

BRIDGING FACTOR AND SPARSE MODELS

BY JIANQING FAN^{1,a} RICARDO P. MASINI^{2,b} AND MARCELO C. MEDEIROS^{3,c}

¹*Department of Operations Research and Financial Engineering, Princeton University, jqfan@princeton.edu*

²*Department of Statistics, University of California, Davis, rmasini@ucdavis.edu*

³*Department of Economics, University of Illinois at Urbana-Champaign, marcelom@illinois.edu*

Factor and sparse models are widely used to impose a low-dimensional structure in high-dimensions. However, they are seemingly mutually exclusive. We propose a lifting method that combines the merits of these two models in a supervised learning methodology that allows for efficiently exploring all the information in high-dimensional datasets. The method is based on a flexible model for high-dimensional panel data with observable and/or latent common factors and idiosyncratic components. The model is called the factor-augmented regression model. It includes principal components and sparse regression as specific models, significantly weakens the cross-sectional dependence, and facilitates model selection and interpretability. The method consists of several steps and a novel test for (partial) covariance structure in high dimensions to infer the remaining cross-section dependence at each step. We develop the theory for the model and demonstrate the validity of the multiplier bootstrap for testing a high-dimensional (partial) covariance structure. A simulation study and applications support the theory.

1. Introduction. With the emergence of new and large datasets in almost all disciplines, the correct characterization of the dependence among variables is of substantial importance. Usually, to achieve this goal, the literature has followed two seemingly orthogonal tracks over the last two decades. On the one hand, factor models have become an essential tool to summarize information in large datasets, assuming that the remaining dependence structure is negligible. For instance, panel factor models are now applied to various problems, ranging from forecasting to causal inference and network analysis. However, on the other hand, there have been significant advances in parameter estimation in ultra high-dimensions under the assumption of sparsity or weak sparsity. That is, a variable depends only on a (very) small subset of the other variables. For an overview of these two topics and their exciting developments, see [Fan et al. \(2020\)](#).

In this paper, we take an alternative route and combine the best of the two worlds described above to better characterize the dependence structure in high-dimensional datasets. More specifically, we consider that the covariance structure of a large set of variables, organized in a panel data format, is characterized as a combination of a factor structure, where factors can be either observed, unobserved, or both, and a weakly-sparse idiosyncratic component. This formulation is general enough to accommodate many data-generating processes of interest in economics, finance, epidemiology, and energy/engineering, for example. The proposed methodology has two ingredients: a multistep estimation procedure and a new test for structure in high dimensional (partial) covariance matrices. The steps of the estimation procedure are as follows. In the first one, we take the original data and remove the effects of any observed factors. These factors can be deterministic terms such as seasonal dummies and/or trends or other observed covariates. The first step can be parametric or nonparametric,

Received August 2022; revised April 2023.

MSC2020 subject classifications. 62H25, 62F03.

Key words and phrases. Factor models, penalized least-squares, high-dimensional inference, (partial) covariance structure, prediction, machine learning, supervised learning.

low or high dimensional. A latent factor model is then estimated using the residuals from the first stage. In the third step, we model the dependence among idiosyncratic terms as a weakly sparse regression estimated by the Least Absolute Shrinkage and Selection Operator (LASSO). A final fourth step can be used to build models for out-of-sample forecasting, considering each of the previous steps' components. At each step, the proposed test for the (partial) covariance structure in high dimensions can test the null hypothesis of no remaining cross-section dependence.

Our approach has many statistical applications. It can enhance high-dimensional prediction, select more interpretable variables, construct counterfactuals for policy evaluations, and depict (partial) correlation networks, among many others.

1.1. Motivation. Let $\mathbf{Y}_t := (Y_{1,t}, \dots, Y_{n,t})'$ be a random vector generated as $Y_{i,t} = \lambda_i' \mathbf{F}_t + U_{i,t}$, for $i \in [n]$, $t \in [T]$, where $\Sigma := \mathbb{E}(\mathbf{U}_t \mathbf{U}_t')$, with $\mathbf{U}_t := (U_{1,t}, \dots, U_{n,t})'$, is not necessarily diagonal. Fix one component of interest $i \in [n]$ as a response variable. Consider the following predictive models:

$$(1.1) \quad \mathcal{M}_1 : \mathbb{E}(Y_{i,t} | \mathbf{Y}_{-i,t}), \quad \mathcal{M}_2 : \mathbb{E}(Y_{i,t} | \mathbf{F}_t), \quad \text{and} \quad \mathcal{M}_3 : \mathbb{E}(Y_{i,t} | \mathbf{F}_t, \mathbf{U}_{-i,t}),$$

where $\mathbf{Y}_{-i,t}$ and $\mathbf{U}_{-i,t}$ are, respectively, vectors with the elements of $\mathbf{Y}_t$ and $\mathbf{U}_t$ without the i th entry. $\mathcal{M}_3$ is a factor augmented regression since it is the same as $\mathbb{E}(Y_{i,t} | \mathbf{F}_t, \mathbf{Y}_{-i,t})$.

Suppose we observe both $\mathbf{F}_t$ and $\mathbf{U}_{-i,t}$. Which of the three models above is best regarding mean square error (MSE) for prediction? Comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$ is not clear since it depends, among other things, on the magnitude of Σ relative to $\Lambda' \Lambda$, where $\Lambda := (\lambda_1, \dots, \lambda_n)'$. However, since the σ -algebras generated by $\mathbf{Y}_{-i,t}$ and $\mathbf{F}_t$ are both included in the σ -algebra generated by $(\mathbf{F}_t, \mathbf{U}_{-i,t})$, it is not surprising that $\text{MSE}(\mathcal{M}_3) \leq \min[\text{MSE}(\mathcal{M}_1), \text{MSE}(\mathcal{M}_2)]$. The same will hold if we replace the models in (1.1) by their best linear projections, which we denote by $\tilde{\mathcal{M}}_j$ for $j \in \{1, 2, 3\}$. Therefore, we can write the “gains” of $\tilde{\mathcal{M}}_3$ when compared to $\tilde{\mathcal{M}}_1$ and $\tilde{\mathcal{M}}_2$:

$$\text{MSE}(\tilde{\mathcal{M}}_3) - \text{MSE}(\tilde{\mathcal{M}}_1) = -\boldsymbol{\vartheta}_i' \Sigma_{-i,-i} \boldsymbol{\vartheta}_i,$$

$$\text{MSE}(\tilde{\mathcal{M}}_3) - \text{MSE}(\tilde{\mathcal{M}}_2) = -\Delta_{1,i} \Delta_{1,i}' - \Delta_{2,i}' \Sigma_{-i,-i} \Delta_{2,i},$$

where $\boldsymbol{\vartheta}_i$ is coefficients of the projection of $U_{i,t}$ onto $\mathbf{U}_{-i,t}$; $\Sigma_{-i,-i}$ is Σ excluding the i th row and column; $\Delta_{1,i} := \Lambda_i - \beta_i' \Lambda_{-i}$; $\Delta_{2,i} := \beta_i - \boldsymbol{\vartheta}_i$; and β_i is the coefficient of the projection of $Y_{i,t}$ onto $\mathbf{Y}_{-i,t}$. From the previous expressions, it becomes evident that both $\tilde{\mathcal{M}}_1$ and $\tilde{\mathcal{M}}_2$ are restrictions on $\tilde{\mathcal{M}}_3$. Broadly speaking, whenever one does not expect to have an *exact* factor model, there are potential gains of considering the contribution of the idiosyncratic components $\mathbf{U}_{-i,t}$. Therefore, we use $\tilde{\mathcal{M}}_3$ as the base model for the estimation methodology described in Section 2.2.

1.2. Main contributions and comparison with the literature. The contributions of this paper are multifold. First, our methodology bridges the gap between two competing methods for high-dimensional modeling; see, for example, the discussion in [Giannone, Lenza and Primiceri \(2021\)](#) and [Fan, Ke and Wang \(2020\)](#). This yields a vast number of potential applications and spin-offs. For instance, in [Fan, Masini and Medeiros \(2022\)](#), we apply the methods developed here to evaluate the effects of interventions and contribute to the literature on synthetic controls and related methods by combining the approaches of [Gobillon and Magnac \(2016\)](#), and [Carvalho, Masini and Medeiros \(2018\)](#). Therefore, a common strong factor structure and weak sparsity can coexist in our setup.

Second, the methodology proposed here contributes to the forecasting literature. For instance, in the second application considered in this paper, we build forecasting models based

on a large cross-section of macroeconomic variables. We call this method the *FarmPredict*. Combining factors and a sparse regression strongly outperforms the traditional principal component regression as in [Stock and Watson \(2002\)](#). Therefore, *FarmPredict* can be an additional contribution to the forecasting and machine learning toolkit. The method can be easily extended to a multivariate setting combining factor-augmented vector autoregressions (FAVAR) as in [Bernanke, Boivin and Elias \(2005\)](#) and sparse vector models as in [Kock and Callot \(2015\)](#) and [Masini, Medeiros and Mendes \(2022\)](#). Our methodology can also be applied in areas beyond economics. For example, it can be useful to construct forecasting models for the spread of infectious diseases, such as COVID-19, by considering co-movements and spillovers among different locations.

Third, we show the consistency of factor estimation based on the residuals of a first-step regression. Our results hold for both the parametric (linear or nonlinear) and nonparametric first stage. A high-dimensional first stage is also allowed. Current results in the literature consider that factors are estimated based on observed data, and our derivations favor a much more flexible and general setup ([Bai \(2003\)](#), [Bai and Ng \(2002, 2006\)](#)). Our methodology allows settings with observed and latent factors and trend-stationary data. In the latter, the trend can be removed by (nonparametric) first-stage regression. Whenever the unobserved factors and the observed covariates are correlated, the method proposed in [Pesaran \(2006\)](#) can be used, and all results follow directly.

Fourth, we contribute to the LASSO literature. LASSO can not be model selection consistent for highly correlated variables. By decomposing covariates into factors and idiosyncratic components, namely the idea of lifting, we decorrelate the variables and make the model selection condition much easier to hold; see, for example, [Fan, Ke and Wang \(2020\)](#). We show the consistency of the estimates based on the residuals of the previous steps. Our results are derived under restrictions on the population covariance matrix of the data and not on the estimated one, as is usual in many papers. See, for example, [van de Geer and Bühlmann \(2009\)](#). Furthermore, we derive our results under much mild conditions than the ones considered in [Fan, Ke and Wang \(2020\)](#).

Fifth, we extend the results in [Chernozhukov, Chetverikov and Kato \(2013, 2019\)](#) to strong-mixing data to construct hypothesis tests for covariance and partial covariance structure in high dimensions.¹ We also establish the consistency of a new estimator of the partial covariance matrix in high-dimensions and strong-mixing data. Our proposed tests can be used to infer if the (partial) covariance matrix of a high-dimensional random vector is diagonal or block-diagonal. More generally, we can test any pre-defined structure.

Furthermore, we show that the test remains valid when we use the residuals from a previous step estimation to compute the covariance matrix. This result allows us to apply the test to the proposed multistage estimation procedure. These are important results for applications in economics, finance, epidemiology, and other areas. For instance, our inference procedure can serve as a diagnostic and misspecification tool. For panel data models with interactive fixed effects as in [Pesaran \(2006\)](#), [Bai \(2009\)](#), [Moon and Weidner \(2015\)](#), and [Bai and Liao \(2017\)](#), our test can be applied to uncover the dependence among cross-sectional units before and after accounting for common factor components. We expect the idiosyncratic covariance matrix to be almost sparse if the factor structure is informative enough. If this is not the case, we may have underestimated the number of factors. One popular application is in asset pricing,

¹[Giessing and Fan \(2020\)](#) also extended [Chernozhukov, Chetverikov and Kato \(2013\)](#). However, their setup differs from ours as they only consider the case of independent and identically distributed data. The results in [Chernozhukov, Chetverikov and Kato \(2013, 2019\)](#) have also been extended to dependent data by [Zhang and Wu \(2017\)](#). However, the dependence measure the authors considered differs from the concept of mixing used in this paper.

as discussed in [Gagliardini, Ossola and Scaillet \(2019\)](#) and the empirical section of this paper. There are a vast number of proposed factors as described in [Feng, Giglio and Xiu \(2020\)](#). We can apply our methodology to test for omitted factors and estimate network connections among firms, as in [Diebold and Yilmaz \(2014\)](#) and [Brownlees, Gudmundsson and Lugosi \(2020\)](#). Finally, as a diagnostic tool, our paper tackles the same problem as [Gagliardini, Ossola and Scaillet \(2019\)](#). However, we take an alternative solution strategy that relies on a different set of hypotheses; see also [Gagliardini, Ossola and Scaillet \(2020\)](#).

Although our results are derived under the assumption that the number of factors is known, simulation results presented in the Supplementary Material provide evidence that the test has good finite-sample properties even when data-driven methods determine the number of factors. Over the past years, many papers proposed different methods to test for covariance structure in high dimensions. See, for example, [Cai \(2017\)](#) and the references therein. To the best of our knowledge, we complement all the previous papers by simultaneously considering high-dimensions, strong-mixing data with mild distributional assumptions, and pre-estimation when constructing tests for both covariance and partial covariance structure.

Finally, we highlight the theoretical challenges we tackle in this paper. First, to derive the properties of our proposed test for (partial) covariance structure, we prove a new high dimensional central limit theorem (CLT) for strong mixing sequences. This CLT is new in the literature and allows us to apply Gaussian approximation results in a much more general framework. Second, as a side result, we show the consistency of kernel-based estimation of a high-dimensional long-run covariance matrix of strong-mixing processes. This is also a new result with significant consequences for the theory of high-dimensional regression with dependent errors. Second, all the nonasymptotic bounds for our multistep estimators are derived under the assumption that the distributions of the random variables in the model have polynomial tails, and the estimation errors of previous steps are also considered. This introduces several difficulties in proving the results but makes our method well suited for applications with fat-tailed data, such as those observed in financial applications. Lastly, the strong mixing coefficient appears explicit in the derived bounds and can grow with the sample size. This not only introduces technical challenges but makes our results very general.

Summarizing, our approach provides the following:

1. A systematic way to unify factor and sparse models to construct statistical specifications which use all available information and that can be applied, for example, to:
 - (a) Forecasting in a high-dimensional setting;
 - (b) Construction of counterfactuals to aggregate data; or
 - (c) Estimation of partial correlation networks;
2. An inferential procedure to test for general structures in covariance and partial covariance matrices which can be useful, among other things, for:
 - (a) Test for misspecification in factor models; or
 - (b) Test for nontrivial links among idiosyncratic units.

1.3. Organization of the paper. In addition to this Introduction, the paper is organized as follows. We present the model setup and assumptions in Section 2. The theoretical results are presented in Section 3. We discuss the empirical application in Section 4. Section 5 concludes. All proofs, additional discussion, simulation, and empirical results are deferred to the Supplementary Material due to space constraints. Tables and figures in the Supplementary Material are referenced with an “S” before the number.

1.4. Notation. All random quantities (real-valued, vectors and matrices) are defined in a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote random variables by an upper case letter, X , and its realization by a lower case letter, $X = x$. The expected value operator is with respect to the $\mathbb{P}$ law, such that $\mathbb{E}(X) := \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$. Matrices and vectors are written in bold letters $\mathbf{X}$. Except for the number of factors, r , and the number of covariates, k , defined below, all other dimensions are allowed to depend on the sample size (T) . However, we omit this dependency throughout the paper to avoid cluttering the notation prematurely. Also, we write $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$ and denote the cardinality of a set $\mathcal{S}$ by $|\mathcal{S}|$.

We use $\|\cdot\|_p$ to denote the ℓ^p norm for $p \in [1, \infty]$, such that for a d -dimensional (possibly random) vector $\mathbf{X} = (X_1, \dots, X_d)'$, we have $\|\mathbf{X}\|_p := (\sum_{i=1}^d |X_i|^p)^{1/p}$ for $p \in [1, \infty)$ and $\|\mathbf{X}\|_{\infty} := \sup_{i \leq d} |X_i|$. If $\mathbf{X}$ is a $(m \times n)$ possibly random matrix, then $\|\mathbf{X}\|_p$ denotes the matrix ℓ^p -induced norm and $\|\mathbf{X}\|_{\max}$ denotes the maximum entry in absolute terms of the matrix $\mathbf{X}$. Note that whenever $\mathbf{X}$ is random, then $\|\mathbf{X}\|_p$ for $p \in [1, \infty]$ and $\|\mathbf{X}\|_{\max}$ are random variables. We also reserve the symbol $\|\cdot\|$ without subscript for the Euclidean norm $\|\cdot\| := \|\cdot\|_2$ for both vectors and matrices. We denote the $\ell^p(\mathbb{P})$ norm of X by $\|X\|_p$ for $p \in [0, \infty]$, that is, $\|X\|_p := (\mathbb{E}|X|^p)^{1/p}$ for $p \in [1, \infty)$ and $\|X\|_{\infty}$ is the essential supremum of X (in respect to $\mathbb{P}$). Also, when $\mathbf{X}$ is a random vector, we define $\|\mathbf{X}\|_p := \sup_{\|u\| \leq 1} \|u' \mathbf{X}\|_p$. Note that $\|\mathbf{X}\|_p$ is random variable while $\|\mathbf{X}\|_p$ is nonrandom scalar for $p \in [0, \infty]$.

For any vector $\mathbf{X}$, $\text{diag}(\mathbf{X})$ is a diagonal matrix whose diagonal is the elements of $\mathbf{X}$. $\mathbb{1}(A)$ is an indicator function on the event A , that is, $\mathbb{1}(A) = 1$ if A is true and 0 otherwise. For any matrix (possibly random) $\mathbf{M} \in \mathbb{R}^{n \times T}$, $M_{i,t} \in \mathbb{R}$ represents the entry for row i and column t , $\mathbf{M}_t \in \mathbb{R}^n$ is the column-vector with all rows of column t and $\mathbf{M}_{i,\cdot} \in \mathbb{R}^T$ is a column-vector with the transpose of all elements of row i . To avoid a cumbersome notation, we do not write $\mathbf{M}_t \in \mathbb{R}^n$ as $\mathbf{M}_{\cdot,t}$. Finally, for nonnegative sequences x_m and y_m , we write $x \lesssim y$ if there is a constant C independent of m such that $x_m \leq Cy_m$ for all m . Also, we write $x \asymp y$ if both $x \lesssim y$ and $y \lesssim x$. Similarly, for nonnegative random sequences X_m and Y_m , we write $X_m \lesssim_{\mathbb{P}} Y_m$ if for every $\epsilon > 0$ there is a constant C independent of m such that $\mathbb{P}(X_m \leq CY_m) \leq \epsilon$ for all m . Also, $X_m \asymp_{\mathbb{P}} Y_m$, if both $X_m \lesssim_{\mathbb{P}} Y_m$ and $Y_m \lesssim_{\mathbb{P}} X_m$.

2. Setup and method.

2.1. Data generating process. We consider a general panel data model rich enough to nest several important cases in economics, finance, and related areas. We define the following Data Generating Process (DGP).

ASSUMPTION 1 (DGP). For $T \geq 4$ and $n \geq 2$, the process $\{Y_{i,t} : i \in [n], t \in [T]\}$ is generated by covariate-adjusted factor model

$$(2.1) \quad Y_{i,t} = \boldsymbol{\gamma}'_i \mathbf{X}_{i,t} + \underbrace{\boldsymbol{\lambda}'_i \mathbf{F}_t + U_{i,t}}_{=: R_{i,t}},$$

where $\mathbf{X}_{i,t}$ is a k -dimensional observable (random) vector which may also include a constant term and is typically used for adjustments of heterogeneity, seasonality, and covariate dependence, $\mathbf{F}_t$ is a r -dimensional vector of common latent factors, $\boldsymbol{\lambda}_i \in \mathbb{R}^r$ is the vector of factor loadings, and $U_{i,t}$ is a zero mean idiosyncratic component.² The unknown parameters are $\boldsymbol{\gamma}_i \in \mathbb{R}^k$, the factor loadings $\boldsymbol{\lambda}_i$, and the covariance matrix of the idiosyncratic components. Finally, we assume that $\mathbf{X}_{i,t}$, $\mathbf{F}_t$, and $U_{i,t}$ are mutually uncorrelated but can be serially autocorrelated.

²For simplicity, we assume that all the units i have the same number of covariates (k). The framework can certainly accommodate situations where k_i depends on i . It also includes cross-sectional regression as a specific example.

REMARK 1. In Assumption 1, we consider that k , the dimension of $\mathbf{X}_{i,t}$, is finite and fixed. Furthermore, the relation between $Y_{i,t}$ and $\mathbf{X}_{i,t}$ is linear. This is for the sake of exposition. As our theoretical results are written in terms of the consistency rate of the first-step estimation, the DGP can be made much more general by just changing the rates.

REMARK 2. The assumption that $\mathbf{X}_{i,t}$, $\mathbf{F}_t$, and $U_{i,t}$ are mutually uncorrelated can be relaxed. Whenever $\mathbf{X}_{i,t}$ is correlated with $\mathbf{F}_t$ and $U_{i,t}$, and the interest lies in the estimation of the parameters $\boldsymbol{\gamma}_i$, $i \in [n]$, the method proposed by Pesaran (2006) can apply in the first stage of the procedure considered in this paper and our theoretical results will follow. Nevertheless, we provide several examples where the assumption that $\mathbf{X}_{i,t}$, $\mathbf{F}_t$, and $U_{i,t}$ are mutually uncorrelated is reasonable.

Our modeling strategy does not stop at (2.1). Instead, we attempt to explain further $U_{i,t}$ and impose dynamics on $\mathbf{F}_t$. The former allows us to use other idiosyncratic components to explain $Y_{i,t}$ and increase the information set. For instance, the latter builds a dynamic model $\mathbf{F}_t$ to facilitate out-of-sample prediction.

For each $i \in [n]$, let $\mathbf{W}_{i,t}$ be a vector whose elements form a (nonempty) subset of $(\mathbf{U}'_{-i,t}, \mathbf{U}'_{t-1}, \dots, \mathbf{U}'_{t-l_U})'$, where l_U is a nonnegative integer (much) smaller than T . This subset of variables explains further $U_{i,t}$ by the following population regression model:

$$(2.2) \quad U_{i,t} = \boldsymbol{\theta}'_i \mathbf{W}_{i,t} + V_{i,t},$$

where $\mathbb{E}(\mathbf{W}'_{i,t} V_{i,t}) = \mathbf{0}$, for $i \in [n]$ and $t \in [T]$. For simplicity of exposition, we assume that the dimension of $\mathbf{W}_{i,t}$ is the same for all $i \in [n]$, which we denote by d_W . Clearly, $d_W \leq n(l_U + 1)$.

Similarly, for a nonnegative integer l_F (much) smaller than T , let $\mathbf{G}_{j,t}$ be a vector whose elements form a (nonempty) subset of $(\mathbf{F}'_t, \mathbf{F}'_{t-1}, \dots, \mathbf{F}'_{t-l_F})'$, for $j \in [r]$, which attempts to explain further the j th latent factor. Consider the following population regression model:

$$F_{j,t} = \boldsymbol{\rho}'_j \mathbf{G}_{j,t} + V_{j,t}^F,$$

where $\mathbb{E}(\mathbf{G}'_{j,t} V_{j,t}^F) = \mathbf{0}$, for $j \in [r]$ and $t \in [T]$. Once again, we assume that the dimension of $\mathbf{G}_{j,t}$ is the same for all $j \in [r]$, which we denote by d_G , where $d_G \leq r(l_F + 1)$. We do not necessarily exclude $\mathbf{F}_t$ from $\mathbf{G}_t$ to contemplate cases when the contemporaneous factor is available in a prediction exercise. For future reference, we write the above equation as

$$(2.3) \quad \mathbf{F}_t = \mathbf{P} \mathbf{G}_t + \mathbf{V}_t^F,$$

where $\mathbf{G}_t := (\mathbf{G}'_{1,t}, \dots, \mathbf{G}'_{r,t})'$, $\mathbf{V}_t^F := (V_{1,t}^F, \dots, V_{r,t}^F)'$, and $\mathbf{P} := (\boldsymbol{\rho}_1 : \dots : \boldsymbol{\rho}_r)'$.³

REMARK 3. The possibility of having contemporaneous factors may seem odd in principle, but this is important for deriving a general theory that will nest applications where the factors are available for the full sample. An example of such a case is the counterfactual estimation as discussed in Fan, Masini and Medeiros (2022) and in Example 4 below. On the other hand, in a pure out-of-sample forecasting exercise $\mathbf{F}_t$ should not be included in $\mathbf{G}_t$.

EXAMPLE 1 (Asset Pricing Models). Suppose $Y_{i,t}$ is the return of an asset i at time t and let $\mathbf{X}_{i,t} := \mathbf{X}_t$ be a set of k observable common risk factors, such as the market returns or Fama-French factors (Fama and French (1993, 2015)). $\mathbf{F}_t$ can be additional, nonobservable risk factors. Several asset pricing models, such as the Capital Asset Pricing Model (CAPM)

³Strategies for selection of l_U and l_F are discussed in Section S.2 of the Supplementary Material (Fan, Masini and Medeiros (2023)).

or the Arbitrage Pricing Theory (APT) model, are nested into this general framework. The idiosyncratic terms, $U_{i,t}$, can be nontrivially correlated across assets, representing links among firms not captured by the common factor structure. In this case, we may be interested in estimating a regression where $W_{i,t} = U_{-i,t}$ such that

$$U_{i,t} = \theta_i' U_{-i,t} + V_{i,t}.$$

θ_i represents the links among firms after controlling for the factors. The structure of the (partial) covariance matrix of $U_t = (U_{1,t}, \dots, U_{n,t})'$ also shed light on these potential links.

EXAMPLE 2 (Networks). Model (2.1) also complements the network specifications discussed in Barigozzi and Hallin (2017) and Barigozzi and Brownlees (2019). Furthermore, as discussed in the previous example, the test proposed here can detect network links as in Diebold and Yilmaz (2014) and Brownlees, Gudmundsson and Lugosi (2020). As another example, $Y_{i,t}$ can be the (realized) volatility of financial assets, and $X_{i,t} := X_t$ can be volatility factors as in Andreou and Ghysels (2021).

EXAMPLE 3 (FAVAR). In the case where the index i represents a different dependent (endogenous) variable and $U_{i,t}$ is a dependent process, model (2.1) turns out to be equivalent to the Factor Augmented Vector Autoregressive (FAVAR) model of Bernanke, Boivin and Elias (2005). $X_{i,t}$ may include a constant and seasonal dummies, and $W_{i,t} = (U'_{t-h}, \dots, U'_{t-l_U})'$ and $G_{j,t} = (F'_{t-h}, \dots, F'_{t-l_F})'$, where h is the forecasting horizon.

EXAMPLE 4 (Panel Data Models). Model (2.1) is the panel model with iterative fixed-effects considered in Gobillon and Magnac (2016), where the authors propose an alternative to the Synthetic Control method of Abadie and Gardeazabal (2003) to evaluate the effects of regional policies. Model (2.1) is also in the heart of the FarmTreat method of Fan, Masini and Medeiros (2022), where the authors set $W_{i,t} = U_{-i,t}$.

Recently, Barigozzi, Cho and Owens (2022) proposed a nice extension of our approach. They consider a generalized dynamic factor model (GDFM) à la Forni et al. (2000) with sparse idiosyncratic second order structure. When the dynamic loadings are filters of finite order then the GDFM can be written as our static factor model having F_t defined as the dynamic factors stacked together with their finite lags. However, if the loadings are autoregressive as in Forni et al. (2017) then the static representation is not possible.

2.2. The method. The method proposed here consists of multiple stages, where the residuals' covariance structure can be tested at the end of each stage.

1. For each $i \in [n]$ run the regression

$$Y_{i,t} = \gamma_i' X_{i,t} + R_{i,t}, \quad t \in [T],$$

and compute $\hat{R}_{i,t} := Y_{i,t} - \hat{\gamma}_i' X_{i,t}$. The first stage may consist of a regression on a constant, a deterministic time trend, seasonal dummies, or, as in Example 1, a regression on observed factors. After removing the contribution from the observables, we can use the test for the null hypothesis of no remaining (partial) covariance structure to check if the (partial) covariance of $R_{i,t}$ is dense or sparse. If it is dense, we move to Step 2. Otherwise, we jump directly to Step 3. This first parametric, low-dimensional step can be replaced by a nonlinear/nonparametric regression or by a high-dimensional model when, for example, the number of observed factors is large. As pointed out in Remark 2, Pesaran's (2006) estimator can also be used whenever correlation between $X_{i,t}$ and $R_{i,t}$ is allowed. This will be discussed more in the subsequent sections.

2. Write $\mathbf{R}_t := (R_{1,t}, \dots, R_{n,t})'$ and $\mathbf{R}_t = \mathbf{\Lambda} \mathbf{F}_t + \mathbf{U}_t$. This step consists of estimating $\mathbf{\Lambda}$ and $\mathbf{F}_t$, for $t \in [T]$, through principal component analysis (PCA)⁴ of $\widehat{\mathbf{R}}_t$ and compute

$$\widehat{\mathbf{U}}_t = \widehat{\mathbf{R}}_t - \widehat{\mathbf{\Lambda}} \widehat{\mathbf{F}}_t.$$

After estimating the factors and loadings, we apply our testing procedure to check for the remaining covariance structure in $\mathbf{U}_t$. The second-step estimation can be carried out also by dynamic factor models. In Section S.2 in the Supplementary Material, we discuss selecting the number of factors.

3. Now, define $\widehat{\mathbf{W}}_{i,t}$ be a nonempty subset of $(\widehat{\mathbf{U}}'_{-i,t}, \widehat{\mathbf{U}}'_{t-1}, \dots, \widehat{\mathbf{U}}'_{t-l_U})'$ where l_U is a nonnegative integer (lag). This includes contemporary regression for association studies and lagged regression for prediction of $U_{i,t}$ as two specific examples. The third step consists of a sparse regression to estimate the following model for each $i \in [n]$:

$$\widehat{U}_{i,t} = \boldsymbol{\theta}'_i \widehat{\mathbf{W}}_{i,t} + V_{i,t}; \quad t \in [T].$$

The regression in Step 3 provides useful augmentation for reducing the error in explaining $Y_{i,t}$ in (2.1) further from $U_{i,t}$ to $V_{i,t}$ and hence the prediction error for $Y_{i,t}$; see (2.5).

4. (For partial covariance analysis only) Estimate the following sparse regression model for each $i, j \in [n]$:

$$\widehat{U}_{i,t} = \boldsymbol{\theta}'_{i,j} \widehat{\mathbf{U}}_{-ij,t} + V_{i,j,t}; \quad t \in [T],$$

where $\widehat{\mathbf{U}}_{-ij,t}$ is the vector $\widehat{\mathbf{U}}_t$ without i th and j th elements. Let $\{\widehat{V}_{i,j,t}\}$ be the residuals. Then compute the partial correlation as the sample correlation of $\{(\widehat{V}_{i,j,t}, \widehat{V}_{j,i,t})\}_{t=1}^T$.

5. (For forecasting only) Define $\widehat{\mathbf{G}}_t := (\widehat{\mathbf{G}}'_{1,t}, \dots, \widehat{\mathbf{G}}'_{r,t})'$ where $\widehat{\mathbf{G}}_{j,t}$ is a nonempty subset of $(\widehat{\mathbf{F}}'_t, \widehat{\mathbf{F}}'_{t-1}, \dots, \widehat{\mathbf{F}}'_{t-l_F})'$ for each $j \in [r]$; and l_F is a nonnegative integer (lag).⁵ This step is multiple linear regression to estimate the following model for each $j \in [r]$:

$$\widehat{F}_{j,t} = \boldsymbol{\rho}'_j \widehat{\mathbf{G}}_{j,t} + V_{j,t}^F, \quad t \in [T].$$

This step aims at establishing a predictive model for latent factors. The estimator $\widehat{\mathbf{P}} := (\widehat{\boldsymbol{\rho}}_1 : \dots : \widehat{\boldsymbol{\rho}}_r)'$ will be used in the predictive model defined in (2.5). Our approach to modeling the factor dynamics is similar to the methodology described in [Stock and Watson \(2016\)](#) in the framework of dynamic factor models.

2.3. Predictive models. In a pure prediction exercise, one is usually interested in the linear projection of $Y_{i,t}$ onto $(\mathbf{X}'_{i,t}, \mathbf{G}'_t, \mathbf{W}'_{i,t})'$ motivated by the discussion in Section 1.1. This results in the factor-augmented regression model (FARM)

$$(2.4) \quad Y_{i,t} = \boldsymbol{\gamma}'_i \mathbf{X}_{i,t} + \boldsymbol{\lambda}'_i \mathbf{P} \mathbf{G}_t + \boldsymbol{\theta}'_i \mathbf{W}_{i,t} + V_{i,t}^Y; \quad i \in [n], t \in [T],$$

in which $\mathbf{P} \mathbf{G}_t$ predicts $\mathbf{F}_t$; see (2.1). This can be used for prediction, following the steps described in Section 2.2, by

$$(2.5) \quad \widehat{Y}_{i,t} := \widehat{\boldsymbol{\gamma}}'_i \mathbf{X}_{i,t} + \widehat{\boldsymbol{\lambda}}'_i \widehat{\mathbf{P}} \widehat{\mathbf{G}}_t + \widehat{\boldsymbol{\theta}}'_i \widehat{\mathbf{W}}_{i,t}; \quad i \in [n], t > T.$$

We call the prediction model `FarmPredict`.

⁴Another approach is to use the joint estimation as in [Agarwal, Negahban and Wainwright \(2012\)](#). The key difference between the two approaches is the optimization-based approach (fully-iterated) and the one-step approach and the different assumptions behind the two approaches. The PCA approach is based on the strong factor assumption with a large eigengap but does not impose the sparse structure on the idiosyncratic component covariance matrix. See [Fan, Liao and Mincheva \(2013\)](#) for additional discussion.

⁵We write at this generality to accommodate the cross-sectional applications in which no latent factor needs to be predicted as in the principal component regression. In this case, $\widehat{\mathbf{G}}_{j,t} = \widehat{\mathbf{F}}_{j,t}$ and this step is not needed.

In the case where it is possible to estimate the factors for the full sample as in [Fan, Masini and Medeiros \(2022\)](#), equation (2.5) becomes⁶

$$\widehat{Y}_{i,t} := \widehat{\gamma}_i' X_{i,t} + \widehat{\lambda}_i' \widehat{F}_t + \widehat{\theta}_i' \widehat{W}_{i,t}; \quad i \in [n], t > T.$$

Note that model (2.4) is equivalent to using the predictors $X_{i,t}$, $Y_{-i,t}$ and F_t , which augment predictors $X_{i,t}$, $Y_{-i,t}$ by using the common factors F_t . The form in (2.4) mitigates the collinearity issues in high dimensions. Model (2.4) also bridges factor regression ($\theta_i = \mathbf{0}$) on one end and (sparse) regression on the other end with $\lambda_i = \Lambda_{-i}' \theta_i$, where Λ_{-i} is the loading matrix without the i th row. In the latter, if we set $W_{i,t} = U_{-i,t}$ and $G_t = F_t$ (hence, $P = I_r$), model (2.4) becomes a (sparse) regression model:

$$Y_{i,t} = \gamma_i' X_{i,t} + \theta_i' R_{-i,t} + V_{i,t}^Y; \quad i \in [n], t \in [T].$$

In this case, (2.4) decorrelates the predictor $R_{-i,t}$, which makes the model selection consistency much easier to satisfy and forms the basis of `FarmSelect` in [Fan, Ke and Wang \(2020\)](#). Our contribution in this specific task is to allow heterogeneity adjustments, resulting in the estimated data R_t . In general, for model (2.4) with sparsity, `FarmPredict` chooses additional idiosyncratic components to enhance the prediction of the factor regression.

EXAMPLE 5 (First-Order FAVAR). Consider the case of model (2.4) where $X_{i,t} = 1$ and $l_U = l_F = 1$, and set $W_{i,t} = U_{t-1}$ and $G_t = F_{t-1}$. Therefore, we can write

$$\begin{aligned} (2.6) \quad Y_t &= (I - \Theta)\gamma + (\Lambda P - \Theta \Lambda)F_{t-1} + \Theta Y_{t-1} + V_t^Y, \\ &= \Theta_0 + \Theta_F F_{t-1} + \Theta Y_{t-1} + V_t^Y, \end{aligned}$$

where $Y_t = (Y_{1t}, \dots, Y_{nt})'$, $\gamma = (\gamma_1, \dots, \gamma_n)$, and $\Theta = (\theta_1 : \dots : \theta_n)'$. Model (2.6) is a first-order FAVAR model.

2.4. Covariance structure and inference. In several applications, the structure of the idiosyncratic components $U_t = (U_{1,t}, \dots, U_{n,t})'$ is the objective of interest. An estimator for $\Sigma := \mathbb{E}(U_t U_t')$ could be simply given by

$$(2.7) \quad \widehat{\Sigma} := \frac{1}{T} \sum_{t=1}^T \widehat{U}_t \widehat{U}_t'.$$

Estimating Σ is well documented in the literature, even in high dimensions; see, for example, [Fan et al. \(2020\)](#), [Ledoit and Wolf \(2021\)](#), and the references therein. Nevertheless, we show that (2.7) can be used within our testing framework.

To understand the (linear) relation between a pair $(U_{i,t}, U_{j,t})$ of U_t , a simple covariance estimate sometimes is not enough. In applications, it is often desirable to directly measure how $U_{i,t}$ and $U_{j,t}$ are connected. By direct connection, we mean the relation between those units removing the contribution of other variables of U_t . For this purpose, we use the partial covariance between $U_{i,t}$ and $U_{j,t}$, defined for any pair $i, j \in [n]$ as $\pi_{i,j} := \mathbb{E}(V_{i,j,t} V_{j,i,t})$, where $V_{i,j,t} := U_{i,t} - \text{Proj}(U_{i,t} | U_{-ij,t})$ and $\text{Proj}(U_{i,t} | U_{-ij,t})$ denotes the linear projection of $U_{i,t}$ onto the space spanned by all the units except i and j , which we denote by $U_{-ij,t}$. As in [Peng, Zhou and Zhu \(2009\)](#), we suggest to estimate the partial covariance matrix $\Pi := (\pi_{i,j})$ by

$$(2.8) \quad \widehat{\Pi} := (\widehat{\pi}_{i,j}) \quad \text{and} \quad \widehat{\pi}_{i,j} := \frac{1}{T} \sum_{t=1}^T \widehat{V}_{i,j,t} \widehat{V}_{j,i,t},$$

⁶See also Remark 3 and Example 4.

where $\widehat{V}_{i,j,t} := \widehat{U}_{i,t} - \widehat{\theta}'_{i,j} \widehat{U}_{-ij,t}$ is the residual of the LASSO regression of $\widehat{U}_{i,t}$ onto $\widehat{U}_{-ij,t}$ obtained in step 4 for $i, j \in [n]$.

We also would like to conduct formal tests on the population structure of U_t . Specifically, we propose a test for the following null hypothesis:

$$(2.9) \quad \mathcal{H}_{\mathcal{D}}^{\Sigma} : \Sigma_{\mathcal{D}} = \Sigma_{\mathcal{D}}^0; \quad \mathcal{D} \subseteq [n]^2,$$

for a given subset $\mathcal{D}$, where, for a given $(n \times n)$ matrix $\mathbf{M}$, the notation $\mathbf{M}_{\mathcal{D}}$ denotes the $d := |\mathcal{D}|$ -dimensional vector of $\text{vec}(\mathbf{M})$ indexed by the elements in $\mathcal{D} \subseteq [n]^2$. We allow d to diverge as $n, T \rightarrow \infty$. For testing the structure of $\Pi_{\mathcal{D}}$, consider the following:

$$(2.10) \quad \mathcal{H}_{\mathcal{D}}^{\Pi} : \Pi_{\mathcal{D}} = \Pi_{\mathcal{D}}^0; \quad \mathcal{D} \subseteq [n]^2.$$

The null hypotheses (2.9) and (2.10) nest cases of interest. The most common would be to test for a diagonal or a block diagonal structure in Σ and/or Π . But it also accommodates other structures.⁷ The challenges for testing (2.9) and (2.10) can be summarized as follows:

1. As we allow for both n and d to diverge as T grows, sometimes at a faster rate, we have a high-dimensional test where some Gaussian approximation result for dependent data must be deployed as we also allow the number covariances to be tested d to diverge. In this case, a high-dimensional long-run covariance matrix must be estimated if one expects to get the (asymptotic) correct test size.

2. We do not observe $\{U_t\}$ or $\{V_{i,j,t}\}$. Instead, we have an estimate of both from a postulated model on observable random variables. Therefore, the estimation error must be considered to claim some asymptotic properties of the test. It is not uncommon to obtain estimates of both $\{U_t\}$ and $\{V_{i,j,t}\}$ from a multistage estimation procedure, as we illustrate later.

We propose to test (2.9) using the statistic

$$(2.11) \quad S_{\mathcal{D}}^{\Sigma} := \|\sqrt{T}(\widehat{\Sigma}_{\mathcal{D}} - \Sigma_{\mathcal{D}}^0)\|_{\max}.$$

A Gaussian bootstrap approximates the quantiles of $S_{\mathcal{D}}^{\Sigma}$. To describe the procedure, let Υ_{Σ} denote the $(d \times d)$ covariance matrix for the vectorized submatrix $(\widetilde{\sigma}_{i,j})_{(i,j) \in \mathcal{D}}$, where $\widetilde{\sigma}_{i,j} := \frac{1}{T} \sum_{t=1}^T U_{i,t} U_{j,t}$. Since the process $\{U_t\}$ might present some form of temporal dependence (refer to Assumption 2(c)), we estimate Υ_{Σ} using a Newey–West-type estimator. For a given $\mathcal{K} \in \mathbb{K}$, where $\mathbb{K}$ is a class of kernel functions described below in (3.6) and bandwidth $h > 0$, Υ_{Σ} is estimated by

$$(2.12) \quad \widehat{\Upsilon}_{\Sigma} := \sum_{|b| < T} \mathcal{K}(b/h) \widehat{\mathbf{M}}_{\Sigma,b} \quad \text{and} \quad \widehat{\mathbf{M}}_{\Sigma,b} := \frac{1}{T} \sum_{t=b+1}^T \widehat{\mathbf{D}}_{\Sigma,t} \widehat{\mathbf{D}}'_{\Sigma,t-b},$$

where $\widehat{\mathbf{D}}_{\Sigma,t}$ is a d -dimensional vector with entries given by $\widehat{U}_{i,t} \widehat{U}_{j,t} - \widehat{\sigma}_{i,j}$ for $(i, j) \in \mathcal{D}$, where $\widehat{\sigma}_{i,j}$ is the (i, j) element of $\widehat{\Sigma}$ defined in (2.7). Finally, let $c_{\Sigma}^*(\tau)$ be the conditional τ -quantile of the Gaussian bootstrap $S_{\mathcal{D}}^* := \|\mathbf{Z}_{\Sigma}^*\|_{\infty}$ where $\mathbf{Z}_{\Sigma}^* | \mathbf{X}, \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \widehat{\Upsilon}_{\Sigma})$, that is,

$$c_{\Sigma}^*(\tau) := \inf\{q \in \mathbb{R} : \mathbb{P}(S_{\mathcal{D}}^* \leq q | \mathbf{X}, \mathbf{Y}) \geq \tau\}$$

Theorem 5 demonstrates the validity of the Gaussian bootstrap procedure described above, that is, it states conditions under which the τ -quantile of the test statistic (2.11) can be approximated by $c_{\Sigma}^*(\tau)$ in the appropriate sense.

⁷With minor changes, the proposed test can also be used to test the null $\mathbf{M} \text{vec}(\Sigma) = \mathbf{m}$ for some $d \times n^2$ matrix $\mathbf{M}$ and d -dimensional vector $\mathbf{m}$ where $d := d_T$ is also a function of T .

Similarly, the test statistic for (2.10) is given by

$$(2.13) \quad S_D^\Pi := \|\sqrt{T}(\widehat{\Pi}_D - \Pi_D^0)\|_{\max}.$$

Let Υ_Π denote the $(d \times d)$ covariance matrix of $(\tilde{\pi}_{i,j})_{(i,j) \in \mathcal{D}}$ where $\tilde{\pi}_{i,j} := \frac{1}{T} \sum_{t=1}^T V_{i,j,t} \times V_{j,i,t}$. Υ_Π is estimated by

$$(2.14) \quad \widehat{\Upsilon}_\Pi := \sum_{|b| < T} \mathcal{K}(b/h) \widehat{M}_{\Pi,b}; \quad \widehat{M}_{\Pi,b} := \frac{1}{T} \sum_{t=b+1}^T \widehat{D}_{\Pi,t} \widehat{D}'_{\Pi,t-b},$$

where $\widehat{D}_{\Pi,t}$ is a d -dimensional vector with entries given by $\widehat{V}_{i,j,t} \widehat{V}_{j,i,t} - \widehat{\pi}_{i,j}$ for $(i, j) \in \mathcal{D}$. Also, let $c_\Pi^*(\tau)$ be the conditional τ -quantile of the Gaussian bootstrap $S_D^* := \|\mathbf{Z}_\Pi^*\|_\infty$ where $\mathbf{Z}_\Pi^* | \mathbf{X}, \mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \widehat{\Upsilon}_\Pi)$. Theorem 6 establish conditions for the validity of Gaussian bootstrap to approximate the quantiles of (2.13).

3. Theoretical results. In this section, we collect all the theoretical guarantees for estimating the model (2.1) by using the proposed multistage method described above. Specifically, Section 3.1 presents nonasymptotic bounds for the (parametric) estimation, Section 3.2 deals with similar results concerning forecasting, and Section 3.3 deals with inference on the (partial) covariance structure of $\mathbf{U}_t$. Here we focus on the case where the random quantities admit polynomial tails. Similar results for the case of geometric mixing and exponential tails can be found in the Supplementary Material, Section S.4.

To present the results, it is convenient to use a compact notation. For each $i \in [n]$, we define the T -dimensional vectors $\mathbf{Y}_{i,\cdot} := (Y_{i,1}, \dots, Y_{i,T})'$ and $\mathbf{U}_{i,\cdot} := (U_{i,1}, \dots, U_{i,T})'$. We also define the $(T \times k)$ matrix of covariates $\mathbf{X}_{i,\cdot} := (\mathbf{X}_{i,1}, \dots, \mathbf{X}_{i,T})'$, for each $i \in [n]$ and the $(T \times r)$ matrix of factors $\mathbf{F} := (\mathbf{F}_1, \dots, \mathbf{F}_T)'$, such that (2.1) can be represented as

$$(3.1) \quad \begin{aligned} \mathbf{Y}_{i,\cdot} &= \mathbf{X}_{i,\cdot} \boldsymbol{\gamma}_i + \mathbf{F} \boldsymbol{\lambda}_i + \mathbf{U}_{i,\cdot}, \quad i = 1, 2, \dots, n, \\ &= \mathbf{X}_{i,\cdot} \boldsymbol{\gamma}_i + \mathbf{R}_{i,\cdot}, \end{aligned}$$

for each cross-sectional unit i , where $\mathbf{R}_{i,\cdot} := \mathbf{F} \boldsymbol{\lambda}_i + \mathbf{U}_{i,\cdot}$.

We define for each $t \in [T]$, the n -dimensional vectors $\mathbf{Y}_t := (Y_{1,t}, \dots, Y_{n,t})'$ and $\mathbf{U}_t := (U_{1,t}, \dots, U_{n,t})'$; and the nk -dimensional vector $\mathbf{X}_t := (\mathbf{X}'_{1,t}, \dots, \mathbf{X}'_{n,t})'$. Also, set the $(n \times nk)$ block diagonal matrix $\boldsymbol{\Gamma}$ whose block diagonal is given by $(\boldsymbol{\gamma}'_1, \dots, \boldsymbol{\gamma}'_n)$ and the $(n \times r)$ loading matrix $\boldsymbol{\Lambda} := (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n)'$. Then, (2.1) can also be represented as panel time series

$$(3.2) \quad \begin{aligned} \mathbf{Y}_t &= \boldsymbol{\Gamma} \mathbf{X}_t + \boldsymbol{\Lambda} \mathbf{F}_t + \mathbf{U}_t, \quad t = 1, 2, \dots, T \\ &= \boldsymbol{\Gamma} \mathbf{X}_t + \mathbf{R}_t, \end{aligned}$$

where $\mathbf{R}_t := \boldsymbol{\Lambda} \mathbf{F}_t + \mathbf{U}_t$.

3.1. Estimation. We start by stating the following assumption.

ASSUMPTION 2 (Moments and Dependency). Consider the following:

- (a) The stochastic process $\{\mathbf{Z}_t := (\mathbf{X}'_{S,t}, \mathbf{F}'_t, \mathbf{U}'_t) : t \in [T]\}$ is weakly stationary for each $T \in \mathbb{N}$, where $\mathbf{X}_{S,t}$ denotes the vector $\mathbf{X}_t$ after excluding all deterministic (nonrandom) components. Furthermore, the strong mixing coefficient of $\mathbf{Z}_t$ is denoted by α_m .
- (b) Define $\mathbf{U}_t := (\mathbf{U}'_t, \mathbf{U}'_{t-1}, \dots, \mathbf{U}'_{t-l})'$ for some integer $l \geq 0$ and let $b > 0$ be a finite constant such that $\lambda_{\min}[\mathbb{E}(\mathbf{U}_t \mathbf{U}_t')] \geq b^2$, where $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of $(\cdot)$.
- (c) Assume there exists an universal constant $C > 0$ such for all $t, s \in [T]$, $T \geq 2$ and $i \in [n]$:

$$(c.1) \quad \|\mathbf{Z}_t\|_{p+\epsilon} \leq C \text{ for some constants } p \geq 8 \text{ and } \epsilon > 0$$

$$(c.2) \quad \|n^{-1/2}[\mathbf{U}'_s \mathbf{U}_t - \mathbb{E}(\mathbf{U}'_s \mathbf{U}_t)]\|_p \leq C$$

$$(c.3) \quad \|n^{-1/2} \sum_{i=1}^n \lambda_{j,i} U_{i,t}\|_p \leq C$$

$$(c.4) \quad \|\|(\mathbf{X}'_i \mathbf{X}_i / T)^{-1}\|\|_p \leq C.$$

Assumption (2.a) excludes the deterministic components of $\mathbf{X}_t$ to accommodate possibly nonrandom nonstationary, but uniformly bounded covariates as in Assumption (2.c). Assumption (2.b) ensures that the parameter in (2.2) is well defined since, by the Cauchy interlacing theorem, we have $\inf_i \lambda_{\min}[\mathbb{E}(\mathbf{W}_{i,t} \mathbf{W}_{i,t}')] \geq \lambda_{\min}[\mathbb{E}(\mathbf{U}_t \mathbf{U}_t')]$. Finally, Assumptions (2.a) and (2.c) allow us to apply a Marcinkiewicz–Zygmund-type inequality for partial sums to deal with polynomial tails (see Lemma S.2 in the Supplementary Material).

For each $i \in [n]$, let $\mathbf{R}_i := \mathbf{F} \lambda_i + \mathbf{U}_i$ denote the unobservable error term in (3.1), $\hat{\mathbf{y}}_i$ the least-squares estimator of $\mathbf{y}_i$ and $\hat{\mathbf{R}}_i := \mathbf{Y}_t - \mathbf{X}_i \hat{\mathbf{y}}_i$ the vector of residuals. Also, set $\hat{\mathbf{R}} := (\hat{\mathbf{R}}_1, \dots, \hat{\mathbf{R}}_n)'$ and $\mathbf{R} := (\mathbf{R}_1, \dots, \mathbf{R}_n)'$. We must control for estimation error in the first step of the proposed methodology. The next result gives a bound for the maximum entry of the $(n \times T)$ matrix $\hat{\mathbf{R}} - \mathbf{R}$ when the first stage is conducted by OLS in a linear setup. We assume that $\mathbf{X}_{i,t}$, $\mathbf{F}_t$, and $\mathbf{U}_{i,t}$ are mutually uncorrelated.

We state the results below in terms of the strong mixing coefficient sequence, whose definition is presented here for convenience. For $m \in \{0, \dots, T-1\}$, define

$$(3.3) \quad \alpha_m := \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{Z}_t^t, B \in \mathcal{Z}_{t+m}^T, t \in [T]\},$$

where $\mathcal{Z}_s^t$ is the σ -algebra generated by $(\mathbf{Z}_s, \dots, \mathbf{Z}_t)$ for $1 \leq s \leq t \leq T$. Note that α_m might depend on both T and n .

THEOREM 1. *Under Assumptions 1 and 2,*

$$\|\hat{\mathbf{R}} - \mathbf{R}\|_{\max} \lesssim_{\mathbb{P}} \frac{\mathcal{R}_\alpha \sqrt{rk}^{1+\frac{3}{p}} n^{4/p}}{T^{1/2-1/p}},$$

where $\mathcal{R}_\alpha := \mathcal{R}_\alpha(T, n) := [\sum_{m=0}^{T-1} (m+1)^{(p/2)-2} \alpha_m^{1-p/(p+\epsilon)}]^{2/p}$ and α_m is defined by (3.3).

REMARK 4. Even though we will treat the number of factors, r , and the number of covariates, k , fixed (not depending on T or n), we kept them explicit in the result above. Furthermore, if we assume that $\alpha_m \leq K(m+1)^{-c}$ for all m , where K is a constant that might depend on n and $c \geq 0$ is an universal constant, then

$$\mathcal{R}_\alpha(T, n) \lesssim K^{\frac{2\epsilon}{p(p+\epsilon)}} \times \begin{cases} 1 & c > \frac{(p-2)(p+\epsilon)}{2\epsilon}, \\ (\log T)^{\frac{2}{p}} & c = \frac{(p-2)(p+\epsilon)}{2\epsilon}, \\ T^{\frac{2}{p}[\frac{(p-2)(p+\epsilon)}{2\epsilon}+1]} & c < \frac{(p-2)(p+\epsilon)}{2\epsilon}. \end{cases}$$

REMARK 5. When the first step of the method involves more complicated estimation, such as Pesaran's (2006) method, instrumental variables, or LASSO, we write $\|\hat{\mathbf{R}} - \mathbf{R}\|_{\max} \lesssim_{\mathbb{P}} \varrho_R$, where ϱ_R is a nonnegative sequence of n and T . This approach is adopted systematically in the following theorems.

Define the $(n \times T)$ matrices $\mathbf{Y} := (\mathbf{Y}_1, \dots, \mathbf{Y}_T)$ and $\mathbf{U} := (\mathbf{U}_1, \dots, \mathbf{U}_T)$; and the $(nk \times T)$ matrix $\mathbf{X} := (\mathbf{X}_1, \dots, \mathbf{X}_T)$. We can write (2.1) in the matrix form as

$$(3.4) \quad \mathbf{Y} = \mathbf{\Gamma} \mathbf{X} + \mathbf{\Lambda} \mathbf{F}' + \mathbf{U}.$$

Notice that $\widehat{\mathbf{R}} = \mathbf{\Lambda} \mathbf{F}' + \widetilde{\mathbf{U}}$ where $\widetilde{\mathbf{U}} := \mathbf{U} + \widehat{\mathbf{R}} - \mathbf{R}$ and $(\mathbf{\Lambda}, \mathbf{F})$ can be estimated by Principal Component Analysis (PCA), which minimizes $\|\widehat{\mathbf{R}} - \mathbf{\Lambda} \mathbf{F}'\|_F^2$ with respect to $\mathbf{\Lambda}$ and $\mathbf{F}$, subject to $\mathbf{F}' \mathbf{F} / T = \mathbf{I}_r$. The solution $\widehat{\mathbf{F}}$ is the matrix whose columns are $\sqrt{T}$ times r eigenvectors of the top r eigenvalues of $\widehat{\mathbf{R}}' \widehat{\mathbf{R}}$ and $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{R}} \widehat{\mathbf{F}} / T$.

Since we do not observe $\mathbf{U}$, in the third step of the method we use $\widehat{\mathbf{U}} := \widehat{\mathbf{R}} - \widehat{\mathbf{\Lambda}} \widehat{\mathbf{F}}'$ instead. Therefore, we must control for the estimation error in the previous steps: $\widehat{\mathbf{U}} - \mathbf{U}$. Also, the loading matrix $\mathbf{\Lambda}$ and the factors $\mathbf{F}$ are not separably identified since $\mathbf{\Lambda} \mathbf{F}_t = \mathbf{\Lambda} \mathbf{H}' \mathbf{H} \mathbf{F}_t$ for any matrix $\mathbf{H}$ such that $\mathbf{H}' \mathbf{H} = \mathbf{I}_r$. If we let $\mathbf{H} := T^{-1} \mathbf{V}^{-1} \widehat{\mathbf{F}}' \mathbf{F} \mathbf{\Lambda}' \mathbf{\Lambda}$, where $\mathbf{V}$ is the $(r \times r)$ diagonal matrix containing the r largest eigenvalues of $\widehat{\mathbf{R}}' \widehat{\mathbf{R}} / T$ in decreasing order, we have that $\mathbf{H} \mathbf{F}_t$ is identified as $\mathbf{\Lambda} \mathbf{F}_t$ is identified.⁸

The result below first appeared in [Bai \(2003\)](#) for the case of (n, T) diverging and was further extended to hold uniformly in $(i \leq n, t \leq T)$ by [Fan, Liao and Mincheva \(2013\)](#). However, both consider the case when the factor model is estimated using the actual data instead of the “estimated” ones (residuals) as in our case. Therefore, the next result is a generalization that considers the pre-estimation error term and quantifies how the error impacts the precision of factor analysis.

ASSUMPTION 3 (Factor Model). Assume:

- (a) $\mathbb{E}(\mathbf{F}_t) = \mathbf{0}$, $\mathbb{E}(\mathbf{F}_t \mathbf{F}_t') = \mathbf{I}_r$, and $\mathbf{\Lambda}' \mathbf{\Lambda}$ is a diagonal matrix;
- (b) All eigenvalues of $\mathbf{\Lambda}' \mathbf{\Lambda} / n$ are bounded away from zero and infinity as $n \rightarrow \infty$;
- (c) $\|\mathbf{\Sigma} - \mathbf{\Lambda} \mathbf{\Lambda}'\| \lesssim 1$; and
- (d) $\|\mathbf{\Lambda}\|_{\max} \lesssim 1$.

REMARK 6. Assumption 3 is standard in the literature. Assumption $\mathbb{E}(\mathbf{F}_t) = \mathbf{0}$ is not restrictive as our approach considers a first-step estimation which may include a constant in the set of regressors. Assumption (3.a) is also needed for identifiability of the factor structure. Assumption (3.c) imposes a strong factor structure.

THEOREM 2. Under Assumptions 1–3, let q_R be a nonnegative sequence of n and T such that $\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max} \lesssim_{\mathbb{P}} q_R$. Then:

- (a) $\max_{t \leq T} \|\widehat{\mathbf{F}}_t - \mathbf{H} \mathbf{F}_t\|_2 \lesssim_{\mathbb{P}} \frac{1}{\sqrt{T}} + \frac{T^{1/p}}{\sqrt{n}} + q_R (nT)^{1/p}$,
- (b) $\max_{i \leq n} \|\widehat{\boldsymbol{\lambda}}_i - \mathbf{H} \boldsymbol{\lambda}_i\|_2 \lesssim_{\mathbb{P}} \frac{\mathcal{R}_\alpha n^{2/p}}{\sqrt{T}} + \frac{1}{\sqrt{n}} + q_R$, and
- (c) $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \lesssim_{\mathbb{P}} \frac{\mathcal{R}_\alpha n^{2/p}}{T^{1/2-1/p}} + \frac{T^{1/p}}{\sqrt{n}} + q_R (nT)^{1/p}$,

provided that $\mathcal{R}_\alpha n^{4/p} / \sqrt{T} + (nT)^{1/p} q_R \lesssim 1$, where $\mathcal{R}_\alpha$ is defined in Theorem 1.

By setting $q_R = 0$, we have the case of no estimation error in the first step. Note that in order to have the error $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max}$ vanishing in probability we must have the pre-estimation error $\|\widehat{\mathbf{R}} - \mathbf{R}\|_{\max}$ of order (in probability) smaller than $(nT)^{-1/p}$.

We have decided not to replace q_R in Theorem 2 with the rate obtained in Theorem 1 as the latter only applies to the least-squares estimator. In some applications, however, the first step of the procedure could be done using a different type of estimator. For instance, a penalized adaptive Huber regression as in [Fan, Li and Wang \(2017\)](#) if the number of features k is comparable or even larger than T and the tail of the distribution of $\mathbf{X}_t$ is heavy.

⁸The definition of $\mathbf{H}$ is the same one appearing in Appendix A of [Bai \(2003\)](#), where $\mathbf{V}$ is defined as the eigenvalues of $\widehat{\mathbf{R}}' \widehat{\mathbf{R}} / nT$.

REMARK 7. The result in Theorem 4 of Fan, Liao and Mincheva (2013), concerning the convergence rate of $\max_{t \leq T} \|\hat{\mathbf{F}}_t - \mathbf{H}\mathbf{F}_t\|_2$, would be recovered by setting $p = 4$ and $\rho_R = 0$ in our Theorem 2(a). However, Assumption 2(c.1) requires $p \geq 8$ to accommodate the case when ρ_R is not zero, and the probability bounds appear in the proof of Theorems 5 and 6. Theorem S.8(a) in the Supplementary Material (with $\rho_R = 0$) differs from Theorem 4 in Fan, Liao and Mincheva (2013). The former shows a better dependency on T in the second term. That results from a stronger condition imposed on the units' dependence (Assumption 4 replacing Assumption 2). In this sense, both results of Theorem 4 of Fan, Liao and Mincheva (2013) are proved under a mix of our Assumptions 2 and 4, that is, exponential tails in the marginals, exponential decay in the mixing coefficient, and polynomial tails with $p = 4$ in Assumption 2(c.2).

REMARK 8. The contrast between Proposition 2 in Bai (2003) and Theorem 2(a) is less straightforward. Both theorems share similar underlying assumptions but are not identical, making a direct comparison cumbersome. For instance, Assumption C.1 in Bai (2003) required a finite eighth moment, and their Assumption C.5, a finite fourth moment. Under these assumptions (and $\rho_R = 0$) we could achieve $T^{-1/2} + T^{1/4}/\sqrt{n}$ in Theorem 2(a) which is an improvement over Proposition 2 in Bai (2003).

By stating Theorem 2 in terms of a generic rate, it is easier to account for the effect of different estimators. By combining Theorems 1 and 2, we have the following corollary.

COROLLARY 1. *Under the same assumptions of Theorems 1 and 2, when OLS is used in the first-stage to obtain $\hat{\mathbf{R}}$, we have*

$$\|\hat{\mathbf{U}} - \mathbf{U}\|_{\max} \lesssim_{\mathbb{P}} \frac{\mathcal{R}_{\alpha} n^{5/p}}{T^{1/2-2/p}} + \frac{T^{1/p}}{\sqrt{n}} =: \varpi_U.$$

We propose estimating (2.3) by LASSO using $\hat{\mathbf{U}}$ instead of $\mathbf{U}$. Specifically, for a regularization parameter $\xi > 0$, we denote by $\hat{\boldsymbol{\theta}}_i$ a minimizer of $\mathcal{Q}_i$ given by

$$(3.5) \quad \mathcal{Q}_i(\mathbf{a}) := \frac{1}{T} \sum_{t \in [T]} (\hat{U}_{i,t} - \mathbf{a}' \hat{\mathbf{W}}_{i,t})^2 + \xi \|\mathbf{a}\|_1; \quad i \in [n].$$

The next result presents nonasymptotic bounds for the (in sample) prediction error and the ℓ_1 -estimation error for the LASSO estimator (3.5) based on the “estimated data” and quantifies how the estimation errors impact on the choice of regularization parameter and rates of convergence.

THEOREM 3. *Let q_U be a nonnegative sequence of n and T such that $\|\hat{\mathbf{U}} - \mathbf{U}\|_{\max} \lesssim_{\mathbb{P}} q_U$ and assume that Assumption 2 holds. For every $\epsilon > 0$ there is a constant $0 < C_{\epsilon} < \infty$ such that if the penalty parameter is set $\xi \geq C_{\epsilon} \xi_0$, then for any minimizer $\hat{\boldsymbol{\theta}}_i$ of (3.5), with probability at least $1 - \epsilon$:*

$$\max_{i \in [n]} [(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)' \mathbb{E}(\mathbf{W}_{i,t} \mathbf{W}_{i,t}') (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \xi \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1] \leq 8 \frac{\xi^2 s_0}{b^2},$$

provided that $\frac{\xi_1 s_0}{b} \leq K_{\epsilon}$ where $s_k := \max_{i \in [n]} \|\boldsymbol{\theta}_i\|_k$ for $k \in \{0, 1, 2\}$, K_{ϵ} is a positive constant only depending on ϵ , and

$$\xi_0 := (1 + s_2) \mathcal{L}_{\alpha} [n(l + 1)]^{2/p} T^{-1/2} + (1 + s_1) [(nT)^{1/p} q_U + q_U^2],$$

$$\xi_1 := \mathcal{L}_{\alpha} [n(l + 1)]^{4/p} T^{-1/2} + [(nT)^{1/p} q_U + q_U^2]$$

with $\mathcal{L}_{\alpha} := \mathcal{L}_{\alpha}(T, n, l) := [\sum_{t=0}^{T-1} (m+1)^{(p/2)-2} \alpha_{(m-l) \vee 0}^{1-p/(p+\epsilon)}]^{2/p}$.

Once again, we purposely avoided replacing ϱ_U in Theorem 3 with the rate of Corollary 1 to make it readily applicable to the case when a different type of factor model was used or, as a matter of fact, any other pre-estimation procedure. We have the next result by setting ϱ_U equal to ϖ_U , the rate of Corollary 1.

COROLLARY 2. *If ϱ_U defined in Theorem 3 is taken to be rate given by Corollary 1 and $l < T$, then under the conditions of Theorem 3:*

$$\max_{i \in [n]} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \lesssim_{\mathbb{P}} \frac{s_0(1+s_1)}{b^2} \left[\frac{\mathcal{L}_\alpha n^{6/p}}{T^{(1/2)-(3/p)}} + \frac{T^{2/p}}{n^{(1/2)-(1/p)}} \right] =: \varpi_\theta.$$

3.2. Forecasting. Recall that in the context of out-of-sample forecasting, our object of interest is $\widetilde{Y}_{i,t} := \boldsymbol{\gamma}_i' \mathbf{X}_{i,t} + \boldsymbol{\lambda}_i' \mathbf{P} \mathbf{G}_t + \boldsymbol{\theta}_i' \mathbf{W}_{i,t}$ for $i \in [n]$ and $t \geq T$, which we estimate using $\widehat{Y}_{i,t} := \widehat{\boldsymbol{\gamma}}_i' \mathbf{X}_{i,t} + \widehat{\boldsymbol{\lambda}}_i' \widehat{\mathbf{P}} \mathbf{G}_t + \widehat{\boldsymbol{\theta}}_i' \widehat{\mathbf{W}}_{i,t}$. The next result bounds (in probability) the prediction error bound in terms of all estimation errors from previous steps.

THEOREM 4. *Under Assumptions 1 and 2, let $\varrho_\gamma, \varrho_U, \varrho_\theta, \varrho_P, \varrho_\lambda$ and ϱ_F be nonnegative sequence on n and T such that, uniformly in $i \in [n]$, $\|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_1 \lesssim_{\mathbb{P}} \varrho_\gamma$, $\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \lesssim_{\mathbb{P}} \varrho_U$, $\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \lesssim_{\mathbb{P}} \varrho_\theta$, $\|\widehat{\mathbf{P}} - \mathbf{P}\|_2 \lesssim_{\mathbb{P}} \varrho_P$, $\|\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{\lambda}_i\|_2 \lesssim_{\mathbb{P}} \varrho_\lambda$, and $\|\widehat{\mathbf{F}}_t - \mathbf{F}_t\|_2 \lesssim_{\mathbb{P}} \varrho_F$, respectively. Then, for every $t \geq 1$,*

$$\max_{i \in [n]} |\widehat{Y}_{i,t} - \widetilde{Y}_{i,t}| \lesssim_{\mathbb{P}} (\varrho_\gamma + \varrho_\theta) n^{1/p} + \varrho_U s_1 + \varrho_P + \varrho_\lambda + \varrho_F,$$

where s_1 is defined in Theorem 3.

REMARK 9. If $\varrho_\gamma, \varrho_U, \varrho_\theta, \varrho_\lambda$ and ϱ_F are taken to be rates appearing in Theorems 1–3, we have $\varrho_\gamma \vee \varrho_\lambda \vee \varrho_F \lesssim \varrho_U$ and $\varrho_U \lesssim \varrho_\theta$. If further, $\mathbf{P}$ is estimated via OLS (for each $j \in [r]$) using $\{\widehat{\mathbf{F}}_t, t \in [T]\}$, we have $\varrho_P \lesssim T^{-1/2} \vee \varrho_F \lesssim \varrho_F$. Therefore, Theorem 4 reduces to

$$\max_{i \in [n]} |\widehat{Y}_{i,t} - \widetilde{Y}_{i,t}| \lesssim_{\mathbb{P}} \varpi_\theta (n^{1/p} + s_1),$$

where ϖ_θ is the rate in Corollary 2.

3.3. Inference on covariance and partial covariance matrices. We now obtain the null distributions of our test statistics for the structures of the covariance and the partial covariance. Recall the setup and notation of Section 2.4. Further define $\widetilde{\boldsymbol{\Sigma}} := T^{-1} \sum_{t=1}^T \mathbf{U}_t \mathbf{U}_t'$ and $\widetilde{\boldsymbol{\Sigma}}_{\mathcal{D}}$ as the element of $\widetilde{\boldsymbol{\Sigma}}$ indexed by $\mathcal{D} \subseteq [n]^2$. Also, for any pair of random vectors $\mathbf{X}, \mathbf{Y}$ of the same dimensional, d , say; define the distance in distribution

$$\rho(\mathbf{X}, \mathbf{Y}) := \sup_{A \in \mathcal{R}} |\mathbb{P}(\mathbf{X} \in A) - \mathbb{P}(\mathbf{Y} \in A)|,$$

where $\mathcal{R}$ is the class of all rectangles in the form $\times_{j=1}^d (a_j, b_j]$ for some $-\infty \leq a_j \leq b_j \leq \infty$ and $j \in [d]$.

We assume that the kernel $\mathcal{K}(\cdot)$ appearing in the covariance estimator defined by (2.12) belongs to the class defined in Andrews (1991) which we reproduce below for convenience

$$(3.6) \quad \mathbb{K} := \left\{ f : \mathbb{R} \rightarrow [-1, 1] : f(0) = 1, f(x) = f(-x), \forall x \in \mathbb{R}, \right. \\ \left. \int f^2(x) dx < \infty, f \text{ is continuous} \right\}.$$

This includes most of the well-known kernels used in the literature. To avoid confusion, it is worth pointing out that our tuning parameter h , also called bandwidth parameter by Andrews (1991), is supposed to diverge, as opposed to the bandwidth in the density kernel estimation setup, which is expected to shrink to zero.

The following result shows how accurately the covariance matrix elements are estimated and validate the bootstrap method.

THEOREM 5. For $\mathcal{D} \subseteq [n]^2$, let $\tilde{\mathbf{J}} := \sqrt{T}(\tilde{\Sigma}_{\mathcal{D}} - \Sigma_{\mathcal{D}})$ and $\mathcal{G}$ be a zero-mean Gaussian vector with the same covariance matrix of $\tilde{\mathbf{J}}$, that is, $\mathcal{G} \sim N(\mathbf{0}, \Upsilon_{\Sigma})$. Under Assumptions 1–3, if further:

- (a) $\{\mathbf{U}_t : t \in [T]\}$ is fourth-order stationary process for each T ;
- (b) The strong mixing coefficient of $\{\mathbf{U}_t : t \in [T]\}$, α_m , obeys $\alpha_m \leq Km^{-r}$ for $r > (\frac{p+\epsilon}{\epsilon})(\frac{p}{4} - 1) \vee \frac{2p}{p-4}$ where $p \geq 8$ and $\epsilon > 0$ are defined in Assumption 2; and a constant K that might depend on d ;
- (c) the minimum eigenvalue of Υ_{Σ} is greater or equal to $\underline{c}$, for some $\underline{c} > 0$, then

$$\begin{aligned} \rho(\tilde{\mathbf{J}}, \mathcal{G}) &\lesssim \frac{\log T}{\underline{c}^2} \left[\frac{\log d}{T^{(1/2)-\kappa}} + \frac{K^{1-4/p} \log d}{T^{(r/2)(1-4/p)-1}} + \frac{(\log d)^{3/2}}{T^{1/4}} + \frac{d^{4/p} (\log d)^2}{T^{1/2-2/p}} \right] \\ &\quad + \frac{[d(\log d)^{(3/4)p-4} \log T \log(dT)]^{\frac{2}{p-4}}}{T^{1/4} \underline{c}^{p/(p-4)}} + \frac{d^{2/(p+2)} \sqrt{\log(dT)} + K}{T^{rk-1/2}}, \end{aligned}$$

where $d := |\mathcal{D}|$, $\kappa := \kappa(p, r) := \frac{1+D_p/2}{2(r+D_p/2)} \wedge 1/2$ and $D_p := \frac{p}{p+2}$.

Let $\hat{\mathbf{J}} := \sqrt{T}(\hat{\Sigma}_{\mathcal{D}} - \Sigma_{\mathcal{D}})$, then

$$\rho(\hat{\mathbf{J}}, \mathcal{G}) \lesssim \rho(\tilde{\mathbf{J}}, \mathcal{G}) + \inf_{\delta > 0} [\delta_1 \sqrt{1 \vee \log(d/\delta)} + \mathbb{P}(\|\hat{\mathbf{J}} - \tilde{\mathbf{J}}\|_{\infty} > \delta)].$$

Let $\tilde{\Upsilon}_{\Sigma}$ be any positive semi definite estimator of Υ_{Σ} and $\mathcal{G}^* | X, Y \sim N(\mathbf{0}, \tilde{\Upsilon}_{\Sigma})$, then

$$\rho(\hat{\mathbf{J}}, \mathcal{G}^*) \lesssim \rho(\hat{\mathbf{J}}, \mathcal{G}) + \inf_{\delta > 0} [\delta \log d (1 \vee |\log d|) + \mathbb{P}(\|\tilde{\Upsilon}_{\Sigma} - \Upsilon_{\Sigma}\|_{\max} > \delta)].$$

REMARK 10. The first result in Theorem 5 bounds the Komolgorov distance between the (unobservable) process $\{\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{U}_t \mathbf{U}_t' - \mathbb{E}(\mathbf{U}_t \mathbf{U}_t')\}_{T \geq 1}$ and a Gaussian process with the same covariance structure. It is a direct consequence of a more general central limit theorem result for high-dimensional alpha mixing sequences (see Theorem S.13 in the Supplementary Material). The second one is similar but controls for the difference between $\mathbf{U}_t - \hat{\mathbf{U}}_t$ and, therefore, considers the estimation error. Finally, the last result ensures a bootstrap validity provided we can estimate the covariance matrix appropriately.

REMARK 11. Theorem 5 seems complicated. However, the results only depend on d , T , p , r , and the “quality” of the estimators $\hat{\mathbf{U}}$ and $\hat{\Upsilon}$. The latter allows a different selection of estimators for any stage and the bootstrap procedure. If we were to specialize Theorem 5 to incorporate the rates obtained in Theorems 1 and 2, and set $\hat{\Upsilon} = \hat{\Upsilon}$ defined by (2.12), we obtain a sufficient condition to ensure the bootstrap validity depending only on n , T , r and p .

COROLLARY 3. Under the same conditions of Theorem 5, if $\tilde{\Upsilon}_{\Sigma} = \hat{\Upsilon}_{\Sigma}$ defined by (2.12) with $\mathcal{K} \in \mathbb{K}$, where $\mathcal{K}$ is defined by (3.6), then, uniformly in $\mathcal{D} \subseteq [n]^2$,

$$\|\hat{\Upsilon}_{\Sigma} - \Upsilon_{\Sigma}\|_{\max} \lesssim_{\mathbb{P}} h [\varpi_U(nT)^{3/p} + n^{8/p}/\sqrt{T}],$$

where $h > 0$ is the bandwidth parameter of the covariance estimator and ϖ_U is the rate appearing in Corollary 1. If further, as $h, n, T \rightarrow \infty$:

- (a) $\varrho_\Sigma = o(1)$, where ϱ_Σ is the rate in the first result of Theorem 5 with d replaced by n^2 ;
- (b) $(\log n)^{3/2}(\sqrt{T}\varpi_U^2 + \frac{n^{9/p}}{\sqrt{T}} + \frac{n^{6/p}}{T^{1/2-1/p}} + \frac{1}{n^{1/2-9/p}}) = o(1)$;
- (c) $(\log n)^3 h[\varpi_U(nT)^{3/p} + d^{8/p}/\sqrt{T}] = o(1)$, then,

$$\sup_{\mathcal{D}} \sup_{\tau \in (0,1)} |\mathbb{P}[S_{\mathcal{D}}^{\Sigma} \leq c_{\Sigma}^*(\tau)] - \tau| = o(1),$$

where the first supremum is over all null hypotheses of the form (2.9) indexed by $\mathcal{D} \subseteq [n]^2$.

The next theorem shows how well the elements of the partial autocovariance matrix are estimated and gives the conditions under which the bootstrap test is valid. In the calculation the partial covariance (2.8), we use the residual $\widehat{V}_{i,j,t}$ of the LASSO regression of $\widehat{U}_{i,t}$ onto $\widehat{U}_{-ij,t}$. Define $\widetilde{\mathbf{\Pi}}_{\mathcal{D}} := (\widetilde{\pi}_{i,j})_{(i,j) \in \mathcal{D}}$ for $\mathcal{D} \subseteq [n]^2$ and recall from Section 2.4 that $\widetilde{\pi}_{i,j} := \frac{1}{T} \sum_{t=1}^T V_{i,j,t} V_{j,i,t}$ for $i, j \in [n]$, $\mathbf{\Pi}_{\mathcal{D}} := \mathbb{E}(\widetilde{\mathbf{\Pi}}_{\mathcal{D}})$, and $\mathbf{\Upsilon}_{\mathbf{\Pi}}$ denotes the covariance of $\mathbf{\Pi}_{\mathcal{D}}$.

THEOREM 6. For $\mathcal{D} \subseteq [n]^2$, let $\widetilde{\mathbf{Q}} := \sqrt{T}(\widetilde{\mathbf{\Pi}}_{\mathcal{D}} - \mathbf{\Pi}_{\mathcal{D}})$ and $\mathcal{H}$ be a zero-mean Gaussian vector with the same covariance matrix of $\widetilde{\mathbf{Q}}$, that is, $\mathcal{H} \sim N(\mathbf{0}, \mathbf{\Upsilon}_{\mathbf{\Pi}})$. Under the same assumptions and notation of Theorem 5, with $\mathbf{\Upsilon}_{\Sigma}$ and $\underline{c}$ replaced by $\mathbf{\Upsilon}_{\mathbf{\Pi}}$ and $\underline{b}$, respectively, in condition (c), we have

$$\begin{aligned} \rho(\widetilde{\mathbf{Q}}, \mathcal{H}) &\lesssim \frac{\log T}{\underline{b}^2} \left[\frac{\log d}{T^{(1/2)-\kappa}} + \frac{K^{1-4/p} \log d}{T^{(r/2)(1-4/p)-1}} + \frac{(\log d)^{3/2}}{T^{1/4}} + \frac{d^{4/p} (\log d)^2}{T^{1/2-2/p}} \right] \\ &\quad + \frac{[d(\log d)^{(3/4)p-4} \log T \log(dT)]^{\frac{2}{p-4}}}{T^{1/4} \underline{b}^{p/(p-4)}} + \frac{d^{1/(p/2+1)} \sqrt{\log d} + K}{T^{r\kappa-1/2}}, \end{aligned}$$

where $d, \kappa := \kappa(p, r)$ are defined in Theorem 5.

Let $\widehat{\mathbf{Q}} := \sqrt{T}(\widehat{\mathbf{\Pi}}_{\mathcal{D}} - \mathbf{\Pi}_{\mathcal{D}})$, then

$$\rho(\widehat{\mathbf{Q}}, \mathcal{H}) \lesssim \rho(\widetilde{\mathbf{Q}}, \mathcal{H}) + \inf_{\delta > 0} [\delta_1 \sqrt{1 \vee \log(d/\delta)} + \mathbb{P}(\|\widehat{\mathbf{Q}} - \widetilde{\mathbf{Q}}\|_{\infty} > \delta)].$$

Let $\widetilde{\mathbf{\Upsilon}}_{\mathbf{\Pi}}$ be any positive semi definite estimator of $\mathbf{\Upsilon}_{\mathbf{\Pi}}$ and $\mathcal{H}^* | X, Y \sim N(\mathbf{0}, \widetilde{\mathbf{\Upsilon}}_{\mathbf{\Pi}})$, then

$$\rho(\widehat{\mathbf{Q}}, \mathcal{H}^*) \lesssim \rho(\widehat{\mathbf{Q}}, \mathcal{H}) + \inf_{\delta > 0} [\delta \log d (1 \vee |\log d|) + \mathbb{P}(\|\widetilde{\mathbf{\Upsilon}}_{\mathbf{\Pi}} - \mathbf{\Upsilon}_{\mathbf{\Pi}}\|_{\max} > \delta)].$$

REMARK 12. Let $\rho_{\mathbf{\Upsilon}} \geq 0$ denote the minimum eigenvalue of $\mathbf{\Upsilon}_{\Sigma}$. A positive lower bound on $\rho_{\mathbf{\Upsilon}}$ is fundamental if we want to ensure an upper bound of the form $f(n)/\sqrt{T}$ (up to $\log T$ terms) in the Central Limit Theorem for independent and identically distributed random vectors, where $f(n)$ depends on the moment conditions. For further details, refer to Theorems 2.1 and 2.2 in Chernozhukov, Chetverikov and Koike (2020) and Theorem 2.1 in Lopes (2022). The former shows that the rate is nearly optimal, up to $\log T$ terms (Proposition 2.1), and further conjectures that the uniform lower bound in γ is necessary.

REMARK 13. The condition $\rho_{\mathbf{\Upsilon}} \geq \underline{c} > 0$ can be dropped at the expense of lowering the convergence rate in Theorem S.13, which results in a lower convergence rate for Theorems 5 and 6. For instance, Chernozhuokov et al. (2022) establish the bound $\log^{5/4}(nT)/T^{1/4}$ while allowing $\rho = 0$, which could replace the bound used in step 1 of Theorem's S.13 proof.

REMARK 14. In terms of the sharpness of the rate of Theorems 5 and 6, it boils down to the sharpness of Theorem S.13. For the univariate case, Tikhomirov (1981) showed the

converge rate to be $\log^2 T / \sqrt{T}$ provided exponential decay of the mixing coefficient and third-moment finiteness. This result can not be recovered from Theorem S.13(b), where the rate would be $T^{-1/4}$ ignoring log terms. As far as we know, there are no results (sharp or otherwise) for high-dimensional central limit theorem approximation for strong-mixing sequences with arbitrary mixing coefficients. Using a different measure of (functional) dependency, Zhang and Wu (2017) proves the strong approximation for normalized mean. Also, Zhang and Cheng (2014) used the concept of graph dependencies to show similar results for the case of m dependency leveraging the results in Chernozhukov, Chetverikov and Kato (2013).

Similar comments as those appearing in Remarks 11 apply to Theorem 6 as well, which results in the following corollary.

COROLLARY 4. *Under the same conditions of Theorem 6, if $\tilde{\Upsilon}_\Pi = \hat{\Upsilon}_\Pi$ defined by (2.14) with $\mathcal{K} \in \mathbb{K}$, where $\mathcal{K}$ is defined by (3.6), then, uniformly in $\mathcal{D} \subset [n]^2$,*

$$\|\hat{\Upsilon}_\Pi - \Upsilon_\Pi\|_{\max} \lesssim_{\mathbb{P}} h \left[((1 + \tilde{s}_1)\varpi_U + \varrho_\chi n^{1/p})(1 + \tilde{s}_1)^3 (nT)^{3/p} + \frac{(1 + \tilde{s}_2)^4 n^{8/p}}{\sqrt{T}} \right],$$

where $h > 0$ is the bandwidth of the covariance estimator, $\tilde{s}_k := \max_{(i,j) \in \mathcal{D}} \|\mathbf{x}_{i,j}\|_k$ for $k \in \{0, 1, 2\}$, ϖ_U is the rate appearing in Corollary 1, and ϱ_χ is the rate appearing in Corollary 2 with s_0 and s_1 replaced by $\tilde{s}_0$ and $\tilde{s}_1$, respectively, and $l = 0$. If further, as $h, n, T \rightarrow \infty$:

- (a) $\varrho_\Pi = o(1)$, where ϱ_Π is the rate in the first result of Theorem 6 with d replaced by n^2 ;
- (b) $(\log n)^{3/2} [(1 + \tilde{s}_1 + \varrho_\chi)^2 (\sqrt{T} \varpi_U^2 + \frac{n^{9/p}}{\sqrt{T}} + \frac{n^{6/p}}{T^{1/2-1/p}} + \frac{1}{n^{1/2-9/p}}) + \varrho_\chi^2 n^{4/p} \sqrt{T}] = o(1)$;
- (c) $(\log n)^3 h \{[(1 + \tilde{s}_1)\varpi_U + \varrho_\chi n^{1/p}](1 + \tilde{s}_1)^3 (nT)^{3/p} + \frac{(1 + M_2)^4 n^{8/p}}{\sqrt{T}}\} = o(1)$, then,

$$\sup_{\mathcal{D}} \sup_{\tau \in (0,1)} |\mathbb{P}[S_{\mathcal{D}}^\Pi \leq c_\Pi^*(\tau)] - \tau| = o(1),$$

where the first supremum is over all null hypotheses of the form (2.9) indexed by $\mathcal{D} \subseteq [n]^2$.

REMARK 15. As opposed to the case of testing covariance, when testing partial covariance in high dimensions, the sparse structure plays a role in terms of $\tilde{s}_0$ appearing in the conditions (b) and (c). Therefore, these assumptions restrict the cases where the proposed partial covariance test has the correct asymptotic size. For instance, in the case of a complete dense partial covariance structure, we are likely to have $\tilde{s}_0$ of the order of n and, therefore, conditions (b) and (c) are not expected to hold.

4. Applications.

4.1. Network structure in asset returns. We illustrate our methodology by studying the factor structure of asset returns. We consider monthly close-to-close excess returns from 9456 firms traded on the New York Stock Exchange. The data starts on November 1991 and runs until December 2018. There are 326 monthly observations in total. In addition to the returns, we also consider 16 monthly factors: Market (MKT), Small-minus-Big (SMB), High-minus-Low (HML), Conservative-minus-Aggressive (CMA), Robust-minus-Weak (RMW), earning/price ratio, cash-flow/price ratio, dividend/price ratio, accruals, market beta, net share

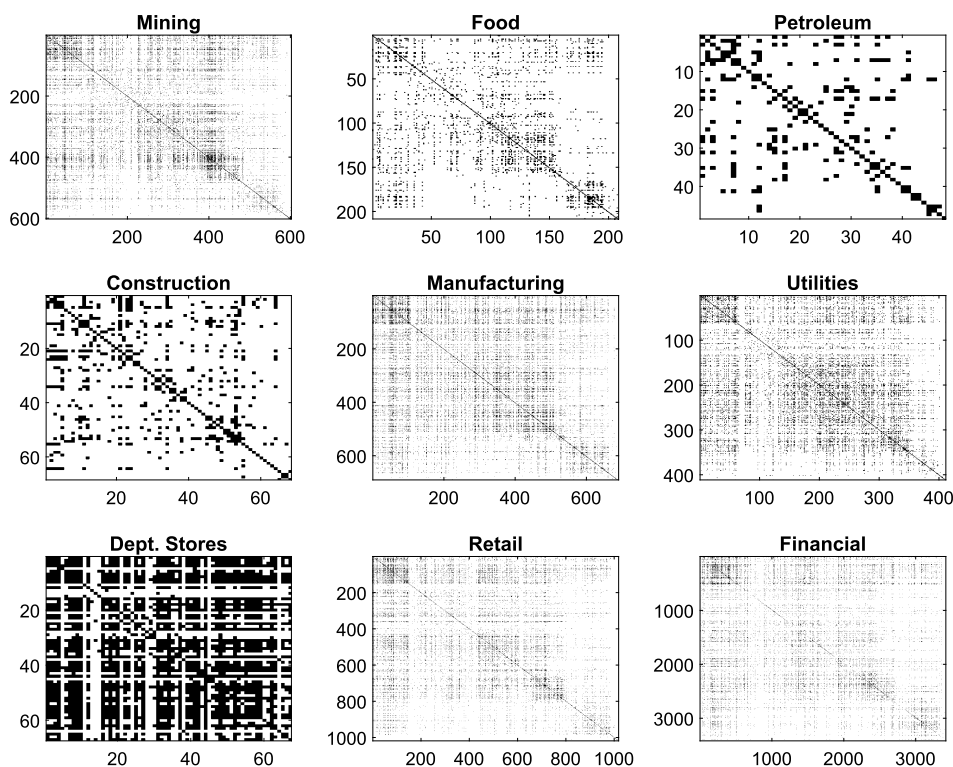


FIG. 1. *Correlations of monthly returns. We estimate the correlations between all pairs of returns from specific sectors. The correlations higher than 0.15 in absolute value are shown as black dots.*

issues, daily variance, daily idiosyncratic variance, 1-month momentum, and 36-month momentum. The firms are grouped according to 20 industry sectors as in Moskowitz and Grinblatt (1999). The following sectors are considered:⁹ Mining (602), Food (208), Apparel (161), Paper (81), Chemical (513), Petroleum (48), Construction (68), Primary Metals (133), Fabricated Metals (186), Machinery (710), Electrical Equipment (782), Transportation Equipment (166), Manufacturing (690), Railroads (25), Other transportation (157), Utilities (411), Department Stores (67), Retail (1018), Financial (3419), and Other (11).

We start by looking at the correlation matrix for monthly returns of a sample of nine sectors: Mining, Food, Petroleum, Construction, Manufacturing, Utilities, Department Stores, Retail, and Financial. Figure 1 plots the correlations larger than 0.15 in absolute value. We test for the null of a diagonal covariance matrix. The null hypothesis is strongly rejected with p -value much lower than 1% for all sectors. To test the covariance matrix, we use the simple sample estimator described in the paper. The correlations in Figure 1 and the subsequent ones are based on the nonlinear shrinkage estimator proposed by Ledoit and Wolf (2020).

We proceed by regressing the daily returns on the observed 16 factors. Figure 2 presents the estimated correlations for the first-stage residuals, namely factor-adjusted returns. We focus on the nine sectors as before. The first-stage regression is efficient in removing the correlation within specific sectors in some cases. The most notable ones are Financial and Retail, followed by Construction, Petroleum, and Manufacturing. On the other hand, Utilities, Department Stores, Mining, and Food still display a dense covariance matrix.

The second step is to conduct a principal component analysis on the residuals of the first stage. The eigenvalue ratio procedure selects two factors, while all four information criteria

⁹The number between parenthesis indicates the number of firms in our sample that belong to each sector.

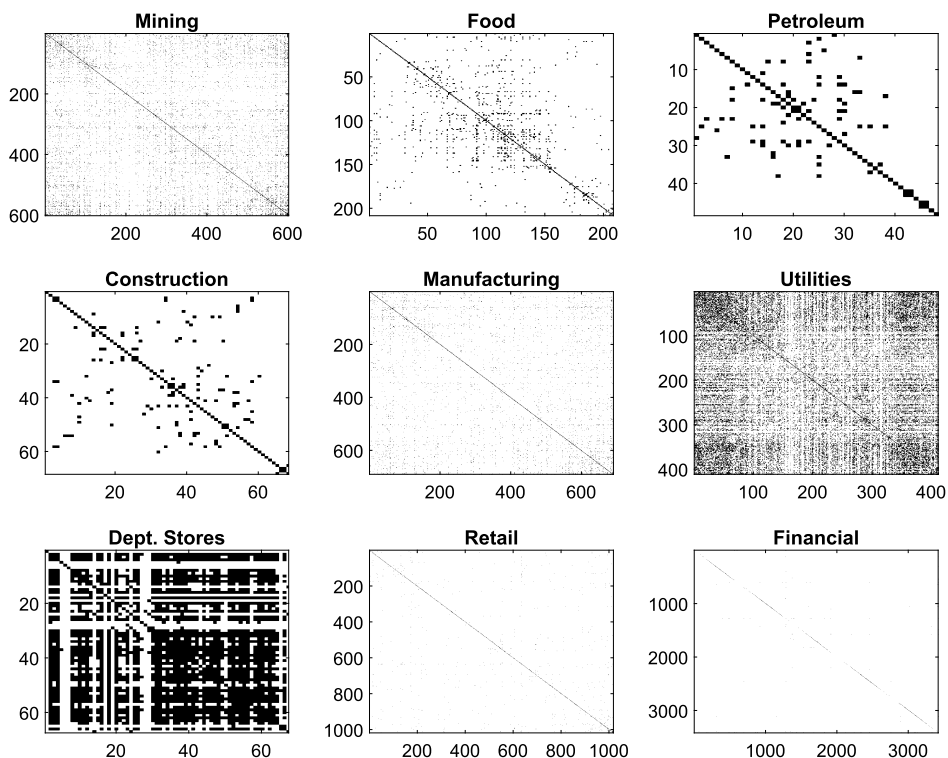


FIG. 2. *Correlations of first-stage residuals. We estimate the correlations between all pairs of residuals from the first-stage OLS regression on 16 observed factors from specific sectors. The correlations that are higher than 0.15 in absolute value are shown as black dots.*

point to a single factor. We proceed with two factors. Note that, by construction, the principal component factors are orthogonal to all the 16 risk factors considered in the first stage. Figure 3 shows the estimated correlations for the second-stage residuals (idiosyncratic component). The latent factor is not able to reduce the correlations within each sector. However, the conclusions are much different when we consider partial correlations. As seen from Figure 4, the partial correlation matrices are (almost) diagonal. In addition, we cannot reject the null of a diagonal covariance matrix at a 5% significance level.

To shed some light on the links among different sectors, we report how often variables from sector i are selected in the third-stage LASSO regression for firms in sector j . The numbers are normalized by the total number of firms in each sector and are presented in Figure 5. The most interesting fact is that covariates from the financial sector are most frequently selected for all the other sectors. Other sectors, such as Mining, Chemical, Machinery, Electrical Equipment, Manufacturing, and Retail, are frequently selected. This may indicate that industry factors, specifically a “financial factor”, are unmodeled in the first two stages. However, if we augment the set of regressors in the first stage by the value-weighted portfolio from the financial sector, although the remaining dependence among firms is attenuated, particularly for Department Stores, we do not get close to an exact factor model. This finding suggests that there are hidden links among firms.

4.2. Forecasting US industrial production. The second application consists of forecasting monthly US industrial production using a large set of monthly macroeconomic variables. We compare four different models: (1) Autoregressive model; (2) Sparse LASSO Regression (SR); (3) Principal Component Regression (PCR); and (4) FarmPredict.

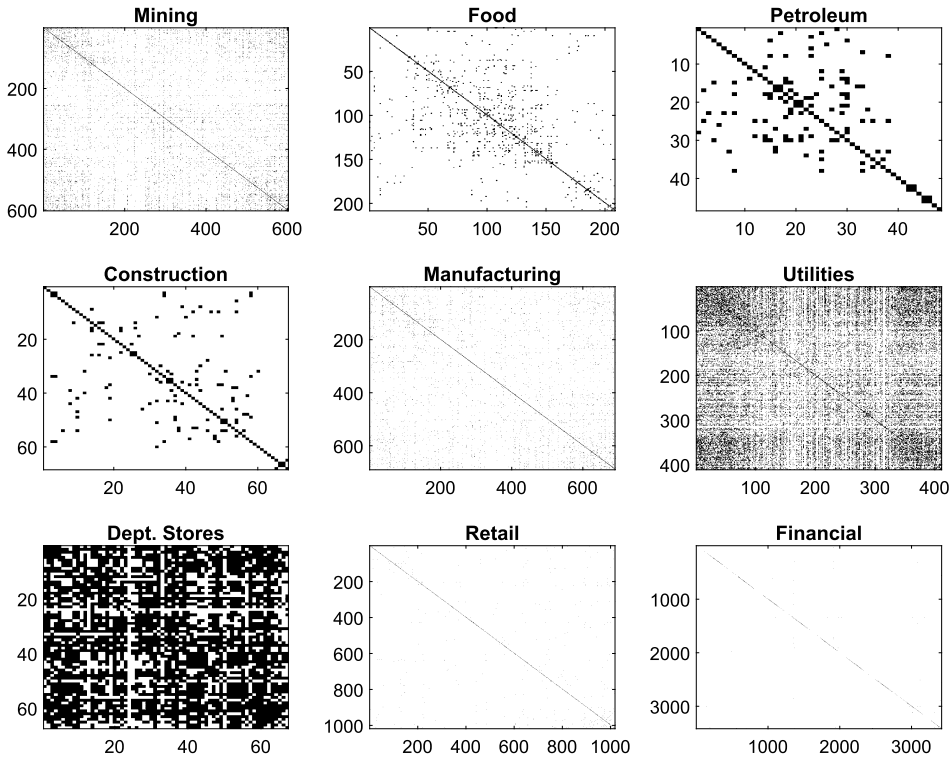


FIG. 3. *Correlations of second-stage residuals. We estimate the correlations between all pairs of residuals from the second-stage principal component analysis from specific sectors. The correlations higher than 0.15 in absolute value are shown as black dots.*

We use variables from the August 2022 vintage of the FRED-MD database, a large monthly macroeconomic dataset designed for empirical analysis in data-rich macroeconomic environments. The dataset is updated in real-time through the FRED database and is available from Michael McCracken’s webpage.¹⁰

Our sample extends from January 1960 to December 2019 (719 observations), and only variables with all observations in the period are used (122 variables). The dataset is divided into eight groups: (i) output and income; (ii) labor market; (iii) housing; (iv) consumption, orders, and inventories; (v) money and credit; (vi) interest and exchange rates; (vii) prices; and (viii) stock market. Finally, all series are transformed to become stationary.

To highlight the gains of exploring all relevant information in the dataset, we construct one-step ahead forecasts for the first-order difference of the logarithm of the monthly industrial production index (IP, growth rate): $Y_{IP,t}$.

We compare the following models:

1. *Autoregressive model (AR):*

$$\widehat{Y}_{IP,t+1|t}^{(AR)} = \widehat{\phi}_0 + \widehat{\phi}_1 \widehat{Y}_{IP,t} + \cdots + \widehat{\phi}_p \widehat{Y}_{IP,t-p+1},$$

where $\widehat{\phi}_0, \widehat{\phi}_1, \dots, \widehat{\phi}_p$ are OLS estimates. The value of p is selected by BIC.

2. *Sparse regression (SR):*

$$\widehat{Y}_{IP,t+1|t}^{(SR)} = \widehat{\beta}_0 + \widehat{\beta}_1' Y_t + \cdots + \widehat{\beta}_p' Y_{t-p+1},$$

¹⁰<https://research.stlouisfed.org/econ/mccracken/fred-databases/>. For further details, we refer to McCracken and Ng (2016).

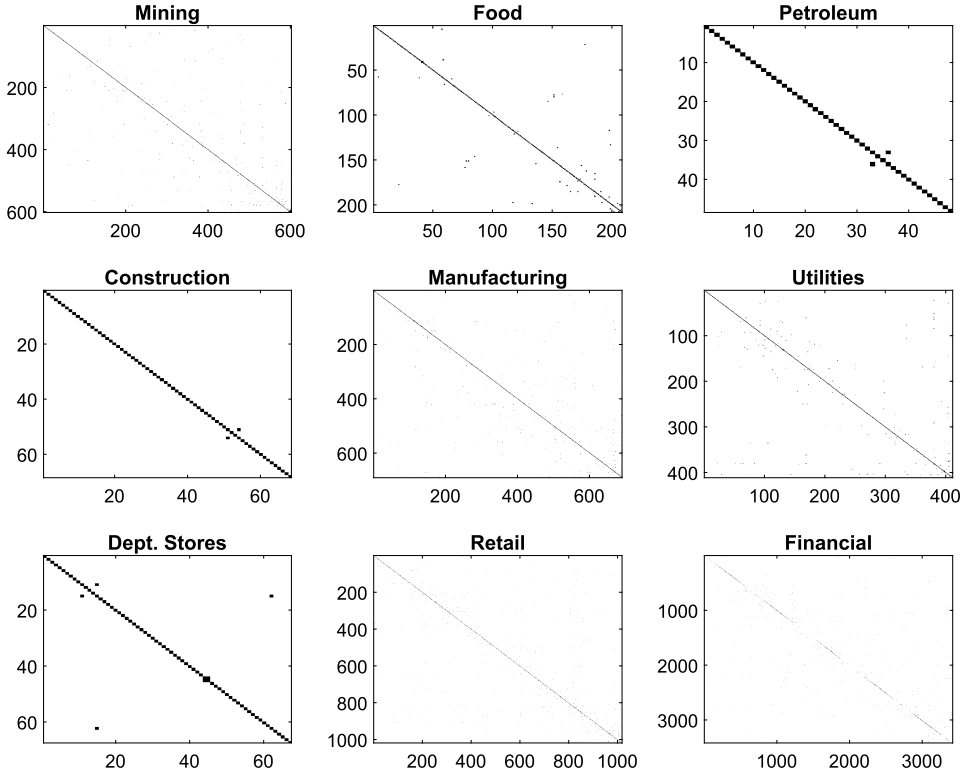


FIG. 4. *Partial correlations of second-stage residuals. We estimate the partial correlations between all pairs of residuals from the second-stage LASSO regression from specific sectors. The correlations higher than 0.15 in absolute value are shown as black dots.*

$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are LASSO estimates and $Y_t = (Y_{1t}, \dots, Y_{nt})$ with $n = 122$. The penalty parameter is selected by modified BIC as in Wang, Li and Leng (2009).

3. Principal Component Regression (PCR):

$$\hat{Y}_{IP,t+1|t}^{(PCR)} = \hat{\pi}_0 + \hat{\pi}'_1 \hat{F}_t + \dots + \hat{\pi}'_q \hat{F}_{t-q+1},$$

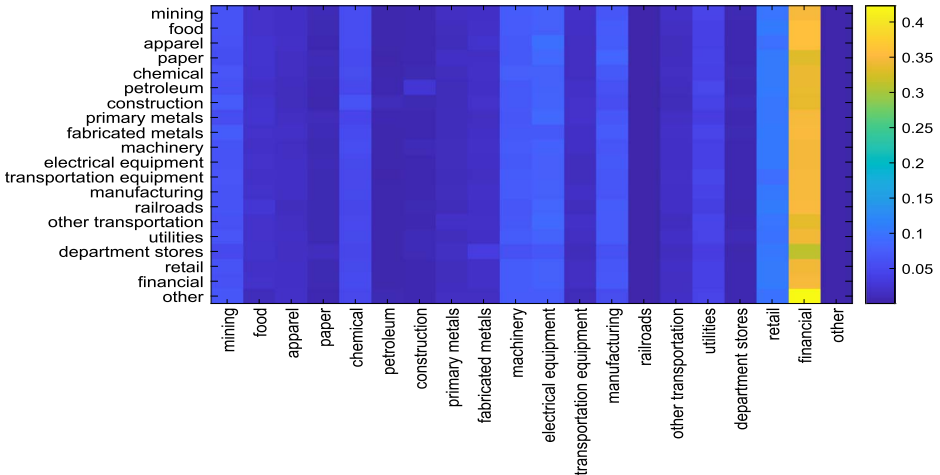


FIG. 5. *Variable Selection Frequency. We report how often the variables from column sectors are selected in the third-stage LASSO regression for firms on row sectors. The numbers are normalized by the total number of firms in each sector.*

where $\hat{\mathbf{F}}_t$ is the estimate of the $(k \times 1)$ vector of factors $\mathbf{F}_t$ given by the first k principal components of $\mathbf{Y}_t - \hat{\boldsymbol{\mu}}$ with $\hat{\boldsymbol{\mu}}$ being the sample average of $\mathbf{Y}_t$. The parameters of the model are computed by OLS regression of $Y_{j,t}$ on a constant and lags of $\hat{\mathbf{F}}_t$. The lag q is selected by BIC.

4. *AR—Principal Component Regression (AR-PCR):*

$$\hat{Y}_{\text{IP},t+1|t}^{(\text{AR-PCR})} = \hat{\alpha}_0 + \hat{\alpha}'_1 Y_{\text{IP},t} + \cdots + \hat{\alpha}'_p Y_{\text{IP},t-p+1} + \hat{\alpha}'_1 \hat{\mathbf{F}}_t + \cdots + \hat{\alpha}'_p \hat{\mathbf{F}}_{t-p+1},$$

where $\hat{\mathbf{F}}_t$ is the estimate of the $(k \times 1)$ vector of factors $\mathbf{F}_t$ given by the first k principal components of $\mathbf{Y}_t - \hat{\boldsymbol{\mu}}$ with $\hat{\boldsymbol{\mu}}$ being the sample average of $\mathbf{Y}_t$. The model's parameters are computed by OLS regression of $Y_{j,t}$ on a constant, its own lags and lags of $\hat{\mathbf{F}}_t$. The lag orders p and q are selected by BIC.

5. *FarmPredict:*

$$\hat{Y}_{i,t+1|t}^{(\text{FarmPredict})} = \hat{\boldsymbol{\mu}}'_j + \hat{\boldsymbol{\lambda}}'_i \hat{\mathbf{P}}'_{j1} \hat{\mathbf{F}}_t + \cdots + \hat{\boldsymbol{\lambda}}'_i \hat{\mathbf{P}}'_{jp} \hat{\mathbf{F}}_{t-p+1} + \hat{\boldsymbol{\theta}}'_{1i} \hat{\mathbf{U}}_t + \cdots + \hat{\boldsymbol{\theta}}'_{pi} \hat{\mathbf{U}}_{t-p+1},$$

where $\hat{\mathbf{U}}_t = (\hat{U}_{1,t}, \dots, \hat{U}_{n,t})'$ and $\hat{U}_{i,t} = Y_{i,t} - \hat{\boldsymbol{\lambda}}'_i \hat{\mathbf{F}}_t$, $i \in [n]$. The estimates $\hat{\theta}_{0i}, \hat{\theta}_{1i}, \dots, \hat{\theta}_{pi}$, $i \in [n]$, are given by LASSO. The penalty parameter is selected by the modified BIC, and the value of p is set to 24.

The forecasts are based on a rolling-window framework of a fixed length of 480 observations, starting in January 1960. Therefore, the forecasts started on January 1990. The last forecasts are for December 2019. Note that the AR model only considers information concerning the own past of the variable of interest. SR and PCR/AR-PCR/expand the information by two opposing routes. While SR uses a sparse combination of the set of variables, PCR and AR-PCR, consider a factor structure (dense model). In the case of AR-PCR, lags of the dependent variable are also included. FarmPredict combines these two approaches and uses the full information available. The number of factors is set to 1.

Figure 6 reports the ratios of cumulative MSE of the FarmPredict model against the cumulative MSE of the other benchmarks over the forecasting period. Several conclusions emerge from the plot. First, FarmPredict outperforms the PCR model over the entire out-of-sample period. It is also generally superior to the AR, SR, and AR-PCR models, apart from 2004 and 2008. During this period, the economy experienced housing and financial crises. Furthermore, the number of out-of-sample forecasting periods was also relatively small. The performance of FarmPredict improves drastically after 2008. Across the sample, the MSE ratio of the FarmPredict model over the AR benchmark is 0.9080, while the SR, PCR, and AR-PCR models have the following ratios, respectively: 0.9217, 1.0249, and 0.9215.

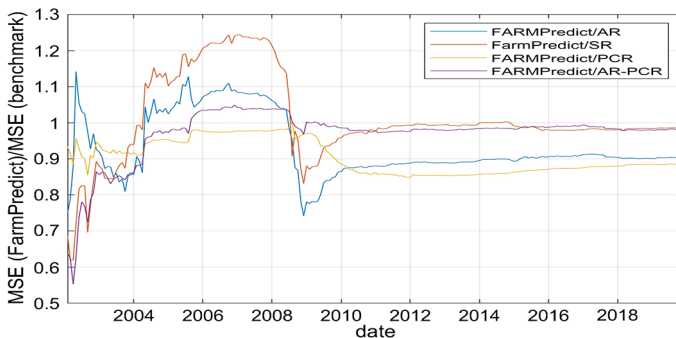


FIG. 6. *Ratios of cumulative MSEs. In the figure we report the cumulative ratios of the mean squared errors (MSE) of the FarmPredict model and the other benchmarks over the rolling windows.*

5. Conclusions. We propose a new methodology that bridges the gap between sparse regressions and factor models and evaluates the gains of increasing the information set via factor augmentation. Our proposal consists of several steps. In the first one, we filter the data for known factors (trends, seasonal adjustments, covariates). In the second step, we estimate a latent factor structure. Finally, in the last part of the procedure, we estimate a sparse regression for the idiosyncratic components. We also propose a new test for remaining structures in high-dimensional and partial covariance matrices. Our test can be used to evaluate the benefits of adding more structure to the model. Our paper also has several important side results. First, we proved the consistency of kernel estimation of long-run covariance matrices in high dimensions where both the number of observations and variables grows. Second, we derive the theoretical properties of factor estimation on the residuals of a first-step process. Third, the proposed test can be used as a diagnostic tool for factor models.

We evaluate our methodology with simulations and real data. The simulations show the test has good size and power properties even when the true number of factors is unknown and must be determined from the data. If the number of factors is underestimated, we observe size distortions. In particular, this is the case when the eigenvalue ratio test is used to determine the number of latent factors. The simulations also show significant informational gains when combining factor models and sparse regressions in a forecasting exercise. Two applications are considered.

Acknowledgments. The authors are grateful to Caio Almeida, Federico Bandi, Matteo Barigozzi, Gilberto Boareto, Gustavo Bulhões, Giuseppe Cavaliere, Frank Diebold, Bruno Ferman, Marcelo Fernandes, Claudio Flores, Conrado Garcia, Eric Ghysels, Alexander Giessing, Nathalie Gimenes, Marcelo J. Moreira, Henrique Pires, Yuri Saporito, and Rodrigo Targino for helpful comments. We also thank seminar participants at the SofiE online seminar series, Princeton University, the University of Amsterdam, the University of Pennsylvania, the University of Illinois at Urbana-Champaign, the Federal University of São Carlos, Rutgers University, the University of North Carolina at Chapel Hill, the University of Chicago, the University of California at Riverside, and Columbia University for several valuable comments. Finally, we are deeply grateful to Michele Lenza, Eduardo F. Mendes, and Michael Wolf for the careful reading of the paper and the many insightful discussions which led to a much-improved version of this manuscript. This manuscript has also been presented at many conferences, and we thank all the participants for their very useful comments.

Funding. The first author was supported by ONR grant N00014-22-1-2340 and NSF grants DMS-2210833, DMS-2053832, and DMS-2052926.

SUPPLEMENTARY MATERIAL

Supplement to “Bridging factor and sparse models” (DOI: [10.1214/23-AOS2304SUPP](https://doi.org/10.1214/23-AOS2304SUPP); .pdf). Supplementary information.

REFERENCES

- ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque country. *Amer. Econ. Rev.* **93** 113–132.
- AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. MR2985947 <https://doi.org/10.1214/12-AOS1000>
- ANDREOU, E. and GHYSELS, E. (2021). Predicting the VIX and the volatility risk premium: The role of short-run funding spreads volatility factors. *J. Econometrics* **220** 366–398. MR4201495 <https://doi.org/10.1016/j.jeconom.2020.04.006>

- ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. [MR1106513](#) <https://doi.org/10.2307/2938229>
- BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. [MR1956857](#) <https://doi.org/10.1111/1468-0262.00392>
- BAI, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279. [MR2547073](#) <https://doi.org/10.3982/ECTA6135>
- BAI, J. and LIAO, Y. (2017). Inferences in panel data with interactive effects using large covariance matrices. *J. Econometrics* **200** 59–78. [MR3683662](#) <https://doi.org/10.1016/j.jeconom.2017.05.014>
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. [MR1926259](#) <https://doi.org/10.1111/1468-0262.00273>
- BAI, J. and NG, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150. [MR2238213](#) <https://doi.org/10.1111/j.1468-0262.2006.00696.x>
- BARIGOZZI, M. and BROWNLEES, C. (2019). NETS: Network estimation for time series. *J. Appl. Econometrics* **34** 347–364. [MR3948470](#) <https://doi.org/10.1002/jae.2676>
- BARIGOZZI, M., CHO, H. and OWENS, D. (2022). FNETS: Factor-adjusted network estimation and forecasting for high-dimensional time series Technical Report No. 2201.06110.
- BARIGOZZI, M. and HALLIN, M. (2017). A network analysis of the volatility of high dimensional financial series. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 581–605. [MR3632343](#) <https://doi.org/10.1111/rssc.12177>
- BERNANKE, B. S., BOIVIN, J. and ELIASZ, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.* **120** 387–422.
- BROWNLEES, C., GUDMUNDSSON, G. S. and LUGOSI, G. (2020). Community detection in partial correlation network models. *J. Bus. Econom. Statist.* forthcoming.
- CAI, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annu. Rev. Stat. Appl.* **4** 4.1–4.24.
- CARVALHO, C., MASINI, R. and MEDEIROS, M. C. (2018). ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *J. Econometrics* **207** 352–380. [MR3868187](#) <https://doi.org/10.1016/j.jeconom.2018.07.005>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. [MR3161448](#) <https://doi.org/10.1214/13-AOS1161>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud.* **86** 1867–1900. [MR4009488](#) <https://doi.org/10.1093/restud/rdy065>
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KOIKE, Y. (2020). Nearly optimal central limit theorem and bootstrap approximations in high dimensions.
- CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. and KOIKE, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *Ann. Statist.* **50** 2562–2586. [MR4500619](#) <https://doi.org/10.1214/22-aos2193>
- DIEBOLD, F. X. and YILMAZ, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econometrics* **182** 119–134. [MR3212765](#) <https://doi.org/10.1016/j.jeconom.2014.04.012>
- FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.
- FAMA, E. F. and FRENCH, K. R. (2015). A five-factor asset pricing model. *J. Financ. Econ.* **116** 1–22.
- FAN, J., KE, Y. and WANG, K. (2020). Factor-adjusted regularized model selection. *J. Econometrics* **216** 71–85. [MR4077382](#) <https://doi.org/10.1016/j.jeconom.2020.01.006>
- FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. [MR3597972](#) <https://doi.org/10.1111/rssb.12166>
- FAN, J., LI, R., ZHANG, C. H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC Press, Boca Raton.
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. [MR3091653](#) <https://doi.org/10.1111/rssb.12016>
- FAN, J., MASINI, R. and MEDEIROS, M. C. (2022). Do we exploit all information for counterfactual analysis? Benefits of factor models and idiosyncratic correction. *J. Amer. Statist. Assoc.* **117** 574–590. [MR4436297](#) <https://doi.org/10.1080/01621459.2021.2004895>
- FAN, J., MASINI, R. P. and MEDEIROS, M. C. (2023). Supplement to “Bridging factor and sparse models.” <https://doi.org/10.1214/23-AOS2304SUPP>

- FENG, G., GIGLIO, S. and XIU, D. (2020). Taming the factor zoo: A test of new factors. *J. Finance* **75** 1327–1370.
- FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econ. Stat.* **82** 540–554.
- FORNI, M., HALLIN, M., LIPPI, M. and ZAFFARONI, P. (2017). Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis. *J. Econometrics* **199** 74–92. [MR3652406 https://doi.org/10.1016/j.jeconom.2017.04.002](https://doi.org/10.1016/j.jeconom.2017.04.002)
- GAGLIARDINI, P., OSSOLA, E. and SCAILLET, O. (2019). A diagnostic criterion for approximate factor structure. *J. Econometrics* **212** 503–521. [MR4001672 https://doi.org/10.1016/j.jeconom.2019.06.001](https://doi.org/10.1016/j.jeconom.2019.06.001)
- GAGLIARDINI, P., OSSOLA, E. and SCAILLET, P. (2020). Estimation of large dimensional conditional factor models in finance. In *Handbook of Econometrics* (S. Durlauf, L. Hansen, J. Heckman and R. Matzkin, eds.) **7A** 219–282.
- GIANNONE, D., LENZA, M. and PRIMICERI, G. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica* **89** 2409–2437.
- GIESSING, A. and FAN, J. (2020). Bootstrapping ℓ_p -statistics in high dimensions.
- GOBILLON, L. and MAGNAC, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Rev. Econ. Stat.* **98** 535–551.
- KOCK, A. B. and CALLOT, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics* **186** 325–344. [MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013](https://doi.org/10.1016/j.jeconom.2015.02.013)
- LEDOIT, O. and WOLF, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Statist.* **48** 3043–3065. [MR4152634 https://doi.org/10.1214/19-AOS1921](https://doi.org/10.1214/19-AOS1921)
- LEDOIT, O. and WOLF, M. (2021). The power of (non)linear shrinking: A review and guide to covariance matrix estimation. *J. Financ. Econom.* forthcoming.
- LOPES, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions: Near $1/\sqrt{n}$ rates via implicit smoothing. *Ann. Statist.* **50** 2492–2513. [MR4505371 https://doi.org/10.1214/22-aos2184](https://doi.org/10.1214/22-aos2184)
- MASINI, R. P., MEDEIROS, M. C. and MENDES, E. F. (2022). Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. *J. Time Series Anal.* **43** 532–557. [MR4446423 https://doi.org/10.1111/jtsa.12627](https://doi.org/10.1111/jtsa.12627)
- MCCRACKEN, M. W. and NG, S. (2016). FRED-MD: A monthly database for macroeconomic research. *J. Bus. Econom. Statist.* **34** 574–589. [MR3547997 https://doi.org/10.1080/07350015.2015.1086655](https://doi.org/10.1080/07350015.2015.1086655)
- MOON, H. R. and WEIDNER, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* **83** 1543–1579. [MR3384227 https://doi.org/10.3982/ECTA9382](https://doi.org/10.3982/ECTA9382)
- MOSKOWITZ, T. J. and GRINBLATT, M. (1999). Do industries explain momentum? *J. Finance* **54** 1249–1290.
- PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. [MR2541591 https://doi.org/10.1198/jasa.2009.0126](https://doi.org/10.1198/jasa.2009.0126)
- PESARAN, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967–1012. [MR2238209 https://doi.org/10.1111/j.1468-0262.2006.00692.x](https://doi.org/10.1111/j.1468-0262.2006.00692.x)
- STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. [MR1951271 https://doi.org/10.1198/016214502388618960](https://doi.org/10.1198/016214502388618960)
- STOCK, J. H. and WATSON, M. W. (2016). Handbook of macroeconomics. In *Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics* (J. B. Taylor and H. Uhlig, eds.) **2** 415–525. Elsevier, Amsterdam.
- TIKHOMIROV, A. N. (1981). On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory Probab. Appl.* **25** 790–809.
- VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. [MR2576316 https://doi.org/10.1214/09-EJS506](https://doi.org/10.1214/09-EJS506)
- WANG, H., LI, B. and LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 671–683. [MR2749913 https://doi.org/10.1111/j.1467-9868.2008.00693.x](https://doi.org/10.1111/j.1467-9868.2008.00693.x)
- ZHANG, D. and WU, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.* **45** 1895–1919. [MR3718156 https://doi.org/10.1214/16-AOS1512](https://doi.org/10.1214/16-AOS1512)
- ZHANG, X. and CHENG, G. (2014). Bootstrapping high dimensional time series.