methods. The $\boldsymbol{\beta}^{PrLS}$, $\boldsymbol{\beta}^{D}$, and $\boldsymbol{\beta}^{Naive}$ are shown to be uniformly unbiased over the set of all coefficients. This, again, is not the case for all the other estimators considered, some of which show erratic behaviors on some coefficients. It is worth mentioning that $\boldsymbol{\beta}^{Naive}$ is subject to regularization bias, as predicted by our theory. This is mainly visible when one looks at Figure A.3b. The bias of the irrelevant coefficients for this estimator is the smallest, meaning that the truly zero coefficients are very shrunken toward zero. On the other hand –as it is visible from Figure A.3a– this too much shrinkage towards zero also affects the relevant coefficients, whose bias is then farther from zero.

Figure A.5a reports the boxplots of the rejection frequencies of the tests for individual significance of the irrelevant coefficient. As one can expect, the test based on the infeasible

(a) Comparison of the boxplot of the sizes of the irrelevant coefficients of different estimators



(b) Comparison of the boxplot of the power of the relevant coefficients of different estimators

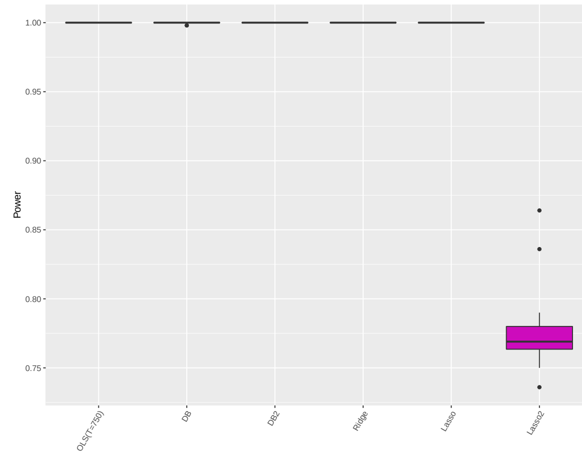


Figure A.4: Rejection frequency of tests of significance

Notes: See Figure A.1 Notes for details about the data generating process. We expect Size= $P(\text{test rejects } H_0 \mid H_0 \text{ is true}) \leq \alpha$ for $H_0 : \beta_j = 0$ $\alpha = 5\%$ to be the significance level.

OLS[6] and Van de Geer et al. (2014) estimators have size around the nominal 5% level. The tests based on the other estimators have size lower than 5%. This is expected, as a PrLS-based test discriminates strongly between relevant and irrelevant variables. In Figure A.5b, we report the boxplot of the rejection frequency of the tests for individual significance of the relevant coefficient. This corresponds to the power of the different tests. One can see that unlike the tests based on the Adamek et al. (2022) estimator, all the tests have power close to one.

Overall, our Monte Carlo simulation result illustrates the good performance of the PrLS estimator. It compares favorably in terms of bias, size, and power to traditional state-of-the-art estimators. Although the reported results are based on i.i.d. normal data, similar results are obtained for ARMA-GARCH processes with skewed student innovations for the covariates and error terms. These results are omitted for the sake of space and are available upon request.

[6]We use a standard formula for i.i.d. data to estimate the variance of the OLS.

A3 Proofs

First, in Remark A3.1 we introduce the concept of physical dependence. This is the key ingredient for the proofs that follow.

Remark A3.1. (Physical Dependence) We use the framework of physical dependence from Zhang et al. (2017). Let $\boldsymbol{e}_t$, $t \in \mathbb{Z}$ be i.i.d. random elements and $\mathcal{F}^t = (\dots, \boldsymbol{e}_{t-1}, \boldsymbol{e}_t)$ be the infinite collection of its past values. Furthermore, given a stationary process $\{\boldsymbol{x}_t\}$, let $\boldsymbol{x}_t = (\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}, \dots, \boldsymbol{x}_{p,t})' = G(\mathcal{F}^t)$, where $G(\cdot)$ is a measurable function taking values in $\mathbb{R}^p$. Any linear process $\boldsymbol{x}_t = \sum_{l=0}^{\infty} \boldsymbol{A}_l \boldsymbol{e}_{t-l}$ for $\boldsymbol{e}_t \overset{iid}{\sim} (0, \mathbb{E}(\boldsymbol{e}_t'\boldsymbol{e}_t) < \infty)$ and $\sum_{t=0}^{\infty} \operatorname{tr}(\boldsymbol{A}_t'\boldsymbol{A}_t) < \infty$ is of this form. The terms $\boldsymbol{e}_t$ can then be seen as the independent inputs of a system, and the outputs are the $\boldsymbol{x}_t$ causally generated by $\boldsymbol{e}_t$ through $G(\cdot)$. As for the notation: for any random variables $\boldsymbol{x}_t$ and $q > 0$, we write $\boldsymbol{x}_t \in \mathcal{L}^q$ if $\|\boldsymbol{x}_t\|_{E,q} := \left(\mathbb{E}\,|\boldsymbol{x}_t|^q\right)^{1/q} < \infty$ where $\mathcal{L}^q$ is the set of Lebesgue-integrable functions of order q. We can then define the *functional dependence measure* as

$$\delta_{q,j,t} = \left\|\boldsymbol{x}_{j,t} - \boldsymbol{x}_{j,t,\{0\}}\right\|_{E,q} = \left\|\boldsymbol{x}_{j,t} - g_j(\mathcal{F}^{t,\{0\}})\right\|_{E,q} = \left\|g_j(\mathcal{F}^t) - g_j(\mathcal{F}^{t,\{0\}})\right\|_{E,q}$$

where $\mathcal{F}^t$ is as defined above and $\mathcal{F}^{t,\{0\}} = (\dots, \boldsymbol{e}_{t-1}, \boldsymbol{e}_0', \boldsymbol{e}_1, \dots, \boldsymbol{e}_t)$ represents the values of a coupled process $\boldsymbol{x}_{j,t,\{0\}} = g_j(\mathcal{F}^{t,\{0\}})$, where $g_j(\cdot)$, $1 \leq j \leq p$, is the j-th coordinate projection of $G(\cdot)$. Intuitively, what $\delta_{q,j,t}$ measures is the dependency of $\boldsymbol{x}_{j,t}$ on $\boldsymbol{e}_0$, i.e., how replacing $\boldsymbol{e}_0$ by an i.i.d. copy $\boldsymbol{e}_0'$ (while keeping all the other $\boldsymbol{e}$'s fixed) affects the output $\boldsymbol{x}_{j,t}$, in the same form as a (nonlinear) impulse-response function.

As a consequence, to account for the dependence in the process $\boldsymbol{x}_{j,\cdot} = \{\boldsymbol{x}_{j,t}\}_{t\in\mathbb{Z}}$, the *dependence adjusted q-norm* for $q \geq 2$ is defined as

$$\|\boldsymbol{x}_{j,\cdot}\|_{q,\alpha} = \sup_{m\geq 0}(m+1)^{\alpha}\Delta_{q,j,m},$$

where $\alpha \geq 0$ and $\Delta_{q,j,m} = \sum_{t=m}^{\infty} \delta_{q,j,t}$ measure the cumulative effect of $\boldsymbol{e}_0$ on $\boldsymbol{x}_{j,t}$, $t \geq m$. Also, as we are working in a high-dimensional setting, let the *uniform* and *overall dependence adjusted norms* of $\boldsymbol{x}_t$ be defined respectively as

$$\Psi_{q,\alpha} = \max_{1 \leq j \leq p} \|\boldsymbol{x}_{j,\cdot}\|_{q,\alpha}, \quad \Gamma_{q,\alpha} = \left(\sum_{j=1}^{p} \|\boldsymbol{x}_{j,\cdot}\|_{q,\alpha}^{q} \right)^{1/q}.$$

Finally, it is also convenient to introduce the $\mathcal{L}^{\infty}$ *physical dependence measure* and its *adjusted norm* i.e., $\omega_{t,q} := \left\| \left\| \boldsymbol{x}_t - \boldsymbol{x}_{t,\{0\}} \right\|_{\infty} \right\|_q$ and $\| \|\boldsymbol{x}_{\cdot}\|_{\infty} \|_{q,\alpha} = \sup_{m \geq 0} (m+1)^{\alpha} \sum_{t=m}^{\infty} \omega_{t,q}$.

Lemma A3.1. Within the framework of physical dependence introduced in Remark A3.1, we are interested in deriving the functional dependence measure of the vector $\boldsymbol{\chi}_t = \text{vech}(\boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t'))$.[7] Following Zhang et al. (2017) Eq.(3.13-3.15), by letting $a = (j,k)$, $j,k \leq p$ such that $\boldsymbol{\chi}_{t,a} = (\boldsymbol{x}_{j,t} \boldsymbol{x}_{k,t}' - \mathbb{E}(\boldsymbol{x}_{j,t} \boldsymbol{x}_{k,t}'))$. The functional dependence measure for the process $(\boldsymbol{\chi}_{t,a})_{t \in \mathbb{Z}}$ is obtained from the following chain of inequalities:

$$\begin{aligned}
\tau_{q/2,a,t} &:= \left\| \boldsymbol{x}_{j,t} \boldsymbol{x}_{k,t} - \mathbb{E}(\boldsymbol{x}_{j,t} \boldsymbol{x}_{k,t}) - \boldsymbol{x}_{j,t,\{0\}} \boldsymbol{x}_{k,t,\{0\}} + \mathbb{E}\left(\boldsymbol{x}_{j,t,\{0\}} \boldsymbol{x}_{k,t,\{0\}} \right) \right\|_{E,q/2} \\
&\leq 2 \left\| \boldsymbol{x}_{j,t} \boldsymbol{x}_{k,t} - \boldsymbol{x}_{j,t,\{0\}} \boldsymbol{x}_{k,t,\{0\}} \right\|_{E,q/2} \\
&\leq 2 \left\| \boldsymbol{x}_{j,t} (\boldsymbol{x}_{k,t} - \boldsymbol{x}_{k,t,\{0\}}) \right\|_{E,q/2} + 2 \left\| (\boldsymbol{x}_{j,t} - \boldsymbol{x}_{j,t,\{0\}}) \boldsymbol{x}_{k,t,\{0\}} \right\|_{E,q/2} \\
&\leq 2 \left\| \boldsymbol{x}_{j,t} \right\|_{E,q} \delta_{q,k,t} + 2 \left\| \boldsymbol{x}_{k,t} \right\|_{E,q} \delta_{q,j,t},
\end{aligned} \tag{A3.1}$$

where the last step follows from Hölder's inequality. An upper bound on the dependence adjusted norm of $\boldsymbol{\chi}_{t,a}$ follows as

$$\begin{aligned}
\|\boldsymbol{\chi}_{a,\cdot}\|_{q/2,\alpha} &:= \sup_{m \geq 0} (m+1)^{\alpha} \sum_{t=m}^{\infty} \tau_{q/2,j,k,t} \\
&\leq 2 \|\boldsymbol{x}_{j,\cdot}\|_{q,0} \|\boldsymbol{x}_{k,\cdot}\|_{q,\alpha} + 2 \|\boldsymbol{x}_{k,\cdot}\|_{q,0} \|\boldsymbol{x}_{j,\cdot}\|_{q,\alpha},
\end{aligned} \tag{A3.2}$$

[7]Note that we now write $\boldsymbol{\chi}_t = \text{vech}(\boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t'))$ in place of $\bar{\boldsymbol{\chi}}_T = T^{-1} \sum_{t=1}^{T} \text{vech}(\boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t'))$. The reason is that by stationarity one has $\left\| T^{-1} \sum_{t=1}^{T} \text{vech}(\boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')) \right\|_{E,q} \leq \|\text{vech}(\boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t'))\|_{E,q}$.

and consequently for the uniform and overall dependence adjusted norms of $\boldsymbol{\chi}_t$:

$$\max_a \|\boldsymbol{\chi}_{a,\cdot}\|_{q/2,\alpha} \leq 4\Psi_{q,0}\Psi_{q,\alpha}, \tag{A3.3}$$

$$\left(\sum_a \|\boldsymbol{\chi}_{a,\cdot}\|_{q/2,\alpha}^{q/2}\right)^{2/q} \leq 4\left(\sum_{j=1}^p \|\boldsymbol{x}_{j,\cdot}\|_{q,0}^{q/2}\right)^{2/q}\left(\sum_{j=1}^p \|\boldsymbol{x}_{j,\cdot}\|_{q,\alpha}^{q/2}\right)^{2/q}.$$

Then, the $\mathcal{L}^\infty$ dependence adjusted norm of $\boldsymbol{\chi}_t$ is as follows:

$$4\Psi_{q,0}\Psi_{q,\alpha} \leq \| \||\boldsymbol{\chi}_\cdot|\|_\infty\|_{q/2,\alpha} \leq 4\,\| \||\boldsymbol{x}_\cdot|\|_\infty\|_{q,0}\,\| \||\boldsymbol{x}_\cdot|\|_\infty\|_{q,\alpha}\,. \tag{A3.4}$$

Note that to prove the statement in Lemma A1.1 it suffices to prove the deviation bound in Proposition 1 which is what we are going to do now.

Proof of Proposition 1. Given $\bar{\boldsymbol{\chi}}_T = T^{-1}\sum_{t=1}^T \boldsymbol{\chi}_t$, for $s \geq 0$ let $\bar{\boldsymbol{\chi}}_{s,T} = T^{-1}\sum_{t=1}^T \boldsymbol{\chi}_{s,t}$ for $\boldsymbol{\chi}_{s,t} = \mathbb{E}(\boldsymbol{\chi}_t|\boldsymbol{e}_{t-s},\ldots,\boldsymbol{e}_t)$ being the s-dependence approximation of $\boldsymbol{\chi}_t$. Note that $\bar{\boldsymbol{\chi}}_{T,s} = T^{-1}\sum_{t=1}^T \mathbb{E}(\mathrm{vec}(\boldsymbol{x}_t\boldsymbol{x}_t' - \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'))|\boldsymbol{e}_{t-s},\ldots,\boldsymbol{e}_t)$. Let $\mathcal{F}_{-s} := (\boldsymbol{e}_{t-s},\ldots,\boldsymbol{e}_t)$. Then, by linearity of the expectation and the vec operator, we obtain that

$$\begin{aligned}\bar{\boldsymbol{\chi}}_{T,s} &= T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{-s}) - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{-s})))\\ &= T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{-s}) - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')),\end{aligned} \tag{A3.5}$$

by law of iterated expectation on the second term. Now, noting that $\bar{\boldsymbol{\chi}}_T$ can also be split by the linearity of the vec operator as

$$\bar{\boldsymbol{\chi}}_T = T^{-1}\sum_{t=1}^T \mathrm{vec}(\boldsymbol{x}_t\boldsymbol{x}_t') - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')) \tag{A3.6}$$

it follows that

$$\bar{\boldsymbol{\chi}}_T - \bar{\boldsymbol{\chi}}_{s,T} = \underbrace{T^{-1}\sum_{t=1}^{T}\text{vec}(\boldsymbol{x}_t\boldsymbol{x}_t')}_{:=\bar{\Gamma}_T^x} - \underbrace{T^{-1}\sum_{t=1}^{T}\text{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t^\top|\mathcal{F}_{-s})}_{:=\bar{\Gamma}_{s,T}^x}. \tag{A3.7}$$

A tail bound on (A3.7) is obtained through Theorem 6.1 of Zhang et al. (2017) by a simple change of variable. Let $s = 1 \vee \log p$. Then, by standard norm properties, we have

$$\mathbb{P}\left(||\bar{\boldsymbol{\Gamma}}_T^x - \bar{\boldsymbol{\Gamma}}_{s,T}^x||_\infty \geq y\right) \leq \mathbb{P}\left(||\bar{\boldsymbol{\Gamma}}_T^x - \bar{\boldsymbol{\Gamma}}_{s,T}^x||_s \geq y\right).$$

Let $L = \lfloor(\log T - \log s)/(\log 2)\rfloor$, $\omega_l = 2^l$ if $1 \leq l \leq L$, $\omega_L = \lfloor T/s\rfloor$, $\tau_l = s\omega_l$ for $1 \leq l < L$, $\tau_0 = s$, and $\tau_L = T$ and define $M_{T,l} := \bar{\boldsymbol{\Gamma}}_{\tau_l,T}^x - \bar{\boldsymbol{\Gamma}}_{\tau_{l-1},T}^x$ for $1 \leq l \leq L$. Then, one can rewrite (A3.7) as

$$\bar{\boldsymbol{\Gamma}}_T^x - \bar{\boldsymbol{\Gamma}}_{s,T}^x = \underbrace{\bar{\boldsymbol{\Gamma}}_T^x - \bar{\boldsymbol{\Gamma}}_{T,T}^x}_{(i)} + \underbrace{\sum_{l=1}^{L} M_{l,T}}_{(ii)}.$$

Note, we can further rewrite (i) as $\sum_{j=T}^{\infty}\bar{\boldsymbol{\Gamma}}_{j+1,T}^x - \bar{\boldsymbol{\Gamma}}_{j,T}^x$. Hence, by union bound

$$\left\|||\bar{\boldsymbol{\Gamma}}_T^x - \bar{\boldsymbol{\Gamma}}_{T,T}^x||_s\right\|_q \leq \sum_{j=T}^{\infty}\left\|||\bar{\boldsymbol{\Gamma}}_{j+1,T}^x - \bar{\boldsymbol{\Gamma}}_{j,T}^x||_s\right\|_q. \tag{A3.8}$$

Let us use the notation $\mathbb{E}_s\bar{\boldsymbol{\Gamma}}_T^x := \mathbb{E}\left[\bar{\boldsymbol{\Gamma}}_T^x|\boldsymbol{e}_s,\ldots,\boldsymbol{e}_t\right]$. Then, by standard norm properties and stationarity

$$\begin{aligned}\left\|||\bar{\boldsymbol{\Gamma}}_{j+1,T}^x - \bar{\boldsymbol{\Gamma}}_{j,T}^x||_s\right\|_q &\leq c\left\|||\bar{\boldsymbol{\Gamma}}_{j+1,T}^x - \bar{\boldsymbol{\Gamma}}_{j,T}^x||_\infty\right\|_q \\ &\leq c\left\|||\mathbb{E}_{t-j-1}\bar{\boldsymbol{\Gamma}}_T^x - \mathbb{E}_{t-j}\bar{\boldsymbol{\Gamma}}_T^x||_\infty\right\|_q \\ &\leq c\left\|||\mathbb{E}_0\bar{\boldsymbol{\Gamma}}_{j+1}^x - \mathbb{E}_1\bar{\boldsymbol{\Gamma}}_{j+1}^x||_\infty\right\|_q.\end{aligned}$$

Then, note that the second term can be rewritten in terms of a coupled version of $\bar{\boldsymbol{\Gamma}}^x_{j+1}$:

$$\begin{aligned}\mathbb{E}_1\bar{\boldsymbol{\Gamma}}^x_{j+1} &= \mathbb{E}\left[\bar{\boldsymbol{\Gamma}}^x_{j+1}|\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1}\right] = \mathbb{E}\left[g_j(\ldots,\boldsymbol{e}_0,\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1})|\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1}\right] \\ &= \mathbb{E}\left[g_j(\ldots,\boldsymbol{e}_0',\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1})|\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1}\right] = \mathbb{E}\left[\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}|\boldsymbol{e}_0,\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{j+1}\right].\end{aligned}$$

where, in the last equality, we can include $\boldsymbol{e}_0$ in the conditioning set as it does not alter the conditional expectation of $\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}$. Then, by first collecting the $\mathbb{E}_0$ terms and then by Jensen's inequality and law of iterated expectation

$$\begin{aligned}c\left\|||\mathbb{E}_0\bar{\boldsymbol{\Gamma}}^x_{j+1}-\mathbb{E}_1\bar{\boldsymbol{\Gamma}}^x_{j+1}||_\infty\right\|_q &\le c\left\|\left|\left|\mathbb{E}_0\left(\bar{\boldsymbol{\Gamma}}^x_{j+1}-\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}\right)\right|\right|_\infty\right\|_q \le c\left\|\mathbb{E}_0||\bar{\boldsymbol{\Gamma}}^x_{j+1}-\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}||_\infty\right\|_q \\ &\le c\left\{\mathbb{E}\left[\mathbb{E}_0||\bar{\boldsymbol{\Gamma}}^x_{j+1}-\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}||^q_\infty\right]^{1/q}\right\} \le c\left[\mathbb{E}||\bar{\boldsymbol{\Gamma}}^x_{j+1}-\bar{\boldsymbol{\Gamma}}^x_{j+1,\{0\}}||^q_\infty\right]^{1/q} \\ &\le c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}\mathrm{vec}(\boldsymbol{x}_t\boldsymbol{x}_t^\top\boldsymbol{x}_t\boldsymbol{x}_t^\top)-(j+1)^{-1}\sum_{t=1}^{(j+1)}\mathrm{vec}(\boldsymbol{x}_t\boldsymbol{x}_{t,\{0\}}^\top\boldsymbol{x}_t\boldsymbol{x}_{t,\{0\}}^\top)\right\|_q.\end{aligned}$$

Now, in the same pretence as the functional dependence measure in (A3.1), we have:

$$\begin{aligned}&c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}\mathrm{vec}(\boldsymbol{x}_{j,t}\boldsymbol{x}_{k,t}^\top)-(j+1)^{-1}\sum_{t=1}^{(j+1)}\mathrm{vec}(\boldsymbol{x}_{j,t,\{0\}}\boldsymbol{x}_{k,t,\{0\}}^\top)\right\|_{q/2} \\ \le& c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}\boldsymbol{x}_{j,t}(\boldsymbol{x}_{k,t}-\boldsymbol{x}_{k,t,\{0\}})\right\|_{q/2}+c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}(\boldsymbol{x}_{j,t}-\boldsymbol{x}_{j,t,\{0\}})\boldsymbol{x}_{k,t,\{0\}}\right\|_{q/2} \\ \le& c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}\boldsymbol{x}_{j,t}\right\|_q\delta_{q,k,t}+c\left\|(j+1)^{-1}\sum_{t=1}^{(j+1)}\boldsymbol{x}_{k,t}\right\|_q\delta_{q,j,t} =: c\left\|\bar{\boldsymbol{x}}^{[j+1]}_{j,t}\right\|_q\delta_{q,k,t}+c\left\|\bar{\boldsymbol{x}}^{[j+1]}_{k,t}\right\|_q\delta_{q,j,t}.\end{aligned}$$

Now,

$$\left\| \left[\sum_{t=1}^{T} \mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_s^2 | \mathcal{F}^i \right] \right]^{1/2} \right\|_q = \left[\left(\mathbb{E}\left\{ \sum_{t=1}^{T} \mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_s^2 | \mathcal{F}^i \right] \right\}^{q/2} \right)^{2/q} \right]^{1/2}$$

$$\overset{(1)}{\leq} \left[\sum_{t=1}^{T} \left\| \mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_s^2 | \mathcal{F}^i \right] \right\|_{q/2} \right]^{1/2} \overset{(2)}{\leq} \left[\sum_{t=1}^{T} \left\| ||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_s^2 \right\|_{q/2} \right]^{1/2}$$

$$\overset{(3)}{\leq} c \left[\sum_{t=1}^{T} \left\| ||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_\infty^2 \right\|_{q/2} \right]^{1/2},$$

where (1), (2), and (3) follow the Minkovski inequality, law of iterated expectations and the fact that $\|\cdot\|_s \leq T^{1/s}\|\cdot\|_\infty = T^{1/\log T}\|\cdot\|_\infty = c\|\cdot\|_\infty$, respectively. Then,

$$c \left[\sum_{t=1}^{T} \left\| ||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_\infty^2 \right\|_{q/2} \right]^{1/2} = c \left[\sum_{t=1}^{T} \left(\mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{j+1,T} - \bar{\boldsymbol{\Gamma}}^{x}_{j,T}||_\infty^2 \right]^q \right)^{2/q} \right]^{1/2}$$

$$= c \left[\sum_{t=1}^{T} \left(\mathbb{E}\left\{ \left|\left| \mathbb{E}\left[\bar{\boldsymbol{\Gamma}}^{x}_{T} - \bar{\boldsymbol{\Gamma}}^{x}_{T,\{t-j-1\}} | \boldsymbol{e}_{t-j-1}, \dots, \boldsymbol{e}_i \right] \right|\right|_\infty \right\}^q \right)^{2/q} \right]^{1/2},$$

hence,

$$\overset{(4)}{\leq} c \left[\sum_{t=1}^{T} \left(\mathbb{E}\left\{ \mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{T} - \bar{\boldsymbol{\Gamma}}^{x}_{T,\{t-j-1\}}||_\infty | \boldsymbol{e}_{t-j-1}, \dots, \boldsymbol{e}_t \right] \right\}^q \right)^{2/q} \right]^{1/2}$$

$$\overset{(5)}{\leq} c \left[\sum_{t=1}^{T} \left(\mathbb{E}\left\{ \mathbb{E}\left[||\bar{\boldsymbol{\Gamma}}^{x}_{T} - \bar{\boldsymbol{\Gamma}}^{x}_{T,\{t-j-1\}}||_\infty^q | \boldsymbol{e}_{t-j-1}, \dots, \boldsymbol{e}_i \right] \right\} \right)^{2/q} \right]^{1/2}$$

$$\overset{(6)}{\leq} c \left[\sum_{t=1}^{T} \left\| ||\bar{\boldsymbol{\Gamma}}^{x}_{T} - \bar{\boldsymbol{\Gamma}}^{x}_{T,\{t-j-1\}}||_\infty \right\|_q^2 \right]^{1/2}$$

$$\overset{(7)}{\leq} c \left[\sum_{t=1}^{T} \left\| ||\bar{\boldsymbol{\Gamma}}^{x}_{t+1} - \bar{\boldsymbol{\Gamma}}^{x}_{t+1,\{0\}}||_\infty \right\|_q^2 \right]^{1/2}$$

$$\leq c \left[\sum_{t=1}^{T} \left(\left\| \bar{\boldsymbol{x}}^{[j+1]}_{j,t} \right\|_q \delta_{q,k,t} + \left\| \bar{\boldsymbol{x}}^{[j+1]}_{k,t} \right\|_q \delta_{q,j,t} \right)^2 \right]^{1/2}$$

$$\leq cT^{1/2} \left(\left\| \bar{\boldsymbol{x}}^{[j+1]}_{j,t} \right\|_q \delta_{q,k,t} + \left\| \bar{\boldsymbol{x}}^{[j+1]}_{k,t} \right\|_q \delta_{q,j,t} \right).$$

where (4), (5), (6), and (7) follow Jensen's inequality, law of iterated expectations and stationarity, respectively. Then, by Markov's inequality, we obtain from (A3.8) and (A3.2),(A3.3),(A3.4) for $\bar{\boldsymbol{\chi}}_t$:

$$\begin{aligned}\mathbb{P}\left(||\bar{\boldsymbol{\Gamma}}^x_T - \bar{\boldsymbol{\Gamma}}^x_{T,T}||_s \geq x\right) &\leq \frac{\left\|||\bar{\boldsymbol{\Gamma}}^x_T - \bar{\boldsymbol{\Gamma}}^x_{T,T}||_s\right\|_q^q}{x^q} \\ &\leq \frac{cT^{q/2}\left(2\left\|\bar{\boldsymbol{x}}^{[j+1]}_{\cdot j}\right\|_{q,0}\left\|\bar{\boldsymbol{x}}^{[j+1]}_{.k}\right\|_{q,\alpha} + 2\left\|\bar{\boldsymbol{x}}^{[j+1]}_{.k}\right\|_{q,0}\left\|\bar{\boldsymbol{x}}^{[j+1]}_{\cdot j}\right\|_{q,\alpha}\right)}{x^q} \\ &\leq \frac{cT^{q/2}4\Psi_{q,0}\Psi_{q,\alpha}}{x^q} \leq \frac{cT^{q/2}\left\|||\bar{\boldsymbol{\chi}}_\cdot||_\infty\right\|_{q/2,\alpha}}{x^q}.\end{aligned} \tag{A3.9}$$

The bound on (ii) follows Zhang et al. (2017), Theorem 6.1. Putting the bounds on (i) and (ii) together gives the claimed bound.

Finally, we only need to verify from (A3.7) that by means of simple manipulation

$$\begin{aligned}&\bar{\boldsymbol{\Gamma}}^x_T - \bar{\boldsymbol{\Gamma}}^x_{s,T} - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')) + T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')) = \\ &= \bar{\boldsymbol{\chi}}_T - \left(T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{-s}) - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'))\right) =: (b)\end{aligned} \tag{A3.10}$$

$$\|(b)\|_2 \leq \|\bar{\boldsymbol{\chi}}_T\|_1 + \left\|T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'|\mathcal{F}_{-s}) - T^{-1}\sum_{t=1}^T \mathrm{vec}(\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'))\right\|_\infty \geq \|\bar{\boldsymbol{\chi}}_T\|_1.$$

Therefore, the claim of Lemma A1.1 on (only) $\bar{\boldsymbol{\chi}}_T$ follows by this last argument and this concludes the proof. □

Remark A3.2. If one would allow for much stronger conditions than in Lemma A1.1, then

$$||\boldsymbol{A}||_\infty \equiv ||\,\mathrm{var}(\boldsymbol{x}_t)^{-1}\boldsymbol{x}_t'\boldsymbol{x}_t/T - \boldsymbol{I}||_\infty \leq C\sqrt{(\log p)/T}.$$

(I) Let $\boldsymbol{x}_t$ be a random matrix directly drawn from the Σ-Gaussian ensemble (i.e., Gaussian rows) and hence $\hat{\boldsymbol{\Sigma}}$ is a Wishart matrix. Then, by letting $\boldsymbol{V} \sim \mathcal{N}(0, \boldsymbol{I})$ and

$\boldsymbol{x} = \boldsymbol{V}\sqrt{\boldsymbol{\Sigma}}$

$$\left\|\boldsymbol{\Theta}\hat{\boldsymbol{\Sigma}} - \boldsymbol{I}\right\|_2 \equiv \left\|\boldsymbol{\Theta}(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\right\|_2 \leq \|\boldsymbol{\Theta}\|_2 \left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2$$

$$\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2 = \left\|\sqrt{\boldsymbol{\Sigma}}\left(T^{-1}\boldsymbol{V}^\top\boldsymbol{V} - \boldsymbol{I}\right)\sqrt{\boldsymbol{\Sigma}}\right\|_2 \leq \|\boldsymbol{\Sigma}\|_2 \underbrace{\left\|T^{-1}\boldsymbol{V}^\top\boldsymbol{V} - \boldsymbol{I}\right\|_2}_{:=G}$$

with probability $1 - 2e^{-T\delta^2/2}$

$$G \leq 2\sqrt{\frac{\log p}{T}} + \frac{\log p}{T} + \delta$$

which follows from the upper deviation inequality over the maximum singular value of $\boldsymbol{x}_t$ (see Wainwright (2019), Theorem 6.1) Therefore, with probability $1 - 2e^{-T\delta^2/2}$ for all $\delta > 0$

$$\frac{\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_2}{\|\boldsymbol{\Sigma}\|_2} \leq 2\sqrt{\frac{\log p}{T}} + 2\delta + \frac{\log p}{T} + \delta^2$$

i.e. the relative error converges to zero as long as $\log p/T \to 0$.

(II) Let the rows of $\boldsymbol{x}_t$ be zero-mean i.i.d. sub-Gaussian with parameter at most σ, i.e. $\mathbb{E}\left[e^{\gamma x_i}\right] \leq e^{\frac{\gamma^2\sigma^2}{2}}$, $\forall \gamma \in \mathbb{R}$. Then, a bound on the moment generating function of $|||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|||_2$ and a consequent tail bound are in order:

$$\mathbb{E}\left[e^{\lambda|||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|||_2}\right] \leq e^{c_0\frac{\gamma^2\sigma^4}{T} + 4p}, \quad \forall|\gamma| \leq \frac{T}{64e^2\sigma^2}$$

$$\mathbb{P}\left[\frac{|||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|||_2}{\sigma^2} \geq \left\{\sqrt{\frac{\log p}{T}}\frac{\log p}{T}\right\} + \delta\right] \leq c_2 e^{-c_3 T\min\{\delta, \delta^2\}}, \quad \forall\delta \geq 0$$

where $\{c_j\}_{j=0}^3$ are universal constants. The proof follows immediately by application of the Chernoff bound; see, e.g., Theorem 6.5 of Wainwright (2019).

Proof of Theorem 1. Assuming that $\hat{\boldsymbol{\Theta}}$ is a consistent estimator of $\boldsymbol{\Theta}$, proof of Theorem 1 follows directly from the results in Lemma A1.1, Lemma A1.2 and by direct application of central limit theorm to the coordinate of $\boldsymbol{\beta}^{Oracle}$. □

Proof of Theorem 2. Recalling Eq. (10), the test statistics can be rewritten for $\tilde{\boldsymbol{Z}}_i \sim N(0, \boldsymbol{\Lambda}^0)$

$$T_i = \frac{\sqrt{T}\boldsymbol{\beta}_i}{\sqrt{\widehat{\boldsymbol{V}}_{ii}}} + \frac{\boldsymbol{A}_{i,.}\boldsymbol{\beta}}{\sqrt{\widehat{\boldsymbol{V}}_{ii}}} + \tilde{\boldsymbol{Z}}_i$$

where $\tilde{\boldsymbol{Z}}_i = \frac{1}{\sqrt{\hat{\boldsymbol{V}}_{ii}}} N(0, \boldsymbol{V}_{ii}) \to N(0, I)$. Let $S_{\geq 0} \equiv \{i \in [p] : \boldsymbol{\beta}_i \geq 0\}$ and $S_{\leq 0} \equiv \{i \in [p] : \boldsymbol{\beta}_i \leq 0\}$. Momentarily consider a fixed $t = \tilde{t}_p$, where we clarify the dependence of $\tilde{t}_p$ on p. In addition, let $\widehat{\text{sign}}_i = \text{sign}(T_i)$ be the estimate of $\text{sign}(\boldsymbol{\beta}_i)$; then, for $c > 0$

$$\begin{aligned}
\mathbb{P}\left(\sum_i \mathbb{I}\left(\widehat{\text{sign}}_i \neq \text{sign}(\boldsymbol{\beta}_i)\right) \geq 1\right) &\leq \underbrace{\mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}\left(T_i \geq \tilde{t}_p\right) \geq 1\right)}_{:=A} \\
&+ \underbrace{\mathbb{P}\left(\sum_{i \in S_{\geq 0}} \mathbb{I}\left(T_i \leq -\tilde{t}_p\right) \geq 1\right)}_{:=B}
\end{aligned}$$

Let us assume that $\hat{\boldsymbol{V}}_{jj} - \boldsymbol{V}_{j,j} = o_p(1)$. For instance, $\hat{\boldsymbol{V}}_{jj}$ can be the batch mean estimator presented in Section 3. Recall that $\lim_{p \to \infty} p^b e^{-p^a} = 0 \ \forall \ b \in \mathbb{R}$ and $a > 0$. First, we consider the term A, then B follows in a symmetric way:

$$\begin{aligned}
A &\leq \mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}\left(\frac{\boldsymbol{A}_{i,.}\boldsymbol{\beta}}{\sqrt{\widehat{\boldsymbol{V}}_{ii}}} + \tilde{\boldsymbol{Z}}_i \geq \tilde{t}_p\right) \geq 1\right) \\
&\leq \mathbb{P}\left(\sum_{i \in S_{\leq 0}} \mathbb{I}\left(\tilde{\boldsymbol{Z}}_i \geq \tilde{t}_p - \frac{\|\boldsymbol{A}\boldsymbol{\beta}\|_\infty}{\sqrt{\widehat{\boldsymbol{V}}_{ii}}}\right) \geq 1\right) \\
&\leq p \max_{i \in [p]} \mathbb{P}\left(\tilde{\boldsymbol{Z}}_i \geq (1-c)\tilde{t}_p - c\right) + \max_{i \in [p]} \mathbb{P}\left(\|\boldsymbol{A}\boldsymbol{\beta}\|_\infty \geq \sqrt{\hat{\boldsymbol{V}}_{ii}} c\right) \\
&\leq p \max_{i \in [p]} \mathbb{P}\left(\tilde{\boldsymbol{Z}}_i \geq (1-c)\tilde{t}_p - c\right) + \max_{i \in [p]} \mathbb{P}\left(\|\boldsymbol{A}\|_\infty \geq \frac{\sqrt{\hat{\boldsymbol{V}}_{ii}} c}{\|\boldsymbol{\beta}\|_1}\right)
\end{aligned}$$

$$
\begin{aligned}
&= p \max_{i\in[p]} \mathbb{P}\left(\tilde{\boldsymbol{Z}}_i \geq (1-c)\tilde{t}_p - c\right) + \max_{i\in[p]}\mathbb{P}\left(\|\boldsymbol{A}\|_\infty \geq \frac{\sqrt{\boldsymbol{V}_{ii}}}{\|\boldsymbol{\beta}\|_1}c - \frac{\sqrt{\boldsymbol{V}_{ii}} - \sqrt{\hat{\boldsymbol{V}}_{ii}}}{\|\boldsymbol{\beta}\|_1}c\right) \\
&\leq p \max_{i\in[p]} \mathbb{P}\left(\tilde{\boldsymbol{Z}}_i \geq (1-c)\tilde{t}_p - c\right) + \max_{i\in[p]}\mathbb{P}\left(\|\boldsymbol{A}\|_\infty \geq \frac{\sqrt{\boldsymbol{V}_{ii}}}{\|\boldsymbol{\beta}\|_1}(1-c_2)c - c_2\right) \\
&+ \max_{i\in[p]}\mathbb{P}\left(\left(\frac{\sqrt{\hat{\boldsymbol{V}}_{ii}}}{\sqrt{\boldsymbol{V}_{ii}}} - 1\right) \geq \frac{c_2\|\boldsymbol{\beta}\|_1}{\boldsymbol{V}_{ii}c}\right) \\
&\leq p \max_{i\in[p]} \mathbb{P}\left(\tilde{\boldsymbol{Z}}_i \geq (1-c)\tilde{t}_p - c\right) + \max_{i\in[p]}\mathbb{P}\left(\|\boldsymbol{A}\|_\infty \geq \frac{\sqrt{\boldsymbol{V}_{ii}}}{\|\boldsymbol{\beta}\|_1}(1-c_2)c - c_2\right) \\
&+ \max_{i\in[p]}\mathbb{P}\left(\left|\sqrt{\frac{\hat{\boldsymbol{V}}_{ii}}{\boldsymbol{V}_{ii}}} - 1\right| \geq \frac{c_2\|\boldsymbol{\beta}\|_1}{c}\right) \\
&\leq p\ \exp\left[-\frac{((1-c)\tilde{t}_p - c)^2}{2}\right] + o_p(1) = o_p(1).
\end{aligned}
$$

□

Proof of Theorem 3. The results of Theorem 3 follow from the following chain of equalities.

$$
\begin{aligned}
\hat{\boldsymbol{\Theta}} &= 2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= 2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= 2\tilde{\boldsymbol{\Theta}} - (\boldsymbol{I} + \boldsymbol{\Delta}\boldsymbol{\Sigma})\tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= 2\tilde{\boldsymbol{\Theta}} - \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Delta}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Delta}\boldsymbol{\Sigma}\tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= \tilde{\boldsymbol{\Theta}} - \boldsymbol{\Delta}(I + \boldsymbol{\Sigma}\boldsymbol{\Delta}) + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= \boldsymbol{\Theta} + \boldsymbol{\Delta} - \boldsymbol{\Delta}(I + \boldsymbol{\Sigma}\boldsymbol{\Delta}) + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\hat{\boldsymbol{\Theta}} &= \boldsymbol{\Theta} - \boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}, \\
\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}} &= \boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta} + \tilde{\boldsymbol{\Theta}}(\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}})\tilde{\boldsymbol{\Theta}}.
\end{aligned}
$$

□

A4 The regression-based modified Cholesky decomposition

In this section, we present the regression-based modified Cholesky decomposition (RBMCD) which is a class of precision matrix estimators based on sequential estimation of linear regression models. The history of the RBMCD originates from Summerfield and Lubin (1951); Dempster (1969) and Hawkins and Eplett (1982) with recent application in Pourahmadi (1999); Boudt et al. (2017); Darolles et al. (2018). Proposition 2 gives the algorithm of RBMCD.

Proposition 2. Recall the definition of $\boldsymbol{\Omega}$ in Remark 1. If $\boldsymbol{\Omega}$ is positive definite then it admits a unique modified Cholesky decomposition. Therefore, it exists a sequence of orthogonal random variables $\varepsilon_{1,t}, \ldots, \varepsilon_{p,t}$ with $\mathbb{V}ar(\varepsilon_{j,t}) = \sigma_j^2 < \infty$, such that

$$\begin{aligned} z_{1,t} &= \varepsilon_{1,t}, \\ z_{j,t} &= b_{j,1} z_{1,t} + b_{j,2} z_{2,t} + \cdots + b_{j,j-1} z_{j-1,t} + \varepsilon_{j,t}, \text{ for each } j = 2, \ldots, p+1 \end{aligned} \tag{A4.1}$$

and $\mathbb{E}(z_{i,t}\epsilon_{j,t}) = 0$ for $1 \leq i < j \leq p$.

Proposition 2 implies that $\boldsymbol{\Omega}^{-1} = \boldsymbol{B}'\boldsymbol{G}^{-1}\boldsymbol{B}$ with

$$
\boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -b_{2,1} & 1 & 0 & \cdots & 0 \\ -b_{3,1} & -b_{3,2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -b_{p+1,1} & -b_{p+1,2} & -b_{p+1,3} & \cdots & 1 \end{bmatrix} \text{ and } \boldsymbol{G} = \begin{bmatrix} {\sigma_1}^2 & 0 & 0 & \cdots & 0 \\ 0 & {\sigma_2}^2 & 0 & \cdots & 0 \\ 0 & 0 & {\sigma_3}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & {\sigma_{p+1}}^2 \end{bmatrix},
$$

can be estimated via a sequence of p independent linear regression models. $\boldsymbol{B}$ is called the Cholesky factor of $\boldsymbol{\Omega}^{-1}$. Proposition 3 below shows how either dense or sparse assumption over the coefficient vector $\boldsymbol{\beta}$ is compatible with a sparsity assumption of the Cholesky factor for the precision matrix of a given permutation of the coordinates of $\boldsymbol{z}_t = (z_{1,t}, \dots, z_{p,t})'$. Such a feature is appealing for estimating $\boldsymbol{\beta}$ in high-dimension via PrLS as estimating $\boldsymbol{B}$ via OLS is not feasible when $p \geq T$.

Proposition 3. Denote $\boldsymbol{z}_t^{\mathcal{N}}$ a vector containing a permutation $\mathcal{N}$ of $\boldsymbol{z}_t$ and $\boldsymbol{B}^{\mathcal{N}}$ the Cholesky factor of its precision matrix. $\boldsymbol{\beta}^{\mathcal{N}}$ is defined accordingly as a permuted version of $\boldsymbol{\beta}$. Note that $\boldsymbol{B}^{\mathcal{N}} \neq \boldsymbol{B}$ if $\boldsymbol{z}_t^{\mathcal{N}} \neq \boldsymbol{z}_t$ and for a given $1 \leq l \leq p+1$, $\boldsymbol{z}_{l,t}^{\mathcal{N}} = y_t$. From Lemma 1 in Peng et al. (2009) we have $\beta_j^{\mathcal{N}} \propto \sum_{k=l}^{p+1} B_{k,l}^{\mathcal{N}} B_{k,j}^{\mathcal{N}}$ and if $\beta_j^{\mathcal{N}} \neq 0$ then, $\exists\, k : l \leq k \leq p+1$ and $B_{k,l}^{\mathcal{N}} B_{k,j}^{\mathcal{N}} \neq 0$.

Proposition 3 implies that if $\boldsymbol{z}_t^{\mathcal{N}} = \boldsymbol{z}_t$ then $\beta_j = 0 \iff B_{p+1,j} = 0$ or equivalently $\beta_j \neq 0 \iff B_{p+1,j} \neq 0$ such that both $\boldsymbol{\beta}$ and $\boldsymbol{B}$ can be sparse. The statement also implies that $\boldsymbol{\beta}$ can be dense while $\boldsymbol{B}^{\mathcal{N}}$ is sparse. For instance, if $\boldsymbol{z}_t^{\mathcal{N}} = (y_t, x_{5,t}, x_{4,t}, x_{3,t}, x_{2,t}, x_{1,t})'$,

$\boldsymbol{G} = \boldsymbol{I}$, and

$$\boldsymbol{B}^{\mathcal{N}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 1 \end{bmatrix}, \text{ then } \boldsymbol{B} = \begin{bmatrix} 2056.7902 & 0 & 0 & 0 & 0 \\ -257.2197 & 0.1289 & 0 & 0 & 0 \\ 32.031 & -0.0089 & 0.1282 & 0 & 0 \\ -4.121 & 0.0124 & -0.0058 & 0.1315 & 0 \\ 0.484 & -0.0307 & -0.0271 & -0.0395 & 0.1118 \end{bmatrix},$$

and $\boldsymbol{\beta} = (-0.1, -0.1, -0.1, -0.1)'$. In fact, such a sparse/dense structure holds whenever $\boldsymbol{z}_{1,t}^{\mathcal{N}} = y_t$, $B_{1,j} \neq 0$ and $B_{i,j} = 0$ for all $i \neq j$ and $1 \leq j \leq p$. By allowing $\boldsymbol{z}_{l,t}^{\mathcal{N}} = y_t$ for a given $1 \leq l \leq p$ and a given $\mathcal{N}$, many other types of sparse/dense structures can be obtained for $\boldsymbol{\beta}$ when $\boldsymbol{B}^{\mathcal{N}}$ is sparse.

Remark A4.1. Choosing $\mathcal{N}$ for estimating sparse $\boldsymbol{B}^{\mathcal{N}}$ is tantamount to imposing some identification conditions. However, in some cases, its choice is straightforward. For instance, if $y_t \sim \mathrm{AR}(p)$ and $\boldsymbol{z}_t^{\mathcal{N}} = (y_{t-p'}, \dots, y_{t-1}, y_t)'$ for any $p' \geq p$ then coefficients in $\boldsymbol{B}^{\mathcal{N}} = \boldsymbol{B}$ are the partial autocorrelation coefficients of y_t and $\boldsymbol{B}$ is (approximately) sparse whenever $p' \gg p$.[8] Likewise, if $\boldsymbol{y}_t \sim \mathrm{VAR}(p)$, with $\boldsymbol{y}_t \in \mathbb{R}^N$ then for $i = 1, \dots, N$ one can let $\boldsymbol{z}_t^{\mathcal{N}} := (\boldsymbol{y}_{t-p}', \dots, \boldsymbol{y}_{t-1}', y_{i,t})'$ and estimate the model parameters equation-by-equation, assuming (approximately) sparse $\boldsymbol{B}^{\mathcal{N}}$ at each step.

For simplicity, let us consider the problem of estimating an approximately sparse $\boldsymbol{B}^{\mathcal{N}}$. Without loss of generality let $\boldsymbol{z}_t^{\mathcal{N}} = \boldsymbol{z}_t$. Recall that the RBMCD is a sequential procedure. At each step, one needs to fit a linear regression model. These models can progressively be of higher dimension. Hereafter, we introduce the refitted adaptive lasso estimator to perform this task as a means of hedging against the curse of dimensionality associated with the use of the OLS estimator. The refitted adaptive lasso estimator is an OLS estimator

[8] A vector $\boldsymbol{\beta}$ is said to be approximately sparse if a small set of coefficients are different from zero and the others are close to it. For a formal definition, see (Belloni et al., 2014).

based on the variables selected by an adaptive lasso procedure. For instance, at step $p+1$, the adaptive lasso takes the following form

$$\begin{aligned} \tilde{\boldsymbol{b}}_{p+1} &= \arg\min \sum_{t=1}^{T} \left(z_{p+1,t} - \boldsymbol{b}_{p+1}{}'\boldsymbol{z}_{1:p,t}\right)^2 + \lambda_{p+1} \sum_{i=1}^{p} \omega_{p+1,i} \left|b_{p+1,i}\right| \\ \Longleftrightarrow \tilde{\boldsymbol{\beta}} &= \arg\min \sum_{t=1}^{T} \left(y_t - \boldsymbol{\beta}'\boldsymbol{x}_t\right)^2 + \lambda_{p+1} \sum_{i=1}^{p} \omega_{p+1,i} \left|\beta_i\right|. \end{aligned} \tag{A4.2}$$

Denote the active set $\mathcal{M}$, i.e., the set of relevant variables among the regressors and $\widehat{\mathcal{M}}$ its estimate via the adaptive lasso. Let $\boldsymbol{b}_{j,\widehat{\mathcal{M}}}$ be the sub-vector of $\boldsymbol{b}_j$ of the estimated relevant coefficient and define $\boldsymbol{z}_{\widehat{\mathcal{M}},t}$ consequently. The post-OLS estimator is the OLS estimator of the regression:

$$\boldsymbol{z}_{j,t} = b'_{j,\widehat{\mathcal{M}}} \boldsymbol{z}_{\widehat{\mathcal{M}},t} + \varepsilon_{j,t}. \tag{A4.3}$$

Recall that, when $\omega_{j,i} = 1$, we obtain the traditional lasso estimator (Tibshirani, 1996). The lasso performs model selection by shrinking the coefficient toward zero and leading to a sparse estimate of $\boldsymbol{b}_j$. It is consistent under standard conditions (Wong et al., 2020; Adamek et al., 2022), but further refinements can be obtained by considering its weighted version called the adaptive lasso (Zou, 2006). There, $\omega_{j,i}$ is equal to the inverse of the OLS estimate of $\boldsymbol{b}_{j,i}$. As the OLS estimator cannot be used when $p/T \not\to 0$, one can instead use a lasso estimate as in Medeiros and Mendes (2017). Namely,

$$\omega_{j,i} = 1/\left|\hat{\boldsymbol{b}}^I_{j,i} + \tilde{\kappa}\right|^{-1} \tag{A4.4}$$

where $\hat{\boldsymbol{b}}^I_j$ is the lasso estimator of $\boldsymbol{b}_j$ and $\tilde{\kappa} \approxeq 0$. In fact, Medeiros and Mendes (2017) show that if the weights are data-dependent and appropriately chosen, then the adaptive lasso enjoys the oracle properties. For an overview, see Fan and Lv (2010). Also, see Fan et al. (2014) for the relationship between the adaptive lasso and the family of folded concave penalized least-squares of Fan and Li (2001). The choice of the regularization parameters λ is usually achieved via cross-validation (Zhang and Yang, 2015) or information criteria (Zhang et al., 2010; Tibshirani et al., 2012). In this paper we use a corrected Akaike information criterion (Hurvich and Tsai, 1993) to select λ.

Note that the adaptive lasso as the lasso estimator can suffer from regularization bias, and to mitigate the effect of this bias, refitting strategies are usually applied, see e.g., Belloni et al. (2014); Chzhen et al. (2019). Namely, one can apply a regularized regression procedure to estimate the set of relevant coefficients and then estimate the corresponding unbiased coefficients by OLS. The resulting estimator is also called "post-OLS".

Proposition 4. The RBMCD estimator is consistent when the sample covariance of $\boldsymbol{z}_t$ converges with high probability to the population covariance matrix and the estimator of the regression coefficients at each step is consistently estimated.

Proposition 4 states that the resulting RBMCD is consistent under two conditions. Section 2 shows that the first condition holds for any stationary data with positive definite (population) covariance matrix. The second condition is met under standard assumption for the adaptive lasso (Medeiros and Mendes, 2017; Chzhen et al., 2019). In fact, the bias expansion of any regression-based modified Cholesky decomposition estimator $\tilde{\boldsymbol{\Theta}}$ is given by

$$\begin{aligned}
\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{G}}^{-1}\tilde{\boldsymbol{B}} &= (\tilde{\boldsymbol{B}} - \boldsymbol{B} + \boldsymbol{B})'(\tilde{\boldsymbol{G}}^{-1} - \boldsymbol{G}^{-1} + \boldsymbol{G}^{-1})(\tilde{\boldsymbol{B}} - \boldsymbol{B} + \boldsymbol{B}) \\
&= (\tilde{\boldsymbol{B}} - \boldsymbol{B})'(\tilde{\boldsymbol{G}}^{-1} - \boldsymbol{G}^{-1})(\tilde{\boldsymbol{B}} - \boldsymbol{B}) + (\tilde{\boldsymbol{B}} - \boldsymbol{B})'(\tilde{\boldsymbol{G}}^{-1} - \boldsymbol{G}^{-1})\boldsymbol{B} \\
&\quad + (\tilde{\boldsymbol{B}} - \boldsymbol{B})'\boldsymbol{G}^{-1}(\tilde{\boldsymbol{B}} - \boldsymbol{B}) + (\tilde{\boldsymbol{B}} - \boldsymbol{B})'\boldsymbol{G}^{-1}\boldsymbol{B} + \boldsymbol{B}'(\tilde{\boldsymbol{G}}^{-1} - \boldsymbol{G}^{-1})(\tilde{\boldsymbol{B}} - \boldsymbol{B}) \\
&\quad + \boldsymbol{B}'(\tilde{\boldsymbol{G}}^{-1} - \boldsymbol{G}^{-1})\boldsymbol{B} + \boldsymbol{B}'\boldsymbol{G}^{-1}(\tilde{\boldsymbol{B}} - \boldsymbol{B}) + \boldsymbol{B}'\boldsymbol{G}^{-1}\boldsymbol{B}.
\end{aligned}$$

One also has,

$$\tilde{\boldsymbol{G}} - \frac{1}{T}\sum_{t=1}^{T} Diag(\boldsymbol{B}\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{B}) = \frac{1}{T}\sum_{t=1}^{T} Diag\left((\tilde{\boldsymbol{B}}-\boldsymbol{B})\boldsymbol{z}_t\boldsymbol{z}_t'(\tilde{\boldsymbol{B}}-\boldsymbol{B})' + \boldsymbol{B}\boldsymbol{z}_t\boldsymbol{z}_t'(\tilde{\boldsymbol{B}}-\boldsymbol{B})' + (\tilde{\boldsymbol{B}}-\boldsymbol{B})\boldsymbol{z}_t\boldsymbol{z}_t'\boldsymbol{B}'\right),$$

with $Diag(\boldsymbol{A})$ a diagonal matrix made up of the diagonal element of $\boldsymbol{A}$.

A5 Details about the empirical application

A5.1 Conditional Granger causality

Granger causality captures predictability given a particular information set. Therefore, the relevant null hypothesis to test for conditional Granger causality from unit i' to i can be expressed as: $\mathbb{H}_{0,i'\to i} : \mathbb{E}\left(y_{i,t} \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left(y_{i,t} \mid \mathcal{F}_{-i',t-1}\right)$, where $\mathcal{F}_{t-1} := \{(y_{i,s}, y_{i',s}, y_{k,s})', s \leq t-1, \forall k \notin \{i,i'\}\}$ and $\mathcal{F}_{-i',t-1} := \{(y_{i,s}, y_{k,s})', s \leq t-1, \forall k \notin \{i,i'\}\}$. In other words, conditional on an information set containing the past returns of all institutions, if the past values of the return of institution i' improve the prediction of the return of institution i at time t, then i' is Granger causal for i. Assuming that $\mathbb{E}\left(y_{i,t} \mid \mathcal{F}_{t-1}\right) = \boldsymbol{x}_{i,t}'\boldsymbol{\beta}_i$ with $x_t = (y_{i,t-1}, \ldots, y_{i,t-L}, y_{i',t-1}, \ldots, y_{i',t-L}, \ldots, y_{k,t-1}, \ldots, y_{k,t-L}, \ldots)'$ for L being the model lag length, we test $\mathbb{H}_{0,i'\to i}$ by testing the significance of the linear regression coefficients $\beta_{i,L+1}, \ldots, \beta_{i,2L}$. $\mathbb{E}\left(y_{i,t} \mid \mathcal{F}_{t-1}\right)$ is a single equation of a VAR(L). For each institution $i = 1, \ldots n$, we estimate the model each year ($T \approx 250$) with $L = 5$ lags.[9] Since $\boldsymbol{\beta}_i \in \mathbb{R}^{440}$, estimation and inference via OLS is infeasible, and a natural solution is to use our PrLS estimator. The refitted adaptive lasso-based modified Cholesky decomposition is used as a consistent estimator of the precision matrices.[10] The number of subseries values is also fixed to $T^{1/3}$. Furthermore, to prevent the accumulation of type I errors due to

[9]For an overview of lag selection procedures, see Hecq et al. (2021). Following their approach, we find that the optimal lag is lower than 5 in the different estimations, but we overspecified the number of optimal lags to avoid potential lag truncation issues. We use 5 to account for the potential dynamic in a week.

[10]See Online Appendix A4 for its presentation and Remark A4.1 for the justification of this choice in this context.

multiple hypothesis testing, we apply a Benjamini and Yekutieli (2001) false-discovery control procedure at each estimation step. Thus, for each year from 2005 to 2020, we obtain an adjacency matrix $\boldsymbol{A}$ containing all the information about the connectedness among the institutions, such that $A_{i,i'} = 1$ if at least one of the parameters $\beta_{i,6}, \ldots, \beta_{i,10}$ has its associated adjusted p-value lower than 20% and zero otherwise.

A5.2 Measuring global systemic risk via modularity

We now introduce the concept of *modularity* from graph theory. In doing so, we closely follow Clauset et al. (2004). Let $A_{i,i'}$ be an element of the adjacency matrix $\boldsymbol{A}$ of the estimated network at a given year, such that $A_{i,i'} = 1$ if vertices i and i' are connected and $A_{i,i'} = 0$ otherwise. Vertices are divided into *communities* in such a way that they form a partition. Let $m = 1/2 \sum_{i,i'} A_{i,i'}$ be the number of edges in the graph and call $k_i = \sum_{i'} A_{i,i'}$ the degree of a vertex i, i.e., the number of edges incident upon such a vertex. We can then define *modularity* (M) as $M := \frac{1}{2m} \sum_{i,i'} \left[A_{i,i'} - \frac{k_i k_{i'}}{2m} \right] \delta\left(c_i, c_{i'}\right)$, for $\delta\left(c_i, c_{i'}\right) = 1$ if $i = i'$ and 0 otherwise. The interpretation of the modularity is straightforward as $k_i k_{i'}/2m$ is the expected fraction of within-community edges for a randomized network. If the number $k_i k_{i'}/2m$ is no different from that observed in the estimated network, then the modularity will be zero. Therefore, nonzero values represent deviations from randomness, where a large value of M indicates significant community structure in a network. We employ the hierarchical agglomeration algorithm of Clauset et al. (2004) to detect communities. The algorithm uses a greedy optimization strategy that starts with each vertex as a unique member of a community of one and repeatedly joins the two communities whose amalgamation produces the largest increase in modularity.

A5.3 Additional figures

Recall Figure 4 in the main paper, which displays the Granger causality networks obtained with the precision least squares estimator between 2006 and 2010. By looking at the causal networks, we observed how financial connectedness substantially decreases when approaching the financial crisis and peaks at its lowest in the crisis year, to only slightly recover in 2009 and fully return to dense in 2010. Prediction in the presence of a structural break as in 2008 (subprime crisis) is very difficult and is reflected by the lack of connections in stock returns during 2008. Interestingly, 2008 does not look empty, meaning that a subset of series can still have predictive power over an unprecedented crisis. In Figures A.7 and A.8, similar patterns can be observed around 2011 (European debt crisis), 2016 (downside macroeconomic concerns), 2018 (Turkish currency and debt crisis), 2020 (COVID-19 pandemic). Along with the connections, Figures A.6, A.7 and A.8 display the communities obtained throughout the different considered years using the greedy optimization algorithm of Clauset et al. (2004) based on modularity. We observe that the number of predictive relationships (communities) decreases (increases) substantially during crisis periods (2007-2009, 2011, 2016, 2018, 2020) and increases (decreases) during the expansion period (2005-2006, 2014). The years between these periods are characterized by a market recovery i.e., a re-connection of the system and a decrease in the number of communities. We can consider such years as “breathing years”.

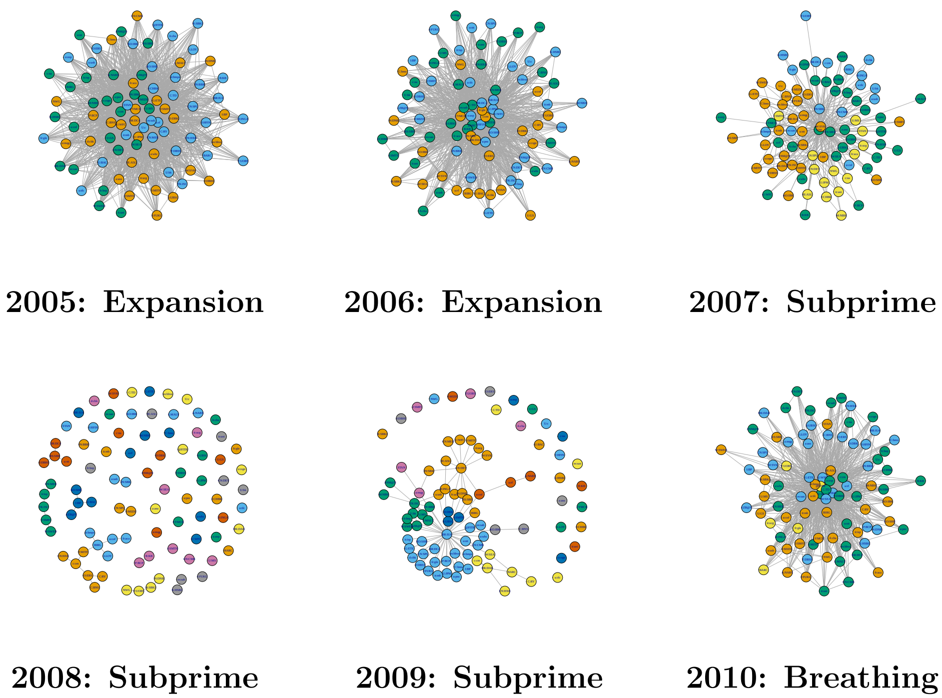


Figure A.6: Network Communities via Modularity (1/3)

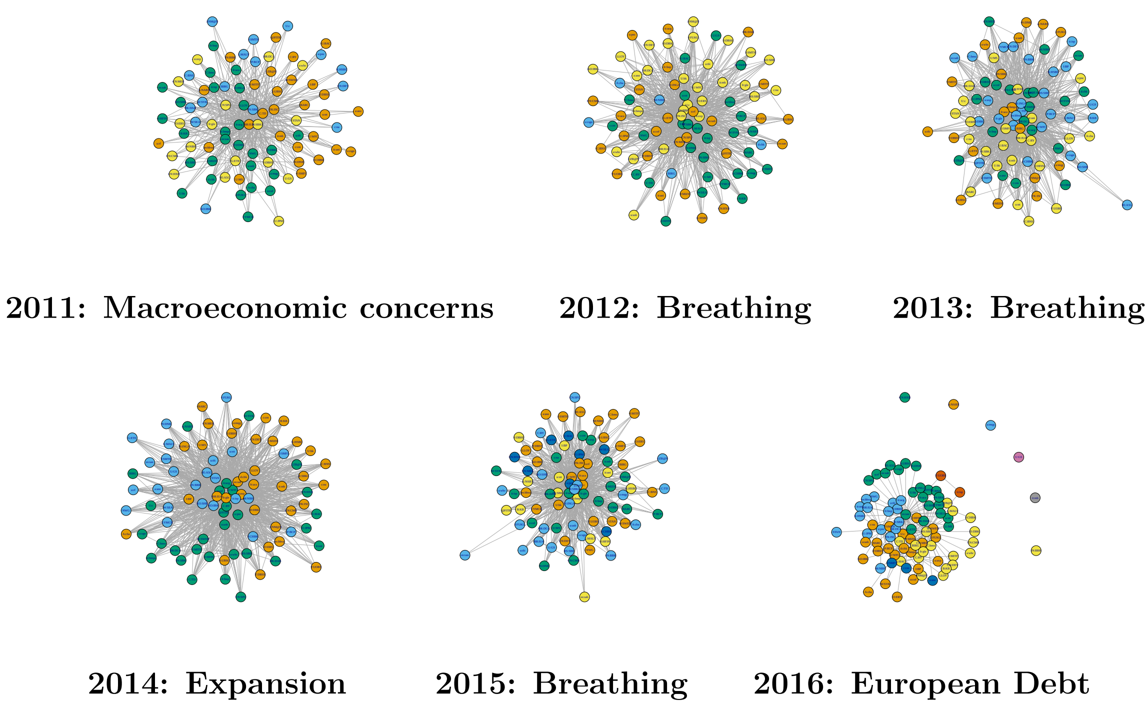
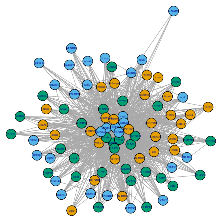
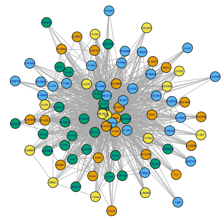


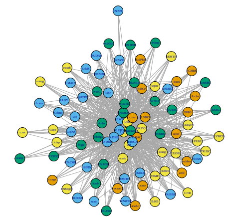
Figure A.7: Network Communities via Modularity (2/3)



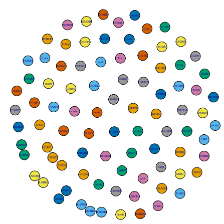
2017: Breathing



2018: Turkish currency and debt



2019: Breathing



2020: COVID-19 Pandemic

Figure A.8: Network Communities via Modularity (3/3)

A5.4 Discussion

Our analysis suggests that crises correspond to a collapse of financial linkages. This finding is consistent with the literature on financial market fragility. " An economy exhibits financial fragility if it possesses a propagation mechanism that allows small exogenous shocks at the initial date to generate financial crises that have large-scale effects on the financial structure and thus on real activity" (Lagunoff and Schreft, 2001). Lagunoff and Schreft (2001) built a model in which agents have portfolios whose returns depend on the portfolio allocations of others. Some agents are subject to shocks that lead them to reallocate their portfolios and consequently cause networks links to break. Two types of crises were proposed. The first type happens gradually as agents do not anticipate possible losses and thus do not instantaneously break links. Losses spread across the network and break more links. The second type of crisis happens instantly, as agents foresee losses and pre-emptively break links to prevent such losses from spreading. We associate this behavior with the fragility of the reputation of international and large financial institutions. According to Ordoñez (2013), reputation concerns are fragile and may suddenly disappear, leading to important changes in aggregate risk-taking, and well-known and reputable firms shift their behavior in response to small and unobvious changes in fundamentals. This translates into a lack of predictability in the sense of Granger, reflecting panic in the global market.

Interestingly, the existing empirical literature suggests the opposite, namely that financial crises correspond to an increase in financial linkages (Billio et al., 2012). If this argument were valid, it would imply that during a crisis, one could better predict the returns of a given financial institution one week ahead using the past returns of the others (along with its own). Such results would be hard to defend. In fact, this result is justified as follows. As the correlation among stock returns increases during a crisis and decreases thereafter, international financial institutions are expected to be more connected during crises. However, this argument does not necessarily imply more predictability during financial crises. Indeed, it is easy to check that if the returns follow a VAR(1) model where each regression equation is an AR(1) and the covariance of the innovation has a Toeplitz structure, one can end up with a system of highly correlated returns but an empty conditional Granger-causal network for a suitable choice of the values of the parameters of the model.

The past literature presents important methodological limits for which the PrLS-based VAR estimator and hypothesis testing framework are robust against. For instance, Billio et al. (2012) used pairwise Granger causality testing, which is known to lead to spurious connections in the presence of indirect links. Following the same framework of Billio et al. (2012), Basu et al. (2019) used a GARCH(1,1) model on the returns to remove GARCH effects. Then, they used a lasso to estimate the network using a VAR model. Basu et al. (2019) assumed i.i.d innovations but tested for conditional Granger causalities on the adjusted returns. Nevertheless, this approach is subject to model risk if the filtering step is based on a misspecified model. Our approach avoids the filtering step by allowing stationary innovations. Another work that could be related to ours is Hué et al. (2019). Nevertheless, this work does not test Granger's conditional causality in high dimensions but only provides a measure of the systemic importance of financial institutions using a computationally expensive leave-one approach.

To the best of our knowledge, our paper is unique in the literature on network analysis using conditional Granger causality among stock returns. The test is valid under the general framework of physical dependence and allows for estimating large systems. It is worth emphasizing that the application success depends on the practitioner ability to estimate the potentially large precision matrix. In our case, we use the regression-based modified Cholesky decomposition as discussed in Online Appendix A4, Remark A4.1 and obtain results in line with different theories in financial economics.

A5.5 Global banking market

Table A5.1: Sample of banks analyzed

Label	Bank name	Label	Bank name
U:JPM	JP MORGAN CHASE & CO.	J:MITF	MITSUBISHI UFJ FINL.GP.
U:BAC	BANK OF AMERICA	J:MIZH	MIZUHO FINL.GP.
U:c	CITIGROUP	J:SMFI	SUMITOMO MITSUI FINL.GP.
U:WFC	WELLS FARGO & CO	J:DBHI	RESONA HOLDINGS
U:GS	GOLDMAN SACHS GP.	J:NM@N	NOMURA HDG.
U:MS	MORGAN STANLEY	J:SMTH	SUMITOMO MITSUI TST.HDG.
U:bk	BANK OF NEW YORK MELLON	J:FUKU	FUKUOKA FINANCIAL GP.
u:usb	US BANCORP	J:CHBK	CHIBA BANK
u:pnc	PNC FINL.SVS.GP.	J:HFIN	HOKUHOKU FINL. GP.
u:cof	CAPITAL ONE FINL.	J:ZB@N	SHIZUOKA BANK
u:stt	STATE STREET	J:YMCB	YAMAGUCHI FINL.GP.
u:tfc	TRUIST FINANCIAL	C:TD	TORONTO-DOMINION BANK
u:axp	AMERICAN EXPRESS	C:RY	ROYAL BANK OF CANADA
@FITB	FIFTH THIRD BANCORP	C:BNS	BK.OF NOVA SCOTIA
u:rf	REGIONS FINL.NEW	C:bmo	BANK OF MONTREAL
CN:CMB	CHINA MINSHENG BANKING 'A'	fr:bnp	BNP PARIBAS
CN:DEV	PING AN BANK 'A'	fr:aca	CREDIT AGRICOLE
CN:HXB	HUAXIA BANK 'A'	fr:gle	SOCIETE GENERALE
K:HSBC	HSBC HOLDINGS	KO:SHB	SHINHAN FINL.GROUP
BARC	BARCLAYS	KO:IBK	INDUSTRIAL BANK OF KOREA
NWG	NATWEST GROUP	S:UBSG	UBS GROUP
LLOY	LLOYDS BANKING GROUP	s:csgn	CREDIT SUISSE GROUP
STAN	STANDARD CHARTERED	B:KB	KBC GROUP
es:san	BANCO SANTANDER	BR:IU4	ITAU UNIBANCO HOLDING PN
es:bbva	BBV.ARGENTARIA	BR:DC4	BANCO BRADESCO PN
es:sab	BANCO DE SABADELL	D:DBK	DEUTSCHE BANK
M:NBH	NORDEA BANK	d:cbk	COMMERZBANK
W:SVK	SVENSKA HANDELSBANKEN A	BIRG	BANK OF IRELAND (LON)
W:SEA	SKANDINAVISKA ENSKILDA BANKEN A	IE:A5G	AIB GROUP
W:SWED	SWEDBANK A	IN:SBK	STATE BANK OF INDIA
c:cm	CANADIAN IMP.BK.COM.	a:mqg	MACQUARIE GROUP
c:na	NATIONAL BANK OF CANADA	CN:MER	CHINA MERCHANTS BANK 'A'
it:ucg	UNICREDIT	CN:SPU	SHAI.PUDONG DEV.BK. 'A'
it:isp	INTESA SANPAOLO	IN:BBR	BANK OF BARODA
it:bmps	BANCA MONTE DEI PASCHI	L:MALY	MALAYAN BANKING
I:BP	BANCO BPM	P:BCP	BANCO COMR.PORTUGUES 'R'
it:uni	UNIPOL GRUPPO FINANZIARI	T:DBSS	DBS GROUP HOLDINGS
it:mb	MEDIOBANCA BC.FIN	T:UOBS	UNITED OVERSEAS BANK
A:NABX	NATIONAL AUS.BANK	O:ERS	ERSTE GROUP BANK
A:cba	COMMONWEALTH BK.OF AUS.	DK:DAB	DANSKE BANK
a:anz	AUS.AND NZ.BANKING GP.	G:ETE	NATIONAL BK.OF GREECE
a:wbc	WESTPAC BANKING	TK:ISC	TURKIYE IS BANKASI 'C'
H:INGA	ING GROEP	R:SBKJ	STANDARD BANK GROUP
N:DNB	DNB	RS:SBE	SBERBANK OF RUSSIA

References

R. Adamek, S. Smeekes, and I. Wilms. Lasso inference for high-dimensional time series. *Journal of Econometrics*, 2022.

S. Basu, S. Das, G. Michailidis, and A. Purnanandam. A system-wide approach to measure connectivity in the financial sector. *Available at SSRN 2816137*, 2019.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104 (3):535–559, 2012.

K. Boudt, S. Laurent, A. Lunde, R. Quaedvlieg, and O. Sauri. Positive semidefinite integrated covariance estimation, factorizations and asynchronicity. *Journal of Econometrics*, 196(2):347–367, 2017.

P. Bühlmann et al. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1):47–70, 2015.

E. Chzhen, M. Hebiri, and J. Salmon. On lasso refitting strategies. *Bernoulli*, 25(4A):3175–3200, 2019.

A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

S. Darolles, C. Francq, and S. Laurent. Asymptotics of cholesky garch models and time-varying conditional betas. *Journal of Econometrics*, 204(2):223–247, 2018.

A. P. Dempster. *Elements of continuous multivariate analysis*. Addison-Wesley Publishing Company, 1969.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

J. Fan, Y. Fan, and E. Barut. Adaptive robust variable selection. *Annals of Statistics*, 42(1):324, 2014.

D. M. Hawkins and W. Eplett. The cholesky factorization of the inverse correlation or covariance matrix in multiple regression. *Technometrics*, 24(3):191–198, 1982.

A. Hecq, L. Margaritella, and S. Smeekes. Granger Causality Testing in High-Dimensional VARs: A Post-Double-Selection Procedure. *Journal of Financial Econometrics*, 2021.

S. Hué, Y. Lucotte, and S. Tokpavi. Measuring network systemic risk contributions: A leave-one-out approach. *Journal of Economic Dynamics and Control*, 100:86–114, 2019.

C. M. Hurvich and C.-L. Tsai. A corrected akaike information criterion for vector autoregressive model selection. *Journal of time series analysis*, 14(3):271–279, 1993.

R. Lagunoff and S. L. Schreft. A model of financial fragility. *Journal of Economic Theory*, 99 (1-2):220–264, 2001.

M. C. Medeiros and E. F. Mendes. Adaptive lasso estimation for ardl models with garch innovations. *Econometric Reviews*, 36(6-9):622–637, 2017.

G. L. Ordoñez. Fragility of reputation and clustering of risk-taking. *Theoretical Economics*, 8(3): 653–700, 2013.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

A. Summerfield and A. Lubin. A square root method of selecting a minimum set of variables in multiple regression: I. the method. *Psychometrika*, 16(3):271–284, 1951.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

R. J. Tibshirani, J. Taylor, et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

K. C. Wong, Z. Li, and A. Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

W. B. Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154, 2005.

D. Zhang and W. B. Wu. Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *The Annals of Statistics*, 49(1):233–254, 2021.

D. Zhang, W. B. Wu, et al. Gaussian approximation for high dimensional time series. *The Annals of Statistics*, 45(5):1895–1919, 2017.

Y. Zhang and Y. Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015.

Y. Zhang, R. Li, and C.-L. Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.