BRIDGING FACTOR AND SPARSE MODELS

JIANQING FAN, RICARDO MASINI, AND MARCELO C. MEDEIROS

ABSTRACT. Factor and sparse models are two widely used methods to impose a low-dimensional structure in high-dimension. 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 to efficiently explore all the information in high-dimensional datasets. The method is based on a flexible model for high-dimensional panel data, called factor-augmented regression (`FarmPredict`) model with both observable or latent common factors, as well as idiosyncratic components. This model not only includes both principal component (factor) regression and sparse regression as specific models but also significantly weakens the cross-sectional dependence and hence facilitates model selection and interpretability. The methodology consists of three steps. At each step, the remaining cross-section dependence can be inferred by a novel test for covariance structure in high-dimensions. We developed asymptotic theory for the `FarmPredict` model and demonstrated the validity of the multiplier bootstrap for testing high-dimensional covariance structure. This is further extended to testing high-dimensional partial covariance structures. The theory is supported by an simulation study and applications to the construction of a partial covariance network of the financial returns and a prediction exercise for a large panel of macroeconomic time series from FRED-MD database.

JEL Codes: C22, C23, C32, C33.

Keywords: Principal component analysis, penalized least-squares, high-dimensional inference, FarmPredict, covariance structure, partial covariance structure, LASSO.

Acknowledgments: Medeiros gratefully acknowledges the partial financial support from CNPq and CAPES. We are grateful to Caio Almeida, Gilberto Boareto, Gustavo Bulhões, Frank Diebold, Bruno Ferman, Marcelo Fernandes, Claudio Flores, Conrado Garcia, Alexander Giessing, Nathalie Gimenes, Marcelo J. Moreira, Henrique Pires, Yuri Saporito, Rodrigo Targino for helpful comments. We also thank seminar participants at Princeton University, the University of Amsterdam and the Federal University of São Carlos for a number of valuable comments. Finally, we are deeply grateful to 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.

1. Introduction

With the emergence of new and large datasets, 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 under the assumption that the remaining dependence structure is negligible. For instance, panel factor models are applied now to a wide variety of important applications, ranging from forecasting (macroeconomic) variables and asset pricing models to causal inference in applied microeconomics and network analysis. On the other hand, there have been major advances on 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 on 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 in order to better characterize the dependence structure of high-dimensional data. 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 in order to accommodate a very large number of data generating processes of interest in economics, finance, and related areas. The proposed methodology has two ingredients: a three-step 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 any other observed covariates. The first step can be parametric or nonparametric, low or high dimensional. A latent factor model is then estimated using the residuals from the first stage. Finally, in a final step we model the dependence among idiosyncratic terms as a weakly sparse regression estimated by the Least Absolute Shrinkage and Selection Operator (LASSO). At each step, the null-hypothesis of no remaining cross-section dependence can be tested by the proposed test for the (partial) covariance structure in high-dimensions.

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

1.1. **Motivation.** Let $\boldsymbol{Y}_t := (Y_{1t}, \ldots, Y_{nt})'$ be a random vector generated as $Y_{it} = \boldsymbol{\lambda}_i' \boldsymbol{F}_t + U_{it}$, for $i = 1, \ldots, n$, $t = 1, \ldots, T$, where $\boldsymbol{\Sigma} := \mathbb{E}(\boldsymbol{U}_t \boldsymbol{U}_t')$, with $\boldsymbol{U}_t := (U_{1t}, \ldots, U_{nt})'$, is not necessarily diagonal. Fix one component of interest $i \in \{1, \ldots, n\}$, which serve as a response variable. Consider the following prediction models:

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

where $\boldsymbol{Y}_{-it}$ and $\boldsymbol{U}_{-it}$ are, respectively, vectors with the elements of $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ excluding the i-th entry. Model $\mathcal{M}_3$ is indeed the factor augmented regression model since it is the same as $\mathbb{E}(Y_{it}|\boldsymbol{F}_t, \boldsymbol{Y}_{-it})$.

Suppose that we observe both $\boldsymbol{F}_t$ and $\boldsymbol{U}_{-it}$. Which one of three models above is best in terms of mean square error (MSE) for prediction? Comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$ is not clear since it depends, among others, on the magnitude of $\boldsymbol{\Sigma}$ relative to $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} := (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)'$. However, since the σ-algebras generated by $\boldsymbol{Y}_{-it}$ and $\boldsymbol{F}_t$ are both included in the σ-algebra generated by $(\boldsymbol{F}_t, \boldsymbol{U}_{-it})$, it is not surprising that $\mathsf{MSE}(\mathcal{M}_3) \leqslant \mathsf{min}[\mathsf{MSE}(\mathcal{M}_1), \mathsf{MSE}(\mathcal{M}_2)]$. The same will hold true if we replace the models in (1.1) by their best linear projections, which we denote by $\widetilde{\mathcal{M}}_j$ for $j \in \{1, 2, 3\}$, since the linear space $\widetilde{\mathcal{M}}_3$ is the largest. In fact, since linear span by $(\boldsymbol{F}_t, \boldsymbol{U}_{-it})$ is the same as that by $(\boldsymbol{Y}_{-it}, \boldsymbol{F}_t)$, we expand the space in the informative factor directions $\boldsymbol{F}_t$. In the linear case, we can explicitly write the "gains" of $\widetilde{\mathcal{M}}_3$ when compared to $\widetilde{\mathcal{M}}_1$ and $\widetilde{\mathcal{M}}_2$:

$$\begin{aligned}
\mathsf{MSE}(\widetilde{\mathcal{M}}_3) - \mathsf{MSE}(\widetilde{\mathcal{M}}_1) &= -\boldsymbol{\theta}_i' \boldsymbol{\Sigma}_{-i,-i} \boldsymbol{\theta}_i \\
\mathsf{MSE}(\widetilde{\mathcal{M}}_3) - \mathsf{MSE}(\widetilde{\mathcal{M}}_2) &= -\boldsymbol{\Delta}_{1i} \boldsymbol{\Delta}_{1i}' - \boldsymbol{\Delta}_{2i}' \boldsymbol{\Sigma}_{-i,-i} \boldsymbol{\Delta}_{2i},
\end{aligned}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$ are the coefficients of the projection of U_{it} onto $\boldsymbol{U}_{-it}$ and the coefficients of the projection of X_{it} onto $\boldsymbol{X}_{-it}$, respectively; $\boldsymbol{\Sigma}_{-i,-i}$ is $\boldsymbol{\Sigma}$ excluding the i-th row and column; $\boldsymbol{\Delta}_{1i} := \boldsymbol{\Lambda}_i - \boldsymbol{\beta}_i' \boldsymbol{\Lambda}_{-i}$ and $\boldsymbol{\Delta}_{2i} := \boldsymbol{\beta}_i - \boldsymbol{\theta}_i$. From the previous expressions, it becomes evident that both $\widetilde{\mathcal{M}}_1$ and $\widetilde{\mathcal{M}}_2$ are restrictions on $\widetilde{\mathcal{M}}_3$. Broadly speaking, whenever one does not expect to have an *exact* factor model, there are potential gains of taking into account the contribution of the idiosyncratic components $\boldsymbol{U}_{-it}$. Therefore, we use $\widetilde{\mathcal{M}}_3$ 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 multi-fold. First, our methodology bridges the gap between two apparently competing methods for high-dimensional modeling; see, for example the discussion in Giannone et al. (2021) and Fan et al. (2020). This yields a vast number of potential applications and spin-offs. For instance, in Fan et al. (2020), we apply the methods developed in here to evaluate the effects of interventions and we contribute to the literature on synthetic controls and related methods by combining the approaches of Gobillon and Magnac (2016) and Carvalho et al. (2018). Therefore, in our setup both a common factor structure and weak sparsity can coexist.[1]

Second, our results can also 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 directly applied to uncover the dependence structure among cross-sectional units before and after accounting for common factor components. If the factor structure is informative enough, we expect the idiosyncratic covariance matrix to be almost sparse. If this is not the case, we may have possibly underestimated the number of factors. One popular application is in asset pricing as discussed in Gagliardini et al. (2019) and in the empirical section of this paper. There are a huge number of proposed factors as described in Feng et al. (2020), Giglio and Xiu (2020), and Gu et al. (2020). We can apply our methodology not only to test for omitted factors, but, as well, to estimate network connections among firms as in Diebold and Yilmaz (2014) and Brownlees et al. (2020). Finally, as a diagnostic tool, our paper tackle the same problem as Gagliardini et al. (2019). However, we take an alternative solution strategy which relies on a much different set of hypothesis; see also Gagliardini et al. (2020).

Third, the methodology proposed here contributes to the forecasting literature. For instance, in the second application considered in this paper, we build forecasting models for a large cross-section of macroeconomic variables. We call this method the `FarmPredict`. We show that the combination of factors and a sparse regression strongly outperforms the traditional principal component regression as in Stock and Watson (2002a,b). 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

[1]Sparsity and factor models can also coexist in the framework of sparse principal components; see Fan et al. (2020).

in Bernanke et al. (2005) with sparse vector models as in Kock and Callot (2015) and Masini et al. (2019). Hence, we contribute to the debate in Giannone et al. (2021).

Fourth, we show consistency of factor estimation based on the residuals of a first-step regression. Our results hold for both parametric (linear or nonlinear) and non-parametric first stage. A high-dimensional first stage is also allowed. Note that, 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 and Ng, 2002, 2003, 2006). More specifically, our methodology favors settings where there are both observed and latent factors, as well as trend-stationary data. In the later, the trend can be first removed by (nonparametric) first-stage regression. In addition, whenever the unobserved factors and the observed covariates are correlated, the method proposed in Pesaran (2006) can be used and all or results follow directly.

Fifth, we contribute to the LASSO literature. LASSO can not be model selection consistent for highly correlated variables. Through the decomposition of 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 et al., 2020). We show consistency of the estimates based on 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 it is usual in many papers. See, for example, van de Geer and Bühlmann (2009). Furthermore, we derive our results under much mild conditions that the ones considered in (Fan et al., 2020).

Finally, we extend the results in Chernozhukov et al. (2013, 2018) to strong-mixing data in order to construct hypothesis tests for covariance and partial covariance structure in high dimensions.[2] This step is necessary for econometrics and financial applications. As side results, in order to develop the test we first show consistency of kernel-based estimation of a high-dimensional long-run covariance matrix of dependent process. This is a new result with important consequences for the theory of high-dimensional regression with dependent errors. We also establish 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

[2]Recently, Giessing and Fan (2020) also extended the results in Chernozhukov et al. (2013). However, their setup is very different from ours and the authors only consider the case of independent and identically distributed data.

when we use the residuals from a previous step estimation to compute the covariance matrix. This result allows us to to apply the test to the three-stage estimation procedure proposed here. Although our results are derived under the assumption that the number of factors is known, simulation results presented in the paper provide evidence that the test have good finite-sample properties even when the number of factors is determined by data-driven methods commonly found in the literature. In addition, due to the factor augmentation, our method is robust to the over-estimation of the number of factors. Over the past years, a vast number of papers proposed different methods to test for covariance structure in high dimensions. See, for example, Ledoit and Wolf (2002), Chen et al. (2010), Onatski et al. (2013), Cai and Ma (2013), Li and Qin (2014), Zheng et al. (2019), Cai et al. (2016), Zheng et al. (2019), and Guo and Tang (2020), among many others.[3] 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.

Summarizing, our approach provides:

(1) A systematic way to unify factor and sparse models in order to construct econometric specifications which use all the available information. These models can be applied to:
   (a) Forecasting in a high-dimensional setting;
   (b) Construction of counterfactuals to aggregate data;
   (c) Estimation of partial correlation networks;
(2) An inferential procedure to test for genereal structures in covariance and partial covariance matrices.

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 with practical guides given in Section 4. We depict the results of a simulation experiment in Section 5 and discuss the empirical application in Section 6. Section 7 concludes. All proofs are deferred to the Supplementary Material.

1.4. **Notation.** All random variables (real-valued scalars, vectors and matrices) are defined in a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We denote random variables by an

[3]For a nice recent review, see Cai (2017).

upper case letter, X for instance, and its realization by a lower case letter, $X = x$. The expected value operator is with respect to the $\mathbb{P}$ law such that $\mathbb{E}X := \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega)$. Matrices and vectors are written in bold letters $\boldsymbol{X}$. Except for the number of factors, r, and 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 clustering the notation prematurely.

We use $\|\cdot\|_p$ to denote the ℓ^p norm for $p \in [1, \infty]$, such that for a $d-$dimensional (possibly random) vector $\boldsymbol{X} = (X_1, \ldots, X_d)'$, we have $\|\boldsymbol{X}\|_p := (\sum_{i=1}^d |X_i|^p)^{1/p}$ for $p \in [1, \infty)$ and $\|\boldsymbol{X}\|_\infty := \mathsf{sup}_{i \leqslant d} |X_i|$. If $\boldsymbol{X}$ is a $(m \times n)$ possibly random matrix then $\|\boldsymbol{X}\|_p$ denotes the matrix ℓ^p-induced norm and $\|\boldsymbol{X}\|_{\mathsf{max}}$ denotes the maximum entry in absolute terms of the matrix $\boldsymbol{X}$. Note that whenever $\boldsymbol{X}$ is random, then $\|\boldsymbol{X}\|_p$ for $p \in [1, \infty]$ and $\|\boldsymbol{X}\|_{\mathsf{max}}$ are random variables. We also reserve the symbol $\|\cdot\|$ without subscript for the Euclidean norm $\|\cdot\| := \|\cdot\|_2$ for both vectors and matrices.

For any convex function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\psi(0) = 0$ and $\psi(x) \to \infty$ as $x \to \infty$ and (real-valued) random variable X, we denote its Orlicz norm by $\|X\|_\psi$, which is defined by $\|X\|_\psi := \mathsf{inf}\left\{C > 0 : \mathbb{E}\left[\psi\left(\frac{|X|}{C}\right)\right] \leqslant 1\right\}$. Since we are only concerned with polynomial and exponential tails we restrict ourselves to Orlicz norm induces by the class of function defined by (3.3). Evidently, as opposed to $\|X\|_p$, $\|X\|_{\psi_p}$ is always a non-negative non-random scalar. We do not abide to any convention to apply Orlicz norm to vector or matrices to avoid confusion.

For any vector $\boldsymbol{X}$, $\mathsf{diag}\,(\boldsymbol{X})$ is a diagonal matrix whose diagonal is the elements of $\boldsymbol{X}$. $\mathbb{1}(A)$ is an indicator function on the event A, i.e, $\mathbb{1}(A) = 1$ if A is true and 0 otherwise. We adopt the Landau big/small O, o notation and the "in probability" O_P and o_p analogues. We say that x is of the same order of y, $x \asymp y$, if both $x = O(y)$ and $y = O(x)$. We write $X \asymp_P Y$ if both $X = O_P(Y)$ and $Y = O_P(X)$. Unless stated otherwise, the asymptotics are taken as $T \to \infty$, where T is the time dimension, and the $o(1)$ and $o_P(1)$ are with respect to the limit as $T \to \infty$. We denote convergence in probability and in distribution by "$\xrightarrow{p}$" and "$\Rightarrow$", respectively.

2. Setup and Method

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

Assumption 1 (DGP)**.** *The process* $\{Y_{it} : 1 \leqslant i \leqslant n, t \geqslant 1\}$ *is generated by*

$$Y_{it} = \boldsymbol{\gamma}_i' \boldsymbol{X}_{it} + \underbrace{\boldsymbol{\lambda}_i' \boldsymbol{F}_t + U_{it}}_{=:R_{it}} \tag{2.1}$$

where $\boldsymbol{X}_{it}$ *is a* k*-dimensional observable (random) vector which may also include a constant term,* $\boldsymbol{F}_t$ *is a* r*-dimensional vector of common latent factors, and* U_{it} *is a zero mean idiosyncratic shock.*[4] *The unknown parameters are* $\boldsymbol{\gamma}_i \in \mathbb{R}^k$*, the factor loadings* $\boldsymbol{\lambda}_i$*, and the covariance matrix of the idiosyncratic shocks. Finally, we assume that* $\boldsymbol{X}_{it}$*,* $\boldsymbol{F}_t$ *and* U_{it} *are mutually uncorrelated.*

Remark 1. *In Assumption 1 we consider that* k*, the dimension of* $\boldsymbol{X}_{it}$ *is finite and fixed. Furthermore, the relation between* Y_{it} *and* $\boldsymbol{X}_{it}$ *is linear. This is for the sake of exposition. However, the theoretical results in this paper are written in terms of the consistency rate of the first-step estimation. Therefore, the DGP can be made much more general by just changing the rates.*

Remark 2. *The assumption that* $\boldsymbol{X}_{it}$*,* $\boldsymbol{F}_t$ *and* U_{it} *are mutually uncorrelated can be relaxed. Whenever* $\boldsymbol{X}_{it}$ *is correlated with* $\boldsymbol{F}_t$ *and* U_{it} *are correlated and the interest lies of the estimation of the parameters* $\boldsymbol{\gamma}_i$*,* $i = 1, \ldots, n$*, the method proposed by Pesaran (2006) can applied in the first-stage of the procedure considered in this paper and our theoretical results will follow. Nevertheless, we provide several examples below where the assumption that* $\boldsymbol{X}_{it}$*,* $\boldsymbol{F}_t$ *and* U_{it} *are mutually uncorrelated is reasonable.*

Example 1 (Asset Pricing Models)**.** *Suppose* Y_{it} *is the return of an asset* i *at time* t *and let* $\boldsymbol{X}_{it} := \boldsymbol{X}_t$ *be a set of* k *observable risk factors, such as the market returns and or Fama-French factors (Fama and French, 1993, 2015).* $\boldsymbol{F}_t$ *can be a set of additional, non observable, risk factors. Several asset pricing models, such as the Capital Asset Pricing Model (CAPM) of the Arbitrage Pricing Theory (APT) model, are nested into this general framework.*

Example 2 (Networks)**.** *Model* (2.1) *also complements the network specifications discussed in Barigozzi and Hallin (2016,2017b) and Barigozzi and Brownlees (2019). Furthermore, the test proposed here can be used to detect networks links as in Diebold and Yilmaz (2014) and Brownlees et al. (2020). For example,* Y_{it} *can be the (realized) volatility of financial assets and* $\boldsymbol{X}_{it} := \boldsymbol{X}_t$ *can be volatility factors as in Brito et al. (2018) and Andreou and Ghysels (2021).*

[4]For simplicity, we assume that all the units i have the same number of covariates (k). The framework can certainly accommodate situations where k_i is a function of i.

Example 3 (FAVAR)**.** *In the case where the index i represents a different dependent (endogenous) variable and U_{it} is a dependent process, model* (2.1) *turns out to be equivalent to the Factor Augmented Vector Autoregressive (FAVAR) model of Bernanke et al. (2005). In this case, $\boldsymbol{X}_{it}$ may also include lagged dependent variables.*

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) and Abadie et al. (2010) to evaluate the effects of regional policies. Model* (2.1) *is also in the heart of the* `FarmTreat` *method of Fan et al. (2020).*

2.2. **Three-Stage Method.** The method proposed here for estimation, inference and prediction consists of three stages where at the end of each stage, the covariance structure of the residuals is tested.

(1) For each $i \in \{1, \dots, n\}$ run the regression:

$$Y_{it} = \boldsymbol{\gamma}_i' \boldsymbol{X}_{it} + R_{it}, \quad t \in \{1, \dots, T\},$$

and compute $\widehat{R}_{it} := Y_{it} - \widehat{\boldsymbol{\gamma}}_i' \boldsymbol{X}_{it}$. The first stage may consist of a regression on a constant, a deterministic time trend and seasonal dummies, for instance, 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_{it} 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, the Pesaran's (2006) estimator can be also used whenever correlation between $\boldsymbol{X}_{it}$ and R_{it} is allowed. This will be discussed more in the subsequent sections.

(2) Write $\boldsymbol{R}_t := (R_{1t}, \dots, R_{nt})'$ and $\boldsymbol{R}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t + \boldsymbol{U}_t$. The second step consists of estimating $\boldsymbol{\Lambda}$ and $\boldsymbol{F}_t$ for $t = 1, \dots, T$ using $\widehat{\boldsymbol{R}}_t$ through principal component analysis (PCA) and compute

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

After estimating the factors and loadings, we apply our testing procedure to check for remaining covariance structure in $\boldsymbol{U}_t$. The second-step estimation

can be carried out also by dynamic factor models as in Barigozzi and Hallin (2016,2017,2020) or Barigozzi et al. (2020). In Section 4 we discuss the determination of the number of factors.

(3) Now, define $\widehat{\boldsymbol{U}}_{-it} := (\widehat{U}_{1t}, \ldots, \widehat{U}_{i-1,t}, \widehat{U}_{i+1,t} \ldots \widehat{U}_{nt})'$. The third step consists of a sparse regression to estimate the following model for each $i \in \{1, \ldots, n\}$:

$$\widehat{U}_{it} = \boldsymbol{\theta}_i' \widehat{\boldsymbol{U}}_{-it} + V_{it}, \qquad t \in \{1, \ldots, T\}.$$

(4) The regression in Step 3 provides useful augmentation for predicting the response variable Y_{it} (if Y_{it} is a variable to be predicted; see (2.3)), which reduces the error further from U_{it} to V_{it}. The sparsity of estimated $\boldsymbol{\theta}_i$ is useful for constructing partial correlation network or graphica model.

At the end of Steps 2 and 3, we can conduct the relevant inference on the structures of the covariance or partial covariance matrices. We can also provide updated prediction future outcomes. We detail those in the next subsection. Also note that the nonzero estimates of $\boldsymbol{\theta}_i$ shed light on the links among idiosyncratic components.[5]

2.3. **Estimators and Inference Procedure.** In a pure prediction exercise one is usually interested in the linear projection of Y_{it} onto $(\boldsymbol{X}_{it}', \boldsymbol{F}_t', \boldsymbol{U}_{-it}')'$, which results in the factor-augmented regression model (`FARM`)

$$Y_{it} = \boldsymbol{\gamma}_i' \boldsymbol{X}_{it} + \boldsymbol{\lambda}_i' \boldsymbol{F}_t + \boldsymbol{\theta}_i' \boldsymbol{U}_{-it} + \varepsilon_{it}, \qquad t \in \{1, \ldots, T\}, \tag{2.2}$$

for each given i, and can be predicted by

$$\widehat{Y}_{it} := \widehat{\boldsymbol{\gamma}}_i' \boldsymbol{X}_{it} + \widehat{\boldsymbol{\lambda}}_i' \widehat{\boldsymbol{F}}_t + \widehat{\boldsymbol{\theta}}_i' \widehat{\boldsymbol{U}}_{-it}; \qquad i \in \{1, \ldots, n\}. \tag{2.3}$$

This will be called `FarmPredict`. Note that model (2.2) is equivalent to using the predictors $X_{it}, \boldsymbol{Y}_{-it}$ and $\boldsymbol{F}_t$, which augment predictors $X_{it}, \boldsymbol{Y}_{-it}$ by using the common factors $\boldsymbol{F}_t$. The form in (2.2) mitigates the collinearity issues in high dimensions.

Model (2.2) also bridges factor regression ($\boldsymbol{\theta}_i = 0$) on one end and (sparse) regression on the other end with $\boldsymbol{\lambda}_i = \boldsymbol{\Lambda}_{-i}' \boldsymbol{\theta}_i$, where $\boldsymbol{\Lambda}_{-i}$ is the loading matrix without the i^{th} row. In the latter case, model (2.2) becomes a (sparse) regression model:

$$Y_{it} = \boldsymbol{\gamma}_i' \boldsymbol{X}_{it} + \boldsymbol{\theta}_i' \boldsymbol{R}_{-it} + \varepsilon_{it}, \qquad t \in \{1, \ldots, T\}. \tag{2.4}$$

[5]The three-stage procedure described here could be replaced by a single-step joint estimation. However, not only the computational burden will be much higher, but also the technical challenges will be greater. I believe that the simplicity of the method is more a blessing than a curse.

In this case, `FARM` specification as in (2.2) decorrelates the variables $\boldsymbol{R}_{-i}$ in (2.4). It makes the model selection consistency much easier to satisfy and forms the basis of `FarmSelect` in (Fan et al., 2020). Our contribution in this specific task is to allow heteroscadestic adjustments, resulting in the estimated data $\boldsymbol{R}_t$. In general, for `FARM` (2.2) with sparsity, `FarmPredict` chooses additional idiosyncratic components to enhance the prediction of the factor regression.

In other cases, the structure of the idiosyncratic components $\boldsymbol{U} = (U_1, \ldots, U_n)'$ is the objective of interest. An estimator for $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{U}_t \boldsymbol{U}_t')$ could be simply given by

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{T} \sum_{t=1}^{T} \widehat{\boldsymbol{U}}_t \widehat{\boldsymbol{U}}_t'. \tag{2.5}$$

In order to proper understand the (linear) relation between a pair (U_{it}, U_{jt}) of $\boldsymbol{U}_t$, a simple covariance estimate sometimes is not enough. In applications, it is often desirable to have a direct measure of how U_{it} and U_{jt} are connected. By direct connection, we meant the relation between those units removing the contribution of other variables of $\boldsymbol{U}_t$. For this purpose, we use the partial covariance between U_{it} and U_{jt}, defined for any pair $i, j \in \{1, \ldots, n\}$ as $\pi_{ij} := \mathbb{E}(V_{ijt} V_{jit})$, where $V_{ijt} := U_{it} - \mathsf{Proj}(U_{it} | \boldsymbol{U}_{-ij,t})$ and $\mathsf{Proj}(U_{it} | \boldsymbol{U}_{-ij,t})$ denotes the linear projection of U_{it} onto the space spanned by all the units except i and j, which we denote by $\boldsymbol{U}_{-ij,t}$. We suggest to estimate the partial covariance matrix $\boldsymbol{\Pi} := (\pi_{ij})$ by

$$\widehat{\boldsymbol{\Pi}} := (\widehat{\pi}_{ij}) \quad \text{and} \quad \widehat{\pi}_{ij} := \frac{1}{T} \sum_{t=1}^{T} \widehat{V}_{ij,t} \widehat{V}_{ji,t}, \tag{2.6}$$

where $\widehat{V}_{ijt}$ is the residual of the LASSO regression of $\widehat{U}_{it}$ onto $\widehat{\boldsymbol{U}}_{-ij,t}$ for $i, j \in \{1, \ldots, n\}$.

We also would like to conduct formal test on the population structure of $\boldsymbol{U}_t$. Specifically, we propose a test for the following null hypothesis:

$$\mathbb{H}_{\mathcal{D}}^{\Sigma} : \boldsymbol{\Sigma}_{\mathcal{D}} = \boldsymbol{\Sigma}_{\mathcal{D}}^0, \qquad \mathcal{D} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}, \tag{2.7}$$

for a given subset $\mathcal{D}$, where $\boldsymbol{\Sigma}_{\mathcal{D}}$ denotes the elements of $\boldsymbol{\Sigma}$ indexed by $\mathcal{D}$ and we allow $d := |\mathcal{D}|$ to diverge as $n, T \to \infty$. For example, to test if $\boldsymbol{\Sigma}$ is diagonal, $\mathcal{D}$ consists of all off diagonal elements and $\boldsymbol{\Sigma}_{\mathcal{D}}^0 = \boldsymbol{0}$. To test if $\boldsymbol{\Sigma}$ is block diagonal, $\mathcal{D}$ can be taken to the corresponding off-diagonal blocks. Similarly, for testing the structure on the partial covariance matrix

$$\mathbb{H}_{\mathcal{D}}^{\Pi} : \boldsymbol{\Pi}_{\mathcal{D}} = \boldsymbol{\Pi}_{\mathcal{D}}^0, \qquad \mathcal{D} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}. \tag{2.8}$$

The null hypotheses (2.7) and (2.8) nest several cases of interest. The most common would be to test for a diagonal or a block diagonal structure in $\mathbf{\Sigma}$ and/or $\mathbf{\Pi}$. But it also accommodates other structures.[6] The task of estimating $\mathbf{\Sigma}$ is well documented in literature even in high-dimensions; see, for example, Ledoit and Wolf (2004,2012,2017,2020), Fan et al. (2008), Lam and Fan (2009), or Fan et al. (2013).[7]

The challenges for testing (2.7) and (2.8) can be summarized as follows:

(1) As we allow for both n and d to diverge to infinite as T grows, sometimes at a faster rate, we have a high-dimensional test where some sort of 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 (asymptotic) correct test size.

(2) We do not observe $\{\boldsymbol{U}_t\}$ or $\{V_{ij,t}\}$. Instead, we have an estimate of both from a postulated model on observable random variables. Therefore, the estimation error must be taken into account to claim some sort of asymptotic properties of the test. In fact, it is not uncommon to obtain estimates of both $\{\boldsymbol{U}_t\}$ and $\{V_{ij,t}\}$ from a multi-stage estimation procedure as we illustrate later.

We propose to test (2.7) using the statistic

$$S_{\mathcal{D}}^{\Sigma} := \|\sqrt{T}(\widehat{\mathbf{\Sigma}}_{\mathcal{D}} - \mathbf{\Sigma}_{\mathcal{D}}^{0})\|_{\mathsf{max}}. \tag{2.9}$$

The quantiles of $S_{\mathcal{D}}^{\Sigma}$ are approximated by a Gaussian bootstrap. To describe the procedure, let $\mathbf{\Upsilon}_{\Sigma}$ denote the $(d \times d)$ covariance matrix for the vectorized submatrix $(\widetilde{\sigma}_{ij})_{(i,j)\in\mathcal{D}}$, where $\widetilde{\sigma}_{ij} := \frac{1}{T}\sum_{t=1}^{T} U_{i,t}U_{j,t}$. Since the process $\{\boldsymbol{U}_t\}$ might present some form of temporal dependence (refer to Assumption 3(c)) we estimate $\mathbf{\Upsilon}_{\Sigma}$ 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 (3.11) and bandwidth $h > 0$, $\mathbf{\Upsilon}_{\Sigma}$ is estimated by

$$\widehat{\mathbf{\Upsilon}}_{\Sigma} := \sum_{|\ell|<T} \mathcal{K}(\ell/h)\widehat{\boldsymbol{M}}_{\Sigma,\ell} \qquad \text{and} \qquad \widehat{\boldsymbol{M}}_{\Sigma,\ell} := \frac{1}{T}\sum_{t=\ell+1}^{T} \widehat{\boldsymbol{D}}_{\Sigma,t}\widehat{\boldsymbol{D}}_{\Sigma,t-\ell}', \tag{2.10}$$

where $\widehat{\boldsymbol{D}}_{\Sigma,t}$ is a d-dimensional vector with entries given by $\widehat{U}_{it}\widehat{U}_{jt} - \widehat{\sigma}_{ij}$ for $(i,j) \in \mathcal{D}$, where $\widehat{\sigma}_{ij}$ is the (i,j) element of $\widehat{\mathbf{\Sigma}}$ defined in (2.5). Finally, let $c_{\Sigma}^{*}(\tau)$ be the τ-quantile of the Gaussian bootstrap $S_{\mathcal{D}}^{*} := \|\boldsymbol{Z}_{\Sigma}^{*}\|_{\infty}$ and $\boldsymbol{Z}_{\Sigma}^{*}|\boldsymbol{X},\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \widehat{\mathbf{\Upsilon}}_{\Sigma})$.

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

[7]See Ledoit and Wolf (2021a) for a recent survey or the book by Fan et al. (2020).

Theorem 4 demonstrates the validity of Gaussian bootstrap procedure described above, i.e., it states conditions under which the τ-quantile of the test statistic (2.9) can be approximated by $c^*_\Sigma(\tau)$ in the appropriate sense.

Similarly, the test statistic for (2.8) is given by

$$S^\Pi_{\mathcal{D}} := \|\sqrt{T}(\widehat{\mathbf{\Pi}}_{\mathcal{D}} - \mathbf{\Pi}^0_{\mathcal{D}})\|_{\mathsf{max}}. \tag{2.11}$$

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

$$\widehat{\mathbf{\Upsilon}}_\Pi := \sum_{|\ell|<T} \mathcal{K}(\ell/h)\widehat{\boldsymbol{M}}_{\Pi,\ell}; \qquad \widehat{\boldsymbol{M}}_{\Pi,\ell} := \frac{1}{T}\sum_{t=\ell+1}^T \widehat{\boldsymbol{D}}_{\Pi,t}\widehat{\boldsymbol{D}}'_{\Pi,t-\ell}, \tag{2.12}$$

where $\widehat{\boldsymbol{D}}_{\Pi,t}$ is a d-dimensional vector with entries given by $\widehat{V}_{ij,t}\widehat{V}_{ji,t} - \widehat{\pi}_{ij}$ for $(i,j) \in \mathcal{D}$. Also, let $c^*_\Pi(\tau)$ be the τ-quantile of the Gaussian bootstrap $S^*_{\mathcal{D}} := \|\boldsymbol{Z}^*_\Pi\|_\infty$ and $\boldsymbol{Z}^*_V|\boldsymbol{X},\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0},\widehat{\mathbf{\Upsilon}}_V)$.

Theorem 5 demonstrate the validity of Gaussian bootstrap, i.e., it states conditions under which the τ-quantile of the test statistic (2.11) can be approximated by $c^*_\Pi(\tau)$ in the appropriate sense.

3. Theoretical Results

In this section we collect all the theoretical guarantees for the estimation of the model (2.1) by using the proposed three-stage method described above. Specifically, Section 3.1 deals with estimation and Section 3.2 with inference on the (partial) covariance structure of $\mathbf{\Pi}$.

To present the next results it is convenient to use a more compact notation. For each $i = 1,\dots,n$, we define the T-dimensional vectors $\boldsymbol{Y}_i := (Y_{i1},\dots,Y_{iT})'$ and $\boldsymbol{U}_i := (U_{i1},\dots,U_{iT})'$. We also define the $(T\times k)$ matrix of covariates $\boldsymbol{X}_i := (\boldsymbol{X}_{i1},\dots,\boldsymbol{X}_{iT})'$ for each $i = 1,\dots,n$ and the $(T\times r)$ matrix of factors $\boldsymbol{F} := (\boldsymbol{F}_1,\dots,\boldsymbol{F}_T)'$ such that (2.1) can be represented as

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\boldsymbol{\gamma}_i + \boldsymbol{F}\boldsymbol{\lambda}_i + \boldsymbol{U}_i, \qquad i = 1,2,\dots,n, \\ &= \boldsymbol{X}_i\boldsymbol{\gamma}_i + \boldsymbol{R}_i, \end{aligned} \tag{3.1}$$

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

When no confusion is likely to arise, we also define for each $t = 1,\dots,T$, the n-dimensional vectors $\boldsymbol{Y}_t := (Y_{1t},\dots,Y_{nt})'$ and $\boldsymbol{U}_t := (U_{1t},\dots,U_{nt})'$; and the nk-dimensional vector $\boldsymbol{X}_t := (\boldsymbol{X}'_{1t},\dots,\boldsymbol{X}'_{nt})'$. Also, set the $(n\times nk)$ block diagonal

matrix $\mathbf{\Gamma}$ whose block diagonal is given by $(\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_n')$ and the $(n \times r)$ loading matrix $\mathbf{\Lambda} := (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)'$. Then, (2.1) can also be represented as panel time series

$$\begin{aligned} \boldsymbol{Y}_t &= \mathbf{\Gamma} \boldsymbol{X}_t + \mathbf{\Lambda} \boldsymbol{F}_t + \boldsymbol{U}_t, \qquad t = 1, 2, \ldots, T \\ &= \mathbf{\Gamma} \boldsymbol{X}_t + \boldsymbol{R}_t, \end{aligned} \tag{3.2}$$

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

3.1. **Estimation.** We consider the following set of assumptions

Assumption 2 (**Factor Model**). *Assume:*

(a) $\mathbb{E}(\boldsymbol{F}_t) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{F}_t \boldsymbol{F}_t') = \boldsymbol{I}_r$ *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* $n \to \infty$*;*
(c) $\|\mathbf{\Sigma} - \mathbf{\Lambda}\mathbf{\Lambda}'\| = O(1)$*; and*
(d) $\|\mathbf{\Lambda}\|_{\mathsf{max}} \leqslant C$.

Remark 3. *Assumption 2 is standard in the literature. Note also that the assumption* $\mathbb{E}(\boldsymbol{F}_t) = \mathbf{0}$ *is not restrictive as our approach considers a first-step estimation which may include a constant in the set of regressors. It is also needed for identifiability.*

In order to present the results in a unified manner for both light and heavy tail distributions, we state the next assumption in terms the Orlicz norm of the random variables. Specifically, since we are only concerned with polynomial and exponential tails we define the following subset of unbounded, convex, real-valued functions that vanish at the origin:

$$\begin{aligned} \Psi := \{\psi_p : \mathbb{R}^+ \to \mathbb{R}^+ : \psi_p(x) = x^p, p \geqslant 6 \\ \text{or } \psi_p(x) = x\mathbb{1}[0 \leqslant x^p < (1-p)/p] + [\mathsf{exp}(x^p) - 1]\mathbb{1}[x^p \geqslant (1-p)/p], p > 0\}. \end{aligned} \tag{3.3}$$

Also, for each $\psi_p \in \Psi$, we define $\psi_{p+}(x) := x^{p+\epsilon}$ for some $\epsilon > 0$ if $\psi_p(x) = x^p$; otherwise (for the exponential case) $\psi_{p+} := \psi_p$.

Assumption 3 (**Moments and Dependency**). *There exists a constant* $C < \infty$ *and function* $\psi_p \in \Psi$ *defined in* (3.3) *such that, for all* $i = 1, \ldots, n$*;* $\ell = 1, \ldots, k$*;* $s, t = 1, \ldots, T$*; and* $j = 1, \ldots, r$*:*

(a) $\|X_{it\ell}\|_{\psi_{p+}} \leqslant C$, $\|U_{it}\|_{\psi_{p+}} \leqslant C$, $\|F_{j,t}\|_{\psi_{p+}} \leqslant C$*;*
(b) $\|\|(\boldsymbol{X}_i'\boldsymbol{X}_i/T)^{-1}\|_{\mathsf{max}}\|_{\psi_{p+}} \leqslant C$*;*

(c) The process $\{(\boldsymbol{X}'_{S,t}, \boldsymbol{F}'_t, \boldsymbol{U}'_t)', t \in \mathbb{Z}\}$ *is weakly stationary with strong mixing coefficient* α *satisfying* $\alpha(m) \leqslant \mathsf{exp}(-2cm)$ *for some* $c > 0$ *and for all* $m \in \mathbb{Z}$*, where* $\boldsymbol{X}_{S,t}$ *denotes the vector* $\boldsymbol{X}_t$ *after excluding all deterministic (non-random) components.*

(d) $\|n^{-1/2}\left[\boldsymbol{U}'_s\boldsymbol{U}_t - \mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)\right]\|_{\psi_{p+}} \leqslant C$*;*

(e) $\|n^{-1/2}\sum_{i=1}^{n}\lambda_{j,i}U_{it}\|_{\psi_{p+}} \leqslant C$*; and*

(f) $\mathsf{log}\, n = o\left(\frac{T^{p/4}}{[\mathsf{log}\, T]^2}\right)$.

A few words about Assumption 3 are in order. Assumptions (3.a) and (3.c) allow us to apply a Marcinkiewicz-Zygmund type inequality for partial sums to deal with the polinomial tails (Rio (1994) and Doukhan and Louhichi (1999)) and a Bernstein inequality (Merlevède et al. (2009) - Theorem 2) to control exponential tails. Moreover, Assumption (3.c) excludes the deterministic component of $\boldsymbol{X}_t$ to accommodate possibly non-random non-stationary (but uniformly bounded by (a)) covariates. Assumption (3.d) is only used to prove results for the first-stage estimation in case it is performed by ordinary least-squares (Theorem 1). Assumption (3.d) controls for the level of cross-sectional dependence among the units. As we allow the number of units to diverge with T, some sort of control on this quantity is necessary which is not implied by (3.*c*). Assumption (3.e) has a similar role to (3.d) but in terms of linear combinations of the idiosyncratic components. Assumption (3.e) only bounds the growth rate of the number of units n to be sub-exponential with respect to T. As a matter of fact, this assumption is only binding in the exponential tail case, otherwise the rate conditions imposed in the theorems below imply (3.e).

For each $i = 1, \ldots, n$, let $\boldsymbol{R}_i := \boldsymbol{F}\boldsymbol{\lambda}_i + \boldsymbol{U}_i$ denote the unobservable error term in (3.1), $\widehat{\boldsymbol{\gamma}}_i$ the least-squares estimator of $\boldsymbol{\gamma}_i$ and $\widehat{\boldsymbol{R}}_i := \boldsymbol{Y}_t - \boldsymbol{X}_t\widehat{\boldsymbol{\gamma}}_i$ the vector of residuals. Also set $\widehat{\boldsymbol{R}} := (\widehat{\boldsymbol{R}}_1, \ldots, \widehat{\boldsymbol{R}}_n)'$ and $\boldsymbol{R} := (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)'$. We must control for the least-squares 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 $\widehat{\boldsymbol{R}} - \boldsymbol{R}$ when the first-stage is conducted by OLS in a linear setup. Note that in this case we assume that $\boldsymbol{X}_{it}$, $\boldsymbol{F}_t$ and U_{it} are mutually uncorrelated.

Theorem 1. *Under Assumption 3(a)-(d)*

$$\max_{i,t}\|\widehat{R}_{it} - R_{it}\|_{\psi_{p/4}} \leqslant \frac{C_{k,\psi}}{\sqrt{T}} \quad \text{and} \quad \|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\mathsf{max}} = O_P\left[\frac{\psi^{-1}_{p/4}(nT)}{\sqrt{T}}\right],$$

where the $C_{k,\psi}$ *is a constant only depending on* k *and* ψ_p*.*

Remark 4. *In case the first step of the method involves more complicated estimation, such as Pesaran's (2006) method, instrumental variables, or LASSO, for example, we write* $\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max} = O_P(\omega)$*, where* $\omega := \omega_{n,T}$ *is a non-negative sequence. This will be used in the next theorems.*

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

$$\boldsymbol{Y} = \boldsymbol{\Gamma}\boldsymbol{X} + \boldsymbol{\Lambda}\boldsymbol{F}' + \boldsymbol{U}. \tag{3.4}$$

Notice that $\widehat{\boldsymbol{R}} = \boldsymbol{\Lambda}\boldsymbol{F}' + \widetilde{\boldsymbol{U}}$ where $\widetilde{\boldsymbol{U}} := \boldsymbol{U} + \widehat{\boldsymbol{R}} - \boldsymbol{R}$ and $(\boldsymbol{\Lambda}, \boldsymbol{F})$ can be estimated by Principal Component Analysis (PCA), which minimizes

$$q(\boldsymbol{\Lambda}, \boldsymbol{F}) := \|\widehat{\boldsymbol{R}} - \boldsymbol{\Lambda}\boldsymbol{F}'\|_F^2, \tag{3.5}$$

with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{F}$, subject to the normalization $\boldsymbol{F}'\boldsymbol{F}/T = \boldsymbol{I}_r$. The solution $\widehat{\boldsymbol{F}}$ is the matrix whose columns are $\sqrt{T}$ times r eigenvectors of the top r eigenvalues of $\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{R}}$ and $\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{R}}\widehat{\boldsymbol{F}}/T$.

Since we do not directly observe $\boldsymbol{U}$, in the third step of our estimation procedure we use $\widehat{\boldsymbol{U}} := \widehat{\boldsymbol{R}} - \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}'$ instead. Therefore, we must control of the estimation error in the factor model given by $(n \times T)$ matrix $\widehat{\boldsymbol{U}} - \boldsymbol{U}$ which is the main purpose of Theorem 2 below. Also, it is well know fact that the loading matrix $\boldsymbol{\Lambda}$ and the factors $\boldsymbol{F}$ are not separably identified since $\boldsymbol{\Lambda}\boldsymbol{F}_t = \boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{F}_t$ for any matrix $\boldsymbol{H}$ such that $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_r$. If we let $\boldsymbol{H} := T^{-1}\boldsymbol{V}^{-1}\widehat{\boldsymbol{F}}'\boldsymbol{F}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$, where $\boldsymbol{V}$ is the $(r \times r)$ diagonal matrix containing the r largest eigenvalues of $\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{R}}/T$ in decreasing order, we have that $\boldsymbol{H}\boldsymbol{F}_t$ is identified as $\boldsymbol{\Lambda}\boldsymbol{F}_t$ is identified.

The result below first appeared in Bai (2003) for the case of fixed (n, T), and was further extended to hold uniformly in $(i \leqslant n, t \leqslant T)$ by Fan et al. (2013). Fan et al. (2020) makes the conditions modular. However, both consider the case when the factor model is estimated using the true data as opposed to an "estimated" one as in our case. Therefore, the next result is a generalization that takes into account that pre-estimation error term and quantifies how the error impact on the precision of factor analysis.

Theorem 2. *Let* $\omega := \omega_{n,T}$ *be a non-negative sequence such that* $\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max} = O_P(\omega)$*. Then, under Assumptions 1 -3 and* $\psi_p^{-1}(n^2)/\sqrt{T} + \psi_p^{-1}(nT)\omega = O(1)$*, we have that*

(a)

$$\max_{t\leqslant T}\|\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t\|_2 = O_P\left[\frac{1}{\sqrt{T}} + \frac{\psi_p^{-1}(T)}{\sqrt{n}} + \omega\psi_{p/2}^{-1}(nT)\right],$$

(b)

$$\max_{i\leqslant n}\|\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i\|_2 = O_P\left[\frac{\psi_{p/2}^{-1}(n)}{\sqrt{T}} + \frac{1}{\sqrt{n}} + \omega\right],$$

(c)

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P\left[\frac{\psi_p^{-1}(n)\psi_p^{-1}(T)}{\sqrt{T}} + \frac{\psi_p^{-1}(T)}{\sqrt{n}} + \omega\psi_{p/2}^{-1}(nT)\right].$$

By setting $\omega = 0$, i.e., no estimation error in the first step, we recover Theorem 4 and Corollary 1 in Fan et al. (2013) under sub-Gaussian assumption. Note that in order to have the error $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max}$ vanishing in probability we must have the pre-estimation error $\|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max}$ of order (in probability) smaller than $1/\psi_{p/2}^{-1}(nT)$.

We have decided not to replace ω in Theorem 2 with the rate obtained in Theorem 1 as the latter only applies to the least square 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 (Fan et al., 2017) if the number of features k is comparable or even larger than T and the tail of the distribution is heavy. By stating the Theorem 2 in terms of a generic rate, it is easier to account for the effect of a different estimator. By combining Theorems 1 and 2 we have the following corollary

Corollary 1. *Under the same assumptions of Theorems 1 and 2, for the OLS used in the first-stage to obtain* $\widehat{\boldsymbol{R}}$*, we have*

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P\left[\frac{\psi_{p/6}^{-1}(nT)}{\sqrt{T}} + \frac{\psi_p^{-1}(T)}{\sqrt{n}}\right].$$

In particular for the sub-Gaussian case ($\psi(x) = \mathsf{exp}(x^2) - 1$) *we have*

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P\left[\frac{[\mathsf{log}(nT)]^3}{\sqrt{T}} + \sqrt{\frac{\mathsf{log}\, T}{n}}\right],$$

and for polynomial tails ($\psi(x) = x^p$)

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P\left[\frac{n^{6/p}}{T^{1/2-6/p}} + \frac{T^{1/p}}{\sqrt{n}}\right].$$

For notational convenience, for each $i \in \{1, \ldots, n\}$, consider the split $\boldsymbol{U}' = (\boldsymbol{U}_i, \boldsymbol{U}_{-i})$ where $\boldsymbol{U}_i$ is a T-dimensional vector and $\boldsymbol{U}_{-i}$ a $T \times (n-1)$-dimensional matrix. Analogously, we split $\widehat{\boldsymbol{U}}' = (\widehat{\boldsymbol{U}}_i, \widehat{\boldsymbol{U}}_{-i})$. Then for a the penalized parameter $\xi \geqslant 0$, the LASSO objective function can be written for each $i \in \{1, \ldots, n\}$

$$L(\boldsymbol{\theta}) + \text{Penalty}(\boldsymbol{\theta}) := \frac{1}{T}\|\widehat{\boldsymbol{U}}_i - \widehat{\boldsymbol{U}}_{-i}\boldsymbol{\theta}\|^2 + \xi\|\boldsymbol{\theta}\|_1. \tag{3.6}$$

To ensure a consistent estimation of $\boldsymbol{\theta}$, a sort of restricted strong convexity of the objective function is required when $n > T$. This in turns is ensured, in the case of a quadratic loss, by bounding the minimum eigenvalue on $\widehat{\boldsymbol{U}}'_{-i}\widehat{\boldsymbol{U}}_{-i}/T$ away from zero restrict to a cone (refer to Negahban et al. (2012) or Fan et al. (2020) for a thorough discussion). Here, we adopt the compatibility constant defined in van de Geer and Bühlmann (2009). For an index $\mathcal{S} \subseteq \{1, \ldots, n\}$ and any n-dimensional vector $\boldsymbol{v}$, let $\boldsymbol{v}_{\mathcal{S}}$ be the vector containing only the elements of the vector $\boldsymbol{v}$ indexed by $\mathcal{S}$. Thus, $\#\boldsymbol{v}_{\mathcal{S}} = \#\mathcal{S}$ and $\mathcal{S}^c := \mathcal{S}\backslash\{1, \ldots, n\}$ is the complement of $\mathcal{S}$.

Definition 1. *For an $n \times n$ matrix $\boldsymbol{M}$, a set $\mathcal{S} \subseteq \{1, \ldots, n\}$ and a scalar $\zeta \geqslant 0$, the compatibility constant is given by*

$$\kappa(\boldsymbol{M}, \mathcal{S}, \zeta) := \inf\left\{ \frac{\|\boldsymbol{x}\|_{\boldsymbol{M}}\sqrt{|\mathcal{S}|}}{\|\boldsymbol{x}_{\mathcal{S}}\|_1} : \boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leqslant \xi\|\boldsymbol{x}_{\mathcal{S}}\|_1 \right\}, \tag{3.7}$$

where $\|\boldsymbol{x}\|_{\boldsymbol{M}} = \boldsymbol{x}'\boldsymbol{M}\boldsymbol{x}$. Moreover, we say that $(\boldsymbol{M}, \mathcal{S}, \zeta)$ satisfies the compatibility condition if $\kappa(\boldsymbol{M}, \mathcal{S}, \zeta) > 0$.

Notice that the square of the compatibility constant is close related to the minimum of the ℓ_1-norm of the eigenvalues of $\boldsymbol{\Sigma}$ restricted to a cone in $\mathbb{R}^n$. Let t $\mathcal{S}_{0,i} := \{j : \theta_{i,j} \neq 0\}$ and $\kappa_i := \kappa\left[\mathbb{E}(\boldsymbol{U}'_{-i}\boldsymbol{U}_{-i})/T), \mathcal{S}_{0,i}, 3\right]$. Set the maximum non-sparsity level and minimum compatibility constant

$$s_0 := \max_{i \leqslant n} |\mathcal{S}_{0,i}| \quad \text{and} \quad \kappa_0 := \min_{i \leqslant n} \kappa_i. \tag{3.8}$$

The next result shows the ℓ_1 and ℓ_2 rate results for the Lasso estimator (3.6) 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 $\eta := \eta_{n,T}$ be a non-negative sequence such that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P(\eta)$. Assume that Assumption 3 holds and*

$$s_0 = O\left[\kappa_0 \left(\eta\left[\psi_p^{-1}(nT) + \eta\right] + \frac{\psi_{p/2}^{-1}(n^2)}{\sqrt{T}}\right)^{-1}\right]. \tag{3.9}$$

For every $\epsilon > 0$ there is a constant $0 < C_\epsilon < \infty$ such that if the penalty parameter is set to

$$\xi = C_\epsilon \left[\frac{\psi_{p/2}^{-1}(n)}{\sqrt{T}} + \eta\psi_p^{-1}(T)\right], \tag{3.10}$$

then for any minimizer $\widehat{\boldsymbol{\theta}}_i$ of (3.6), with probability at least $1 - \epsilon$:

$$T^{-1}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\boldsymbol{U}'_{-i}\boldsymbol{U}_{-i}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) + \xi\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \leqslant 8\frac{\xi^2 s_0}{\kappa_0}. \quad i \in \{1, \ldots, n\}.$$

Remark 5. *Notice that we apply the compatibility condition on the non-random covariance matrix $\mathbb{E}(\boldsymbol{U}'_{-i}\boldsymbol{U}_{-i})/T$ instead of the estimated random covariance matrix $\widehat{\boldsymbol{U}}'_{-i}\widehat{\boldsymbol{U}}_{-i}/T$ or the "unobservable" random matrix $\boldsymbol{U}'_{-i}\boldsymbol{U}_{-i}/T$. Careful review of the proofs reveals that the same is true for the gradient of the objective function that defines our parameter via a first order condition.*

Once again, we purposely avoided to replace η in Theorem 3 with the rate of Corollary 1 to make it readily applicable to the case when a different type of factor models was used or, as a matter of fact, any other pre-estimation procedure. By plugging the rate of Corollary 1 into η we have the next corollary

Corollary 2. *If η defined in Theorem 3 is taken to be rate given by Corollary 1 and the compatibility condition holds, i.e.: $\kappa_0 \geqslant C > 0$ then under the conditions of the Theorem 3:*

$$\max_{i \leqslant n} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 = O_P\left[\left(\frac{\psi_p^{-1}(T)\psi_{p/6}^{-1}(nT)}{\sqrt{T}} + \frac{\psi_{p/2}^{-1}(T)}{\sqrt{n}}\right) s_0\right].$$

3.2. **Inference on Covariance and Precision 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 2.3. In particular, we consider the kernel $\mathcal{K}(\cdot)$ appearing in the covariance estimator defined by (2.10) belongs to the

class defined in Andrews (1991) which we reproduce below for convenience

$$\mathbb{K} := \{f : \mathbb{R} \to [-1,1] : f(0) = 1, f(x) = f(-x), \forall x \in \mathbb{R}, \int f^2(x)dx < \infty, f \text{ is continuous}\}. \tag{3.11}$$

This includes most of the well-known kernel used in the literature. To avoid confusion, it is worth to point 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 next result shows how accurately the elements of the covariance matrix are estimated and validates the bootstrap method.

Theorem 4. *Let $\eta := \eta_{n,T}$ and $\nu := \nu_{n,T}$ be non-negative sequence such that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathsf{max}} = O_P(\eta)$ and $\mathsf{max}_{i,t} \|\widehat{R}_{it} - R_{it}\|_{\psi_p} = O(\nu)$ and $\mathcal{K}(\cdot) \in \mathbb{K}$. Under Assumptions 1–3, if further*

(a) $\{\boldsymbol{U}_t\}$ is fourth-order stationary process

(b) $\|\mathsf{diag}(\boldsymbol{\Upsilon}_\Sigma)\|_\infty \geqslant \underline{c}$ for some $\underline{c} > 0$

(c) As $h, n, T \to \infty$:

(c.1) $\frac{(\log n)^{7/6}\psi_{p/2}^{-1}(n)}{T^{1/6}} + \frac{\sqrt{\log T \log n}\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(T^{1/4})}{T^{1/4}} = o(1)$

(c.2) $(\log n)^3 h\left[\eta(\psi_p^{-1}(nT))^3 + \psi_{p/4}^{-1}(n^4)/\sqrt{T}\right] = o(1)$

(c.3) $(\log n)^3\left(\sqrt{T}\eta^2 + \frac{r_1}{\sqrt{T}} + \frac{r_2}{\sqrt{n}} + r_3\nu\right) = o(1)$,

where the rates r_1, r_2, r_3 are defined in Lemma C.10 and $h > 0$ is the bandwidth parameter of the covariance estimator defined in (2.10); then

$$\|\widehat{\boldsymbol{\Upsilon}}_\Sigma - \boldsymbol{\Upsilon}_\Sigma\|_{\mathsf{max}} = O_P\left[h\left(\eta\left[\psi_p^{-1}(nT)\right]^3 + \psi_{p/4}^{-1}(n^4)/\sqrt{T}\right)\right] = o(1),$$

and $\mathsf{sup}_{\mathcal{D}}\,\mathsf{sup}_{\tau\in(0,1)} |\mathbb{P}\left[S_{\mathcal{D}}^{\Sigma} \leqslant c_{\Sigma}^{*}(\tau)\right] - \tau| = o(1)$, where the first supremum is over all null hypotheses of the form (2.7) indexed by $\mathcal{D} \in \{1:n\} \times \{1:n\}$.

Remark 6. *The rate assumptions (c.1)–(c.3) in Theorem 4 seem over complicated. However, they are a direct consequence of having the first and second step estimation error rates, ν and η respectively, explicitly appearing in the final rate and the general tail condition through the $\psi_p(\cdot)$ function. It allows the practitioner to directly adjust the final rate should (s)he prefer to employ different intermediate estimators. For instance, a LASSO estimator in the first step in case the number of covariates k is large enough or estimate the factor model by PCA variants.*

Remark 7. *Careful review of Theorem 4's proof reveals that (c.1) traces back to the Gaussian Approximation of the (unobservable) process* $\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{U}_t\boldsymbol{U}_t' - \mathbb{E}\boldsymbol{U}_t\boldsymbol{U}_t'\right\}_{T\geqslant 1}$*; whereas (c.3) controls for the difference between* $\boldsymbol{U}_t - \widehat{\boldsymbol{U}}_t$ *and, therefore, takes into account the estimation error of the first and second steps. Note the presence of* ν *and* η *in (c.3) which are absent in (c.1). Finally, (c.2) make sure that the bootstrap constructed in terms of the estimated covariance matrix is close to the bootstrap based in the true covariance. Note the presence of the bandwidth parameter* h *in (c.2).*

If we were to specialized Theorem 4 to the sub-Gaussian case and incorporate the rates obtain in Theorem 1 and Corollary 1 we have the following cleaner Corollary.

Corollary 3. *Consider the sub-Gaussian where* $\psi_2(x) = \mathsf{exp}(x^2)$*. Suppose that the Assumptions 1-3 and conditions* (a) *and* (b) *of Theorem 4 hold. If the rates* ν *and* η *are set to be rates given by Theorem 1 and Corollary 1, respectively, then the conclusion of Theorem 4 holds provided that as* $h, n, T \to \infty$*:*

(a) $\mathsf{log}\, n = o(T^{1/18})$
(b) $h\left[\frac{(\mathsf{log}\, n)^{15/2}}{\sqrt{T}} + \frac{(\mathsf{log}\, n)^5}{\sqrt{n}}\right] = o(1)$
(c) $\frac{(\mathsf{log}\, n)^3(\mathsf{log}\, T)\sqrt{T}}{n} = o(1)$.

Remark 8. *In order to establish the rate of convergence in the last result of Theorem 4 we need an upper bound on the tails of the pre-estimation error namely* $\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|_{max}$*. In fact, we need to control the tails of the factor model estimation to establish uniform bounds on* $\|\widehat{U}_{it} - U_{it}\|_\psi$*, which translate into obtain bounds on* $\mathsf{max}_{jt}\,\|\widehat{F}_{jt} - F_{jt}\|_\psi$ *and* $\mathsf{max}_{ji}\,\|\widehat{\lambda}_{ji} - \lambda_{ji}\|_\psi$*.*

The next theorem shows how well the elements of the partial autocovariance matrix is estimated and gives the conditions under which the bootstrap test works. Note that in calculation the partial covariance (2.6), we used the residual $\widehat{V}_{ijt}$ of the LASSO regression of $\widehat{U}_{it}$ onto $\widehat{\boldsymbol{U}}_{-ij,t}$ for $i, j \in \{1, \ldots, n\}$. This adds extra technical challenges to the proof. Let $\widetilde{\mathcal{S}}_{0,ij}$ denote the active set of the projection of U_{it} onto $\boldsymbol{U}_{-ij,t}$ for $i, j \in \{1, \ldots, n\}$ and $\widetilde{\kappa}_{ij} := \kappa\left[\mathbb{E}(\boldsymbol{U}_{-ij}'\boldsymbol{U}_{-ij})/T), \widetilde{\mathcal{S}}_{0,ij}, 3\right]$. Set the maximum non-sparsity level and minimum compatibility constant

$$\widetilde{s}_0 := \underset{1\leqslant i\neq j\leqslant n}{\mathsf{max}} |\widetilde{\mathcal{S}}_{0,ij}| \quad \text{and} \quad \widetilde{\kappa}_0 := \underset{1\leqslant i\neq j\leqslant n}{\mathsf{min}} \widetilde{\kappa}_{ij}. \tag{3.12}$$

We assume the same sparsity structure as in Theorem 3.

Theorem 5. *Let $\eta := \eta_{n,T}$ and $\nu := \nu_{n,T}$ be non-negative sequence such that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P(\eta)$ and $\max_{i,t} \|\widehat{R}_{it} - R_{it}\|_{\psi_p} = O(\nu)$, $\mathcal{K} \in \mathbb{K}$ where $\mathcal{K}$ appears in (2.12) and $\mathbb{K}$ defined by (3.11), the LASSO regularization parameter $\widetilde{\xi}$ set as in (3.10), and the sparsity level obeying (3.9) with s_0 and κ_0 replaced by $\widetilde{s}_0$ and $\widetilde{\kappa}_0$, respectively. Under Assumptions 2-4, if further*

(a) $\{\boldsymbol{U}_t\}$ is fourth-order stationary process

(b) $\|\mathrm{diag}(\boldsymbol{\Upsilon}_\Pi)\|_\infty \geqslant \underline{c}$ for some $\underline{c} > 0$

(c) As $n, T \to \infty$:

(c.1) $\frac{(\log n)^{7/6}\psi_{p/2}^{-1}(n)}{T^{1/6}} + \frac{\sqrt{\log T \log n}\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(T^{1/4})}{T^{1/4}} = o(1)$

(c.2) $(\log n)^3 h \left(\widetilde{s}_0 \left[\eta + \xi\psi_p^{-1}(n)\right] \left[\widetilde{s}_0 \psi_p^{-1}(nT)\right]^3 + \widetilde{s}_0 \frac{\psi_{p/4}^{-1}(n^4)}{\sqrt{T}} \right)] = o(1)$

(c.3) $(\log n)^3 \left(\widetilde{s}_0^2 \left\{ \frac{r_1}{\sqrt{T}} + \frac{r_2}{\sqrt{n}} + r_3\nu + \widetilde{\xi}\psi_p^{-1}(n) + \sqrt{T}\left[\eta + \xi\psi_p^{-1}(n)\right]^2 \right\} \right) = o(1)$,

where the rates r_1, r_2, r_3 are defined in Lemma C.10 and $h > 0$ is the bandwidth parameter of the covariance estimator defined in (2.12); then

$$\|\widehat{\boldsymbol{\Upsilon}}_\Pi - \boldsymbol{\Upsilon}_\Pi\|_{\max} = O_P\left(h \left\{ \widetilde{s}_0[\eta + \widetilde{\xi}\psi_p^{-1}(n)][\widetilde{s}_0\psi_p^{-1}(nT)]^3 + \widetilde{s}_0 \frac{\psi_{p/4}^{-1}(n^4)}{\sqrt{T}} \right\} \right) = o(1)$$

and $\sup_{\mathcal{D}} \sup_{\tau\in(0,1)} |\mathbb{P}(S_{\mathcal{D}}^\Pi \leqslant c_\Pi^*(\tau)) - \tau| = o(1)$ under $\mathbb{H}_0^\Pi$, where the first supremum is over all null hypotheses of the form (2.8) indexed by $\mathcal{D} \in \{1 \times n\} \times \{1 \times n\}$.

Similar comments as in Remarks 6-8 and Corollary 3 apply to Theorem 5 as well.

Remark 9. *As opposed to the case of testing covariance, when testing partial covariance in high-dimensional setup, the sparse structure plays a role in terms of $\widetilde{s}_0$ appearing in the rates (c.2) and (c.3). Therefore, these assumptions restricts the cases when the proposed partial covariance test has the correct asymptotic size. For instance, in the case of a complete dense partial covariance structure, i.e, all the regressors are active in all LASSO regressions we are likely to have $\widetilde{s}_0$ of order of n and, therefore, (c.2) and (c.3) are not expected to hold.*

4. Guide to Practice

The methodology in this paper involves three steps. The first step consists of identifying known covariates that we may want to control for. It may involve the removal of deterministic trends and seasonal effects, for instance. This can be done either by parametric or nonparametric regressions. It is important to notice, however,

that the convergence rates of the estimations in the subsequent steps will be influenced by the convergence rate of the estimation in the first part of the procedure.

After the data is filtered in the first step, one can test for remaining covariance structure. For instance, if the covariance matrix of the filtered data is (almost) diagonal, there is no need to estimate a latent factor structure and the practitioner may jump directly to the third step of the method.

On the other hand, if the covariance of the first-step filtered data is dense, a latent factor model should be considered and the number of factors must be determined. To determine the number of factors we consider either the eigenvalue ratio test of Horenstein (2013) or the information criteria put forward in Bai and Ng (2002). The factors can be estimated by the usual methods.

The last step involves a sparse regression in order to estimate any remaining links between idiosyncratic components. Before running the last step, we may test for a diagonal covariance matrix of the idiosyncratic terms. If the null is not rejected, there is no need for additional estimation. In case of rejection, we can proceed with a LASSO regression. We recommend that the penalty term is selected by Bayesian Information Criterion (BIC) as advocated by Medeiros and Mendes (2016).

Finally, concerning the estimation of the long-run matrices, the usual methods discussed in the literature can be used here to select the kernel and the bandwidth. We use the simple Bartlett kernel with bandwidth given as $\lfloor T/3 \rfloor$.

5. Simulation

In this section we report simulation results divided into two parts. In the first one, we evaluate the finite-sample properties of the test for remaining covariance structure. In the second part, we highlight the informational gains when considering both the common factors and the idiosyncratic component. We simulate 1,000 replications of the following model for various combinations of sample size (T) and number of variables (n):

$$\boldsymbol{Y}_{it} = \boldsymbol{\Lambda}_i' \boldsymbol{F}_t + W_{it}, \tag{5.1}$$

$$\boldsymbol{F}_t = 0.8\boldsymbol{F}_{t-1} + \boldsymbol{E}_t, \tag{5.2}$$

$$W_{it} = \phi W_{it-1} + U_{it}, \tag{5.3}$$

$$U_{it} = I(i=1)(\theta_{12}U_{2t} + \theta_{13}U_{3t} + \theta_{14}U_{4t} + \theta_{15}U_{5t}) + V_{it} \tag{5.4}$$

where $I(\cdot)$ is the indicator function, $\{V_{it}\}$ is a sequence of independent Gaussian random variables with zero mean and variance equal to 0.25, and $\{\boldsymbol{E}_t\}$ is a sequence of r-dimensional independent random vectors normally distributed with zero mean and identity covariance. Furthermore, $\{V_{it}\}$ and $\{\boldsymbol{E}_t\}$ are mutually independent for all time periods, factors and variables. For each Monte Carlo replication, the vector of loadings is sampled from a Gaussian distribution with mean -6 and standard deviation 0.2 for $i = 1$ and mean 2 and unit variance for $i = 2, \ldots, n$. The value of ϕ is either 0 or 0.5. The coefficients θ_{12}, θ_{13}, θ_{14}, and θ_{15} are equal to zero or 0.8, 0.9, -0.7, and 0.5, respectively. We set $r = 3$.

5.1. **Test for Remaining Covariance Structure.** We start by reporting results for the test of no remaining structure on the covariance matrix of $\boldsymbol{U}_t = (U_{1t}, \ldots, U_{nt})'$. The null hypothesis considered is that all the covariances between the first variable ($i = 1$) and the remaining ones are all zero. For size simulations we set $\theta_{12} = \theta_{13} = \theta_{14} = \theta_{15} = 0$ in the DGP. In order to evaluate the effects of factor estimation as well as the methods in selecting the number of factors, we consider the following scenarios: (1) factors are known and there is no estimation involved; (2) factors are estimated by principal components but the number of factors is known; (3) the number of factors is determined by the eigenvalue ratio procedure of Horenstein (2013); (4)-(7) the number of factors is determined by one of the four information criteria proposed by Bai and Ng (2002) as defined by

$$\begin{aligned} \mathrm{IC}_1 &= \log[S(r)] + r\tfrac{n+T}{nT}\log\left(\tfrac{nT}{n+T}\right) & \mathrm{IC}_2 &= \log[S(r)] + r\tfrac{n+T}{nT}\log C_{nT}^2 \\ \mathrm{IC}_3 &= \log[S(r)] + r\tfrac{\log C_{nT}^2}{C_{nT}^2} & \mathrm{IC}_4 &= \log[S(r)] + r\tfrac{(n+T-k)\log(nT)}{nT}. \end{aligned}$$

where $S(r) = \frac{1}{nT}\|\boldsymbol{R} - \widehat{\boldsymbol{\Lambda}}_r\widehat{\boldsymbol{F}}_r\|_2^2$ and $C_{nT} := \sqrt{\min(n,T)}$.

Tables 1 and 5 report the results of the empirical size of test for different significance levels. We consider the case of $\phi = 0$ in Table 1 and $\phi = 0.5$ in Table 5 in the Supplementary Material. The tables present the results when the factors are known in panel (a), the factors are unknown but the number of factors is known in panel (b), or the number of factors are estimated either by the information criterion IC_1 in panel (c) or the eigenvalue ratio procedure in panel (d). Table 7 in the Supplementary Material shows the results when the number of factors are determined by $\mathrm{IC}_2 - \mathrm{IC}_4$.

A number of facts emerge from the inspection of the results in the Table 1. First, size distortions are small when the factors are known. In this case, the test is undersized when the pair (n, T) is small. When the factor are not known but the true

number of factors is available, the size distortions are high only when $T = 100$ and $n = 50$ due to inaccurate estimation of factors. However, the distortions disappear when the pair (T, n) grows. In this case, the empirical size is similar to the situation reported in Panel (a). The finite performance of test in the case where the number of factors is selected by information criterion IC_1 is almost indistinguishable to the case reported in Panel (b). However, the results with the eigenvalue ratio procedure are much worse when $T = 100$ and $n = 50$. In this case, the procedure selects less factors than true number $r = 3$. For instance, the procedure selects 2 or less factors in 36% of the replications. Just as comparison, for $T = 100$ and $n = 50$, IC_1 underdetermines the number of factors only in 3.10% of the cases. The latter also confirms that over estimation of the number of factors will not have big adversorial effect. For all the other combinations of T and n all the data-driven methods selects the correct number of factors in almost all replications.

When the idiosyncratic components are autocorrelated the size distortions are higher, as reported in Table 5. This is mainly caused by the well-known difficulties in the estimation of the long-run covariance matrix.

Tables 2 and 6 report the results of the empirical power with $\beta_{12} = 0.8$, $\beta_{13} = 0.9$, $\beta_{14} = -0.7$, and $\beta_{15} = -0.5$ in the DGP. When the factors are known, the test always rejects the null and the empirical power is one for any significance level. On the other hand, when factors must be estimated but the number of factors are known, the power decreases as depicted in panel (b) in the tables. Nevertheless for $T = 500, 700$ the power is reasonably high, specially when test is conducted at a 10% significance level. For $T = 100$, the performance deteriorates as n grows. The results are similar when data-driven procedures are used to determine the number of factors and the conclusions are mostly the same if $\phi = 0$ or $\phi = 0.5$.

The main message of the simulation exercise is that the finite-sample performance of the proposed tests depend on the correct selection of factors. Nevertheless, for the DGP considered here, the usual data-driven methods available in the literature to determine the true number of factors seem to work reasonably well.

5.2. **Informational Gains.** The goal of this simulation is to compare, in a prediction environment, the three-stage method developed in the paper by evaluating the information gains in predicting Y_{1t} by three different methods. First, the predictions are computed from a LASSO regression of Y_{1t} on all the other $n - 1$ variables. This is the Sparse Regression (SR) approach. Second, we consider a principal component

regression (PCR), i.e., an ordinary least squares (OLS) regression of the variable of interest on factors computed from the pool of other variables. Finally, we consider predictions constructed from the method proposed here, the `FarmPredict` methodology. Table 3 presents the results. The table presents the average mean squared error (MSE) over 5-fold cross-validation (CV) subsamples. As in the size and power simulations, we consider different combinations of T and n. We report results for the case where $\theta_{12} = 0.8$, $\theta_{13} = 0.9$, $\theta_{14} = -0.7$, and $\theta_{15} = -0.5$ in the DGP.

According to the DGP, the theoretical MSE is 0.25 when all the information is used. When just a factor is used, the MSE is 2.21. From the table is clear that there are significant informational gains when we consider both factors and the cross-dependence between idiosyncratic components. Several conclusions emerge from the table. First, it is clear that when the sample size increases the MSE reduces. This is expected. Second, the PCR's MSE and `FarmPredict`'s MSE are close to their theoretical values of 2.21 and 0.25 when the sample increases. The performance of the `FarmPredict` is quite remarkable when $T = 500$ or $T = 700$ and is always superior to Sparse Regression and PCR.

6. Applications

In this section we consider two applications with real data to illustrate the benefits of the methodology developed in the paper.

6.1. **Factor Models and Network Structure in Asset Returns.**

6.1.1. *The Dataset.* We illustrate the methodology developed in this paper by studying the factor structure of asset returns. We consider monthly close-to-close excess returns from a cross-section of 9,456 firms traded in 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 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:[8] Mining (602), Food

[8]The number between parenthesis indicate the number of firms in our sample that belong to each sector.

(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).

6.1.2. *Results.* We start the analysis by looking at the correlation matrix of a sample of nine different sectors, namely: Mining, Food, Petroleum, Construction, Manufacturing, Utilities, Department Stores, Retail, and Financial. Figure 1 plots the correlations that are larger than 0.15 in absolute value. We also test for the null of diagonal covariance matrix. The null hypothesis is strongly rejected with p-value much lower than 1%. To conduct the test of the covariance matrix we use the simple sample estimator as described in the paper. However, the correlations plotted in Figure 1 and in 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. These three factors explain most of the variation of the returns. Figure 2 shows the empirical distribution of the OLS estimates of factor loadings over the 9,456 regressions. Figure 3 presents the estimated correlations for the first-stage residuals. We focus on the nine sectors as before. The first-stage regression as 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. Nevertheless, the tests for diagonal covariance matrix reject the null even in these specific cases.

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 points 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 4 shows the estimated correlations for the residuals of the second-stage. The latent factor are not able to reduce the correlations within each sector. However, when we consider the partial correlations the conclusions are much different. As can be seen from Figure 5 that the partial correlation matrices are (almost) diagonal. In addition, we are not able to 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 6. The most interesting fact is that covariates from the financial sector are the ones most frequently selected for all the other sectors. This may indicate that there is a "financial factor" that was unmodeled in the first two stages.

The results presented here can be useful in applications where forecasting future returns is the goal, for instance. The results indicate that the inclusion of the returns of firms belonging to the financial sector may improve the performance of forecasting models. For example, if we run a regression of the residuals of the first-stage regression of firms that do not belong to the financial sector on the first principal component computed with the first-stage residuals only from the financial sector, we find a statistically significant coefficient in 28% of the cases.

6.2. **Macroeconomic Forecasting.** The second application consists of forecasting of 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`.

6.2.1. *The Dataset.* Our data consist of variables from the FRED-MD database, which is 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 McCraken's webpage.[9] For further details, we refer to McCracken and Ng (2016).

We use the vintage as of October 2020. Our sample extends from January 1960 to December 2019 (719 monthly observations), and only variables with all observations in the sample period are used (119 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 in order to become approximately stationary as in McCracken and Ng (2016).

6.2.2. *Setup and Methodology.* In order to highlight the gains of exploring all relevant information in the the dataset, we construct one-step forecasts for each one of the 119 variables in the dataset according to the following models:

[9]https://research.stlouisfed.org/econ/mccracken/fred-databases/.

(1) **Autoregressive model (`AR`):**

$$\widehat{Y}^{(\texttt{AR})}_{i,t+1|t} = \widehat{\phi}_{i0} + \widehat{\phi}_{i1}\widehat{Y}_{i,t} + \ldots + \widehat{\phi}_{ip}\widehat{Y}_{i,t-p+1}, \quad i = 1, \ldots, n,$$

where $\widehat{\phi}_{i0}, \widehat{\phi}_{i1}, \ldots, \widehat{\phi}_{ip}$, $i = 1, \ldots, n$, are OLS estimates. This model will be also the first-stage model in our methodology.

(2) **AR + Sparse regression (`SR`):**

$$\widehat{Y}^{(\texttt{SR})}_{i,t+1|t} = \widehat{Y}^{(\texttt{AR})}_{i,t+1|t} + \widehat{R}_{i,t+1|t},$$

where $\widehat{R}_{i,t} = Y_{i,t} - \widehat{Y}^{(\texttt{AR})}_{i,t|t-1}$, $i = 1, \ldots, n$, $\widehat{\boldsymbol{R}}_t = \left(\widehat{R}_{1,t}, \ldots, \widehat{R}_{n,t}\right)$, and

$$\widehat{R}_{i,t+1|t} = \widehat{\beta}_{0i} + \widehat{\boldsymbol{\beta}}'_{1i}\widehat{\boldsymbol{R}}_t + \ldots + \widehat{\boldsymbol{\beta}}'_{pi}\widehat{\boldsymbol{R}}_{t-p+1}, \quad i = 1, \ldots, n,$$

$\widehat{\beta}_{0i}, \widehat{\boldsymbol{\beta}}_{1i} \ldots, \widehat{\boldsymbol{\beta}}_{pi}$, $i = 1, \ldots, n$, are LASSO estimates. The parameters are estimated equation-wise for each one of the 119 variables in the dataset. The penalty parameter is selected by BIC as discussed in Section 4.

(3) **AR + Principal Component Regression (`PCR`):**

$$\widehat{Y}^{(\texttt{PCR})}_{i,t+1|t} = \widehat{Y}^{(\texttt{AR})}_{i,t+1|t} + \widehat{\boldsymbol{\lambda}}'_i\widehat{\boldsymbol{F}}_t,$$

where $\widehat{\boldsymbol{F}}_t$ is the estimate of the $(k \times 1)$ vector of factors $\boldsymbol{F}_t$ given by principal component analysis of $\widehat{\boldsymbol{R}}_t$, the residuals of the first-stage regression. The parameter $\boldsymbol{\lambda}_i$ is computed by OLS regression of $\widehat{R}_{i,t}$ on $\widehat{\boldsymbol{F}}_t$ in-sample.

(4) **AR + Full Information (`FarmPredict`):**

$$\widehat{Y}^{(\texttt{FarmPredict})}_{i,t+1|t} = \widehat{Y}^{(\texttt{PCR})}_{i,t+1|t} + \widehat{U}_{i,t+1|t}$$

where

$$\widehat{U}_{i,t+1|t} = \widehat{\theta}_{0i} + \widehat{\boldsymbol{\theta}}'_{1i}\widehat{\boldsymbol{U}}_t + \ldots + \widehat{\boldsymbol{\theta}}'_{pi}\widehat{\boldsymbol{U}}_{t-p+1},$$

$\widehat{\boldsymbol{U}}_t = \left(\widehat{U}_{1,t}, \ldots, \widehat{U}_{n,t}\right)'$ and $\widehat{U}_{i,t} = Y_{i,t} - \widehat{Y}^{(\texttt{PCR})}_{i,t|t-1}$, $i = 1, \ldots, n$. The estimates $\widehat{\theta}_{0i}, \widehat{\boldsymbol{\theta}}_{1i} \ldots, \widehat{\boldsymbol{\theta}}_{pi}$, $i = 1, \ldots, n$, are given by LASSO.

The forecasts are based on a rolling-window framework of fixed length of 480 observations, starting in January 1960. Therefore, the forecasts start 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` expand the information by two opposing routes. While `SR` uses a sparse combination of the set of variables, `PCR` considers only a factor structure (dense model). `FarmPredict` combines these two approaches and uses the full information available.

6.2.3. *Brief In-Sample Analysis.* We start by looking at the full sample in order to analyze the structure of dependence among the variables. We first estimate an autoregressive model of order 4, AR(p), for each transformed series. Panel (a) in Figure 7 reports the empirical distribution of the OLS estimators of the AR coefficients and Panel (b) in the figure shows the distribution of the absolute value of the sum of the estimates. This gives an idea of the persistence of each series. Only one series has estimated persistence above one. This is the case for *NONBORRES: Reserves of Depositary Institutions*, which belongs to group (v): Money and Credit. The reason for such high persistence if due to a major structural break present in the second half of the series. However, 82.35% of the series have estimated persistence below 0.9.[10]

We continue by estimating the number of factors when the full sample is used. We consider two different situations. In the first, we do not include any lag in the basket of variables used to compute the factors. In the second approach, we include four lags of each variable as well. The eigenvalue ratio procedure selects either two (no lags) or a single factor (with lags). The four information criteria of Bai and Ng (2002) estimate for the case with no lags (with lags) the following number of factors: six (one), five (one), nine (one), and one (one). Note that the factors are estimated for the residuals of the first-step AR filter. If we remove the *NONBORRES* variable from the sample the results to not change for the eigenvalue ratio procedure. On the other hand, the new numbers of factors selected by the information criteria are as follows: seven (one), six (one), eleven (one), and one (one).

Finally, we apply the testing approach developed in this paper to check for remaining (partial) covariance structure in the data. The tests strongly reject the null of a diagonal matrix when applied to the residuals either of the first or the second stages of the methodology. This serves as evidence that `FarmPredict` may be a useful modeling approach for this macroeconomic dataset.

6.2.4. *Forecasting Results.* For each of the models above, we report a number of performance metrics in Table 4. The table presents the frequency each model has the best performance among the four alternatives. Numbers between parentheses indicates the frequency each model is the second, third, and fourth best. We report the results for each one of the eight sectors as well as for the set of all 119 variables. We show the results for two methods to determine the number of factors. Panel (a) reports the results for the eigenvalue ratio method while Panel (d) presents the

[10]Conventional unit-root tests also reject the null of unit-root for all but one of the series.

results for the information criterion IC_4. Criteria IC_1, IC_2, and IC_3 select a very large number of factors. Panels (c) and (d) in the table show the results for the cases where the number of factors are kept fixed ($r = 1$ or $r = 2$).

`FarmPredict` is the model which is ranked first more frequently when all the series are considered. It is also the best model for the following groups: output and income, labor market, housing, and consumption, orders and inventories. The `AR` model is best for the following groups: money and credit and stock market. The sparse regression is superior also for two groups: interest and exchange rates and prices.

7. Conclusions

In this paper we propose a new methodology which bridges the gap between sparse regressions and factor models and evaluate the gains of increasing the information set via factor augmentation. Our proposal consists in 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 structure in both high-dimensional covariance and partial covariance matrices. Our test can be used to evaluate the benefits of adding more structure in the model. Our paper has also a number of important side results. First, we proved 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. This is specially the case when the eigenvalue ratio test is used to determine the number of latent factors. The simulations also show that there are major informational gains when combining factor models and sparse regressions in a forecasting exercise. Two applications are considered.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control

program. *Journal of the American Statistical Association 105*, 493–505.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review 93*, 113–132.

Andreou, E. and E. Ghysels (2021). Predicting the VIX and the volatility risk premium: The role of short-run funding spreads volatility factors. *Journal of Econometrics 220*, 366–398.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 2135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*, 1229–1279.

Bai, J. and Y. Liao (2017). Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics 200*, 59–78.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

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

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor augmented regressions. *Econometrica 74*, 1133–1155.

Barigozzi, M. and C. Brownlees (2019). NETS: Network estimation for time series. *Journal of Applied Econometrics 34*, 347–364.

Barigozzi, M. and M. Hallin (2016). Generalized dynamic factor models and volatilities: Recovering the market volatility shocks. *Econometrics Journal 19*, C33–C60.

Barigozzi, M. and M. Hallin (2017a). Generalized dynamic factor models and volatilities: Estimation and forecasting. *Journal of Econometrics 201*, 307–321.

Barigozzi, M. and M. Hallin (2017b). A network analysis of the volatility of high-dimensional financial series. *Journal of the Royal Statistical Society - series C 66*, 581–605.

Barigozzi, M. and M. Hallin (2020). Generalized dynamic factor models and volatilities: consistency, rates, and prediction intervals. *Journal of Econometrics 116*, 4–34.

Barigozzi, M., M. Hallin, S. Soccorsi, and R. von Sachs (2020). Time-varying general dynamic factor models and the measurement of financial connectedness. *Journal*

of Econometrics. forthcoming.

Bernanke, B., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics 120*, 387–422.

Brito, D., M. Medeiros, and R. Ribeiro (2018). Forecasting large realized covariance matrices: The benefits of factor models and shrinkage. Technical Report 3163668, SSRN.

Brownlees, C., G. Gudmundsson, and G. Lugosi (2020). Community detection in partial correlation network models. *Journal of Business & Economic Statistics*. forthcoming.

Cai, T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and its Application 4*, 4.1–4.24.

Cai, T. and Z. Ma (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli 19*, 2359–2388.

Cai, T., Z. Ren, and H. Zhou (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics 10*, 1–59.

Carvalho, C., R. Masini, and M. Medeiros (2018). Arco: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics 207*, 352–380.

Chen, S., L.-X. Zhang, and P.-S. Zhong (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association 105*, 810–819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics 41*, 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2018). Inference on causal and structural parameters using many moment inequalities.

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

Doukhan, P. and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications 84*, 313–342.

Fama, E. and K. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*, 3–56.

Fama, E. and K. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*, 1–22.

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

Fan, J., Y. Ke, and K. Wang (2020). Factor-adjusted regularized model selection. *Journal of Econometrics 216*, 71–85.

Fan, J., Q. Li, and Y. Wang (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B 79*, 247–265.

Fan, J., R. Li, C.-H.Zhang, and H. Zou (2020). *Statistical Foundations of Data Science*. CRC Press.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*, 603–680.

Fan, J., R. Masini, and M. Medeiros (2020). Do we exploit all information for counterfactual analysis? benefits of factor models and idiosyncratic correction. Working paper, Princeton University.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *Journal of Finance 75*, 1327–1370.

Gagliardini, P., E. Ossola, and P. Scaillet (2019). A diagnostic criterion for approximate factor structure. *Journal of Econometrics 212*, 503–521.

Gagliardini, P., E. Ossola, and P. Scaillet (2020). Estimation of large dimensional conditional factor models in finance. In S. Durlauf, L. Hansen, J. Heckman, and R. Matzkin (Eds.), *Handbook of Econometrics*, Volume Volume 7A, pp. 219–282.

Giannone, D., M. Lenza, and G. Primiceri (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*. forthcoming.

Giessing, A. and J. Fan (2020). Bootstrapping ℓ_p-statistics in high dimensions.

Giglio, S. and D. Xiu (2020). Asset pricing with omitted factors. *Journal of Political Economy*. forthcoming.

Gobillon, L. and T. Magnac (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics 98*, 535–551.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*, 2223–2273.

Guo, X. and C. Tang (2020). Specification tests for covariance structures in high-dimensional statistical models. *Biometrika*. forthcoming.

Horenstein, S. A. A. (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*, 1203–1227.

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

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics 37*, 4254–4278.

Ledoit, O. and M. Wolf (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics 30*, 1081–1102.

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

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics 40*, 1024–1060.

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

Ledoit, O. and M. Wolf (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Annals of Statistics*. forthcoming.

Ledoit, O. and M. Wolf (2021a). The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics*. forthcoming.

Ledoit, O. and M. Wolf (2021b). Quadratic shrinkage for lage covariance matrices. *Bernoulli*. forthcoming.

Li, W. and Y. Qin (2014). Hypothesis testing for high-dimensional covariance matrices. *Journal of Multivariate Analysis 128*, 108–119.

Masini, R. and M. Medeiros (2021). Counterfactual analysis with artificial controls: Inference, high dimensions and nonstationarity. *Journal of the American Statistical Association*. forthcoming.

Masini, R., M. Medeiros, and E. Mendes (2019). Regularized estimation of high-dimensional vector autoregressions with weakly dependent innovations. Technical

Report 1912.09002, arxiv.

McCracken, M. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*, 574–589.

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

Merlevède, F., M. Peligrad, and E. Rio (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In C. Houdré, V. Koltchinskii, D. Mason, and M. Peligrad (Eds.), *High Dimensional Probability V: The Luminy Volume*, Volume Volume 5, pp. 273–292. Institute of Mathematical Statistics.

Moon, R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*, 1543–1579.

Moskowitz, T. and M. Grinblatt (1999). Do industries explain momentum? *Journal of Finance 54*, 1249–1290.

Negahban, S., P. Ravikumar, M. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*, 538–557.

Onatski, A., M. Moreira, and M. Hallin (2013). Asymptotic power of sphericity tests for high-dimensional data. *Annals of Statistics 41*, 1204–1231.

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

Rio, E. (1994). Inégalités de moments pour les suites stationnaires et fortement mélangeantes. *Comptes rendus Acad. Sci. Paris, Série I 318*, 355–360.

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

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

van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Zheng, S., Z. Chen, H. Cui, and R. Li (2019). Hypothesis testing on linear structures of high-dimensional covariance matrix. *Annals of Statistics 47*, 3300–3334.

Zheng, S., G. Cheng, J. Guo, and H. Zhu (2019). Test for high-dimensional correlation matrices. *Annals of Statistics 47*, 2887–2921.

TABLE 1. **Simulation Results: Size with $\phi = 0$.**

The table reports the empirical size of the test of remaining covariance structure. Panel (a) reports the case where the factors are known, whereas Panel (b) considers that the factors are unknown but the number of factors is known. Panels (c) and (d) present the results when the number of factors are determined, respectively, by the eigenvalue ratio test and the information criterion IC_1. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10. The table reports the results for the case where $\phi = 0$ in (5.3).

Panel(a): Known factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.08	0.03	0.01	0.10	0.05	0.01	0.09	0.04	0.01
$n = 1 \times T$	0.06	0.02	0.00	0.07	0.03	0.01	0.10	0.05	0.01
$n = 2 \times T$	0.07	0.02	0.00	0.07	0.02	0.00	0.08	0.04	0.00
$n = 3 \times T$	0.05	0.01	0.00	0.08	0.04	0.01	0.07	0.04	0.01

Panel(b): Known number of factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.23	0.13	0.02	0.14	0.06	0.02	0.11	0.05	0.01
$n = 1 \times T$	0.13	0.06	0.01	0.09	0.04	0.01	0.12	0.05	0.01
$n = 2 \times T$	0.09	0.04	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n = 3 \times T$	0.06	0.02	0.00	0.08	0.04	0.01	0.07	0.03	0.01

Panel(c): Information criterion (IC_1)

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.24	0.14	0.03	0.14	0.06	0.02	0.11	0.05	0.01
$n = 1 \times T$	0.14	0.07	0.02	0.09	0.04	0.01	0.12	0.05	0.01
$n = 2 \times T$	0.10	0.05	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n = 3 \times T$	0.07	0.03	0.01	0.08	0.04	0.01	0.07	0.03	0.01

Panel(d): Eigenvalue ratio

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.47	0.38	0.25	0.14	0.06	0.02	0.11	0.05	0.01
$n = 1 \times T$	0.14	0.07	0.02	0.09	0.04	0.01	0.12	0.05	0.01
$n = 2 \times T$	0.09	0.04	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n = 3 \times T$	0.06	0.02	0.00	0.08	0.04	0.01	0.07	0.03	0.01

TABLE 2. **Simulation Results: Power ($\phi = 0$).**

The table reports the empirical power of the test of remaining covariance structure. Panel (a) reports the case where the factors are known, whereas Panel (b) considers that the factors are unknown but the number of factors is known. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10.

Panel(a): Known factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	1	1	1	1	1	1	1	1	1
$n = 1 \times T$	1	1	1	1	1	1	1	1	1
$n = 2 \times T$	1	1	1	1	1	1	1	1	1
$n = 3 \times T$	1	1	1	1	1	1	1	1	1

Panel(b): Known number of factors

	$T = 100$			$T = 500$				$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01		0.10	0.05	0.01
$n = 0.5 \times T$	0.33	0.18	0.03	0.99	0.97	0.83	0.99	0.99	0.94	
$n = 1 \times T$	0.20	0.08	0.01	0.84	0.60	0.13	0.95	0.81	0.33	
$n = 2 \times T$	0.16	0.07	0.01	0.83	0.55	0.11	0.94	0.82	0.34	
$n = 3 \times T$	0.08	0.03	0.00	0.82	0.56	0.10	0.95	0.81	0.34	

Panel(c): Eigenvalue ratio

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.15	0.09	0.02	0.99	0.97	0.83	0.99	0.99	0.94
$n = 1 \times T$	0.19	0.07	0.01	0.84	0.60	0.13	0.95	0.81	0.33
$n = 2 \times T$	0.16	0.07	0.01	0.83	0.55	0.11	0.94	0.82	0.34
$n = 3 \times T$	0.08	0.03	0.00	0.82	0.56	0.10	0.95	0.81	0.34

Panel(d): Information criterion (IC_1)

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.15	0.09	0.02	0.99	0.97	0.83	0.99	0.99	0.94
$n = 1 \times T$	0.19	0.07	0.01	0.84	0.60	0.13	0.95	0.81	0.33
$n = 2 \times T$	0.16	0.07	0.01	0.83	0.55	0.11	0.94	0.82	0.34
$n = 3 \times T$	0.08	0.03	0.00	0.82	0.56	0.10	0.95	0.81	0.34

TABLE 3. **Simulation Results: Informational Gains**

The table reports the average mean squared error (MSE) of three different prediction models over 5-fold cross-validation subsamples. The goal is to predict the first variable using information from the remaining $n-1$. Panel (a) considers the case of Sparse Regression (SR) where Y_{1t} is LASSO-regressed on all the other variables. Panel (b) shows the results of Principal Component Regression (PCR). Finally, Panel (c) presents the results of `FarmPredict`. "N/A" means "not available". Note that there is no factor selection for Sparse Regression. "Known Number" means that the number of factors is known.

Panel(a): Sparse Regression (SR)

	Known Number			Eigenvalue Ratio			Information Criterion (IC_1)		
	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$
$n=0.5\times T$	0.57	0.35	0.34	N/A	N/A	N/A	N/A	N/A	N/A
$n=1\times T$	0.40	0.36	0.32	N/A	N/A	N/A	N/A	N/A	N/A
$n=2\times T$	0.39	0.33	0.31	N/A	N/A	N/A	N/A	N/A	N/A
$n=3\times T$	0.35	0.32	0.30	N/A	N/A	N/A	N/A	N/A	N/A

Panel(b): Principal Component Regression (PCR)

	Known Number			Eigenvalue Ratio			Information Criterion (IC_1)		
	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$
$n=0.5\times T$	3.82	3.12	3.01	4.69	3.12	3.01	3.26	3.04	2.34
$n=1\times T$	3.09	2.35	2.34	4.05	3.35	3.34	3.22	3.02	2.32
$n=2\times T$	3.14	2.97	2.21	4.13	3.97	2.21	3.29	3.21	2.27
$n=3\times T$	3.83	3.00	2.33	3.83	3.00	2.33	3.12	3.00	2.28

Panel(c): FarmPredict

	Known Number			Eigenvalue Ratio			Information Criterion (IC_1)		
	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$	$T=100$	$T=500$	$T=700$
$n=0.5\times T$	0.50	0.33	0.31	0.52	0.33	0.31	0.50	0.34	0.30
$n=1\times T$	0.32	0.29	0.28	0.37	0.29	0.28	0.53	0.28	0.27
$n=2\times T$	0.27	0.27	0.26	0.28	0.27	0.26	0.32	0.28	0.28
$n=3\times T$	0.22	0.21	0.21	0.22	0.21	0.21	0.34	0.27	0.27

TABLE 4. **Forecasting Results.**

The table reports the frequency each model is ranked the first, second, third and fourth best model among the four alternatives. Panel (a) considers the case when the factors are selected by the eigenvalue ratio procedure. Panel (b) presents the results when factors are selected by the information criterion IC_1. Panels (c) and (d) consider the cases when the number of factors are pre-specified as either one or two. We present the results for each individual group of variables as well as for the full set of macroeconomic variables.

Panel (a): Optimal Factor Selection (eigenvalue ratio)

	AR				SR				PCR				FarmPredict			
Group	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
(i) output and income	0.125	0.000	0.250	0.625	0.000	0.125	0.625	0.250	0.375	0.500	0.125	0.000	0.500	0.375	0.000	0.125
(ii) labor market	0.032	0.097	0.290	0.581	0.226	0.065	0.516	0.194	0.194	0.516	0.097	0.194	0.548	0.323	0.097	0.032
(iii) housing	0.100	0.100	0.300	0.500	0.400	0.400	0.100	0.100	0.000	0.200	0.400	0.400	0.500	0.300	0.200	0.000
(iv) consumption, orders and inventories	0.000	0.000	0.333	0.667	0.000	0.000	0.667	0.333	0.333	0.667	0.000	0.000	0.667	0.333	0.000	0.000
(v) money and credit	0.429	0.357	0.143	0.071	0.214	0.214	0.357	0.214	0.214	0.286	0.357	0.143	0.143	0.143	0.143	0.571
(vi) interest and exchange rates	0.368	0.211	0.263	0.158	0.526	0.316	0.158	0.000	0.053	0.263	0.211	0.474	0.053	0.211	0.368	0.368
(vii) prices	0.150	0.150	0.600	0.100	0.650	0.100	0.200	0.050	0.050	0.200	0.100	0.650	0.150	0.550	0.100	0.200
(viii) stock market	0.667	0.000	0.000	0.333	0.000	0.000	0.667	0.333	0.000	1.000	0.000	0.000	0.333	0.000	0.333	0.333
(ix) all	0.185	0.134	0.311	0.370	0.311	0.160	0.378	0.151	0.160	0.387	0.168	0.286	0.345	0.319	0.143	0.193

Panel (b): Optimal Factor Selection (IC_4)

	AR				SR				PCR				FarmPredict			
Group	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
(i) output and income	0.125	0.125	0.188	0.563	0.063	0.250	0.500	0.188	0.250	0.375	0.313	0.063	0.563	0.250	0.000	0.188
(ii) labor market	0.032	0.097	0.258	0.613	0.226	0.032	0.548	0.194	0.226	0.581	0.065	0.129	0.516	0.290	0.129	0.065
(iii) housing	0.000	0.000	0.400	0.600	0.200	0.500	0.100	0.200	0.200	0.100	0.500	0.200	0.600	0.400	0.000	0.000
(iv) consumption, orders and inventories	0.000	0.000	0.333	0.667	0.167	0.000	0.500	0.333	0.167	0.667	0.167	0.000	0.667	0.333	0.000	0.000
(v) money and credit	0.571	0.286	0.071	0.071	0.143	0.429	0.357	0.071	0.143	0.286	0.429	0.143	0.143	0.000	0.143	0.714
(vi) interest and exchange rates	0.316	0.105	0.105	0.474	0.368	0.158	0.368	0.105	0.158	0.263	0.474	0.105	0.158	0.474	0.053	0.316
(vii) prices	0.100	0.150	0.650	0.100	0.500	0.300	0.150	0.050	0.100	0.150	0.100	0.650	0.300	0.400	0.100	0.200
(viii) stock market	0.667	0.000	0.000	0.333	0.000	0.667	0.333	0.000	0.000	0.333	0.667	0.000	0.333	0.000	0.000	0.667
(ix) all	0.176	0.118	0.277	0.429	0.252	0.227	0.378	0.143	0.176	0.353	0.269	0.202	0.395	0.303	0.076	0.227

Panel (c): Fixed Number of Factors ($r = 1$)

	AR				SR				PCR				FarmPredict			
Group	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
(i) output and income	0.125	0.125	0.188	0.563	0.063	0.250	0.500	0.188	0.250	0.375	0.313	0.063	0.563	0.250	0.000	0.188
(ii) labor market	0.032	0.097	0.258	0.613	0.226	0.032	0.548	0.194	0.226	0.581	0.065	0.129	0.516	0.290	0.129	0.065
(iii) housing	0.000	0.000	0.400	0.600	0.200	0.500	0.100	0.200	0.200	0.100	0.500	0.200	0.600	0.400	0.000	0.000
(iv) consumption, orders and inventories	0.000	0.000	0.333	0.667	0.167	0.000	0.500	0.333	0.167	0.667	0.167	0.000	0.667	0.333	0.000	0.000
(v) money and credit	0.571	0.286	0.071	0.071	0.143	0.429	0.357	0.071	0.143	0.286	0.429	0.143	0.143	0.000	0.143	0.714
(vi) interest and exchange rates	0.316	0.105	0.105	0.474	0.368	0.158	0.368	0.105	0.158	0.263	0.474	0.105	0.158	0.474	0.053	0.316
(vii) prices	0.100	0.150	0.650	0.100	0.500	0.300	0.150	0.050	0.100	0.150	0.100	0.650	0.300	0.400	0.100	0.200
(viii) stock market	0.667	0.000	0.000	0.333	0.000	0.667	0.333	0.000	0.000	0.333	0.667	0.000	0.333	0.000	0.000	0.667
(ix) all	0.176	0.118	0.277	0.429	0.252	0.227	0.378	0.143	0.176	0.353	0.269	0.202	0.395	0.303	0.076	0.227

Panel (d): Fixed Number of Factors ($r = 2$)

	AR				SR				PCR				FarmPredict			
Group	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
(i) output and income	0.063	0.125	0.250	0.563	0.063	0.063	0.625	0.250	0.250	0.625	0.063	0.063	0.625	0.188	0.063	0.125
(ii) labor market	0.065	0.129	0.226	0.581	0.226	0.097	0.516	0.161	0.226	0.452	0.097	0.226	0.484	0.323	0.161	0.032
(iii) housing	0.200	0.200	0.000	0.600	0.500	0.400	0.100	0.000	0.000	0.100	0.700	0.200	0.300	0.300	0.200	0.200
(iv) consumption, orders and inventories	0.167	0.167	0.167	0.500	0.167	0.167	0.500	0.167	0.333	0.333	0.333	0.000	0.333	0.333	0.000	0.333
(v) money and credit	0.500	0.357	0.071	0.071	0.214	0.357	0.143	0.286	0.143	0.286	0.429	0.143	0.143	0.000	0.357	0.500
(vi) interest and exchange rates	0.316	0.368	0.000	0.316	0.368	0.263	0.263	0.105	0.105	0.316	0.263	0.316	0.211	0.053	0.474	0.263
(vii) prices	0.100	0.100	0.100	0.700	0.500	0.150	0.250	0.100	0.200	0.150	0.550	0.100	0.200	0.600	0.100	0.100
(viii) stock market	0.667	0.000	0.000	0.333	0.000	0.667	0.333	0.000	0.000	0.333	0.667	0.000	0.333	0.000	0.000	0.667
(ix) all	0.193	0.193	0.126	0.487	0.286	0.202	0.361	0.151	0.176	0.345	0.311	0.168	0.345	0.261	0.202	0.193

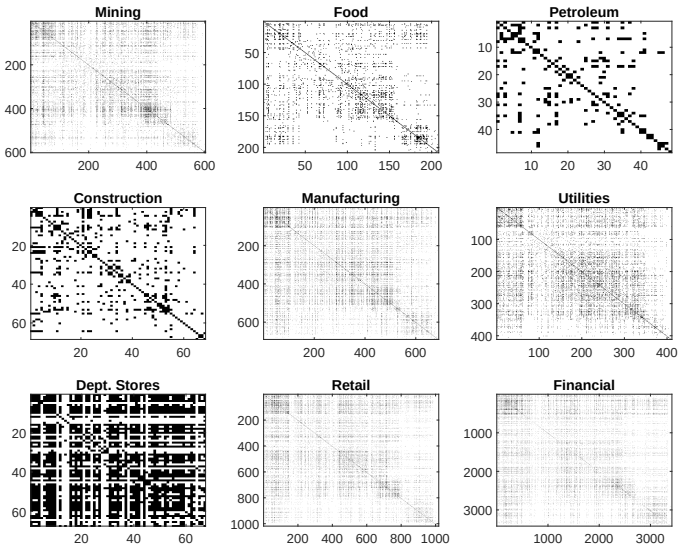


FIGURE 1. Correlations of returns.
We estimate the correlations between all pairs of returns from specific sectors. The correlations that are higher than 0.15 in absolute value are shown as black dots. We consider the sectors: mining, food, petroleum, construction, manufacturing, utilities, department stores, retail, and financial.

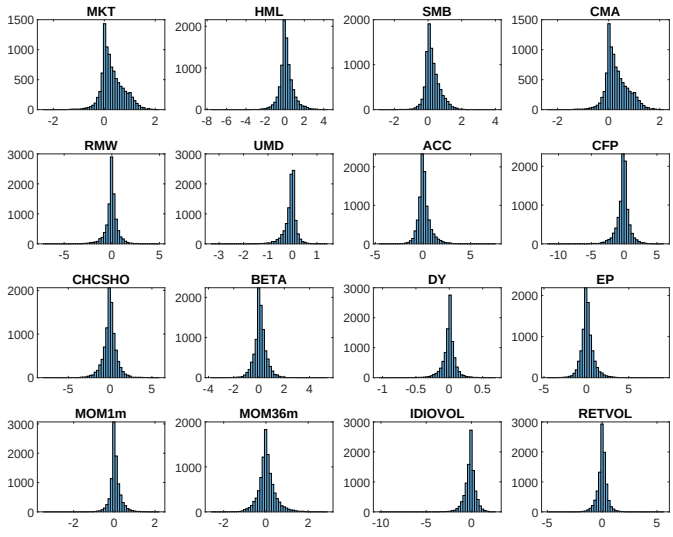


FIGURE 2. First-stage coefficient estimates.
The figure shows the empirical distribution of the first-stage regression where each excess returns are linearly regressed on 16 risk factors.

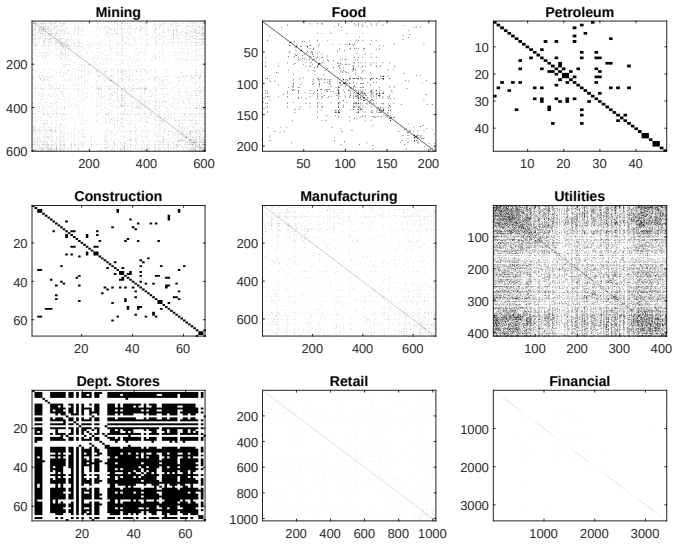


FIGURE 3. Correlations of first-stage residuals.
We estimate the correlations between all pairs of residuals from the first-stage OLS regression on 16 observed risk factors from specific sectors. The correlations that are higher than 0.15 in absolute value are shown as black dots. We consider the sectors: mining, food, petroleum, construction, manufacturing, utilities, department stores, retail, and financial.

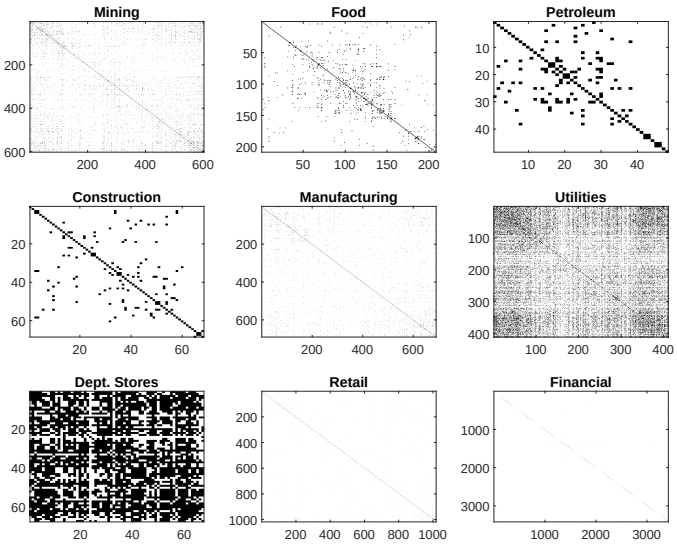


FIGURE 4. Correlations of second-stage residuals.
We estimate the correlations between all pairs of residuals from the second-stage principal component analysis from specific sectors. The correlations that are higher than 0.15 in absolute value are shown as black dots. We consider the sectors: mining, food, petroleum, construction, manufacturing, utilities, department stores, retail, and financial.

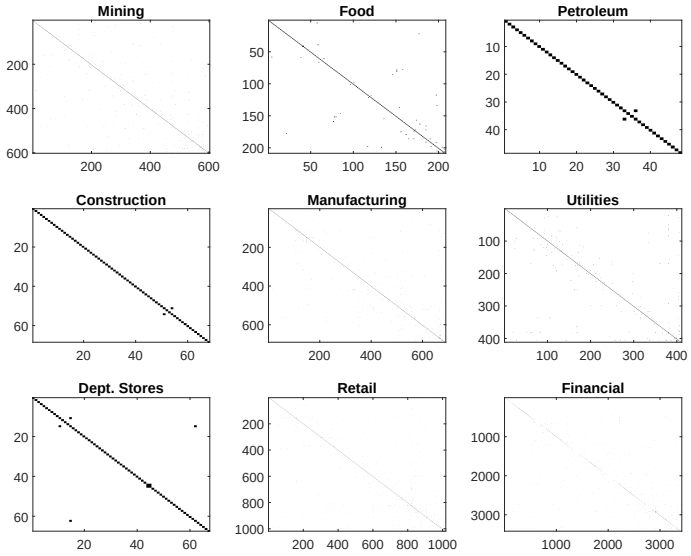


FIGURE 5. Partial correlations of second-stage residuals.
We estimate the partial correlations between all pairs of residuals from the second-stage LASSO regression from specific sectors. The correlations that are higher than 0.15 in absolute value are shown as black dots. We consider the sectors: mining, food, petroleum, construction, manufacturing, utilities, department stores, retail, and financial.

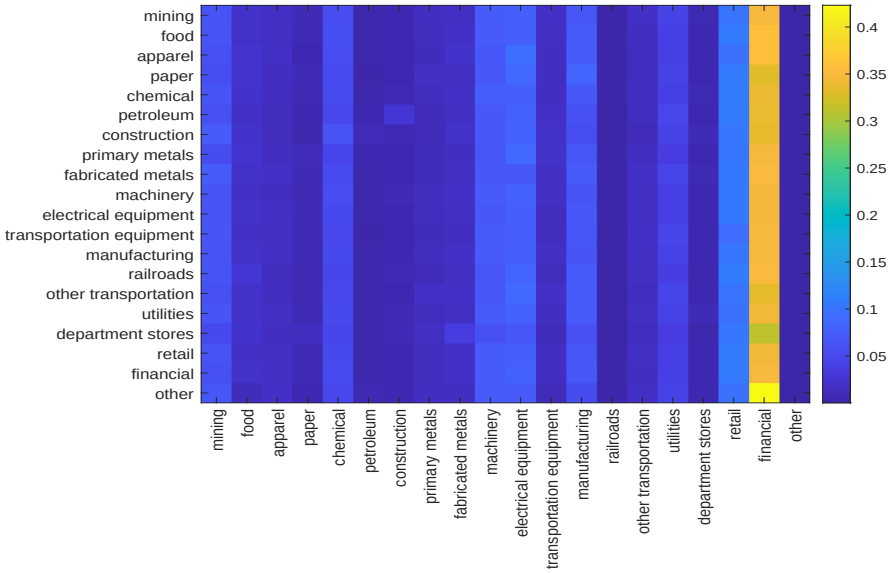
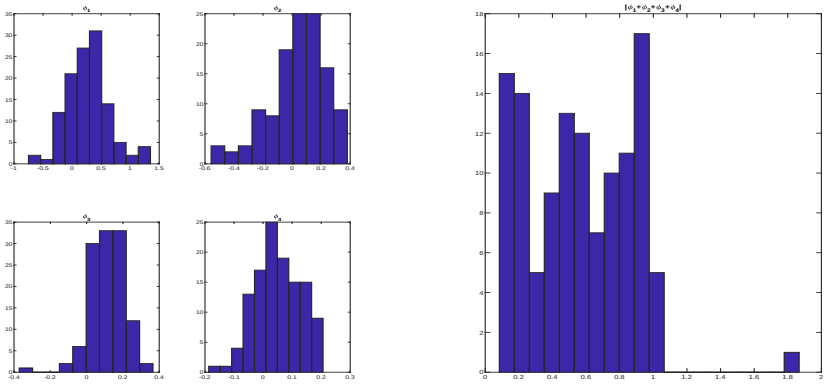


FIGURE 6. Variable Selection Frequency.
We report how often that variables from column sectors are selected in the third-stage LASSO regression for firms on row sectors . The numbers are normalized by the total number of firms in each sector.



(a) AR coefficients. (b) Absolute sum of AR coefficients.

FIGURE 7. Distribution of AR estimates

Panel (a) illustrates the empirical distribution of the ordinary least squares (OLS) estimation of the coefficients of an fourth-order autoregressive, AR(4), model across the 119 macroeconomic time series. Panel (b) illustrates the empirical distribution of the absolute sum of the AR(4) coefficients across the 119 series.

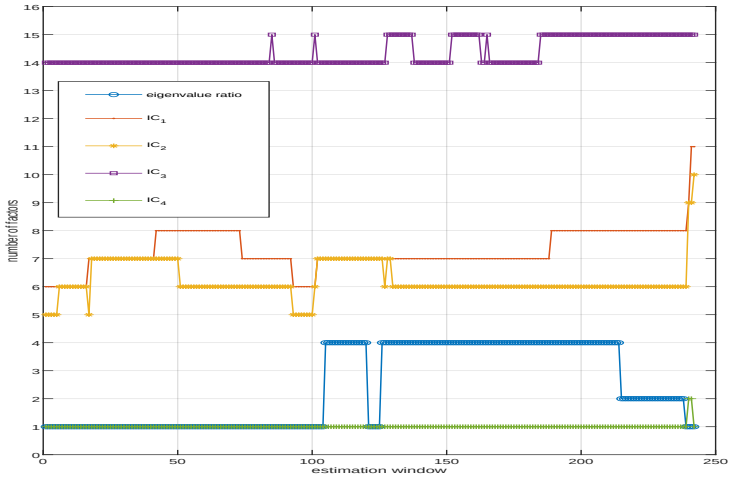


FIGURE 8. Estimated number of factors.

The figure illustrates the number of selected factors over the estimation windows. The figure reports the results for the eigenvalue ratio procedure and the four information criteria discussed in the paper.

Appendix A. Additional Simulation Results

Tables 5 and 6 show, respectively, size and power simulations with $\phi = 0.5$. Table 7 presents additional size results. The table reports the empirical size of the test of remaining covariance structure. Panels (a)–(c) consider that the number of factors are determined by information criteria IC_2, IC_3 and IC_4. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10.

Appendix B. Proof of the Theorems

Throughout the proofs we use the equivalence $\|X\|_{\psi_p} < \infty \iff \mathbb{P}(|X| > x) = O[\psi_p^{-1}(x)]$, as $x \to \infty$, for any random variable X and $\psi_p \in \Psi$ (refer to Lemma C.17 below) combined with Lemma 6 in Carvalho et al. (2018) and Lemma 1 in Masini and Medeiros (2021). The key ingredients of the lemmas are a Marcinkiewicz-Zygmund type inequality for strong mixing sequences to deal with the polynomial tails (Rio, 1994; Doukhan and Louhichi, 1999) and a Bernstein inequality under strong mixing conditions to control exponential tails (Merlevède et al. (2009) - Theorem 2).

B.1. **Proof of Theorem 1.** We first upper bound $\|\widehat{R}_{it} - R_{it}\|_{\psi}$. By subsequent application of Hölder's inequality we have

$$\begin{aligned}|\widehat{R}_{it} - R_{it}| &= |(\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i)'\boldsymbol{X}_{it}| \leqslant \|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_1 \|\boldsymbol{X}_{it}\|_\infty \\ &= \|\widehat{\boldsymbol{\Sigma}}_i^{-1}\widehat{\boldsymbol{v}}_i\|_1 \|\boldsymbol{X}_{it}\|_\infty \leqslant k^2 \|\widehat{\boldsymbol{\Sigma}}_i^{-1}\|_{\mathsf{max}} \|\widehat{\boldsymbol{v}}_i\|_\infty \|\boldsymbol{X}_{it}\|_\infty,\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_i := \boldsymbol{X}_i'\boldsymbol{X}_i/T$ and $\widehat{\boldsymbol{v}}_i := \boldsymbol{X}_i'\boldsymbol{U}_i/T$. Then, by the Cauchy-Schwartz conjugate

$$\|\widehat{R}_{it} - R_{it}\|_{\psi_{p/4}} \leqslant k^2 \|\|\widehat{\boldsymbol{\Sigma}}_i^{-1}\|_{\mathsf{max}}\|_{\psi_p} \|\|\widehat{\boldsymbol{v}}_i\|_\infty\|_{\psi_{p/2}} \|\|\boldsymbol{X}_{it}\|_\infty\|_{\psi_p}.$$

The first term is bounded by Assumption 3(b). For the second term we have: $\|X_{it\ell}U_{it}\|_{\psi_{p/2}} \leqslant \|X_{it\ell}\|_{\psi_p}\|U_{it}\|_{\psi_p} \leqslant C^2$ by Assumption 3(a). Then, $\{X_{it\ell}U_{it}\}_{t>0}$ is a zero-mean strong mixing sequence with exponential decay (Assumption 3(c)) with bounded $\psi_{p/2}$-norm. Therefore, $\|\|\widehat{\boldsymbol{v}}_i\|_\infty\|_{\psi_{p/2}} = O(1/\sqrt{T})$, uniformly in $i \leqslant n$. The last term is bounded by the maximal inequality (van der Vaart and Wellner (1996) - Lemma 2.2.2) and Assumption 3(a). The result follows.

B.2. **Proof of Theorem 2.** The proof is an adaption of the proof of Theorem 4 and Corollary 1 in Fan et al. (2013), henceforth FLM, to include the estimation error in the sample covariance matrix. For part (a), we use expression (A.1) in Bai (2003) to

TABLE 5. **Simulation Results: Size with $\phi = 0.5$.**

The table reports the empirical size of the test of remaining covariance structure. Panel (a) reports the case where the factors are known, whereas Panel (b) considers that the factors are unknown but the number of factors is known. Panels (c) and (d) present the results when the number of factors are determined, respectively, by the eigenvalue ratio test and the information criterion IC_1. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10. The table reports the results for the case where $\phi = 0.5$ in (5.3).

Panel(a): Known factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.09	0.03	0.01	0.11	0.06	0.01	0.10	0.05	0.01
$n = 1 \times T$	0.07	0.03	0.00	0.07	0.03	0.01	0.11	0.06	0.01
$n = 2 \times T$	0.08	0.02	0.00	0.08	0.03	0.00	0.09	0.05	0.00
$n = 3 \times T$	0.05	0.02	0.00	0.09	0.04	0.01	0.07	0.04	0.01

Panel(b): Known number of factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.24	0.14	0.02	0.15	0.07	0.02	0.12	0.06	0.01
$n = 1 \times T$	0.13	0.07	0.01	0.09	0.04	0.01	0.14	0.06	0.02
$n = 2 \times T$	0.09	0.04	0.01	0.08	0.04	0.01	0.09	0.05	0.00
$n = 3 \times T$	0.07	0.02	0.00	0.08	0.04	0.01	0.08	0.03	0.01

Panel(c): Information criterion (IC_1)

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.49	0.42	0.29	0.15	0.07	0.02	0.11	0.06	0.01
$n = 1 \times T$	0.15	0.09	0.02	0.10	0.04	0.01	0.14	0.06	0.01
$n = 2 \times T$	0.09	0.04	0.01	0.09	0.04	0.01	0.09	0.05	0.00
$n = 3 \times T$	0.07	0.03	0.00	0.10	0.04	0.01	0.08	0.03	0.01

Panel(d): Eigenvalue ratio

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.25	0.14	0.04	0.14	0.07	0.02	0.13	0.06	0.01
$n = 1 \times T$	0.15	0.07	0.02	0.10	0.04	0.01	0.13	0.06	0.02
$n = 2 \times T$	0.11	0.05	0.01	0.08	0.05	0.01	0.10	0.05	0.00
$n = 3 \times T$	0.08	0.03	0.01	0.09	0.04	0.01	0.08	0.03	0.01

TABLE 6. **Simulation Results: Power ($\phi = 0.5$).**

The table reports the empirical power of the test of remaining covariance structure. Panel (a) reports the case where the factors are known, whereas Panel (b) considers that the factors are unknown but the number of factors is known. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10.

Panel(a): Known factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	1	1	1	1	1	1	1	1	1
$n = 1 \times T$	1	1	1	1	1	1	1	1	1
$n = 2 \times T$	1	1	1	1	1	1	1	1	1
$n = 3 \times T$	1	1	1	1	1	1	1	1	1

Panel(b): Known number of factors

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.36	0.18	0.03	1.00	1.00	0.91	1.00	1.00	1.00
$n = 1 \times T$	0.20	0.09	0.02	0.89	0.69	0.13	1.00	0.92	0.39
$n = 2 \times T$	0.18	0.07	0.01	0.98	0.59	0.13	1.00	0.96	0.36
$n = 3 \times T$	0.10	0.03	0.00	0.91	0.66	0.11	1.00	0.92	0.39

Panel(c): Eigenvalue ratio

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.15	0.11	0.03	1.00	1.00	0.96	1.00	1.00	1.00
$n = 1 \times T$	0.22	0.08	0.01	0.99	0.70	0.16	1.00	0.95	0.38
$n = 2 \times T$	0.17	0.07	0.01	0.94	0.62	0.12	0.98	0.88	0.37
$n = 3 \times T$	0.09	0.03	0.00	0.89	0.59	0.11	1.00	0.87	0.40

Panel(d): Information criterion (IC_1)

	$T = 100$			$T = 500$			$T = 700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n = 0.5 \times T$	0.15	0.11	0.03	1.00	1.00	0.96	1.00	1.00	1.00
$n = 1 \times T$	0.22	0.08	0.01	0.99	0.70	0.16	1.00	0.95	0.38
$n = 2 \times T$	0.17	0.07	0.01	0.94	0.62	0.12	0.98	0.88	0.37
$n = 3 \times T$	0.09	0.03	0.00	0.89	0.59	0.11	1.00	0.87	0.40

obtain the following identity

$$\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{F}_t = \left(\frac{\boldsymbol{V}}{n}\right)^{-1}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\frac{\mathbb{E}(\boldsymbol{U}_s'\boldsymbol{U}_t)}{n} + \frac{1}{T}\sum_{s=1}^{T}\left(\widehat{\boldsymbol{f}}_s\widetilde{\zeta}_{st} + \widehat{\boldsymbol{f}}_s\widetilde{\eta}_{st} + \widehat{\boldsymbol{f}}_s\widetilde{\xi}_{st}\right)\right], \quad \text{(B.1)}$$

TABLE 7. **Simulation Results: Size (Additional Results).**

The table reports the empirical size of the test of remaining covariance structure. Panels (a)–(c) consider that the number of factors are determined by information criteria IC_2, IC_3 and IC_4. Factors are estimated by the usual principal component algorithm. Three nominal significance levels are considered: 0.01, 0.05, and 0.10.

Panel(a): Information criterion (IC_2)

	$T=100$			$T=500$			$T=700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n=0.5\times T$	0.25	0.14	0.03	0.14	0.06	0.02	0.11	0.05	0.01
$n=1\times T$	0.13	0.07	0.02	0.09	0.04	0.01	0.12	0.05	0.01
$n=2\times T$	0.10	0.04	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n=3\times T$	0.07	0.03	0.01	0.08	0.04	0.01	0.07	0.03	0.01

Panel(b): Information criterion (IC_3)

	$T=100$			$T=500$			$T=700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n=0.5\times T$	0.24	0.13	0.03	0.14	0.06	0.02	0.11	0.06	0.01
$n=1\times T$	0.13	0.07	0.02	0.09	0.04	0.01	0.12	0.05	0.01
$n=2\times T$	0.10	0.04	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n=3\times T$	0.07	0.03	0.00	0.08	0.04	0.01	0.07	0.03	0.01

Panel(c): Information criterion (IC_4)

	$T=100$			$T=500$			$T=700$		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
$n=0.5\times T$	0.26	0.15	0.04	0.14	0.06	0.02	0.11	0.05	0.01
$n=1\times T$	0.14	0.08	0.03	0.09	0.04	0.01	0.12	0.05	0.01
$n=2\times T$	0.10	0.04	0.01	0.08	0.04	0.01	0.09	0.04	0.00
$n=3\times T$	0.07	0.03	0.01	0.08	0.04	0.01	0.07	0.03	0.01

where $\widetilde{\zeta}_{st}, \widetilde{\eta}_{st}$ and $\widetilde{\xi}_{st}$ are defined before Lemma C.3.

By Assumptions 2(d) and 3(a) and the maximal inequality we have $\|\boldsymbol{R}\|_{\mathsf{max}} \leqslant r\|\boldsymbol{\Lambda}\|_{\mathsf{max}}\|\boldsymbol{F}\|_{\mathsf{max}} + \|\boldsymbol{U}\|_{\mathsf{max}} = O_P[\psi^{-1}(nT)]$. Applying Lemma C.14 we conclude that $\|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\mathsf{max}} = O_P\{\omega[\psi^{-1}(nT) + \omega]\} = O_P(1)$. Finally, $\psi^{-1}(n^2)/\sqrt{T} = O(1)$ also by assumption. Then, $\|\frac{\boldsymbol{V}}{n}\|^{-1} = O_P(1)$ by Lemma C.6. Using the results (a)-(d) of Lemma C.5 we can bound in probability each of the terms in brackets of (B.1) in ℓ_2 norm, uniformly in $t \leqslant T$. Result (a) follows.

For part (b) we use the fact that $\widehat{\boldsymbol{\Lambda}} := \widehat{\boldsymbol{R}}\widehat{\boldsymbol{F}}/T$ and set $\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}} = \boldsymbol{I}_r$ to write

$$\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{H}\boldsymbol{F}_t\widetilde{U}_{it} + \frac{1}{T}\sum_{t=1}^{T}\widehat{R}_{it}(\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t) + \boldsymbol{H}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t\boldsymbol{F}_t' - \boldsymbol{I}_r\right)\boldsymbol{\lambda}_i. \quad \text{(B.2)}$$

The first term can be upper bounded in ℓ_2 norm, uniformly in $i \leqslant n$, by

$$\sqrt{r}\|\boldsymbol{H}\| \max_{i\leqslant n}\max_{j\leqslant r}\left|\frac{1}{T}\sum_{t=1}^{T}F_{jt}\widetilde{U}_{it}\right| = O_P(1)O_P\left[\psi_{p/2}^{-1}(n)/\sqrt{T} + \omega\right],$$

where the equality follows from Lemma B.6(b) and (e). The ℓ_2 norm of the second term is upper bounded uniformly in $i \leqslant n$ by

$$\left(\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^{T}\widehat{R}_{it}^2\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t\|^2\right)^{1/2} = \left\{O_P(1)O_P\left[\frac{1}{T} + (1/\sqrt{n} + \omega)^2\right]\right\}^{1/2},$$

where the first term after the equality follows from Lemma C.6(d) together with theorem's assumption and the second term from Lemma C.4(e). Finally, the last term of (B.2) is upper bounded by

$$\|\boldsymbol{H}\|\max_{i\leqslant n}\|\boldsymbol{\lambda}_i\|\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}_t\boldsymbol{F}_t' - \boldsymbol{I}_r\right\| = O_P(1)O(1)O_P(1/\sqrt{T}),$$

where the last term is $O_P(1/\sqrt{T})$ by the maximum inequality and Assumption 3. Plugging the last three displays back into (B.2) yields result (b).

For part (c) we use we have $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = \|\boldsymbol{\Lambda}\boldsymbol{F}' - \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}' + \widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max} \leqslant \|\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{F}}' - \boldsymbol{\Lambda}\boldsymbol{F}'\|_{\max} + \|\widehat{\boldsymbol{R}} - \boldsymbol{R}\|_{\max}$. The last term is $O_P(\omega)$ by assumption. For the first term we use the decomposition

$$\begin{aligned}\widehat{\boldsymbol{\lambda}}_i'\widehat{\boldsymbol{F}}_t - \boldsymbol{\lambda}_i'\boldsymbol{F}_t = {}& (\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i)'(\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t) + (\boldsymbol{H}\boldsymbol{\lambda}_i)'(\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t) \\ & + (\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i)'\boldsymbol{H}\boldsymbol{F}_t + \boldsymbol{\lambda}_i'(\boldsymbol{H}'\boldsymbol{H} - \boldsymbol{I}_r)\boldsymbol{F}_t. \quad \text{(B.3)}\end{aligned}$$

Therefore, we can upper bound the left hand side as

$$\begin{aligned}|\widehat{\boldsymbol{\lambda}}_i'\widehat{\boldsymbol{F}}_t - \boldsymbol{\lambda}_i'\boldsymbol{F}_t| \leqslant {}& \|\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i\|\|\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t\| + \|\boldsymbol{H}\boldsymbol{\lambda}_i\|\|\widehat{\boldsymbol{F}}_t - \boldsymbol{H}\boldsymbol{F}_t\| \\ & + \|\widehat{\boldsymbol{\lambda}}_i - \boldsymbol{H}\boldsymbol{\lambda}_i\|\|\boldsymbol{H}\boldsymbol{F}_t\| + \|\boldsymbol{\lambda}_i\|\|\boldsymbol{F}_t\|\|\boldsymbol{H}'\boldsymbol{H} - \boldsymbol{I}_r\|.\end{aligned}$$

Now we bound in probability, uniformly in $i \leqslant n$ and $t \leqslant T$, each of the four terms above. The first one is given by part (a) and (b). $\max_{i\leqslant n}\|\boldsymbol{H}\boldsymbol{\lambda}_i\| \leqslant \|\boldsymbol{H}\|\max_{i\leqslant n}||\boldsymbol{\lambda}_i|| \leqslant O_P(1)r\|\boldsymbol{\Lambda}\|_{\max} = O_P(1)$ by Lemma C.6(b) and Assumption 2(d). Thus, the second term is bounded by part (a). For the third term, $\max_{t\leqslant T}\|\boldsymbol{H}\boldsymbol{F}_t\| \leqslant \|\boldsymbol{H}\|\max_{t\leqslant T}||\boldsymbol{F}_t|| = O_P(1)O_P\left[\psi^{-1}(T)\right] = O_P\left[\psi^{-1}(T)\right]$ by Lemma C.6(b) and Assumption 2(a). Finally, $\|\boldsymbol{H}'\boldsymbol{H} - \boldsymbol{I}_r\| = O_P(1/\sqrt{T} + 1/\sqrt{n} + \omega)$ by Lemma C.6(c). Hence, the last term is $O_P[\psi^{-1}(T)(1/\sqrt{T} + 1/\sqrt{n} + \omega)]$ by Assumptions 2(d) and 3(a).

B.3. **Proof of Theorem 3.** We have that $L(\widehat{\boldsymbol{\theta}}_\xi)+\xi\|\widehat{\boldsymbol{\theta}}_\xi\|_1 \leqslant L(\boldsymbol{\theta})+\xi\|\boldsymbol{\theta}\|_1$ for all $\boldsymbol{\theta}\in\mathbb{R}^n$ by definition of $\widehat{\boldsymbol{\theta}}_\xi$, where $L(\boldsymbol{\theta}) := \|\widehat{\boldsymbol{u}}_y - \boldsymbol{\theta}'\widehat{\boldsymbol{U}}_x\|_2^2/T$. Also, since $L(\boldsymbol{\theta})$ is a quadratic function, it implies that $(\widehat{\boldsymbol{\theta}}_\xi-\boldsymbol{\theta})'\nabla^2 L(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_\xi-\boldsymbol{\theta}) \leqslant -\nabla L(\boldsymbol{\theta})'(\widehat{\boldsymbol{\theta}}_\xi-\boldsymbol{\theta})+\xi(\|\boldsymbol{\theta}\|_1-\|\widehat{\boldsymbol{\theta}}_\xi\|_1)$. By Holder's inequality we have $|\nabla L(\boldsymbol{\theta})'(\widehat{\boldsymbol{\theta}}_\xi-\boldsymbol{\theta})| \leqslant \|\nabla L(\boldsymbol{\theta})\|_\infty\|\widehat{\boldsymbol{\theta}}_\xi-\boldsymbol{\theta}\|_1$. By assumption $\xi \geqslant 2\|\nabla L(\boldsymbol{\theta})\|_\infty$. Therefore,

$$(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta})'\nabla^2 L(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}) \leqslant \xi/2\|\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}\|_1 + \xi(\|\boldsymbol{\theta}\|_1 - \|\widehat{\boldsymbol{\theta}}_\xi\|_1). \tag{B.4}$$

For any index set $\mathcal{S}\in[n]$, by the decomposability of the ℓ_1 norm (refer to Definition 1 in Negahban et al. (2012)) followed by the triangle inequality we have $\|\widehat{\boldsymbol{\theta}}_\xi\|_1 = \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}}\|_1 + \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}^c}\|_1 \geqslant \|\boldsymbol{\theta}_\mathcal{S}\|_1 - \|\widehat{\boldsymbol{\theta}}_{\xi.\mathcal{S}} - \boldsymbol{\theta}_\mathcal{S}\|_1 + \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}^c}\|_1$ and $\|\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}\|_1 = \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}} - \boldsymbol{\theta}_\mathcal{S}\|_1 + \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}^c} - \boldsymbol{\theta}_{\mathcal{S}^c}\|_1 \leqslant \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}} - \boldsymbol{\theta}_\mathcal{S}\|_1 + \|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}^c} - \boldsymbol{\theta}_{\mathcal{S}^c}\|_1$. Plugging it back in (B.4) yields

$$2(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta})'\nabla^2 L(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}) + \xi\|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}^c} - \boldsymbol{\theta}_{\mathcal{S}^c}\|_1 \leqslant 3\xi\|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}} - \boldsymbol{\theta}_\mathcal{S}\|_1 + 4\xi\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_1. \tag{B.5}$$

We then conclude that any minimizer $\widehat{\boldsymbol{\theta}}_\xi$ of (3.6) and $\boldsymbol{\theta}\in\mathbb{R}^n$ obeys $\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta} \in \mathbb{C}(\mathcal{S},\boldsymbol{\theta}) := \{\boldsymbol{x}\in\mathbb{R}^n : \|\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leqslant 3\|\boldsymbol{x}_\mathcal{S}\|_1 + 4\|\boldsymbol{\theta}_{\mathcal{S}^c}\|_1\}$. If we take $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ and $\mathcal{S}=\mathcal{S}_0 := \{i : \theta_{0,i}\neq 0\}$ then $\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}_0 \in \mathbb{C}_0 := \mathbb{C}(\mathcal{S}_0,\boldsymbol{\theta}_0)$. $\mathbb{C}_0$ is a cone in $\mathbb{R}^n$ that does not depend on $\boldsymbol{\theta}_0$ as $\|\boldsymbol{\theta}_{0,\mathcal{S}_0^c}\| = 0$. Moreover, by definition of the compatibility constant $\kappa := \kappa(\widehat{\boldsymbol{U}}_x\widehat{\boldsymbol{U}}_x/T,\mathcal{S}_0,3)$ we have that $\|\widehat{\boldsymbol{\theta}}_{\xi,\mathcal{S}} - \boldsymbol{\theta}_\mathcal{S}\|_1 \leqslant (\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta})'\nabla^2 L(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta})\sqrt{|\mathcal{S}_0|}/\kappa$. Apply this inequality to (B.5) and use the fact that $4ab < a^2 + 4b^2$ for non-negative $a,b\in\mathbb{R}$ to obtain $(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta})'\nabla^2 L(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}) + \xi\|\widehat{\boldsymbol{\theta}}_\xi - \boldsymbol{\theta}\|_1 \leqslant 4\xi^2|\mathcal{S}_0|/\kappa$. Finally, we have by assumption that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathsf{max}} \leqslant C_1$, $\|\boldsymbol{U}\|_{\mathsf{max}} \leqslant C_2$ and $C_1(2C_2 + C_1) \leqslant \frac{\kappa_0}{32|\mathcal{S}|}$ which, in turn fulfills the assumptions of Lemma C.14 with $\zeta = 3$ and $\alpha = 1/2$. Therefore, we conclude that $\kappa \geqslant \kappa_0/2$ and we have the result.

B.4. **Proof of Theorem 4.** We use in this proof the following additional notation for short: For every random vector $\boldsymbol{X}$, we denote by $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ its covariance matrix, $d_{\boldsymbol{X}}$ the diagonal of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and $\sigma^2_{\boldsymbol{X}} := \|d_{\boldsymbol{X}}\|_\infty$. Also, $\boldsymbol{X}_G$ denotes zero-mean Gaussian random vector defined in the same probability space, independent of $\boldsymbol{X}$ and with the same covariance matrix of $\boldsymbol{X}$. Finally, for every pair of random vectors $\boldsymbol{X},\boldsymbol{Y}$ of the same dimension and scalar $s>0$ define

$$\rho(\boldsymbol{X},\boldsymbol{Y}) := \sup_{t\in\mathbb{R}} |\mathbb{P}(\|\boldsymbol{X}\|_\infty \leqslant t) - \mathbb{P}(\|\boldsymbol{Y}\|_\infty \leqslant t)|$$

$$\Delta(\boldsymbol{X},s) := \sup_{t\in\mathbb{R}} \mathbb{P}(t \leqslant \|\boldsymbol{X}\|_\infty \leqslant t+s)$$

Combining equations (83)–(86) in Giessing and Fan (2020) yields

$$|\mathbb{P}\left[S \leqslant c^*(\tau)\right] - \tau| \leqslant \rho(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{Q}}_G) + \inf_{\delta_1>0}\left\{\sqrt{\delta_1}\frac{\log n}{\omega_{\max}} + \mathbb{P}(\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\max} > \delta_1)\right\} + \inf_{\delta_2>0}\left\{\delta_2\frac{\sqrt{\log n}}{\omega_{\max}} + \mathbb{P}(\|\boldsymbol{Q} - \widetilde{\boldsymbol{Q}}\|_\infty > \delta_2)\right\} \tag{B.6}$$

where $\widetilde{\boldsymbol{Q}}$ is defined below.

We start by Bounding the first term to the right-hand side of (B.6). We adapt the classical "big block-small block" which was used in the proof of Theorem E.1 in Chernozhukov et al. (2018). Consider two sequences of non-negative integers $a := a_T$ and $b := b_T$ such that $b < a$, $a + b \leqslant T$, $\min\{a, b\} \to \infty$, $a = o(T)$ and $b = o(a)$ as $T \to \infty$. Let $m := [T/(a+b)]$ and define for $j \in \{1, \ldots, m\}$ consecutive blocks of size a and b with index set $\mathcal{A}_j := \{((j-1)(a+b)+1, \ldots, (j-1)(a+b)+a\}$ and $\mathcal{B}_j := \{(j-1)(a+b)+a+1, \ldots j(a+b)\}$. Finally, set $\mathcal{C} := \{m(a+b)+1 \ldots, T\}$, which might be empty. Thus,

$$\boldsymbol{A}_j := \frac{1}{\sqrt{a}}\sum_{t\in\mathcal{A}_j}\widetilde{\boldsymbol{D}}_t, \qquad \boldsymbol{B}_j = \frac{1}{\sqrt{b}}\sum_{t\in\mathcal{B}_j}\widetilde{\boldsymbol{D}}_t, \qquad \text{and} \qquad \boldsymbol{C} = \frac{1}{\sqrt{|\mathcal{C}|}}\sum_{t\in\mathcal{C}}\widetilde{\boldsymbol{D}}_t,$$

such that

$$\widetilde{\boldsymbol{Q}} := \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{\boldsymbol{D}}_t = \sqrt{\frac{ma}{T}}\underbrace{\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\boldsymbol{A}_j\right)}_{=:\boldsymbol{V}} + \sqrt{\frac{mb}{T}}\underbrace{\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\boldsymbol{B}_j\right)}_{=:\boldsymbol{L}} + \sqrt{\frac{T-m(a+b)}{T}}\boldsymbol{C}$$

Let $\widetilde{\boldsymbol{V}} := \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\widetilde{\boldsymbol{A}}_j$ where $\{\widetilde{\boldsymbol{A}}_t, 1 \leqslant t \leqslant m\}$ is an independent sequence such that $\boldsymbol{A}_t$ and $\widetilde{\boldsymbol{A}}_t$ have the same distribution for all $1 \leqslant t \leqslant m$. Similarly define $\widetilde{\boldsymbol{L}} := \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\widetilde{\boldsymbol{B}}_j$. Lemma C.7 give us for any scalar $s > 0$

$$\rho\left(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{Q}}_G\right) \leqslant \rho\left(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{V}}_G\right) + \rho\left(\sqrt{\frac{ma}{T}}\widetilde{\boldsymbol{V}}_G, \widetilde{\boldsymbol{Q}}_G\right) + \Delta\left(\sqrt{\frac{ma}{T}}\widetilde{\boldsymbol{V}}_G, s\right) + \mathbb{P}\left(\sqrt{\frac{mb}{T}}\|\widetilde{\boldsymbol{L}}\|_\infty > s\right) + \rho\left(\boldsymbol{V}, \widetilde{\boldsymbol{V}}\right) + \rho\left(\boldsymbol{L}, \widetilde{\boldsymbol{L}}\right). \tag{B.7}$$

Notice that any measurable $A \subseteq \mathbb{R}^2$ we have $|\mathbb{P}[(\boldsymbol{A}_1, \boldsymbol{A}_2) \in A] - \mathbb{P}[\widetilde{\boldsymbol{A}}_1, \widetilde{\boldsymbol{A}}_2,]| \leqslant \alpha_b$ where $\{\alpha_n, n \in \mathbb{N}\}$ denote the α-mixing coefficient of the sequence $(\widetilde{\boldsymbol{D}}_t)$ which is the same of the sequence $(\boldsymbol{U}_t)$. Then, the last two terms in (B.7) can be upper bounded by $(m-1)\alpha_b$ and $(m-1)\alpha_a$ respectively by induction. Since α_n is non-increasing in n and $a \geqslant b$ we have that

$$\rho\left(\boldsymbol{V}, \widetilde{\boldsymbol{V}}\right) + \rho\left(\boldsymbol{L}, \widetilde{\boldsymbol{L}}\right) \leqslant 2(m-1)\alpha_b \leqslant 2T\exp(-2cb). \tag{B.8}$$

where we use Assumption 3(c) to obtain the last inequality.

For the fourth term we have by the maximal and Markov's inequalities that

$$\mathbb{P}\left(\sqrt{\frac{mb}{T}}\|\widetilde{\boldsymbol{L}}\|_\infty > s\right) \leqslant \left\{\psi\left[\frac{s\sqrt{T}}{C_\psi \psi_{p/2}^{-1}(n)\sqrt{mb}}\right]\right\}^{-1}.$$

By the anti-concentration inequality for Gaussian random vectors (Theorem 7 in Giessing and Fan (2020) with $p=\infty$) we have that

$$\Delta\left(\sqrt{\frac{ma}{T}}\widetilde{\boldsymbol{V}}_G, s\right) \lesssim \frac{Ts\sqrt{\log n}}{ma\sigma_{\widetilde{\boldsymbol{V}}}}.$$

Set $s = \frac{C_\psi \psi_{p/2}^{-1}(n)\sqrt{mb}}{\sqrt{T}}\psi_{p/2}^{-1}(T^\gamma)$ for some $\gamma > 0$ then

$$\Delta\left(\sqrt{\frac{ma}{T}}\widetilde{\boldsymbol{V}}_G, s\right) + \mathbb{P}\left(\sqrt{\frac{mb}{T}}\|\widetilde{\boldsymbol{L}}\|_\infty > s\right) \lesssim \frac{T}{ma}\sqrt{\frac{mb}{T}}\frac{\sqrt{\log n}\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(T^\gamma)}{\sigma_{\widetilde{\boldsymbol{V}}}} + \frac{1}{T^\gamma}. \tag{B.9}$$

For the second term we have from Rio (2013) that, for every $\epsilon > 0$,

$$|[\boldsymbol{M}_\ell]_{ij}| = |\mathsf{Cov}(\widetilde{D}_{it}, \widetilde{D}_{j,t-\ell})| \leqslant 2\alpha_\ell^{\epsilon/(2+\epsilon)}\|\widetilde{D}_{it}\|_{2+\epsilon}\|\widetilde{D}_{jt-\ell}\|_{2+\epsilon}.$$

Hence, from Assumption 3 we have that $\|\boldsymbol{M}_\ell\|_{\max} \lesssim \exp(-2c\frac{\epsilon}{2+\epsilon}\ell)$ and

$$\begin{aligned}
\left\|\frac{ma}{T}\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}_G} - \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}_G}\right\|_{\max} &\leqslant \left(1-\frac{ma}{T}\right)\|\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}}\|_{\max} + \|\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}} - \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}}\|_{\max} \\
&\leqslant \left(\frac{b}{a+b} + \frac{a}{T}\right)\|\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}}\|_{\max} + \frac{1}{a}\sum_{|\ell|<a}|\ell|\|\boldsymbol{M}_\ell\|_{\max} + \sum_{a\leqslant|\ell|<T}\|\boldsymbol{M}_\ell\|_{\max} \\
&\lesssim \frac{b}{a} + \frac{a}{T} + \frac{1}{a} + T\exp\left(-2c\frac{\epsilon}{2+\epsilon}a\right),
\end{aligned}$$

where we use the fact that $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}_G} = \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}} = \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{A}}_j} = \boldsymbol{\Sigma}_{\boldsymbol{A}_j} = \sum_{|\ell|<a}(1-|\ell|/a)\boldsymbol{M}_\ell$, $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}_G} = \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}} = \sum_{|\ell|<T}(1-|\ell|/T)\boldsymbol{M}_\ell$, $\sum_{|\ell|<a}|\ell|\|\boldsymbol{M}_\ell\|_{\max} \leqslant c$ for some $c<\infty$ and $\sum_{a\leqslant|\ell|<T}\|\boldsymbol{M}_\ell\|_{\max} \lesssim T\exp(-2c\frac{\epsilon}{2+\epsilon}a)$. Finally, we can bound the second term using Theorem 8 in Giessing and Fan (2020). In particular, for $p=\infty$ it implies that

$$\begin{aligned}
\rho\left(\sqrt{\frac{ma}{T}}\widetilde{\boldsymbol{V}}_G, \widetilde{\boldsymbol{Q}}_G\right) &\lesssim \frac{\log n\sqrt{\|\frac{ma}{T}\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{V}}_G} - \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}_G}\|_{\max}}}{\sqrt{\frac{ma}{T}}\sigma_{\widetilde{\boldsymbol{V}}} \vee \sigma_{\widetilde{\boldsymbol{Q}}}} \\
&\lesssim \frac{\sqrt{\frac{T}{ma}}\log n\sqrt{\frac{b}{a}+\frac{a}{T}+\frac{1}{a}+T\exp(-\frac{2c\epsilon}{2+\epsilon}a)}}{\sigma_{\widetilde{\boldsymbol{V}}} \vee \sigma_{\widetilde{\boldsymbol{Q}}}}.
\end{aligned}$$

For the first term we have that $\|\widetilde{D}_{it}\|_{\psi_{p/2}}$ is uniformly (upper) bounded by Assumption 3(a). Then, so is $\|\widetilde{A}_{it}\|_{\psi_{p/2}} = \|A_{it}\|_{\psi_{p/2}} = \|\frac{1}{\sqrt{a}}\sum_{s\in\mathcal{A}_t}\widetilde{D}_{is}\|_{\psi_{p/2}}$. Also $(\mathbb{E}(\max_i|\widetilde{A}_{it}|)^3)^{1/3} \lesssim \|\max_i|\widetilde{A}_{it}|\|_{\psi_{p/2}} \lesssim \psi_{p/2}^{-1}(n)\max_i\|\widetilde{A}_{it}\|_{\psi_{p/2}} \lesssim \psi_{p/2}^{-1}(n)$. Since $\{\widetilde{A}_t, 1\leqslant t\leqslant m\}$ is an iid

sequence of random vectors, Theorem 5 in Giessing and Fan (2020) implies

$$\rho(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{V}}_G) \lesssim \frac{(\log n)^{7/6}\psi_{p/2}^{-1}(n)}{T^{1/6}\sigma_{\widetilde{\boldsymbol{V}}}}. \tag{B.10}$$

By the triangle inequality we have that $\sigma^2_{\widehat{\boldsymbol{V}}} \geqslant \sigma^2_{\widetilde{\boldsymbol{Q}}} - \|d_{\widetilde{\boldsymbol{Q}}} - d_{\widehat{\boldsymbol{V}}}\|_{\max} \geqslant \underline{c} - \|\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{Q}}} - \boldsymbol{\Sigma}_{\widehat{\boldsymbol{V}}}\|_{\max} \gtrsim c - \frac{1}{a} - T\exp(-2c\frac{\epsilon}{2+\epsilon}a)$. By setting $a = [\sqrt{T}]$, we conclude that $\sigma^2_{\widehat{\boldsymbol{V}}}$ is eventually bounded away from zero for large enough T. If we further set $b = [\log T/c]$ and $\gamma = 1/4$ and apply (B.8)-(B.10) to bound the right-hand side of (B.6), we obtain

$$\rho\left(\widetilde{\boldsymbol{Q}}, \widetilde{\boldsymbol{Q}}_G\right) = O\left[\frac{(\log n)^{7/6}\psi_{p/2}^{-1}(n)}{T^{1/6}} + \frac{\sqrt{\log T \log n}\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(T^{1/4})}{T^{1/4}}\right].$$

Finally, we bound the last two terms appearing in (B.6). Let γ_1 and γ_2 be positive sequences depending on n and T such that $\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\max} = O_P(\gamma_1)$ and $\|\boldsymbol{Q} - \widetilde{\boldsymbol{Q}}\|_\infty = O_P(\gamma_2)$. Suppose we can state conditions under which

$$\log^3 n(\gamma_1 \vee \gamma_2) = o(1) \quad T, n \to \infty. \tag{B.11}$$

Then, the last two terms vanish in probability if we set $\delta_1 = \gamma_1 \log n$ and $\delta_2 = \gamma_2 \log n$ in (B.6). Lemmas C.8 and C.10 give us expressions for γ_1 and γ_2, respectively, which combined with the rate assumptions in the theorem implies (B.11).

Appendix C. Additional Lemmas

Lemma C.1. *Let a_j and b_j denote the j-th eigenvalue in decreasing order of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ respectively. Then, under Assumption 2(b) and (c): (a) $b_j \asymp n$ for $1 \leqslant j \leqslant r$; (b) $\max_{j\leqslant n}|a_j - b_j| = O(1)$; and (c) $a_j \asymp n$ for $1 \leqslant j \leqslant r$.*

Proof. Result (a) follows from the fact that the r eigenvalues of $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ are also (the only r non-zero) eigenvalues of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and Assumption 2(b). Part (b) follows from Weyl's inequality that implies $\max_{j\leqslant n}|a_j - b_j| \leqslant \|\boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}'\| = O(1)$, where the last equality follows from Assumption 2(c). Finally, result (c) follows from part (a) and (b) and the (reverse) triangle inequality. □

Recall that $\boldsymbol{\Sigma}$ is the $(n \times n)$ covariance matrix of $\boldsymbol{U}_t = \boldsymbol{Z}_t - \boldsymbol{\Gamma}\boldsymbol{W}_t$. Let $\widetilde{\boldsymbol{\Sigma}} := \frac{1}{T}\sum_{t=1}^T \boldsymbol{U}_t\boldsymbol{U}_t'$ and $\widehat{\boldsymbol{\Sigma}}$ the same as $\widetilde{\boldsymbol{\Sigma}}$ but with $\boldsymbol{\Gamma}$ replaced by the estimator $\widehat{\boldsymbol{\Gamma}}$. Let $\widehat{a}_j$ denote the j-th eigenvalue in decreasing order of $\widehat{\boldsymbol{\Sigma}}$.

Lemma C.2. *Let ω_1 be a non-negative sequence of n and T such that $\|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} = O_P(\omega_1)$. Then, under the Assumptions 2 and 3: (a) $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P[\omega_1 + \psi_{p/2}^{-1}(n^2)/\sqrt{T}]$;*

(b) $\max_{j\leqslant n}|\widehat{a}_j - a_j| = O_P[n(\omega_1 + \psi_{p/2}^{-1}(n^2)/\sqrt{T})]$*; and (c)* $\widehat{a}_j \asymp_P n$ *for* $j \leqslant r$ *provided that* $\omega_1 + \psi_{p/2}^{-1}(n^2)/\sqrt{T} = O_P(1)$.

Proof. Part (a) follows by the triangle and maximal inequalities, since $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leqslant \|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} + \|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_P(\omega_1) + O_P\left[\psi_{p/2}^{-1}(n^2)/\sqrt{T}\right]$. Part (b) follows from Weyl's inequality, the fact that $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leqslant n\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}$ and part (a). Part (c) follows from the triangle inequality combined with part (b) and Lemma C.1(c). □

The Lemmas C.3-C.6 below are an adaption of Lemmas 8–10 in Fan et al. (2013), henceforth FLM, to include the estimation error in the sample covariance matrix. To avoid confusion and make it easier for the reader to follow through the changes we use the same notation adopted in FLM. In particular, if δ_{it} denotes the (i,t) element of $\boldsymbol{\Delta} := \widehat{\boldsymbol{R}} - \boldsymbol{R}$ then $\widetilde{U}_{it} = U_{it} + \delta_{it}$ for $i \in [n]$ and $t \in [T]$. We consider that $\|\boldsymbol{\Delta}\|_{\max} = O_P(\omega)$ for some non-negative sequence ω depending on n and T.

Define:

$$\widetilde{\zeta}_{st} := \frac{\widetilde{\boldsymbol{U}}'_s\widetilde{\boldsymbol{U}}_t}{n} - \frac{\mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)}{n} = \left(\frac{\boldsymbol{U}'_s\boldsymbol{U}_t}{n} - \frac{\mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)}{n}\right) + \left(\frac{\boldsymbol{U}'_s\boldsymbol{\delta}_t}{n} + \frac{\boldsymbol{\delta}'_s\boldsymbol{U}_t}{n} + \frac{\boldsymbol{\delta}'_s\boldsymbol{\delta}_t}{n}\right) =: \zeta_{st} + \zeta^*_{st}$$

$$\widetilde{\eta}_{st} := \frac{\boldsymbol{f}'_s\sum_{i=1}^n \boldsymbol{\lambda}_i\widetilde{U}_{it}}{n} = \frac{\boldsymbol{f}'_s\sum_{i=1}^n \boldsymbol{\lambda}_i U_{it}}{n} + \frac{\boldsymbol{f}'_s\sum_{i=1}^n \boldsymbol{\lambda}_i\delta_{it}}{n} =: \eta_{st} + \eta^*_{st}$$

$$\widetilde{\xi}_{st} := \frac{\boldsymbol{F}'_t\sum_{i=1}^n \boldsymbol{\lambda}_i\widetilde{U}_{is}}{n} = \frac{\boldsymbol{F}'_t\sum_{i=1}^n \boldsymbol{\lambda}_i U_{is}}{n} + \frac{\boldsymbol{F}'_t\sum_{i=1}^n \boldsymbol{\lambda}_i\delta_{is}}{n} = \xi_{st} + \xi^*_{st}.$$

Lemma C.3. *Under Assumption 3: (a)* $\zeta_{st} = O_P(1/\sqrt{n})$*; (b)* $\eta_{st} = O_P(1/\sqrt{n})$*; (c)* $\xi_{st} = O_P(1/\sqrt{n})$*; (d)* $\zeta^*_{st} = O_P(\omega + \omega^2)$ *and* $\max_{s,t\leqslant T}\zeta^*_{st} = O_P\left[\psi^{-1}(nT)\omega + \omega^2\right]$*; (e)* $\eta^*_{st} = O_P(\omega)$*; and (f)* $\xi^*_{st} = O_P(\omega)$.

Proof. Parts $(a) - -(c)$ are straightforward. For (d) we have that $\frac{1}{n}\boldsymbol{U}'_s\boldsymbol{U}_t = O_P(1)$ and $\frac{1}{n}\boldsymbol{\delta}'_s\boldsymbol{\delta}_t \leqslant \|\boldsymbol{\Delta}\|^2_{\max} = O_P(\omega^2)$. Then, the other two terms in parentheses in the definition of ζ^*_{st} are $O_P(\omega)$ by the Cauchy-Schwartz inequality. Part (e) and (f) follows by similar arguments. Therefore,

$$\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^T\left(\frac{1}{n}\boldsymbol{\delta}'_s\boldsymbol{U}_t\right)^2 = \max_{t\leqslant T}\frac{1}{n^2}\boldsymbol{U}'_t\left(\frac{1}{T}\sum_{s=1}^T\boldsymbol{\delta}_s\boldsymbol{\delta}'_s\right)\boldsymbol{U}_t \leqslant \|\boldsymbol{\Delta}\|^2_{\max}\left(\max_{t\leqslant T}\|\boldsymbol{U}_t\|_1/n\right)^2$$

and

$$\zeta^*_{st} \leqslant \|\boldsymbol{U}_s\|_\infty\|\boldsymbol{\delta}_t\|_\infty + \|\boldsymbol{U}_t\|_\infty\|\boldsymbol{\delta}_s\|_\infty + \|\boldsymbol{\delta}_t\|_\infty\|\boldsymbol{\delta}_s\|_\infty \leqslant 2\|\boldsymbol{U}\|_{\max}\|\boldsymbol{\Delta}\|_{\max} + \|\boldsymbol{\Delta}\|^2_{\max}.$$

□

Lemma C.4. *Under Assumption 3: (a)* $\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{nT}\sum_{s=1}^{T}\widehat{f}_{js}\mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)\right]^2 = O_P(1/T)$*; (b)* $\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\zeta}_{st}\right]^2 = O_P\left[(1/\sqrt{n}+\omega+\omega^2)^2\right]$*; (c)* $\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\eta}_{st}\right]^2 = O_P\left[(1/\sqrt{n}+\omega)^2\right]$*; (d)* $\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\xi_{st}\right]^2 = O_P\left[(1/\sqrt{n}+\omega)^2\right]$*; and finally, (e)* $\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t\right\|^2 = O_P\left[1/T + (1/\sqrt{n}+\omega+\omega^2)^2\right]$.

Proof. Part (a) is unaltered by the presence of a pre-estimation, so it follows directly from Lemma 8(a) in FLM. For part (b), we have that for $s,l \in [n]$ and $j \in [r]$ by Cauchy-Schwartz inequality

$$\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\zeta}_{st}\right]^2 \leqslant \left[\frac{1}{T^2}\sum_{s,l=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T}\widetilde{\zeta}_{st}\widetilde{\zeta}_{lt}\right)^2\right]^{1/2}.$$

Since $\widetilde{\zeta}_{st} = \zeta_{st} + \zeta^*_{st} = O_P(1/\sqrt{n}+\omega+\omega^2)$ by Lemma C.3, the term in parentheses is $O_P\left[(1/\sqrt{n}+\omega+\omega^2)^2\right]$. Result (*b*) follows. For (c), by the triangle inequality and Lemma 8(c) in FLM, we have that $\|\sum_{i=1}^{n}\lambda_{ji}\widetilde{u}_{it}\| \leqslant \|\sum_{i=1}^{n}\lambda_{ji}U_{it}\| + \|\sum_{i=1}^{n}\lambda_{ji}\delta_{it}\| = O_P(\sqrt{n}) + O_P(n\omega)$. Then, we conclude

$$\frac{1}{T}\sum_{t=1}^{T}[\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\widetilde{\eta}_{st}]^2 \leqslant \frac{1}{Tn^2}\sum_{t=1}^{T}\|\sum_{i=1}^{n}\boldsymbol{U}_{it}\lambda_i\|^2 = O_P(1/n+\omega/\sqrt{n}+\omega^2).$$

The proof of part (d) is analogous to (c) and is omitted. For (e), let $[\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t]_j$ denote the j-th entry of $\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t$. Since $\boldsymbol{V}/n$ is bounded away from zero by Lemma C.2(c), the fact that $(a+b+c+d)^2 \leqslant 4(a^2+b^2+c^2+d^2)$ and using (B.1), we have that $\max_{j\leqslant r} T^{-1}\sum_t[\widehat{\boldsymbol{f}}_t - \boldsymbol{H}\boldsymbol{f}_t]_j$ is upper bounded by some constant $C<\infty$ times

$$\left\{\max_{j\leqslant r}\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\frac{\mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)}{n}\right]^2 + \max_{j\leqslant r}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\zeta}_{st}\right)^2\right.$$
$$\left. + \max_{j\leqslant r}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\eta}_{st}\right)^2 + \max_{j\leqslant r}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{s=1}^{T}\widehat{f}_{js}\widetilde{\xi}_{st}\right)^2\right\}.$$

The result follows by applying the bounds from (a)–(d) to each of the terms above.

□

Lemma C.5. *Under Assumption 2: (a)* $\max_{t\leqslant T}\|\frac{1}{nT}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\mathbb{E}(\boldsymbol{U}'_s\boldsymbol{U}_t)\| = O_P(1/\sqrt{T})$*; (b)* $\max_{t\leqslant T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\widetilde{\zeta}_{st}\| = O_P\left[\sqrt{\psi^{-1}_{p/2}(T)/n} + \psi^{-1}(nT)\omega + \omega^2\right]$*; (c)* $\max_{t\leqslant T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\widetilde{\eta}_{st}\| = O_P\left[\psi^{-1}(T)/\sqrt{n}+\omega\right]$*; and (d)* $\max_{t\leqslant T}\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\xi_{st}\| = O_P\left[\psi^{-1}(T)(1/\sqrt{n}+\omega)\right]$.

Proof. Part (a) is unaltered by pre-estimation, so it follows directly from Lemma 9(a) in FLM. For part (b), from the Cauchy-Schwartz inequality we have

$$\max_{t\leqslant T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\widetilde{\zeta}_{st}\right\| \leqslant \left(\frac{1}{T}\sum_{s=1}^{T}\|\widehat{\boldsymbol{f}}_s\|^2\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\widetilde{\zeta}_{st}^2\right)^{1/2}.$$

The first summation inside the parentheses equal r due to the normalization. For the second summation, by the triangle inequality, we have $\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\widetilde{\zeta}_{st}^2 \leqslant \max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2 + 2\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}\zeta_{st}^* + \max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^{*\,2}$. For the first term, the maximum inequality followed by Assumption 2(e) yields

$$\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\zeta_{st}^2 = O_P\left[\psi_{p/2}^{-1}(T)\max_{s,t}\|\zeta^2\|_{\psi_{p/2}}\right] = O_P\left[\psi_{p/2}^{-1}(T)\max_{s,t}\|\zeta\|_{\psi}^2\right] = O_P\left[\frac{\psi_{p/2}^{-1}(T)}{n}\right].$$

The last one is $O_P\left[(\psi^{-1}(nT)w+\omega^2)^2\right]$ by Lemma C.3(d). Then, by Cauchy Schwartz we have that $\max_{t\leqslant T}\frac{1}{T}\sum_{s=1}^{T}\widetilde{\zeta}_{st}^2 = O_P\left[(\sqrt{\psi_{p/2}^{-1}(T)/n}+\psi^{-1}(nT)w+\omega^2)^2\right]$ and result (b) follows.

For (c), by the triangle inequality we have that $\max_{t\leqslant T}\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i\widetilde{U}_{it}\| \leqslant \max_{t\leqslant T}\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i U_{it}\| + \max_{t\leqslant T}\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i\delta_{it}\|$. For the first term, the maximum inequality followed by Assumption 2(f) yields

$$\max_{t\leqslant T}\left\|\frac{1}{n}\boldsymbol{\Lambda}'\boldsymbol{U}_t\right\| = O_P\left[\frac{\psi^{-1}(T)}{\sqrt{n}}\max_{t}\left\|\frac{1}{\sqrt{n}}\boldsymbol{\Lambda}'\boldsymbol{U}_t\right\|\right] = O_P\left[\psi^{-1}(T)/\sqrt{n}\right].$$

The second term is upper bounded by $\mathrm{r}\|\boldsymbol{\Lambda}\|_{\max}\|\boldsymbol{\Delta}\|_{\max} = O_P(\omega)$ by Assumption 2(d). We obtain the result since

$$\max_{t\leqslant T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\widetilde{\eta}_{st}\right\| \leqslant \left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\boldsymbol{f}_s'\right\|\max_{t\leqslant T}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i\widetilde{U}_{it}\right\| = O_P\left[\frac{\psi^{-1}(T)}{\sqrt{n}}+\omega\right]. \tag{C.1}$$

By the triangle inequality, $\|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i\widetilde{U}_{is}\widehat{\boldsymbol{f}}_s\| \leqslant \|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i U_{is}\widehat{\boldsymbol{f}}_s\| + \|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i\delta_{is}\widehat{f}_{is}\|$. Lemma 9(d) of FLM shows that the first term is $O_P(1/\sqrt{n})$. For the second term, for each $j\in[r]$:

$$\left\|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i\delta_{is}\widehat{f}_{js}\right\|^2 \leqslant \left(\frac{1}{T}\sum_{s=1}^{n}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\lambda}_i\delta_{is}\right\|^2\widehat{f}_{js}\right)\left(\frac{1}{T}\sum_{s=1}^{n}\widehat{f}_{js}^2\right) = O_P(\omega^2).$$

Thus, $\|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i\widetilde{U}_{it}\widehat{\boldsymbol{f}}_s\| = O_P(1/\sqrt{n}+\omega)$ and by Cauchy-Schwartz inequality:

$$\max_{t\leqslant T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widehat{\boldsymbol{f}}_s\xi_{st}\right\| \leqslant \max_{t\leqslant T}\|\boldsymbol{F}_t\|\left\|\frac{1}{nT}\sum_s\sum_i\boldsymbol{\lambda}_i\widetilde{U}_{it}\widehat{\boldsymbol{f}}_s\right\| = O_P\left[\psi^{-1}(T)(1/\sqrt{n}+\omega)\right]. \tag{C.2}$$

□

Lemma C.6. *Let $\omega_1+\psi^{-1}(n^2)/\sqrt{T}=O(1)$ where ω_1 is defined in Lemma C.2. Then, under Assumption 3 we have: (a) $\|\boldsymbol{V}^{-1}\|=O_P(1/n)$; (b) $\|\boldsymbol{H}\|=O_P(1)$; (c) $\|\boldsymbol{H}'\boldsymbol{H}-\boldsymbol{I}_r\|_F=O_P(1/\sqrt{T}+1/\sqrt{n}+\omega)$; (d) $\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T\widehat{R}_{it}^2=O_P\left[w(\psi^{-1}(nT)+\omega)+\psi_{p/2}^{-1}(n)/\sqrt{T}+1\right]$; and (e) $\max_{i\leqslant n}\max_{j\leqslant r}\frac{1}{T}\sum_{t=1}^T F_{jt}\widetilde{U}_{it}=O_P\left[\psi_{p/2}^{-1}(n)/\sqrt{T}+\omega\right]$.*

Proof. We have that $\boldsymbol{V}^{-1}=\mathsf{diag}\,(1/\widehat{a}_1,\ldots,1/\widehat{a}_r)$ and $1/\widehat{a}_j\asymp_P 1/n$ for $j\leqslant r$ by Lemma C.2(c). The result (a) then follows. The normalization tell us $\|\widehat{\boldsymbol{F}}\|=\sqrt{T}$, Lemma 11(a) in FLM give us $\|\boldsymbol{F}\|=O_P(\sqrt{T})$, $\|\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\|=\widetilde{a}_1\asymp n$ by Lemma C.1(a) and from part (b) we have $\|\boldsymbol{V}^{-1}\|=O_P(1/n)$. Result (b) then follows since by definition $\boldsymbol{H}:=T^{-1}\boldsymbol{V}^{-1}\widehat{\boldsymbol{F}}'\boldsymbol{F}\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$. For (c) we have by the triangle inequality

$$\|\boldsymbol{H}'\boldsymbol{H}-\boldsymbol{I}_r\|_F\leqslant\|\boldsymbol{H}'\boldsymbol{H}-\boldsymbol{H}'\boldsymbol{F}'\boldsymbol{F}/T\boldsymbol{H}\|_F+\|\boldsymbol{H}'\boldsymbol{F}'\boldsymbol{F}/T\boldsymbol{H}-\boldsymbol{I}_r\|_F$$

For the first term we have

$$\|\boldsymbol{H}'(\boldsymbol{I}_r-\boldsymbol{F}'\boldsymbol{F}/T)\boldsymbol{H}\|_F\leqslant\|\boldsymbol{H}\|^2\|\boldsymbol{I}_r-\boldsymbol{F}'\boldsymbol{F}/T\|_F=O_P(1)O_P(1/\sqrt{T}).$$

The second term is equal to $\|\boldsymbol{H}'\boldsymbol{F}'\boldsymbol{F}/T\boldsymbol{H}-\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}}/T\|_F$.

For (d) we have

$$\begin{aligned}\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T\widehat{R}_{it}^2&\leqslant\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T(\widehat{R}_{it}^2-R_{it}^2)+\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T R_{it}^2-\mathbb{E}(R_{it}^2)+\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T\mathbb{E}(R_{it}^2)\\&\leqslant\max_{i,t}|\widehat{R}_{it}^2-R_{it}^2|+\max_{i\leqslant n}\frac{1}{T}\sum_{t=1}^T R_{it}^2-\mathbb{E}(R_{it}^2)+\max_{i,t}\mathbb{E}(R_{it}^2).\end{aligned}$$

The last term is $O(1)$ by Assumption 3(a), the middle term $O_P(\psi_{p/2}^{-1}(n)/\sqrt{T})$. The first term is no larger then $\|\boldsymbol{\Delta}\|_{\max}(2\|\boldsymbol{R}\|_{\max}+\|\boldsymbol{\Delta}\|_{\max})=O_P(\omega(\psi^{-1}(nT)+\omega))$. The result (d) then follows.

For (e) we have for each $j\leqslant r$:

$$\begin{aligned}\left|T^{-1}\sum_t F_{jt}\widetilde{U}_{it}\right|&\leqslant\left|T^{-1}\sum_t F_{jt}U_{it}\right|+\left|T^{-1}\sum_t F_{jt}\delta_{it}\right|\\&\leqslant\left|T^{-1}\sum_t F_{jt}U_{it}\right|+\left(T^{-1}\sum_t F_{jt}^2T^{-1}\sum_t\delta_{it}^2\right)^{1/2}\end{aligned}$$

The first term is $O_P\left[\psi_{p/2}^{-1}(n)/\sqrt{T}\right]$ by the maximum inequality and Assumption 3 and the second is $O_P(\omega)$. □

Lemma C.7. *For every $s > 0$:*

$$\rho_p(S,Z) \leqslant \rho_p\left(\widetilde{T},\widetilde{Z}\right) + \Delta_p\left(\sqrt{\frac{mq}{n}}\widetilde{Z}, s\right) + \rho_p\left(\sqrt{\frac{mq}{n}}\widetilde{Z}, Z\right) + \mathbb{P}\left(\sqrt{\frac{mr}{n}}\|\widetilde{U}\|_p > s\right) + \rho_p\left(T,\widetilde{T}\right) + \rho_p\left(U,\widetilde{U}\right).$$

Proof. We start by showing that for every pair of random variables X and Ydefined in the same probability space taking values in the normed space $(S, \|\cdot\|)$ and pair of non-negative reals t, s, we have

$$\mathbb{P}(\|X\| \leqslant t-s) - \mathbb{P}(\|Y\| > s) \leqslant \mathbb{P}(\|X+Y\| \leqslant t) \leqslant \mathbb{P}(\|X\| \leqslant t+s) + \mathbb{P}(\|Y\| > s). \quad \text{(C.3)}$$

Indeed, for the right hand side inequality we use $\|X+Y\| = \|X-(-Y)\| \geqslant \|X\| - \|Y\|$. Hence, for any $t, s > 0$:

$$\begin{aligned}\mathbb{P}(\|X+Y\| \leqslant t) &\leqslant \mathbb{P}(\|X\| \leqslant t + \|Y\|) \leqslant \mathbb{P}(\|X\| \leqslant t + \|Y\|, \|Y\| \leqslant s) + \mathbb{P}(\|Y\| > s) \\ &\leqslant \mathbb{P}(\|X\| \leqslant t+s) + \mathbb{P}(\|Y\| > s).\end{aligned}$$

For the other side we use $\|X+Y\| \leqslant \|X\| + \|Y\|$ to write

$$\mathbb{P}(\|X+Y\| \leqslant t) \geqslant \mathbb{P}(\|X\| \leqslant t - \|Y\|) \geqslant \mathbb{P}(\|X\| \leqslant t - \|Y\|) + \mathbb{P}(\|Y\| > s) - \mathbb{P}(\|Y\| > s).$$

Now replace X and Y by $\sqrt{\frac{mq}{n}}T$ and $\sqrt{\frac{mr}{n}}U$ in (C.3), respectively and set $\|\cdot\| = \|\cdot\|_p$. The right hand side of the resulting expression can be upper bounded by $\mathbb{P}(\sqrt{\frac{mq}{n}}\|\widetilde{T}\|_p \leqslant t+s) + \mathbb{P}(\sqrt{\frac{mr}{n}}\|\widetilde{U}\| > s) + \rho_p(T,\widetilde{T}) + \rho_p(U,\widetilde{U})$, whereas the left hand side can be lower bounded by $\mathbb{P}(\sqrt{\frac{mq}{n}}\|\widetilde{T}\| \leqslant t-s) - \mathbb{P}(\sqrt{\frac{mq}{r}}\|\widetilde{U}\| > s) - \rho_p(T,\widetilde{T}) - \rho_p(U,\widetilde{U})$. Therefore,

$$\begin{aligned}&\mathbb{P}\left(\sqrt{\frac{mq}{n}}\|\widetilde{T}\|_p \leqslant t-s\right) - \mathbb{P}\left(\sqrt{\frac{mr}{n}}\|\widetilde{U}\|_p > s\right) - \rho_p\left(T,\widetilde{T}\right) - \rho_p\left(U,\widetilde{U}\right) \\ &\quad\leqslant \mathbb{P}\left(\|S\|_p \leqslant t\right) \leqslant \mathbb{P}\left(\sqrt{\frac{mq}{n}}\|\widetilde{T}\|_p \leqslant t+s\right) + \mathbb{P}\left(\sqrt{\frac{mr}{n}}\|\widetilde{U}\|_p > s\right) + \rho_p\left(T,\widetilde{T}\right) + \rho_p\left(U,\widetilde{U}\right).\end{aligned}$$

Then, for the right-hand side,

$$\begin{aligned}\mathbb{P}\left(\sqrt{\frac{mq}{n}}\|\widetilde{T}\|_p \leqslant t+s\right) &\leqslant \mathbb{P}\left(\sqrt{\frac{mq}{n}}\|\widetilde{Z}\|_p \leqslant t+s\right) + \rho_p\left(\widetilde{T},\widetilde{Z}\right) \\ &\leqslant \mathbb{P}\left(\sqrt{\frac{mq}{n}}\|\widetilde{Z}\|_p \leqslant t\right) + \Delta_p\left(\sqrt{\frac{mq}{n}}\widetilde{Z}, s\right) + \rho_p\left(\widetilde{T},\widetilde{Z}\right) \\ &\leqslant \mathbb{P}\left(\|Z\|_p \leqslant t\right) + \rho_p\left(\sqrt{\frac{mq}{n}}\widetilde{Z}, Z\right) + \Delta_p\left(\sqrt{\frac{mq}{n}}\widetilde{Z}, s\right) + \rho_p\left(\widetilde{T},\widetilde{Z}\right).\end{aligned}$$

Similarly for the left-hand side and the proof is completed. □

By the triangle inequality $\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\mathsf{max}} \leqslant \|\widehat{\boldsymbol{\Upsilon}} - \widetilde{\boldsymbol{\Upsilon}}\|_{\mathsf{max}} + \|\widetilde{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\mathsf{max}}$ where $\widetilde{\boldsymbol{\Upsilon}}$ is the sample covariance matrix of $\widetilde{\boldsymbol{D}}_t := U_{1t}\boldsymbol{U}_{-1t}$. The second term is $O(\psi_{p/4}^{-1}(n^2)/\sqrt{T})$ while for the first $\|\widehat{\boldsymbol{\Upsilon}} - \widetilde{\boldsymbol{\Upsilon}}\|_{\mathsf{max}} \leqslant \|\boldsymbol{D} - \widetilde{\boldsymbol{D}}\|_{\mathsf{max}}(2\|\widetilde{\boldsymbol{D}}\|_{\mathsf{max}} + \|\boldsymbol{D} - \widetilde{\boldsymbol{D}}\|_{\mathsf{max}})$. The first term in parentheses is $O\left[\psi_*^{-1}(nT)\right]$ and the second can be upper bounded by $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathsf{max}}(2\|\boldsymbol{U}\|_{\mathsf{max}} + \|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\mathsf{max}})$ which is shown to be $O_P\left[\eta(n,T)\psi^{-1}(nT)\right]$ in the proof of Lemma C.16. Therefore, we conclude that

$$\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\mathsf{max}} = O_P\left(\eta(n,T)\psi^{-1}(nT)\psi_{p/2}^{-1}(nT) + \psi_{p/4}^{-1}(n^2)/\sqrt{T}\right).$$

To leverage on the results of Gaussian approximation, in particular on the work of Giessing and Fan (2020) we would like to establish some sort of asymptotic linearity namely

$$\boldsymbol{Q}_T = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{D}_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\widetilde{\boldsymbol{D}}_t + \boldsymbol{R}_T =: \widetilde{\boldsymbol{Q}}_T + \boldsymbol{R}_T. \tag{C.4}$$

such that $\|\boldsymbol{R}_t\|_\infty$ vanishes in probability at an appropriate rate as $n, T \to \infty$. Then we can approximate the distribution of $S = \|\boldsymbol{Q}\|_\infty$ by the distribution of $\widetilde{S} := \|\widetilde{\boldsymbol{Q}}\|_p$, which in turn can be approximated by the distribution of $S^* := \|\boldsymbol{Q}^*\|_\infty$ with high probability.

For some $\epsilon > 0$ we might set

$$\delta_1 = h[\eta(n,T)(\psi^{-1}(nT))^3 + \psi_{p/4}^{-1}(n^2)/\sqrt{T}]$$
$$\delta_2 = \eta^{1-\epsilon}[\psi^{-1}(n) + \sqrt{T}\eta]$$

Lemma C.8. $\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\mathsf{max}} = O_P\left(h[\eta(\psi_p^{-1}(nT))^3 + \psi_{p/4}^{-1}(n^4)/\sqrt{T}]\right)$

Proof. Let $\boldsymbol{i} := (i_1, i_2, i_3, i_4)$ be a multi-index where $i_1, i_2, i_3, i_4 \in [n]$. Define for $\boldsymbol{i}$ and $|\ell| < T$:

$$\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} := \frac{1}{T}\sum_{t=|\ell|+1}^{T} U_{i_1,t}U_{i_2,t}U_{i_3,t-|\ell|}U_{i_4,t-|\ell|}; \qquad \gamma_{\boldsymbol{i}}^{\ell} := \mathbb{E}\widetilde{\gamma}_{\boldsymbol{i}},$$

and $\widehat{\gamma}_{\boldsymbol{i}}^{\ell}$ as $\widetilde{\gamma}_{\boldsymbol{i}}^{\ell}$ with U's replaced by $\widehat{U}$'s. Also define

$$\widetilde{\upsilon}_{\boldsymbol{i}} := \sum_{|\ell|<T} k(\ell/h)\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} \qquad \upsilon_{\boldsymbol{i}} := \sum_{|\ell|<T}\gamma_{\boldsymbol{i}}^{\ell},$$

and $\widehat{\upsilon}_{\boldsymbol{i}}$ as $\widetilde{\upsilon}_{\boldsymbol{i}}$ with U's replaced by $\widehat{U}$'s. Then we write

$$\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}} = \sum_{|\ell|<T} k(\ell/h)(\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}) + \sum_{|\ell|<T}(k(\ell/h) - 1)\gamma_{\boldsymbol{i}}^{\ell}. \tag{C.5}$$

Since $\|\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}\|_{\psi_{p/4}} = O(\sqrt{T-|\ell|}/T) = O(1/\sqrt{T})$, the $\psi_{p/4}$-Orlicz norm of the first term is bounded by

$$h \sum_{|\ell|<T} |h^{-1}k(\ell/h)| \|\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}\|_{\psi_{p/4}} = O\left(\frac{h}{\sqrt{T}} \int |k(u)|du\right) = O(h/\sqrt{T}),$$

whereas the second term is deterministic and is shown to be $O(h/\sqrt{T})$ by Andrews (1991). Thus $\|\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}\|_{\psi_{p/4}} = O(h/\sqrt{T})$ uniformly in $\boldsymbol{i} \in [n]^4$. Thus, by the maximal inequality followed by Markov's inequality we conclude that

$$\max_{\boldsymbol{i}} |\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}| = O_P(\psi_{p/4}^{-1}(n^4) \max_{\boldsymbol{i}} \|\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}\|_{\psi_{p/4}}) = O_P[\psi_{p/4}^{-1}(n^4)h/\sqrt{T}]. \tag{C.6}$$

We now use the fact that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^q$ we have $|\prod_{i=1}^q x_i - \prod_{i=1}^q y_i| = O(\sum_{i=0}^{q-1} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{n-i} \|\boldsymbol{y}\|_\infty^i)$ combined with the fact that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = o(1)$ to obtain

$$\begin{aligned}
\max_{\boldsymbol{i},\ell} |\widehat{\gamma}_{\boldsymbol{i}}^{\ell} - \widetilde{\gamma}_{\boldsymbol{i}}^{\ell}| &\leqslant \max_{\boldsymbol{i},t,\ell} |\widehat{U}_{i_1,t}\widehat{U}_{i_2,t}\widehat{U}_{i_3,t-|\ell|}\widehat{U}_{i_4,t-|\ell|} - U_{i_1,t}U_{i_2,t}U_{i_3,t-|\ell|}U_{i_4,t-|\ell|}| \\
&= O(\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} \|\boldsymbol{U}\|_{\max}^3) \\
&= O_P[\eta[\psi_p^{-1}(nT)]^3]
\end{aligned}$$

Therefore we conclude

$$\max_{\boldsymbol{i}} |\widehat{\upsilon}_{\boldsymbol{i}} - \widetilde{\upsilon}_{\boldsymbol{i}}| \leqslant \max_{\boldsymbol{i},\ell} |\widehat{\gamma}_{\boldsymbol{i}}^{\ell} - \widetilde{\gamma}_{\boldsymbol{i}}^{\ell}| \sum_{|\ell|<T} |k(\ell/h)| = O_P\left(h\eta[\psi_p^{-1}(nT)]^3 \int |k(u)|du\right) = O_P(h\eta[\psi_p^{-1}(nT)]^3). \tag{C.7}$$

The result then follows from the triangle inequality $\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\max} \leqslant \max_{\boldsymbol{i}} |\widehat{\upsilon}_{\boldsymbol{i}} - \widetilde{\upsilon}_{\boldsymbol{i}}| + \max_{\boldsymbol{i}} |\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}|$, expression (C.10) and (C.11). □

Lemma C.9. *If $\|\delta_{it}\|_{\psi_p} \leqslant C < \infty$ where $\delta_{it} := \widehat{R}_{it} - R_{it}$ then*

$$\|\|(\boldsymbol{V}/n)(\boldsymbol{F}_t - \boldsymbol{H}\boldsymbol{F}_t)\|_2\|_{\psi_p} = O(\frac{1}{\sqrt{T}} + \frac{\psi_{p/2}^{-1}(T)}{\sqrt{n}} + \psi_{p/2}^{-1}(T)C).$$

Proof. In this proof we use the fact that for any (possibly random) A_{st}, by Cauchy-Schwartz inequality and the normalization $\widehat{\boldsymbol{F}}\widehat{\boldsymbol{F}}/T = \boldsymbol{I}_r$, we have $\|\frac{1}{T}\sum_{s=1}^T \widehat{\boldsymbol{F}}_s A_{st}\| \leqslant \sqrt{r}\left(\frac{1}{T}\sum_{s=1}^T A_{st}^2\right)^{1/2}$. Thus

$$g(A_{st}) := \left\| \|\frac{1}{T}\sum_{s=1}^T \widehat{\boldsymbol{F}}_s A_{st}\| \right\|_\psi = O\left[\left\| \left(\frac{1}{T}\sum_{s=1}^T A_{st}^2\right)^{1/2} \right\|_\psi\right].$$

(a) Set $A_{st} = \mathbb{E}(\boldsymbol{U}_s'\boldsymbol{U}_t)/n$, then $g(A_{st}) = O(1/\sqrt{T})$.

(b) Set $A_{st} = \widetilde{\zeta}_{st} := (\boldsymbol{U}_s'\boldsymbol{U}_t - \mathbb{E}(\boldsymbol{U}_s'\boldsymbol{U}_t))/n$, then by maximal inequality $g(A_{st}) = O(\|\max_{s\leqslant T} |\widetilde{\zeta}_{st}|\|_\psi) = O(\psi^{-1}(T) \max_{s\leqslant T} \|\widetilde{\zeta}_{st}\|_\psi)$. By the triangle inequality

$\|\widetilde{\zeta}_{st}\|_\psi \leqslant \|\zeta_{st}\|_\psi + \|\zeta^*_{st}\|_\psi$. The first term is $O(1/\sqrt{n})$ by Assumption 3(d). The second can be upper bounded by $\|\boldsymbol{U}'_s\boldsymbol{\delta}_t/n\|_\psi + \|\boldsymbol{\delta}'_s\boldsymbol{U}_t/n\|_\psi + \|\boldsymbol{\delta}'_s\boldsymbol{\delta}_t/n\|_\psi = O(\|U_{is}\|_{\psi_{p/2}}\|\delta_{it}\|_{\psi_{p/2}}) + O(\|\delta_{it}\|^2_{\psi_{p/2}})$. Thus $g(A_{st}) = O(\psi^{-1}(T)(1/\sqrt{n} + C + C^2))$.

(c) Set $A_{st} = \widetilde{\eta}_{st} := \boldsymbol{F}'_s \sum_{i=1}^n \boldsymbol{\lambda}_i(U_{it} + \delta_{it})/n$, then apply Cauchy-Schwartz twice to obtain

$$g(A_{st}) = O(\|(\frac{1}{T}\sum_{s=1}^T \|\boldsymbol{F}_s\|^2)^{1/2}\|_{\psi_{p/2}}\| \sum_{i=1}^n \boldsymbol{\lambda}_i \frac{U_{it}+\delta_{it}}{n}\|_{\psi_{p/2}}) = O(1)O(\|\sum_{i=1}^n \boldsymbol{\lambda}_i\frac{U_{it}}{n}\|_{\psi_{p/2}} + \|\sum_{i=1}^n \boldsymbol{\lambda}_i\frac{\delta_{it}}{n}\|_{\psi_{p/2}}).$$

The first term in square brackets is $O(1/\sqrt{n})$ by Assumption 2(d) and 3(e); the second is $O(C)$. Hence $g(A_{st}) = O(\frac{1}{\sqrt{n}} + C)$.

(d) Set $A_{st} = \widetilde{\xi}_{st} := \boldsymbol{F}'_t \sum_{i=1}^n \boldsymbol{\lambda}_i(U_{is} + \delta_{is})/n$, then apply Cauchy-Schwartz twice followed by the maximal inequality to obtain

$$\begin{aligned} g(A_{st}) &= O(\|\|\boldsymbol{F}_t\|\|_{\psi_{p/2}}\|(\frac{1}{T}\sum_{s=1}^T \|\sum_{i=1}^n \boldsymbol{\lambda}_i \frac{U_{is}+\delta_{is}}{n}\|^2)^{1/2}\|_{\psi_{p/2}})) \\ &= O(1)O(\psi^{-1}(T)[\|\sum_{i=1}^n \boldsymbol{\lambda}_i\frac{U_{is}}{n}\|_{\psi_{p/2}} + \|(\sum_{i=1}^n \boldsymbol{\lambda}_i\frac{\delta_{is}}{n}\|_{\psi_{p/2}}]). \end{aligned}$$

The first term in square brackets is $O(1/\sqrt{n})$ by Assumption 2(d) and 3(e); the second is $O(C)$. Hence $g(A_{st}) = O(\psi^{-1}_{p/2}(T)[\frac{1}{\sqrt{n}} + C])$.

Finally, use the identity (B.1), the triangle inequality twice and the bounds $(a) - (d)$ to obtain the result. □

Lemma C.10. *If* $\max_{i,t}\|\delta_{it}\|_\psi = O(C)$ *and* $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = O_P(\eta)$ *then*

$$\left\|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}' - \boldsymbol{U}\boldsymbol{U}')\right\|_{\max} = O_P\left(\sqrt{T}\eta^2 + \frac{r_1}{\sqrt{T}} + \frac{r_2}{\sqrt{n}} + r_3 C\right)$$

where

$$\begin{aligned} r_1 &:= \psi_p^{-1}(n)\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(n^2) \\ r_2 &:= \psi_p^{-1}(n)\psi_{p/4}^{-1}(T) + \psi_{p/2}^{-1}(n) \\ r_3 &:= \psi_p^{-1}(n)\psi_{p/4}^{-1}(T) + \psi_p^{-1}(nT)\psi_{p/2}^{-1}(n). \end{aligned}$$

Proof. By the triangle inequality we have

$$\left\|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}' - \boldsymbol{U}\boldsymbol{U}')\right\|_{\max} \leqslant \left\|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{U}} - \boldsymbol{U})(\widehat{\boldsymbol{U}} - \boldsymbol{U})'\right\|_{\max} + 2\left\|\frac{1}{\sqrt{T}}\boldsymbol{U}(\widehat{\boldsymbol{U}} - \boldsymbol{U})'\right\|_{\max}.$$

For the first term we have

$$\left\|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{U}} - \boldsymbol{U})(\widehat{\boldsymbol{U}} - \boldsymbol{U})'\right\|_{\max} \leqslant \sqrt{T}\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|^2_{\max} = O_P(\sqrt{T}\eta^2).$$

For the second term we use decomposition (B.3) to write

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}(\widehat{U}_{jt}-U_{jt}) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}(\widehat{\boldsymbol{\lambda}}_j'\widehat{\boldsymbol{F}}_t-\boldsymbol{\lambda}_j'\boldsymbol{F}_t+\widehat{R}_{jt}-R_{jt})$$

$$= \Big[(\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}\boldsymbol{\lambda}_j)+\boldsymbol{H}\boldsymbol{\lambda}_j\Big]'\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}(\widehat{\boldsymbol{F}}_t-\boldsymbol{H}\boldsymbol{F}_t)$$

$$+\Big[(\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}\boldsymbol{\lambda}_j)+(\boldsymbol{H}'\boldsymbol{H}-I_r)\boldsymbol{\lambda}_j\Big]'\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}\boldsymbol{F}_t+(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j)'\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}\boldsymbol{W}_{jt}$$

Apply Cauchy-Schwartz inequality in each term followed by the triangle inequality we obtain

$$\left\|\frac{1}{\sqrt{T}}\boldsymbol{U}(\widehat{\boldsymbol{U}}-\boldsymbol{U})'\right\|_{\max} \leqslant \left[\max_{j\leqslant n}\|\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}\boldsymbol{\lambda}_j\|+\sqrt{r}\|\boldsymbol{H}\|\|\boldsymbol{\Lambda}\|_{\max}\right]\max_{i\leqslant n}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}(\widehat{\boldsymbol{F}}_t-\boldsymbol{H}\boldsymbol{F}_t)\right\|$$

$$+\left[\max_{j\leqslant n}\|\widehat{\boldsymbol{\lambda}}_j-\boldsymbol{H}\boldsymbol{\lambda}_j\|+\sqrt{r}\|\boldsymbol{H}'\boldsymbol{H}-I_r\|\|\boldsymbol{\Lambda}\|_{\max}\right]\max_{i\leqslant n}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}\boldsymbol{F}_t\right\|$$

$$+\max_{j\leqslant n}\|\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\|\max_{i,j\leqslant n}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}U_{it}\boldsymbol{W}_{jt}\right\|.$$

The first term is $O_P(1)O_P(\psi^{-1}(n)[\frac{1}{\sqrt{T}}+\frac{\psi_{p/2}^{-1}(T)}{\sqrt{n}}+\psi_{p/2}^{-1}(T)C])$ due to Lemma C.6(a), Lemma C.9 and the maximal inequality; the second term is $O_P(\frac{\psi_{p/2}^{-1}(n)}{\sqrt{T}}+\frac{1}{\sqrt{n}}+\psi^{-1}(nT)C)O_P(\psi_{p/2}^{-1}(n))$ since, by the maximal inequality, we might take $\omega=\psi^{-1}(nT)C$ in Theorem 2(b). The last term is $O_P(\psi^{-1}(n)\psi_{p/2}^{-1}(n)/\sqrt{T})O_P(\psi_{p/2}^{-1}(n^2))$. Thus, $\left\|\frac{1}{\sqrt{T}}\boldsymbol{U}(\widehat{\boldsymbol{U}}-\boldsymbol{U})'\right\|_{\max}=O_P(r_4)$ where

$$r_4 := \frac{\psi_p^{-1}(n)\psi_{p/2}^{-1}(n)\psi_{p/2}^{-1}(n^2)}{\sqrt{T}}+\frac{\psi_p^{-1}(n)\psi_{p/4}^{-1}(T)+\psi_{p/2}^{-1}(n)}{\sqrt{n}}+(\psi_p^{-1}(n)\psi_{p/4}^{-1}(T)+\psi_p^{-1}(nT)\psi_{p/2}^{-1}(n))C. \tag{C.8}$$

The result then follows.

□

Lemma C.11. *Let* $\|\widehat{\boldsymbol{U}}-\boldsymbol{U}\|=O_P(\eta)$ *then* $\max_{i,j,t}|\widehat{V}_{ij,t}-V_{ij,t}|=O_P(s_0[\eta+\xi\psi^{-1}(n)])$.

Proof. By the triangle inequality we have

$$|\widehat{V}_{ij,t}-V_{ij,t}|\leqslant|\widehat{U}_{i,t}-U_{i,t}|+|\widehat{\boldsymbol{\theta}}_i'\widehat{\boldsymbol{U}}_{-ij,t}-\boldsymbol{\theta}_i'\boldsymbol{U}_{-ij,t}|.$$

Using Hölder's inequality, the second term can be further bounded as

$$\begin{aligned}|\widehat{\boldsymbol{\theta}}_i'\widehat{\boldsymbol{U}}_{-ij,t} - \boldsymbol{\theta}_i'\boldsymbol{U}_{-ij,t}| &\leqslant |\widehat{\boldsymbol{\theta}}_i'(\widehat{\boldsymbol{U}}_{-ij,t} - \boldsymbol{U}_{-ij,t})| + |(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)'\boldsymbol{U}_{-ij,t}| \\ &\leqslant \|\widehat{\boldsymbol{\theta}}_i\|_1\|\widehat{\boldsymbol{U}}_{-ij,t} - \boldsymbol{U}_{-ij,t}\|_\infty + \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1\|\boldsymbol{U}_{-ij,t}\|_\infty \\ &\leqslant (\|\boldsymbol{\theta}_i\|_1 + \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1)\|\widehat{\boldsymbol{U}}_{-ij,t} - \boldsymbol{U}_{-ij,t}\|_\infty + \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1\|\boldsymbol{U}_{-ij,t}\|_\infty.\end{aligned}$$

Combining the last two expressions with the fact that $\|\boldsymbol{\theta}_i\|_1 \leqslant s_0\|\boldsymbol{\theta}_i\|_\infty \leqslant Cs_0$ and $\|\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_1 = O_P(\xi s_0) = O_P(1)$ by Assumption 3(f) and the the maximum inequality yields the result □

Lemma C.12. *Let* $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\| = O_P(\eta)$ *then*

$$\max_{i,j}|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\widehat{V}_{ij,t}\widehat{V}_{ji,t} - V_{ij,t}V_{ij,t)}| = O_P\{s_0^2[r_4 + \xi\psi^{-1}(n) + \sqrt{T}(\eta + \xi\psi^{-1}(n))^2]\}$$

.

Proof. By the triangle inequality

$$\max_{i,j}|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{V}}'_{ij}\widehat{\boldsymbol{V}}_{ji} - \boldsymbol{V}'_{ij}\boldsymbol{V}_{ji})| \leqslant \max_{i,j}|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{V}_{ij})(\widehat{\boldsymbol{V}}_{ji} - \boldsymbol{V}_{ji})| + 2\max_{i,j}|\frac{1}{\sqrt{T}}\boldsymbol{V}'_{ij}(\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{V}_{ij})|.$$

The first term can be bounded using Lemma C.11 since

$$\max_{i,j}|\frac{1}{\sqrt{T}}(\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{V}_{ij})(\widehat{\boldsymbol{V}}_{ji} - \boldsymbol{V}_{ji})| \leqslant \sqrt{T}[\max_{i,j,t}(\widehat{V}_{ijt} - V_{ijt})]^2 = O_P(\sqrt{T}[s_0(\eta + \xi\psi^{-1}(n))]^2).$$

The second term can be upper bounded by

$$\begin{aligned}\max_{i,j}|\frac{1}{\sqrt{T}}\boldsymbol{V}'_{ij}(\widehat{\boldsymbol{U}}_i - \boldsymbol{U}_i)| + \max_{i,j}\|\widehat{\boldsymbol{\theta}}_{ij}\|_1 \max_{i,j}\|\frac{1}{\sqrt{T}}\boldsymbol{V}'_{ij}(\widehat{\boldsymbol{U}}_{-ij} - \boldsymbol{U}_{-ij})\|_\infty \\ + \max_{i,j}\|\widehat{\boldsymbol{\theta}}_{ij} - \boldsymbol{\theta}_{ij}\|_1 \max_{i,j}|\frac{1}{\sqrt{T}}\boldsymbol{V}'_{ij}\boldsymbol{U}_{-ij}|.\end{aligned}$$

Recall the rate r_4 appearing in (C.8).Then the first term is $O_P(s_0 r_4)$, the second $O_P(s_0^2 r_4)$ and the last term is $O_P(\xi s_0^2\psi^{-1}(n))$. Thus $\max_{i,j}|\frac{1}{\sqrt{T}}\boldsymbol{V}'_{ij}(\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{V}_{ij})| = O_P[s_0^2(r_4 + \xi\psi^{-1}(n))]$. The result then follows.

□

Lemma C.13. $\|\widehat{\boldsymbol{\Upsilon}}_V - \boldsymbol{\Upsilon}_V\|_{\max} = O_P\left(h[s_0[\eta + \xi\psi_p^{-1}(n)](s_0\psi_p^{-1}(nT))^3 + s_0\frac{\psi_{p/4}^{-1}(n^4)}{\sqrt{T}}]\right)$

Proof. The proof is similar to the proof of Lemma C.8, refer to it for details. It suffices to bound in probability $\|\widehat{\boldsymbol{V}} - \boldsymbol{V}\|_{\max}$ and $\|\boldsymbol{V}\|_{\max}$, where $\boldsymbol{V}$ is $(n^2 \times T)$ matrix whose entries are $V_{ij,t}$ for $i, j \in [n]$ and $t \in [T]$. Similar for $\widehat{\boldsymbol{V}}$ with $V_{ij,t}$ replaced $\widehat{V}_{ij,t}$. Lemma C.11 bounds the former, for the later we have $\|\boldsymbol{V}\|_{\max} \leqslant \max_{i,j}\|\boldsymbol{\theta}_{ij}\|_1\|\boldsymbol{U}\|_{\max} = O(s_0\psi^{-1}(nT))$.

Let $\boldsymbol{i} := (i_1, i_2, i_3, i_4)$ be a multi-index where $i_1, i_2, i_3, i_4 \in [n]$. Define for $\boldsymbol{i}$ and $|\ell| < T$:

$$\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} := \frac{1}{T} \sum_{t=|\ell|+1}^{T} U_{i_1,t} U_{i_2,t} U_{i_3,t-|\ell|} U_{i_4,t-|\ell|}; \qquad \gamma_{\boldsymbol{i}}^{\ell} := \mathbb{E}\widetilde{\gamma}_{\boldsymbol{i}},$$

and $\widehat{\gamma}_{\boldsymbol{i}}^{\ell}$ as $\widetilde{\gamma}_{\boldsymbol{i}}^{\ell}$ with U's replaced by $\widehat{U}$'s. Also define

$$\widetilde{\upsilon}_{\boldsymbol{i}} := \sum_{|\ell|<T} k(\ell/h)\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} \qquad \upsilon_{\boldsymbol{i}} := \sum_{|\ell|<T} \gamma_{\boldsymbol{i}}^{\ell},$$

and $\widehat{\upsilon}_{\boldsymbol{i}}$ as $\widetilde{\upsilon}_{\boldsymbol{i}}$ with U's replaced by $\widehat{U}$'s. Then we write

$$\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}} = \sum_{|\ell|<T} k(\ell/h)(\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}) + \sum_{|\ell|<T} (k(\ell/h) - 1)\gamma_{\boldsymbol{i}}^{\ell}. \tag{C.9}$$

Since $\|\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}\|_{\psi_{p/4}} = O(\sqrt{T - |\ell|}/T) = O(1/\sqrt{T})$, the $\psi_{p/4}$-Orlicz norm of the first term is bounded by

$$h \sum_{|\ell|<T} |h^{-1}k(\ell/h)| \|\widetilde{\gamma}_{\boldsymbol{i}}^{\ell} - \gamma_{\boldsymbol{i}}^{\ell}\|_{\psi_{p/4}} = O\left(\frac{h}{\sqrt{T}} \int |k(u)|du\right) = O(h/\sqrt{T}),$$

whereas the second term is deterministic and is shown to be $O(h/\sqrt{T})$ by Andrews (1991). Thus $\|\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}\|_{\psi_{p/4}} = O(h/\sqrt{T})$ uniformly in $\boldsymbol{i} \in [n]^4$. Thus, by the maximal inequality followed by Markov's inequality we conclude that

$$\max_{\boldsymbol{i}} |\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}| = O_P(\psi_{p/4}^{-1}(n^4) \max_{\boldsymbol{i}} \|\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}\|_{\psi_{p/4}}) = O_P[\psi_{p/4}^{-1}(n^4)h/\sqrt{T}]. \tag{C.10}$$

We now use the fact that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^q$ we have $|\prod_{i=1}^q x_i - \prod_{i=1}^q y_i| = O(\sum_{i=0}^{q-1} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{n-i} \|\boldsymbol{y}\|_\infty^i)$ combined with the fact that $\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} = o(1)$ to obtain

$$\begin{aligned}
\max_{\boldsymbol{i},\ell} |\widehat{\gamma}_{\boldsymbol{i}}^{\ell} - \widetilde{\gamma}_{\boldsymbol{i}}^{\ell}| &\leqslant \max_{\boldsymbol{i},t,\ell} |\widehat{U}_{i_1,t}\widehat{U}_{i_2,t}\widehat{U}_{i_3,t-|\ell|}\widehat{U}_{i_4,t-|\ell|} - U_{i_1,t}U_{i_2,t}U_{i_3,t-|\ell|}U_{i_4,t-|\ell|}| \\
&= O(\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\|_{\max} \|\boldsymbol{U}\|_{\max}^3) \\
&= O_P[\eta[\psi^{-1}(nT)]^3]
\end{aligned}$$

Therefore we conclude

$$\max_{\boldsymbol{i}} |\widehat{\upsilon}_{\boldsymbol{i}} - \widetilde{\upsilon}_{\boldsymbol{i}}| \leqslant \max_{\boldsymbol{i},\ell} |\widehat{\gamma}_{\boldsymbol{i}}^{\ell} - \widetilde{\gamma}_{\boldsymbol{i}}^{\ell}| \sum_{|\ell|<T} |k(\ell/h)| = O_P\left(h\eta[\psi^{-1}(nT)]^3 \int |k(u)|du\right) = O_P(h\eta[\psi^{-1}(nT)]^3). \tag{C.11}$$

The result then follows from the triangle inequality $\|\widehat{\boldsymbol{\Upsilon}} - \boldsymbol{\Upsilon}\|_{\max} \leqslant \max_{\boldsymbol{i}} |\widehat{\upsilon}_{\boldsymbol{i}} - \widetilde{\upsilon}_{\boldsymbol{i}}| + \max_{\boldsymbol{i}} |\widetilde{\upsilon}_{\boldsymbol{i}} - \upsilon_{\boldsymbol{i}}|$, expression (C.10) and (C.11). □

Lemma C.14. *Let $\boldsymbol{U}, \boldsymbol{V}$ be $T \times n$ matrices such that $\|\boldsymbol{U}-\boldsymbol{V}\|_{\max} \leqslant C_1$ and $\|\boldsymbol{V}\|_{\max} \leqslant C_2$, then*

$$\|\boldsymbol{\Sigma}_{\boldsymbol{U}} - \boldsymbol{\Sigma}_{\boldsymbol{V}}\|_{\max} \leqslant C_3 := C_1(2C_2 + C_1),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{U}} := \boldsymbol{U}'\boldsymbol{U}/T$ and $\boldsymbol{\Sigma}_{\boldsymbol{V}} := \boldsymbol{V}'\boldsymbol{V}/T$. Furthermore, if $C_3 \leqslant \alpha\kappa(\boldsymbol{\Sigma}_{\boldsymbol{V}}, \mathcal{S}, 3)/(|\mathcal{S}|(1+\zeta)^2)$ for $\mathcal{S} \subseteq [n]$, $\zeta > 0$ and $\alpha \in [0,1]$, then

$$(1-\alpha)\kappa(\boldsymbol{\Sigma}_{\boldsymbol{V}}, \mathcal{S}, \zeta) \leqslant \kappa(\boldsymbol{\Sigma}_{\boldsymbol{U}}, \mathcal{S}, \zeta) \leqslant (1+\alpha)\kappa(\boldsymbol{\Sigma}_{\boldsymbol{V}}, \mathcal{S}, \zeta)$$

.

Proof. By the (reverse) triangle inequality we have $\|\boldsymbol{U}\|_{\max} - \|\boldsymbol{V}\|_{\max} \leqslant \|\boldsymbol{U} - \boldsymbol{V}\|_{\max}$, from which we conclude that $\|\boldsymbol{U}\|_{\max} \leqslant \|\boldsymbol{U} - \boldsymbol{V}\|_{\max} + \|\boldsymbol{V}\|_{\max} \leqslant C_1 + C_2$. Now $\|\boldsymbol{\Sigma}_{\boldsymbol{U}} - \boldsymbol{\Sigma}_{\boldsymbol{V}}\|_{\max} = \max_{1\leqslant i,j\leqslant n} |T^{-1}\sum_{t=1}^{T} U_{it}U_{jt} - V_{it}V_{ijt}| \leqslant \max_{i,j,t} |U_{it}U_{jt} - V_{it}V_{jt}|$ and

$$|U_{it}U_{jt} - V_{it}V_{jt}| \leqslant |(U_{it} - V_{it})U_{jt} + (U_{jt} - V_{jt})V_{it}| \leqslant$$
$$\|\boldsymbol{U} - \boldsymbol{V}\|_{\max}(\|\boldsymbol{U}\|_{\max} + \|\boldsymbol{V}\|_{\max}) \leqslant C_1(2C_2 + C_1).$$

For the second part of the lemma notice that for any $\boldsymbol{x} \in \mathbb{R}^n$ we have $|\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{U}}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}| = |\boldsymbol{x}'(\boldsymbol{\Sigma}_{\boldsymbol{U}} - \boldsymbol{\Sigma}_{\boldsymbol{V}})\boldsymbol{x}| \leqslant \|\boldsymbol{\Sigma}_{\boldsymbol{U}} - \boldsymbol{\Sigma}_{\boldsymbol{V}}\|_{\max}\|\boldsymbol{x}\|_1^2 \leqslant C_3\|\boldsymbol{x}\|_1^2$ by the first part. Also, if $\|\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leqslant \zeta\|\boldsymbol{x}_{\mathcal{S}}\|_1$ we have that $\|\boldsymbol{x}\|_1 = \|\boldsymbol{x}_{\mathcal{S}}\|_1 + |\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leqslant (1+\zeta)\|\boldsymbol{x}_{\mathcal{S}}\|_1 \leqslant (1+\zeta)\sqrt{\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}|\mathcal{S}|}/\kappa(\boldsymbol{\Sigma}_{\boldsymbol{V}}, \mathcal{S}, \zeta)$ where the last inequality follows from the definition of compatibility condition. Thus $|\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{U}}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}| \leqslant C_3(1+\zeta)^2\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}|\mathcal{S}|/\kappa(\boldsymbol{\Sigma}_{\boldsymbol{V}}, \mathcal{S}, \zeta) \leqslant \boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}/2$, where the last inequality follows from the definition of compatibility condition. Therefore, we have that $(1-\alpha)\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x} \leqslant \boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{U}}\boldsymbol{x} \leqslant (1+\alpha)\boldsymbol{x}'\boldsymbol{\Sigma}_{\boldsymbol{V}}\boldsymbol{x}$ whenever $\|\boldsymbol{x}_{\mathcal{S}^c}\|_1 \leqslant \zeta\|\boldsymbol{x}_{\mathcal{S}}\|_1$. Take the infimum to conclude. □

Lemma C.15. *Let $\boldsymbol{W} := (\boldsymbol{U}, \boldsymbol{V})$ and $\boldsymbol{Z} := (\boldsymbol{X}, \boldsymbol{Y})$ be $T \times (n+1)$ matrices such that $\|\boldsymbol{W} - \boldsymbol{Z}\|_{\max} \leqslant C_1$ and $\|\boldsymbol{Z}\|_{\max} \leqslant C_2$, then for any $\boldsymbol{\delta} \in \mathbb{R}^n$ we have*

$$\|\boldsymbol{U}'(\boldsymbol{V} - \boldsymbol{U}\boldsymbol{\delta})/T - \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\delta})/T\|_\infty \leqslant (1 + \|\boldsymbol{\delta}\|_1)C_1(2C_2 + C_1)$$

Proof. For convenience let $q := V - U\delta \in \mathbb{R}^T$ and $r := Y - X\delta \in \mathbb{R}^T$, then Hölder's inequality gives us $\|r\|_\infty \leqslant (1 + \|\delta\|_1)\|Z\|_{\max} \leqslant (1 + \|\delta\|_1)C_2$ and $\|q - r\|_\infty \leqslant (1 + \|\delta\|_1)\|W - Z\|_{\max} \leqslant (1 + \|\delta\|_1)C_1$. From the (reverse) triangle inequality we obtain $\|q\|_\infty \leqslant \|q - r\|_\infty + \|r\|_\infty \leqslant (1 + \|\delta\|_1)(C_1 + C_2)$. Now, following the same steps in the proof of previous Lemma, we can upper bound the right hand side of the display by $\|U - X\|_{\max}\|q\|_\infty + \|q - r\|_\infty\|X\|_{\max}$, which in turn can be upper bounded by the left hand size of the display. □

Lemma C.16. *Under the same conditions of Theorems 1 and 2*

$$\|\nabla L(\boldsymbol{\theta}_1) - \nabla L_0(\boldsymbol{\theta}_1)\|_\infty = O_P\left[\frac{\psi^{-1}(T)\psi^{-1}(nT)\psi^{-1}(n)\psi_{p/2}^{-1}(n)}{T^{1/4}} + \frac{\psi^{-1}(T)T^{1/4}}{\sqrt{n}}\right]$$

$$\|\nabla^2 L(\boldsymbol{\theta}) - \nabla^2 L_0(\boldsymbol{\theta})\|_{\mathsf{max}} = O_P\left[\eta(n,T)\left[\psi^{-1}(nT) + \eta(n,T)\right] + \frac{\psi_{p/2}^{-1}(n^2)}{\sqrt{T}}\right],$$

where $\nabla L_0(\boldsymbol{\theta}) := 2\mathbb{E}[\boldsymbol{U}_{-1t}(U_{1t} - \boldsymbol{\theta}'\boldsymbol{U}_{-1t})]$ *and* $\nabla^2 L_0(\boldsymbol{\theta}) := \mathbb{E}\boldsymbol{U}_{-1t}\boldsymbol{U}'_{-1t}$.

Proof. By the triangle inequality we have

$$\begin{aligned}\frac{1}{2}\|\nabla L(\boldsymbol{\theta}) - \nabla L_0(\boldsymbol{\theta})\|_\infty &= \|(\widehat{\boldsymbol{U}}_x - \boldsymbol{U}_x + \boldsymbol{U}_x)'\boldsymbol{V}/T - \mathbb{E}(\boldsymbol{U}'_x V/T)\|_\infty \\ &\leqslant \|\boldsymbol{U}'_x\boldsymbol{V}/T - \mathbb{E}(\boldsymbol{U}'_x V/T)\|_\infty + \|\widehat{\boldsymbol{U}}_x - \boldsymbol{U}_x\|_{\mathsf{max}}\|\boldsymbol{V}\|_\infty.\end{aligned}$$

Similarly, using Lemma 5.B

$$\begin{aligned}\|\nabla^2 L(\boldsymbol{\theta}) - \nabla^2 L_0(\boldsymbol{\theta})\|_{\mathsf{max}} &\leqslant \|\widehat{\boldsymbol{U}}'_x\widehat{\boldsymbol{U}}_x/T - \boldsymbol{U}'_x\boldsymbol{U}_x/T\|_{\mathsf{max}} + \|\boldsymbol{U}'_x\boldsymbol{U}_x/T - \mathbb{E}(\boldsymbol{U}'_x\boldsymbol{U}_x/T)\|_{\mathsf{max}} \\ &\leqslant \|\widehat{\boldsymbol{U}}_x - \boldsymbol{U}_x\|_{\mathsf{max}}(2\|\boldsymbol{U}_x\|_{\mathsf{max}} + \|\widehat{\boldsymbol{U}}_x - \boldsymbol{U}_x\|_{\mathsf{max}}) \\ &\quad + \|\boldsymbol{U}'_x\boldsymbol{U}_x/T - \mathbb{E}(\boldsymbol{U}'_x\boldsymbol{U}_x/T)\|_{\mathsf{max}}.\end{aligned}$$

By Corollary 1 and Assumption 3 we can bound in probability each of those terms

$$\begin{aligned}\|\boldsymbol{U}'_x\boldsymbol{V}/T - \mathbb{E}(\boldsymbol{U}'_x\boldsymbol{V}/T)\|_\infty &= O_P\left[\frac{\psi_{p/2}^{-1}(n)}{\sqrt{T}}\right] \\ \|\widehat{\boldsymbol{U}}_x - \boldsymbol{U}_x\|_{\mathsf{max}} &= O_P\left[\frac{\psi^{-1}(nT)\psi^{-1}(n)\psi_{p/2}^{-1}(n)}{T^{1/4}} + \frac{T^{1/4}}{\sqrt{n}}\right] =: O_P[\eta(n,T)] \\ \|\boldsymbol{V}\|_\infty &= \psi^{-1}(T) \\ \|\boldsymbol{U}_x\|_{\mathsf{max}} &= O_P[\psi^{-1}(nT)] \\ \|\boldsymbol{U}'_x\boldsymbol{U}_x/T - \mathbb{E}(\boldsymbol{U}'_x\boldsymbol{U}_x/T)\|_{\mathsf{max}} &= O_P\left[\frac{\psi_{p/2}^{-1}(n^2)}{\sqrt{T}}\right].\end{aligned}$$

Therefore

$$\|\nabla L(\boldsymbol{\theta}) - \nabla L_0(\boldsymbol{\theta})\|_\infty = O_P\left[\frac{\psi_{p/2}^{-1}(n)}{\sqrt{T}} + \frac{\psi^{-1}(T)\psi^{-1}(nT)\psi^{-1}(n)\psi_{p/2}^{-1}(n)}{T^{1/4}} + \frac{\psi^{-1}(T)T^{1/4}}{\sqrt{n}}\right]$$

and

$$\|\nabla^2 L(\boldsymbol{\theta}) - \nabla^2 L_0(\boldsymbol{\theta})\|_{\mathsf{max}} = O_P\left[\eta(n,T)\left[\psi^{-1}(nT) + \eta(n,T)\right] + \frac{\psi_{p/2}^{-1}(n^2)}{\sqrt{T}}\right]$$

□

Lemma C.17. *For $p > 0$, let $\psi_p : \mathbb{R}_+ \to \mathbb{R}_+$ defined by $\psi_p(x) := x\mathbb{1}_{\{0 \leqslant x < t\}} + (\exp(x^p) - 1)\mathbb{1}\{x \geqslant t\}$ where $t := \left(\frac{1-p}{p}\right)^{1/p}$. If $\|X\|_{\psi_p} < \infty$ then there exist constants $C_1 > 0$ and $C_2 > 0$ such that*

$$\mathbb{P}(|X| > x) \leqslant C_1 \exp(-x^p/C_2) \qquad x > 0.$$

In particular, if $0 < \|X\|_{\psi_p} < \infty$ we might take $C_1 = 2 + \mathbb{1}0 < p < 1 \exp((1-p)/p)$ and $C_2 = \|X\|_{\psi_p}^p$. Conversely, if there exist constants $C_1 > 0$ and $C_2 > 0$ such that $\mathbb{P}(|X| > x) \leqslant C_1 \exp(-x^p/C_2)$ for $x > 0$, then

$$\|X\|_{\psi_p} \leqslant \begin{cases} [(2C_1 + 1)C_2]^{1/p} \vee 2C_1 C_2^{1/p} p^{-1}\Gamma(1/p) & ; 0 < p < 1 \\ [(C_1 + 1)C_2]^{1/p} & ; p \geqslant 1, \end{cases}$$

where $\Gamma(\cdot)$ denotes the Gamma function.

Proof. If $\|X\|_{\psi_p} = 0$ then $X = 0$ a.s and the inequality holds for any choice of $C_1, C_2 > 0$. For the case when $0 < \|X\|_{\psi_p} < \infty$ we have by Markov inequality and the fact that $x \mapsto \exp a|x|^p$ is non-decreasing for $a > 0$

$$\mathbb{P}(|X| \geqslant x) = \mathbb{P}(\exp(a|X|^p) \geqslant \exp(ax^p)) \leqslant \exp(-ax^p)\mathbb{E}\exp(a|X|^p).$$

Also

$$\begin{aligned} \mathbb{E}\exp(a|X|^p) &= \mathbb{E}\exp(a|X|^p)\mathbb{1}a^{1/p}|X| < t + \mathbb{E}\exp(a|X|^p)\mathbb{1}a^{1/p}|X| \geqslant t \\ &\leqslant \exp((1-p)/p)\mathbb{1}0 < p < 1 + \mathbb{E}\psi_p(a^{1/p}|X|) + 1. \end{aligned}$$

Set $a = \|X\|_{\psi_p}^{-p}$ to conclude that the middle term is less or equal to 1.

For the converse we have for $a > 0$, by Fubini's Theorem

$$\begin{aligned} \mathbb{E}\exp(|aX|^p) - 1 &= \int \int_0^{|x|^p} a^p \exp(a^p y) dy \mathbb{P}(dx) \\ &= a^p \int_0^\infty \mathbb{P}(|X| \geqslant x^{1/p}) \exp(a^p x) dx. \end{aligned}$$

Since $\mathbb{P}(|X| > x) \leqslant C_1 \exp(-x^p/C_2)$, for $a < C_2^{-1/p}$, we have

$$\mathbb{E}\exp[(|aX|)^p] - 1 \leqslant a^p C_1 \int_0^\infty \exp\left[-x(\tfrac{1}{C_2} - a^p)\right] dx \leqslant \tfrac{a^p C_1}{C_2^{-1} - a^p}.$$

Also

$$\mathbb{E}|X| = \int_0^\infty \mathbb{P}(|X| > x) dx \leqslant C_1 \int_0^\infty \exp(-x^p/C_2) dx = \tfrac{C_1}{pC_2^{-1/p}}\Gamma(1/p).$$

Therefore using the last two displays we have, for $0 < a < C_2^{-1/p}$

$$\begin{aligned}\mathbb{E}\psi_p(a|X|) &\leqslant \mathbb{E}|aX|\mathbb{1}0 < p < 1 + \mathbb{E}\,\mathsf{exp}[(|aX|)^p] - 1\\ &\leqslant \tfrac{aC_1}{pC_2^{-1/p}}\Gamma(1/p)\mathbb{1}0 < p < 1 + \tfrac{a^pC_1}{C_2^{-1}-a^p}.\end{aligned}$$

When $p \geqslant 1$ the right hand side is less or equal than 1 for $a \leqslant [(1+C_1)C_2]^{-1/p}$ hence $\|X\|_{\psi_p} \leqslant [(1+C_1)C_2]^{1/p}$. For $0 < p < 1$, the right hand side is less or equal than 1 for $a \leqslant \{C_2^{-1/p}[(2C_1+1)^{-1/p} \wedge 2C_1p^{-1}\Gamma(1/p)]\}^{-1}$ then $\|X\|_{\psi_p} \leqslant (2C_1+1)C_2]^{1/p} \vee 2C_1C_2^{1/p}p^{-1}\Gamma(1/p)$.

□

Lemma C.18. *For $p > 0$, there is a constant C_p only depending on p such that*

$$\|XY\|_{\psi_{p/2}} \leqslant C_p\left(\|X\|_{\psi_p} \vee \|Y\|_{\psi_p}\right),$$

where ψ_p defined as per Lemma (C.17)*.*

Proof. If $\|X\|_{\psi_p} = 0$ or $\|Y\|_{\psi_p} = 0$ then $XY = 0$ a.s and the inequality hold trivially. Similarly if $\|X\|_{\psi_p} = \infty$ or $\|Y\|_{\psi_p} = \infty$. So we assume that $0 < \|X\|_{\psi_p} < \infty$ and $0 < \|Y\|_{\psi_p} < \infty$ and from Lemma (C.17) we have for $x > 0$

$$\begin{aligned}\mathbb{P}(|X| > x) &\leqslant K_p\,\mathsf{exp}[-(x/\|X\|_{\psi_p})^p]\\ \mathbb{P}(|Y| > x) &\leqslant K_p\,\mathsf{exp}[-(x/\|X\|_{\psi_p})^p],\end{aligned}$$

where $K_p := 2 + \mathbb{1}0 < p < 1\,\mathsf{exp}((1-p)/p)$. Then

$$\begin{aligned}\mathbb{P}(|XY| \geqslant x) &\leqslant \mathbb{P}(|X| \geqslant \sqrt{x}) + \mathbb{P}(|Y| \geqslant \sqrt{x}) \leqslant K_p\,\mathsf{exp}(-z^{p/2}/\|X\|_{\psi_p}^p) + K_p\,\mathsf{exp}(-z^{p/2}/\|Y\|_{\psi_p}^p)\\ &\leqslant 2K_p\,\mathsf{exp}(-z^{p/2}/D_p^p)\end{aligned}$$

where $D_p := \|X\|_{\psi_p} \vee \|Y\|_{\psi_p}$. Apply once again Lemma (C.17) to conclude that $\|XY\|_{\psi_{p/2}} \leqslant C_pD_p$ where $C_p = (2K_p+1)^{1/p}$ for $p \geqslant 1$ otherwise $(4K_p+1)^{1/p} \vee 8K_p\Gamma(2/p)p^{-1}$. □

(J. Fan) Department of Operations Research and Financial Engineering, Princeton University

Email address: jqfan@princeton.edu

(R.P. Masini) Center for Statistics and Machine Learning, Princeton University

Email address: rmasini@princeton.edu

(M.C. Medeiros) Department of Economics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil.

Email address: mcm@econ.puc-rio.br