HHS Public Access

Author manuscript

Econometrica. Author manuscript; available in PMC 2016 January 15.

Published in final edited form as:

Econometrica. 2015 July 1; 83(4): 1497–1541. doi:10.3982/ECTA12749.

Power Enhancement in High Dimensional Cross-Sectional Tests

Jianqing Fan[*,†], **Yuan Liao**[‡], and **Jiawei Yao**[*]

Yuan Liao: yuanliao@umd.edu; Jiawei Yao: jiaweiy@princeton.edu

[*]Department of Operations Research and Financial Engineering, Princeton University [†]Bendheim Center for Finance, Princeton University [‡]Department of Mathematics, University of Maryland

Abstract

We propose a novel technique to boost the power of testing a high-dimensional vector $H : \theta = 0$ against sparse alternatives where the null hypothesis is violated only by a couple of components. Existing tests based on quadratic forms such as the Wald statistic often suffer from low powers due to the accumulation of errors in estimating high-dimensional parameters. More powerful tests for sparse alternatives such as thresholding and extreme-value tests, on the other hand, require either stringent conditions or bootstrap to derive the null distribution and often suffer from size distortions due to the slow convergence. Based on a screening technique, we introduce a "power enhancement component", which is zero under the null hypothesis with high probability, but diverges quickly under sparse alternatives. The proposed test statistic combines the power enhancement component with an asymptotically pivotal statistic, and strengthens the power under sparse alternatives. The null distribution does not require stringent regularity conditions, and is completely determined by that of the pivotal statistic. As specific applications, the proposed methods are applied to testing the factor pricing models and validating the cross-sectional independence in panel data models.

Keywords

sparse alternatives; thresholding; large covariance matrix estimation; Wald-test; screening; cross-sectional independence; factor pricing model

1 Introduction

High-dimensional cross-sectional models have received growing attentions in both theoretical and applied econometrics. These models typically involve a structural parameter, whose dimension can be either comparable or much larger than the sample size. This paper addresses testing a high-dimensional structural parameter:

$$H_0: \theta = 0,$$

[†]Address: Department of Operations Research and Financial Engineering, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA. Department of Mathematics, University of Maryland, College Park, MD 20742, USA. jqfan@princeton.edu.

JEL code: C12, C33, C58

where $N = \dim(\boldsymbol{\theta})$ is allowed to grow faster than the sample size T. We are particularly interested in boosting the power in *sparse* alternatives under which $\boldsymbol{\theta}$ is approximately a sparse vector. This type of alternative is of particular interest, as the null hypothesis typically represents some economic theory and violations are expected to be only by some exceptional individuals.

A showcase example is the factor pricing model in financial economics. Let y_{it} be the excess return of the i-th asset at time t, and $\mathbf{f}_t = (f_{1t},\ldots, f_{Kt})'$ be the K-dimensional observable factors. Then, the excess return has the following decomposition:

$$y_{it}=\theta_i+\mathbf{b}_i'\mathbf{f}_t+u_{it},\ \ i=1,\ldots,N,\ \ t=1,\ldots,T,$$

where $\mathbf{b}_i = (b_{i1},\ldots, b_{iK})'$ is a vector of factor loadings and u_{it} represents the idiosyncratic error. The key implication from the multi-factor pricing theory is that the intercept θ_i should be zero, known as the "mean-variance efficiency" pricing, for any asset i. An important question is then if such a pricing theory can be validated by empirical data, namely we wish to test the null hypothesis $H_0 : \boldsymbol{\theta} = 0$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ is the vector of intercepts for all N financial assets. As the factor pricing model is derived from theories of financial economics (Ross, 1976), one would expect that inefficient pricing by the market should only occur to a small fractions of exceptional assets. Indeed, our empirical study of the constituents in the S&P 500 index indicates that there are only a couple of significant nonzero-alpha stocks, corresponding to a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market. Therefore, it is important to construct tests that have high power when $\boldsymbol{\theta}$ is sparse.

Most of the conventional tests for $H_0 : \boldsymbol{\theta} = 0$ are based on a quadratic form:

$$W=\hat{\boldsymbol{\theta}}'\mathbf{V}\hat{\boldsymbol{\theta}}.$$

Here $\hat{\boldsymbol{\theta}}$ is an element-wise consistent estimator of $\boldsymbol{\theta}$, and $\mathbf{V}$ is a high-dimensional positive definite weight matrix, often taken to be the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ (e.g., the Wald test). After a proper standardization, the standardized W is asymptotically pivotal under the null hypothesis. In high-dimensional testing problems, however, various difficulties arise when using a quadratic statistic. First, when $N > T$, estimating $\mathbf{V}$ is challenging, as the sample analogue of the covariance matrix is singular. More fundamentally, tests based on W have low powers under sparse alternatives. The reason is that the quadratic statistic accumulates high-dimensional estimation errors under H_0, which results in large critical values that can dominate the signals in the sparse alternatives. A formal proof of this statement will be given in Section 3.3.

To overcome the aforementioned difficulties, this paper introduces a novel technique for high-dimensional cross-sectional testing problems, called the "power enhancement". Let J_1 be a test statistic that has a correct asymptotic size (e.g., Wald statistic), which may suffer from low powers under sparse alternatives. Let us augment the test by adding a *power enhancement component* $J_0 \geq 0$, which satisfies the following three properties:

Power Enhancement Properties

a. Non-negativity: $J_0 \geq 0$ almost surely.

b. No-size-distortion: Under H_0, $P(J_0 = 0|H_0) \to 1$.

c. Power-enhancement: J_0 diverges in probability under some specific regions of alternatives H_a.

Our constructed power enhancement test takes the form

$$J = J_0 + J_1.$$

The non-negativity property of J_0 ensures that J is at least as powerful as J_1. Property (b) guarantees that the asymptotic null distribution of J is determined by that of J_1, and the size distortion due to adding J_0 is negligible, and property (c) guarantees significant power improvement under the designated alternatives. The *power enhancement principle* is thus summarized as follows: Given a standard test statistic with a correct asymptotic size, its power is substantially enhanced with little size distortion; this is achieved by adding a component J_0 that is asymptotically zero under the null, but diverges and dominates J_1 under some specific regions of alternatives.

An example of such a J_0 is a *screening statistic*:

$$J_0 = \sqrt{N} \sum_{j \in \hat{S}} \hat{\theta}_j^2 \hat{v}_j^{-1} = \sqrt{N} \sum_{j=1}^{N} \hat{\theta}_j^2 \hat{v}_j^{-1} 1\{|\widehat{\theta_j}| > \hat{v}_j^{1/2} \delta_{N,T}\},$$

where $\hat{S} = \{j \leq N : |\widehat{\theta_j}| > \hat{v}_j^{1/2} \delta_{N,T}\}$, and $\widehat{v_j}$ denotes a data-dependent normalizing factor, taken as the estimated asymptotic variance of $\widehat{\theta_j}$. The critical value $\delta_{N,T}$, depending on (N, T), is a high-criticism threshold, chosen to be slightly larger than the noise level $\max_{j \leq_N} |\hat{\theta}_j - \theta_j| / \hat{v}_j^{1/2}$ so that under H_0, $J_0 = 0$ with probability approaching one. In addition, we take J_1 as a pivotal statistic, e.g., standardized Wald statistic or other quadratic forms such as the sum of the squared marginal *t*-statistics (Bai and Saranadasa, 1996; Chen and Qin, 2010; Pesaran and Yamagata, 2012). The screening set $\hat{S}$ also captures indices where the null hypothesis is violated.

One of the major differences of our test from most of the thresholding tests (Fan, 1996; Hansen, 2005) is that, it enhances the power substantially by adding a screening statistic, which does not introduce extra difficulty in deriving the asymptotic null distribution. Since $J_0 = 0$ under H_0, it relies on the pivotal statistic J_1 to determine its null distribution. In contrast, the existing tests such as thresholding, extreme value, and higher criticism tests (e.g., Hall and Jin (2010)) often require stringent conditions to derive their asymptotic null distributions, making them restrictive in econometric applications, due to slow rates of convergence. Moreover, the asymptotic null distributions are inaccurate at finite sample. As pointed out by Hansen (2003), these statistics are non-pivotal even asymptotically, and require bootstrap methods to simulate the null distributions.

As a specific application, in addition to testing the aforementioned factor pricing model, this paper also studies the tests for cross-sectional independence in mixed effect panel data models:

$$y_{it}=\alpha+\mathrm{x}_{it}^{'}\boldsymbol{\beta}+\mu_i+u_{it},\ \ i\leq n, t\leq T.$$

Let ρ_{ij} denote the correlation between u_{it} and u_{jt}, assumed to be time invariant. The "cross-sectional independence" test is concerned about the following null hypothesis:

$$H_0{:}\rho_{ij}=0, \text{for all } i\neq j,$$

that is, under the null hypothesis, the $n \times n$ covariance matrix Σ_u of $\{u_{it}\}_{i\leq n}$ is diagonal. In empirical applications, weak cross-sectional correlations are often present, which results in a sparse covariance Σ_u with just a few nonzero off-diagonal elements. Namely, the vector $\boldsymbol{\theta}=(\rho_{12}, \rho_{13},\ldots, \rho_{n-1,n})$ is sparse and should be incorporated to improve power of the test. The dimensionality $N = n(n-1)/2$ can be much larger than the number of observations. Therefore, the power enhancement in sparse alternatives is very important to the testing problem. By choosing $\delta_{N,T}$ to dominate $\max_{i,j\leq n}\{|\hat{\rho}_{ij}|/\hat{v}_{ij}^{1/2}{:}\rho_{ij}=0\}$ as detailed in Section 5, under the sparse alternative, the set $\hat{S}$ "screens out" most of the estimation noises, and contains only a few indices of the nonzero off-diagonal entries. Therefore, $\hat{S}$ not only reveals the sparse structure of Σ_u, but also determines the nonzero off-diagonal entries with an over-whelming probability.

There has been a large literature on high-dimensional cross-sectional tests. For instance, the literature on testing the factor pricing model is found in Gibbons et al. (1989), MacKinlay and Richardson (1991), Beaulieu et al. (2007) and Pesaran and Yamagata (2012), all in quadratic forms. Gagliardini et al. (2011) studied estimation of the risk premia in a CAPM and its associated testing problem, which is closely related to our work. While we also study a large panel of stock returns as a specific example and double asymptotics (as $N, T \to \infty$), the problems and approaches being considered are very different. This paper addresses a general problem of enhancing powers under high-dimensional sparse alternatives.

For the mixed effect panel data model, most of the testing statistics are based on the sum of squared residual correlations, which also accumulates many off-diagonal estimation errors in estimating the covariance matrix of $(u_{1t}, \ldots, u_{nt})$. See, for example, Breusch and Pagan (1980), Pesaran et al. (2008), and Baltagi et al. (2012). Our problem is also related to the test with a restricted parameter space, previously considered by Andrews (1998), who improves the power by directing towards the "relevant" alternatives; see also Hansen (2003) for a related idea. Chernozhukov et al. (2013) proposed a high-dimensional inequality test, and employed an extreme value statistic, whose critical value is determined through applying the moderate deviation theory on an upper bound of the rejection probability. In contrast, the asymptotic distribution of our proposed power enhancement statistic is determined through the pivotal statistic J_1, and the power is improved via the contributions of sparse alternatives that survive the screening process.

The remainder of the paper is organized as follows. Section 2 sets up the preliminaries and highlights the major differences from existing tests. Section 3 presents the main result of power enhancement test. As applications to specific cases, Section 4 and Section 5 respectively study the factor pricing model and test of cross-sectional independence. Section 6 presents simulation results are empirical evidence of sparse alternatives based on the real data. Section 7 provides further discussions. Proofs are given in the supplementary material.

Throughout the paper, for a symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent its minimum and maximum eigenvalues. Let $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_1$ denote its operator norm and l_1-norm respectively, defined by $\|\mathbf{A}\|_2=\lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$ and $\max_i \Sigma_j |\mathbf{A}_{ij}|$. For a vector $\boldsymbol{\theta}$, define $\|\boldsymbol{\theta}\|=\left(\sum_j \theta_j^2\right)^{1/2}$ and $\|\boldsymbol{\theta}\|_{\max} = \max_j |\theta_j|$. For two deterministic sequences a_T and b_T, we write $a_T \ll b_T$ (or equivalently $b_T \gg a_T$) if $a_T = o(b_T)$. Also, $a_T \overset{\cup}{\cap} b_T$ if there are constants $C_1, C_2 > 0$ so that $C_1 b_T \le a_T \le C_2 b_T$ for all large T. Finally, we denote $|S|_0$ as the number of elements in a set S.

2 Power Enhancement in high dimensions

This section introduces power enhancement techniques and provides heuristics to justify the techniques. The differences from related methods in the literature are also highlighted.

2.1 Power enhancement

Consider a testing problem:

$$H_0:\boldsymbol{\theta}=0, \quad H_a:\boldsymbol{\theta} \in \Theta_a,$$

where $\Theta_a \subset \mathbb{R}^N \backslash \{\mathbf{0}\}$ is an alternative set in $\mathbb{R}^N$. A typical example is $\Theta_a = \{\boldsymbol{\theta}: \boldsymbol{\theta} \neq 0\}$.

Suppose we observe a stationary process $\mathbf{D}=\{\mathbf{D}_t\}_{t=1}^T$ of size T. Let $J_1(\mathbf{D})$ be a certain test statistic, which will also be written as J_1. Often J_1 is constructed such that under H_0, it has a non-degenerate limiting distribution F : As $T, N \to \infty$,

$$J_1|H_0 \to^d F. \quad (2.1)$$

For the significance level $q \in (0, 1)$, let F_q be the critical value for J_1. Then the critical region is taken as $\{\mathbf{D}: J_1 > F_q\}$ and satisfies

$$\lim_{T,N\to\infty} P(J_1 > F_q|H_0)=q.$$

This ensures that J_1 has a correct asymptotic size. In addition, it is often the case that J_1 has high power against H_0 on a subset $\Theta(J_1) \subset \Theta_a$, namely,

$$\lim_{T,N\to\infty} \inf_{\boldsymbol{\theta}\in\Theta(J_1)} P(J_1 > F_q|\boldsymbol{\theta}) \to 1.$$

Typically, $\Theta(J_1)$ consists of those $\boldsymbol{\theta}$'s, whose l_2-norm is relatively large, as J_1 is normally an omnibus test (e.g. Wald test).

In a data-rich environment, econometric models often involve high-dimensional parameters in which $\dim(\boldsymbol{\theta}) = N$ can grow fast with the sample size T. We are particularly interested in *sparse alternatives* $\Theta_s \subset \Theta_a$ under which H_0 is violated only on a couple of exceptional components of $\boldsymbol{\theta}$. Specifically, $\Theta_s \in \mathbb{R}^N$ is a subset of Θ_a, and when $\boldsymbol{\theta} \in \Theta_s$, the number of non-vanishing components is much less than N. As a result, its l_2-norm is relatively small. Therefore, under sparse alternative Θ_s, the omnibus test J_1 typically has a lower power, due to the accumulation of high-dimensional estimation errors. Detailed explanations are given in Section 3.3 below.

We introduce a *power enhancement principle* for high-dimensional sparse testing, by bringing in a data-dependent component J_0 that satisfies the **Power Enhancement Properties** defined in Section 1. The component J_0 does not serve as a test statistic on its own, but is added to a classical statistic J_1 that is often pivotal (e.g., Wald-statistic), so the proposed test statistic is defined by

$$J = J_0 + J_1.$$

Our introduced "power enhancement principle" is explained as follows.

1. Under mild conditions, $P(J_0 = 0|H_0) \to 0$ by construction. Hence when (2.1) is satisfied, we have

$$\lim_{T,N\to\infty} sup P(J > F_q|H_0) = q.$$

   Therefore, adding J_0 to J_1 does not affect the size of the standard test statistic asymptotically. Both J and J_1 have the same limiting distribution under H_0.

2. The critical region of J is defined by

$$\{\mathbf{D}: J > F_q\}.$$

   As $J_0 \geq 0$, $P(J > F_q|\boldsymbol{\theta}) \geq P(J_1 > F_q|\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta_a$. Hence the power of J is at least as large as that of J_1.

3. When $\boldsymbol{\theta} \in \Theta_s$ is a sparse high-dimensional vector under the alternative, the "classical" test J_1 may have low power as $\|\boldsymbol{\theta}\|$ is typically relatively small. On the other hand, for $\boldsymbol{\theta} \in \Theta_s$, J_0 stochastically dominates J_1. As a result, $P(J > F_q|\boldsymbol{\theta}) > P(J_1 > F_q|\boldsymbol{\theta})$ strictly holds, so the power of J_1 over the set Θ_s is enhanced after adding J_0. Often J_0 diverges fast under sparse alternatives Θ_s, which ensures $P(J > F_q|\boldsymbol{\theta}) \to 1$ for $\boldsymbol{\theta} \in \Theta_s$. In contrast, the classical test only has $P(J_1 > F_q|\boldsymbol{\theta}) < c < 1$ for some $c \in (0, 1)$ and $\boldsymbol{\theta} \in \Theta_s$, and when $\|\boldsymbol{\theta}\|$ is sufficiently small, $P(J_1 > F_q|\boldsymbol{\theta})$ is approximately q.

It is important to note that the power is enhanced without sacrificing the size asymptotically. In fact the power enhancement principle can be asymptotically fulfilled under a weaker condition $J_0|H_0 \to^p 0$. However, we construct J_0 so that $P(J_0 = 0|H_0) \to 1$ to ensure a good finite sample size.

2.2 Construction of power enhancement component

We construct a specific power enhancement component J_0 that satisfies (a)–(c) of the power enhancement properties, and identify the sparse alternatives in Θ_s. Such a component can be constructed via *screening* as follows. Suppose we have a consistent estimator $\hat{\boldsymbol{\theta}}$ such that $\max_{j\le N}|\hat{\theta}_j - \theta_j| = o_P(1)$. For some slowly growing sequence $\delta_{N,T} \to \infty$ (as $T, N \to \infty$), define a screening set:

$$\hat{S} = \{j : |\hat{\theta}_j| > \hat{v}_j^{1/2}\delta_{N,T}, j = 1, \ldots, N\}, \quad (2.2)$$

where $\hat{v}_j > 0$ is a data-dependent normalizing constant, often taken as the estimated asymptotic variance of $\hat{\theta}_j$. The sequence $\delta_{N,T}$, called "high criticism", is chosen to dominate the maximum-noise-level, satisfying: (recall that Θ_a denotes the alternative set)

$$\inf_{\boldsymbol{\theta}\in\Theta_a\cup\{0\}} P(\max_{j\le N}|\hat{\theta}_j - \theta_j|/\hat{v}_j^{1/2} < \delta_{N,T}|\boldsymbol{\theta}) \to 1 \quad (2.3)$$

for $\boldsymbol{\theta}$ under both null and alternate hypotheses. The screening statistic J_0 is then defined as

$$J_0 = \sqrt{N}\sum_{j\in\hat{S}}\hat{\theta}_j^2\hat{v}_j^{-1} = \sqrt{N}\sum_{j=1}^{N}\hat{\theta}_j^2\hat{v}_j^{-1}1\{|\hat{\theta}_j| > \hat{v}_j^{1/2}\delta_{N,T}\}.$$

By (2.2) and (2.3), under $H_0 : \boldsymbol{\theta} = 0$,

$$P(J_0 = 0|H_0) = P(\hat{S} = \varnothing|H_0) = P(\max_{j\le N}|\hat{\theta}_j|/\hat{v}_j^{1/2} \le \delta_{N,T}|H_0) \to 1.$$

Therefore J_0 satisfies the non-negativeness and no-size-distortion properties.

Let $\{v_j\}_{j\le N}$ be the population counterpart of $\{\hat{v}_j\}_{j\le N}$. To satisfy the power-enhancement property, define

$$S(\boldsymbol{\theta}) = \left\{j : |\theta_j| > 3v_j^{1/2}\delta_{N,T}, j = 1, \ldots, N\right\}, \quad (2.4)$$

and in particular $S(\mathbf{0}) = \varnothing$. We shall show in Theorem 3.1 below that $P(S(\boldsymbol{\theta}) \subset \hat{S}|\boldsymbol{\theta}) \to 1$, for all $\boldsymbol{\theta} \in \Theta_a \cup \{0\}$. Thus all the significant signals are contained in $\hat{S}$ with a high probability. If $S(\boldsymbol{\theta}) \neq \varnothing$, then by the definition of $\hat{S}$ and $\delta_{N,T} \to \infty$, we have

$$P(J_0 > \sqrt{N} | S(\boldsymbol{\theta}) \neq \varnothing) \geq P(\sqrt{N} \sum_{j \in \hat{S}} \delta_{N,T}^2 > \sqrt{N} | S(\boldsymbol{\theta}) \neq \varnothing) \to 1.$$

Thus, the power of J_1 is *enhanced* on the subset

$$\Theta_s \equiv \{\boldsymbol{\theta} \in \mathrm{R}^N : S(\boldsymbol{\theta}) \neq \varnothing\} = \{\boldsymbol{\theta} \in \mathrm{R}^N : \max_{j \leq N} \frac{|\theta_j|}{v_j^{1/2}} > 3\delta_{N,T}\}.$$

Furthermore, $\hat{S}$ not only reveals the sparse structure of $\boldsymbol{\theta}$ under the alternative, but also determines the nonzero entries with an over-whelming probability.

The introduced J_0 can be combined with any other test statistic with an accurate asymptotic size. Suppose J_1 is a "classical" test statistic. Our power enhancement test is simply

$$J = J_0 + J_1.$$

For instance, suppose we can consistently estimate the asymptotic inverse covariance matrix of $\hat{\boldsymbol{\theta}}$, denoted by $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}})^{-1}$, then J_1 can be chosen as the standardized Wald-statistic:

$$J_1 = \frac{\hat{\boldsymbol{\theta}}' \widehat{\mathrm{var}}(\widehat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\theta}} - N}{\sqrt{2N}}.$$

As a result, the asymptotic distribution of J is $\mathcal{N}(0, 1)$ under the null hypothesis.

In sparse alternatives where $\|\boldsymbol{\theta}\|$ may not grow fast enough with N but $\boldsymbol{\theta} \in \Theta_s$, the combined test $J_0 + J_1$ can be very powerful. In contrast, we will formally show in Theorem 3.4 below that the conventional Wald test J_1 can have very low power on its own. On the other hand, when the alternative is "dense" in the sense that $\|\boldsymbol{\theta}\|$ grows fast with N, the conventional test J_1 itself is consistent. In this case, J is still as powerful as J_1. Therefore, if we denote $\Theta(J_1) \subset \mathbb{R}^N / \{\mathbf{0}\}$ as the set of alternative $\boldsymbol{\theta}$'s against which the classical J_1 test has power converging to one, then the combined $J = J_0 + J_1$ test has power converging to one against $\boldsymbol{\theta}$ on

$$\Theta_s \cup \Theta(J_1).$$

We shall show in Section 3 that the power is enhanced uniformly over $\boldsymbol{\theta} \in \Theta_s \cup \Theta(J_1)$. In addition, the set $\hat{S}$ indicates which components may violate the null hypothesis.

2.3 Comparisons with thresholding and extreme-value tests

One of the fundamental differences between our power enhancement component J_0 and existing tests with good power under sparse alternatives is that, existing test statistics have a non-degenerate distribution under the null, and often require either bootstrap or strong conditions to derive the null distribution. Such convergences are typically slow and the

serious size distortion appears in finite sample. In contrast, our screening statistic J_0 uses "high criticism" sequence $\delta_{N,T}$ to make $P(J_0 = 0|H_0) \to 1$, hence does not serve as a test statistic on its own. The asymptotic null distribution is determined by that of J_1, which usually not hard to derive As we shall see in sections below, the required regularity condition is relatively mild, which makes the power enhancement test widely applicable to many econometric problems.

In the high-dimensional testing literature, there are mainly two types of statistics with good power under sparse alternatives: extreme value test and thresholding test respectively. The test based on extreme values studies the maximum deviation from the null hypothesis across the components of $\hat{\boldsymbol{\theta}}=(\widehat{\theta_1},\ldots,\widehat{\theta_N})$, and forms the statistic based on $\max_{j\le N}|\frac{\hat{\theta}_j}{w_j}|^\delta$ for some $\delta > 0$ and a weight w_j (e.g., Cai et al. (2013), Chernozhukov et al. (2013)). Such a test statistic typically converges slowly to its asymptotic counterpart. An alternative test is based on thresholding: for some $\delta > 0$ and pre-determined threshold level t_N,

$$R=\sqrt{T}\sum_{j=1}^{N}|\frac{\hat{\theta}_j}{w_j}|^\delta 1\{|\hat{\theta}_j|>t_N w_j\} \quad (2.5)$$

For example, when t_T is taken slightly less than $\max_{j\le N}|\hat{\theta}_j|/w_j$, R becomes the extreme statistic. When t_T is small (e.g. 0), R becomes a traditional test, which is not powerful in detecting sparse alternatives, though it can have good size properties. The accumulation of estimation errors is prevented due to the threshold $1\{|\hat{\theta}_j|>t_N w_j\}$ for sufficiently large t_N (see, e.g., Fan (1996) and Zhong et al. (2013)). In a low-dimensional setting, Hansen (2005) suggested using a threshold to enhance the power in a similar way.

Although (2.5) looks similar to J_0, the ideas behind are very different. Both the extreme value test and the thresholding test require regularity conditions that may be restrictive in econometric applications. For instance, it can be difficult to employ the central limit theorem directly on (2.5), as it requires the covariance between $\widehat{\theta_j}$ and $\hat{\theta}_{j+k}$ decay fast enough as $k \to \infty$ (Zhong et al., 2013). In cross-sectional testing problems, this essentially requires an explicit ordering among the cross-sectional units which is, however, often unavailable in panel data applications. In addition, as (2.5) involves effectively limited terms of summations due to thresholding, the asymptotic theory does not provide adequate approximations, resulting in size-distortion in applications. We demonstrate the size-distortion in the simulation study.

3 Asymptotic properties

3.1 Main results

This section presents the regularity conditions and formally establishes the claimed power enhancement properties. Below we use $P(\cdot|\boldsymbol{\theta})$ to denote the probability measure defined from the sampling distribution with parameter $\boldsymbol{\theta}$. Let $\Theta \subset \mathbb{R}^N$ be the parameter space of $\boldsymbol{\theta}$ that covers the union of $\{\mathbf{0}\}$ and the alternative set Θ_a. When we write $\inf_{\theta\in\Theta} P(\cdot|\boldsymbol{\theta})$, the infimum is taken in the space that covers the union of both null and alternative space.

We begin with a high-level assumption. In specific applications, they can be verified with primitive conditions.

Assumption 3.1—As $T, N \to \infty$, the sequence $\delta_{N,T} \to \infty$, and the estimators $\{\widehat{\theta}_j, \widehat{v}_j\}_{j \le N}$ are such that

i. $\inf_{\boldsymbol{\theta} \in \Theta} P(\max_{j \le N} |\widehat{\theta}_j - \theta_j| / \hat{v}_j^{1/2} < \delta_{N,T} | \boldsymbol{\theta}) \to 1$;

ii. $\inf_{\boldsymbol{\theta} \in \Theta} P(4/9 < \widehat{v}_j / v_j < 9/4, \forall j = 1, \ldots, N | \boldsymbol{\theta}) \to 1$.

The normalizing constant v_j is often taken as the asymptotic variance of $\widehat{\boldsymbol{\theta}}_j$, with $\widehat{v}_j$ being its consistent estimator. The constants 4/9 and 9/4 in condition (ii) are not optimally chosen, as this condition only requires $\{\widehat{v}_j\}_{j \le N}$ be *not-too-bad* estimators of their population counterparts.

In many high-dimensional problems with strictly stationary data that satisfy strong mixing conditions, following from the large-deviation theory, typically, $\max_{j \le N} |\widehat{\theta}_j - \theta_j| / \hat{v}_j^{1/2} = O_P(\sqrt{\log N})$. Therefore, we shall fix

$$\delta_{N,T} = \log(\log T) \sqrt{\log N}, \quad (3.1)$$

which is a high criticism that slightly dominates the standardized noise level (it may be useful to recall that the maximum of N i.i.d. Gaussian noises with a bounded variance behaves as $\sqrt{\log N}$ asymptotically). We shall provide primitive conditions for this choice of $\delta_{N,T}$ in the subsequent sections, so that Assumption 3.1 holds.

Recall that $\hat{S}$ and $S(\boldsymbol{\theta})$ are defined by (2.2) and (2.4) respectively. In particular, $S(\boldsymbol{\theta}) = \left\{ j : |\theta_j| > 3 v_j^{1/2} \delta_{N,T}, j = 1, \ldots, N \right\}$, so under $H_0 : \boldsymbol{\theta} = 0$, $S(\boldsymbol{\theta}) = \emptyset$. The following theorem characterizes the asymptotic behavior of $J_0 = \sqrt{N} \sum_{j \in \hat{S}} \widehat{\theta}_j^2 \hat{v}_j^{-1}$ under both the null and alternative hypotheses.

Theorem 3.1—Let Assumption 3.1 hold. As $T, N \to \infty$, we have under $H_0 : \theta = 0$, $P(\hat{S} = \emptyset | H_0) \to 1$. Hence

$$P(J_0 = 0 | H_0) \to 1 \quad \text{and} \quad \inf_{\{\boldsymbol{\theta} \in \Theta : S(\boldsymbol{\theta}) \neq \emptyset\}} P(J_0 > \sqrt{N} | \boldsymbol{\theta}) \to 1.$$

In addition,

$$\inf_{\boldsymbol{\theta} \in \Theta} P(S(\boldsymbol{\theta}) \subset \hat{S} | \boldsymbol{\theta}) \to 1.$$

Besides the asymptotic behavior of J_0, Theorem 3.1 also establishes a "sure screening" property of $\hat{S}$, which means it selects all the significant components whose indices are in $S(\boldsymbol{\theta})$. This result is achieved uniformly in $\boldsymbol{\theta}$ under both the null and alternative hypotheses.

Remark 3.1—Under additional mild assumptions, it can be further shown that $P(\hat{S}=S(\boldsymbol{\theta})|\boldsymbol{\theta}) \to 1$ uniformly in $\boldsymbol{\theta}$. Hence the selection is consistent. While the selection consistency is not a requirement of the power enhancement principle, we refer to our earlier manuscript Fan et al. (2014) for technical details.

We are now ready to formally show the power enhancement argument. The enhancement is achieved uniformly on the following set:

$$\Theta_s=\{\boldsymbol{\theta} \in \Theta:\max_{j\le N}\frac{|\theta_j|}{v_j^{1/2}}>3\delta_{N,T}\}. \quad (3.2)$$

In particular, if $\widehat{\theta}_j$ is $\sqrt{T}$-consistent, and $v_j^{1/2}$ is the asymptotic standard deviation of $\widehat{\theta}_j$, then $\sigma_j=\sqrt{Tv_j}$ is bounded away from both zero and infinity. Using (3.1), we have

$$\Theta_s=\left\{\boldsymbol{\theta} \in \Theta:\max_{j\le N}|\theta_j|/\sigma_j>3log(logT)\sqrt{\frac{logN}{T}}\right\}.$$

This is a relatively weak condition on the strength of the maximal signal in order to be detected by J_0.

A test is said to have a high power uniformly on a set $\Theta^* \subset \mathbb{R}^N \setminus \{\mathbf{0}\}$ if

$$\inf_{\boldsymbol{\theta}\in\Theta*} P(\text{reject } H_0 \text{ by the test}|\boldsymbol{\theta}) \to 1.$$

For a given distribution function F, let F_q denote its qth quantile.

Theorem 3.2—Let Assumptions 3.1 hold. Suppose there is a test J_1 such that

- **i.** it has an asymptotic non-degenerate null distribution F, and the critical region takes the form $\{\mathbf{D} : J_1 > \mathrm{F_q}\}$ for the significance level $q \in (0, 1)$,
- **ii.** it has a high power uniformly on some set $\Theta(J_1) \subset \Theta$,
- **iii.** there is $c > 0$ so that $\inf_{\boldsymbol{\theta}\in\Theta_s} P(c\sqrt{N}+J_1>F_q|\boldsymbol{\theta}) \to 1$, as $T, N \to \infty$.

Then the power enhancement test $J = J_0 + J_1$ has the asymptotic null distribution F, and has a high power uniformly on the set $\Theta_s \cup \Theta(J_1)$: as $T, N \to \infty$

$$\inf_{\boldsymbol{\theta}\in\Theta_s\cup\Theta(J_1)} P(J>F_q|\boldsymbol{\theta}) \to 1.$$

The three required conditions for J_1 are easy to understand: Conditions (i) and (ii) respectively require the size and power conditions for J_1. Condition (iii) requires J_1 be dominated by J_0 under Θ_s. This condition is not restrictive since J_1 is typically standardized (e.g., Donald et al. (2003)).

Theorem 3.2 also shows that J_1 and J have the critical regions $\{\mathbf{D} : J_1 > F_q\}$ and $\{\mathbf{D} : J > F_q\}$ respectively, but the high power region is enhanced from $\Theta(J_1)$ to $\Theta_s \cup \Theta(J_1)$. In high-dimensional testing problems, $\Theta_s \cup \Theta(J_1)$ can be much larger than $\Theta(J_1)$. As a result, the power of J_1 can be substantially enhanced after J_0 is added.

3.2 Power enhancement for quadratic tests

As an example of J_1, we consider the widely used quadratic test statistic, which is asymptotically pivotal:

$$J_Q = \frac{T\hat{\boldsymbol{\theta}}' \mathbf{V}\hat{\boldsymbol{\theta}} - N(1+\mu_{N,T})}{\xi_{N,T}\sqrt{N}},$$

where $\mu_{N,T}$ and $\xi_{N,T}$ are deterministic sequences that satisfy $\mu_{N,T} \to 0$ and $\xi_{N,T} \to \xi \in (0, \infty)$. The weight matrix $\mathbf{V}$ is positive definite, whose eigenvalues are bounded away from both zero and infinity. Here $T\mathbf{V}$ is often taken to be the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. Other popular choices are $\mathrm{V} = \mathrm{diag}(\sigma_1^{-2}, \cdots, \sigma_N^{-2})$ with $\sigma_j = \sqrt{Tv_j}$ (Bai and Saranadasa, 1996; Chen and Qin, 2010; Pesaran and Yamagata, 2012) and $\mathbf{V} = \mathbf{I}_N$, the $N \times N$ identity matrix. We set $J_1 = J_Q$, whose power enhancement version is $J = J_0 + J_Q$. For the moment, we shall assume $\mathbf{V}$ to be known, and just focus on the power enhancement properties. We will deal with unknown $\mathbf{V}$ for testing the factor pricing problem in the next section.

Assumption 3.2

- **i.** There is a non-degenerate distribution F so that under H_0, $J_Q \to^d F$.
- **ii.** The critical value $F_q = O(1)$ and the critical region of J_Q is $\{\mathbf{D} : J_Q > F_q\}$,
- **iii.** $\mathbf{V}$ is positive definite, and there exist two positive constants C_1 and C_2 such that $C_1 \le \lambda_{\min}(\mathbf{V}) \le \lambda_{\max}(\mathbf{V}) \le C_2$.
- **iv.** $C_3 \le Tv_j \le C_4$, $j = 1, \ldots, N$ for positive constants C_3 and C_4.

Analyzing the power properties of J_Q and applying Theorem 3.2, we obtain the following theorem. Recall that $\delta_{N,T}$ and Θ_s are defined by (3.1) and (3.2).

Theorem 3.3—Under Assumptions 3.1–3.2, the power enhancement test $J = J_0 + J_Q$ satisfies: as $T, N \to \infty$,

- **i.** under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$, $J \to^d F$,
- **ii.** there is $C > 0$ so that J has high power uniformly on the set

$$\Theta_s \cup \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta}\|^2 > C\delta_{N,T}^2 N/T\} \equiv \Theta_s \cup \Theta(J_Q);$$

that is, $\inf_{\boldsymbol{\theta} \in \Theta_s \cup \Theta(J_Q)} P(J > F_q | \boldsymbol{\theta}) \to 1$ for any $q \in (0,1)$.

3.3 Low power of quadratic statistics under sparse alternatives

When J_Q is used on its own, it can suffer from a low power under sparse alternatives if N grows much faster than the sample size, even though it has been commonly used in the econometric literature. Mainly, $T\hat{\boldsymbol{\theta}}'\mathbf{V}\hat{\boldsymbol{\theta}}$ aggregates high-dimensional estimation errors under H_0, which become large with a non-negligible probability and potentially override the sparse signals under the alternative. The following result gives this intuition a more precise description.

To simplify our discussion, we shall focus on the Wald-test with $T\mathbf{V}$ being the inverse of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, assumed to exist. Specifically, we assume the standardized $T\hat{\theta}'\mathbf{V}\hat{\theta}$ to be asymptotically normal under H_0:

$$\frac{T\hat{\theta}'\mathbf{V}\hat{\boldsymbol{\theta}} - N}{\sqrt{2N}} | H_0 \to^d \mathcal{N}(0,1). \quad (3.3)$$

This is one of the most commonly seen cases in various testing problems. The diagonal entries of $\frac{1}{T}\mathrm{V}^{-1}$ are given by $\{v_j\}_{j \le N}$.

Theorem 3.4—Suppose that (3.3) holds with $\|\mathbf{V}\|_1 < C$ and $\|\mathbf{V}^{-1}\|_1 < C$ for some $C > 0$. Under Assumptions 3.1–3.2, $T = o(\sqrt{N})$ and $\log N = o(T^{1-\gamma})$ for some $0 < \gamma < 1$, the quadratic test J_Q has low power at the sparse alternative Θ_c for any $c > 0$, given by

$$\Theta_c = \{\boldsymbol{\theta} \in \Theta : \sum_{j=1}^{N} 1\{\theta_j \neq 0\} = o(\sqrt{N}/T), \|\boldsymbol{\theta}\|_{\max} < c\}.$$

In other words, $\forall c > 0$, $\forall \boldsymbol{\theta} \in \Theta_c$, for any significance level q,

$$\lim_{T,N \to \infty} P(J_Q > z_q | \boldsymbol{\theta}) = q,$$

where z_q is the *qth* upper quantile of standard normal distribution.

In the above theorem, the alternative is a sparse vector. However, using the quadratic test itself, the asymptotic power of the test is as low as q. This is because the signals in the sparse alternative are dominated by the aggregated high-dimensional estimation errors: $T\sum_{j:\theta_j=0} \hat{\theta}_j^2$. In contrast, the non-vanishing components of $\boldsymbol{\theta}$ (fixed constants) are actually detectable by using J_0. The power enhancement test $J_0 + J_Q$ takes this into account, and has a substantially improved power.

4 Application: Testing Factor Pricing Models

4.1 The model

The multi-factor pricing model, motivated by the Arbitrage Pricing Theory (APT) by Ross (1976), is one of the most fundamental results in finance. It postulates how financial returns are related to market risks, and has many important practical applications. Let y_{it} be the excess return of the i-th asset at time t and $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$ be the observable factors. Then, the excess return has the following decomposition:

$$y_{it}=\theta_i+\boldsymbol{b}_i'\mathbf{f}_t+u_{it},\ \ i=1,\ldots,N,\ \ t=1,\ldots,T,$$

where $\mathbf{b}_i = (b_{i1}, \ldots, b_{iK})'$ is a vector of factor loadings and u_{it} represents the idiosyncratic error. To make the notation consistent, we stick to use $\boldsymbol{\theta}$ to represent the commonly used "alpha" in the finance literature.

While the APT does not necessarily deliver an observable factor model, the specification of an observable factor structure is of considerable interest and is often the case in empirical analyses. The key implication from the multi-factor pricing theory for tradable factors is that the intercept θ_i should be zero for any asset i. Such an *exact* feature of factor pricing can also be motivated from Connor (1984), who presented a competitive equilibrium version of the APT. An important question is then testing the null hypothesis

$$H_0:\boldsymbol{\theta}=0, \quad (4.1)$$

namely, whether the factor pricing model is consistent with empirical data, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ is the vector of intercepts for all N financial assets. One typically picks five-year monthly data, because the factor pricing model is technically a one-period model whose factor loadings can be time-varying; see Gagliardini et al. (2011) on how to model the time-varying effects using firm characteristics and market variables. As the theory of the factor pricing model applies to all tradable assets, rather than a handful selected portfolios, the number of assets N should be much larger than T. This ameliorates the selection biases in the construction of testing portfolios. On the other hand, our empirical study on the S&P500 index provides empirical evidence of sparse alternatives: there are only a few significantly nonzero components of $\boldsymbol{\theta}$, corresponding to a small portion of mis-priced stocks, instead of systematic mispricing of the whole market.

Most existing tests to the problem (4.1) are based on the quadratic statistic $W=T\hat{\boldsymbol{\theta}}'\mathbf{V}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the OLS estimator for $\boldsymbol{\theta}$, and $\mathbf{V}$ is some positive definite matrix. Prominent examples are given by Gibbons et al. (1989), MacKinlay and Richardson (1991) and Beaulieu et al. (2007). When N is possibly much larger than T, Pesaran and Yamagata (2012) showed that, under regularity conditions (Assumption 4.1 below),

$$J_1=\frac{a_{f,T}T\hat{\boldsymbol{\theta}}'\sum_u^{-1}\hat{\boldsymbol{\theta}}-N}{\sqrt{2N}}\to^d\mathcal{N}(0,1).$$

where $a_{f,T} > 0$, given in the next subsection, is a constant that depends only on factors' empirical moments, and $\mathbf{\Sigma}_u$ is the $N \times N$ covariance matrix of $\mathbf{u}_t = (u_{1t}, \ldots, u_{Nt})'$, assumed to be time-invariant.

Recently, Gagliardini et al. (2011) proposed a novel approach to modeling and estimating time-varying risk premiums using two-pass least-squares method under asset pricing restrictions. Their problems and approaches differ substantially from ours, though both papers study similar problems in finance. As a part of their model validation, they develop test statistics against the asset pricing restrictions and weak risk factors. Their test statistics are based on a weighted sum of squared residuals of the cross-sectional regression, which, like all classical test statistics, have power only when there are many violations of the asset pricing restrictions. They do not consider the issue of enhancing the power under sparse alternatives, nor do they involve a Wald statistic that depends on a high-dimensional covariance matrix. In fact, their testing power can be enhanced by using our techniques.

4.2 Power enhancement component

We propose a new statistic that depends on (i) the power enhancement component J_0, and (ii) a feasible Wald component based on a consistent covariance estimator for $\sum_u^{-1}$, which controls the size under the null even when $N/T \to \infty$.

Define $\overline{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$ and $\mathbf{w} = \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t'\right)^{-1} \overline{\mathbf{f}}$. Also define

$$a_{f,T} = 1 - \overline{\mathbf{f}}' \mathbf{w}, \text{ and } a_f = 1 - E\mathbf{f}_t' \left(E\mathbf{f}_t \mathbf{f}_t'\right)^{-1} E\mathbf{f}_t.$$

The OLS estimator of $\boldsymbol{\theta}$ can be expressed as

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_N)', \quad \hat{\theta}_j = \frac{1}{T a_{f,T}} \sum_{t=1}^{T} y_{jt}(1 - \mathbf{f}_t' \mathbf{w}). \quad (4.2)$$

When cov($\mathbf{f}_t$) is positive definite, under mild regularity conditions (Assumption 4.1 below), $a_{f,T}$ consistently estimates a_f, and $a_f > 0$. In addition, without serial correlations, the conditional variance of $\hat{\theta}_j$ (given $\{\mathbf{f}_t\}$) converges in probability to

$$v_j = \text{var}(u_{jt}) / (T a_f),$$

which can be estimated by $\hat{v}_j$ based on the residuals of OLS estimator:

$$\hat{v}_j = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_{jt}^2 / (T a_{f,T}), \quad \text{where } \hat{u}_{jt} = y_{jt} - \hat{\theta}_j - \hat{\mathbf{b}}_j' \mathbf{f}_t.$$

We show in Proposition 4.1 below that $\max_{j\le N}|\hat{\theta}_j-\theta_j|/\hat{v}_j^{1/2}=O_P\left(\sqrt{logN}\right)$. Therefore, $\delta_{N,T}=log(logT)\sqrt{logN}$ slightly dominates the maximum estimation noise. The screening set and the power enhancement component are defined as

$$\hat{S}=\{j:|\widehat{\theta_j}|>\hat{v}_j^{1/2}\delta_{N,T},j=1,\ldots,N\},$$

and

$$J_0=\sqrt{N}\sum_{j\in\hat{S}}\hat{\theta}_j^2\hat{v}_j^{-1}.$$

4.3 Feasible Wald test in high dimensions

Assuming no serial correlations among $\{\mathbf{u}_t\}_{t=1}^T$ and conditional homoskedasticity (Assumption 4.1 below), given the observed factors, the conditional covariance of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\Sigma}_u/(Ta_{f,T})$. If the covariance matrix $\boldsymbol{\Sigma}_u$ of $\mathbf{u}_t$ were known, the standardized Wald test statistic is

$$\frac{Ta_{f,T}\hat{\boldsymbol{\theta}}'\sum_u^{-1}\hat{\boldsymbol{\theta}}-N}{\sqrt{2N}}. \quad (4.3)$$

Under $H_0:\boldsymbol{\theta}=0$, it converges in distribution to $\mathcal{N}(0,1)$.

Note that factor models are often only justified as being “approximate” (Chamberlain and Rothschild (1983)), where $\boldsymbol{\Sigma}_u$ is a non-diagonal covariance matrix of cross-sectionally correlated idiosyncratic errors (u_{1t}, …, u_{Nt}). When $N/T\to\infty$, it is difficult to consistently estimate $\sum_u^{-1}$, as there are $O(N^2)$ off-diagonal parameters. Without parametrizing the off-diagonal elements, we assume $\Sigma_u=\text{cov}(\mathbf{u}_t)$ a sparse matrix. This assumption is natural for large covariance estimations for factor models, and was previously considered by Fan et al. (2011). Since the common factors dictate preliminarily the co-movement across the whole panel, a particular asset’s idiosyncratic shock is usually correlated significantly only with a few of other assets. For example, some shocks only exert influences on a particular industry, but are not pervasive for the whole economy (Connor and Korajczyk, 1993). Recently, Gagliardini et al. (2011) also obtained a feasible test by using a similar thresholding technique as to be introduced below to estimate the asymptotic covariance matrix. They showed that the sparsity approach for estimating covariance matrices covers the block diagonal case, which is expected to be present in the factor modeling of stocks (as elaborated in Gagliardini et al. (2011), a typical example of the sparsity of Σ_u is due to the presence of remaining industry sector effects).

Following the approach of Bickel and Levina (2008), we can consistently estimate $\sum_u^{-1}$ via thresholding: let $s_{ij}=\frac{1}{T}\sum_{t=1}^T\hat{u}_{it}\hat{u}_{jt}$. Define the covariance estimator as

$$(\hat{\sum}_u)_{ij} = \begin{cases} s_{ii}, & \text{if } i=j, \\ h_{ij}(s_{ij}), & \text{if } i \neq j, \end{cases}$$

where $h_{ij}(\cdot)$ is a generalized thresholding function (Antoniadis and Fan, 2001; Rothman et al., 2009), with threshold value $\tau_{ij} = C\left(s_{ii}s_{jj}\frac{logN}{T}\right)^{1/2}$ for some constant $C > 0$, designed to keep only the sample correlation whose magnitude exceeds $C\left(\frac{logN}{T}\right)^{1/2}$. The hard-thresholding function, for example, is $h_{ij}(x) = x1\{|x| > \tau_{ij}\}$, and many other thresholding functions such as soft-thresholding and SCAD (Fan and Li, 2001) are specific examples. In general, $h_{ij}(\cdot)$ should satisfy:

- **i.** $h_{ij}(z) = 0$ if $|z| < \tau_{ij}$;
- **ii.** $|h_{ij}(z) - z| \leq \tau_{ij}$;
- **iii.** there are constants $a > 0$ and $b > 1$ such that $|h_{ij}(z) - z| \leq a\tau_{ij}^2$ if $|z| > b\tau_{ij}$.

The thresholded covariance matrix estimator sets most of the off-diagonal estimation noises in $\left\{\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}\right\}_{i,j\leq N}$ to zero. As studied in Fan et al. (2013), the constant C in the threshold can be chosen in a data-driven manner so that $\hat{\sum}_u$ is strictly positive definite in finite sample even when $N > T$.

With $\hat{\sum}_u^{-1}$, we are ready to define the *feasible standardized Wald statistic*:

$$J_{wald} = \frac{Ta_{f,T}\hat{\boldsymbol{\theta}}'\hat{\sum}_u^{-1}\hat{\boldsymbol{\theta}} - N}{\sqrt{2N}},$$

whose power can be enhanced under sparse alternatives by:

$$J = J_0 + J_{wald}.$$

Remark 4.1—The thresholding approach described here can be modified to take advantages of the block structure of $\boldsymbol{\Sigma}_u$. The covariance matrix can first be divided into blocks according to the industries, and then estimated block-by-block. The estimation procedure and theoretical analysis will be similar to the block-thresholding of Cai and Yuan (2012).

4.4 Does the thresholded covariance estimator affect the size?

A natural but technical question to address is that when $\boldsymbol{\Sigma}_u$ indeed admits a sparse structure, is the thresholded estimator $\hat{\sum}_u^{-1}$ accurate enough so that the feasible J_{wald} is still

asymptotically normal? The answer is affirmative if $N(\log N)^4 = o(T^2)$, and still we can allow $N/T \to \infty$. However, such a simple question is far more technically involved than anticipated, as we now explain.

When $\mathbf{\Sigma}_u$ is a sparse matrix, under regularity conditions (Assumption 4.2 below), Fan et al. (2011) showed that

$$\|\sum\nolimits_u^{-1} - \hat{\sum}\nolimits_u^{-1}\|_2 = O_P\left(\sqrt{\frac{logN}{T}}\right). \quad (4.4)$$

This convergence rate is minimax optimal for the sparse covariance estimation, by the lower bound derived by Cai et al. (2010). On the other hand, when replacing $\sum_u^{-1}$ in (4.3) by $\hat{\sum}_u^{-1}$, one needs to show that the effect of such a replacement is asymptotically negligible, namely, under H_0,

$$T\hat{\boldsymbol{\theta}}'(\sum\nolimits_u^{-1} - \hat{\sum}\nolimits_u^{-1})\hat{\boldsymbol{\theta}}/\sqrt{N} = o_P(1). \quad (4.5)$$

However, when $\boldsymbol{\theta} = 0$, it can be shown that $\|\hat{\boldsymbol{\theta}}\|^2 = O_P(N/T)$. Using this and (4.4), by the Cauchy-Schwartz inequality, we have

$$|T\hat{\boldsymbol{\theta}}'(\sum\nolimits_u^{-1} - \hat{\sum}\nolimits_u^{-1})\hat{\boldsymbol{\theta}}|/\sqrt{N} = O_P\left(\sqrt{\frac{NlogN}{T}}\right).$$

Thus, it requires $N \log N = o(T)$ to converge, which is basically a low-dimensional scenario.

The above simple derivation uses, however, a Cauchy-Schwartz bound, which is too crude for a large N. In fact, $\hat{\boldsymbol{\theta}}'(\sum_u^{-1} - \hat{\sum}_u^{-1})\hat{\boldsymbol{\theta}}$ is a weighted estimation error of $\sum_u^{-1} - \hat{\sum}_u^{-1}$, where the weights $\hat{\boldsymbol{\theta}}$ "average down" the accumulated estimation errors in estimating elements of $\sum_u^{-1}$, and result in an improved rate of convergence. The formalization of this argument requires further regularity conditions and novel technical arguments. These are formally presented in the following subsection.

4.5 Regularity conditions

We are now ready to present the regularity conditions. These conditions are imposed for three technical purposes: (i) Achieving the uniform convergence for $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ as required in Assumption 3.1, (ii) defining the sparsity of $\mathbf{\Sigma}_u$ so that $\hat{\sum}_u^{-1}$ is consistent, and (iii) showing (4.5), so that the errors from estimating $\sum_u^{-1}$ do not affect the size of the test.

Let $\mathrm{F}_{-\infty}^{0}$ and F_{T}^{∞} denote the σ-algebras generated by $\{\mathbf{f}_t: -\infty \le t \le 0\}$ and $\{\mathbf{f}_t: T \le t \le \infty\}$ respectively. In addition, define the α-mixing coefficient

$$\alpha(T)=\sup_{A\in} \mathrm{F}_{-\infty}^{0}, B \in \mathrm{F}_{T}^{\infty}|P(A)P(B)-P(AB)|.$$

Assumption 4.1

i. $\{\mathbf{u}_t\}_{t\le T}$ is i.i.d. $\mathcal{N}(0, \sum_u)$, where both $\|\mathbf{\Sigma}_u\|_1$ and $\|\sum_u^{-1}\|_1$ are bounded;

ii. $\{\mathbf{f}_t\}_{t\le T}$ is strictly stationary, independent of $\{\mathbf{u}_t\}_{t\le T}$, and there are r_1, $b_1 > 0$ so that

$$\max_{i\le K} P(|f_{it}|>s) \le \exp(-(s/b_1)^{r_1}).$$

iii. There exists $r_2 > 0$ such that $r_1^{-1}+r_2^{-1}>0.5$ and $C > 0$, for all $T \in \mathbb{Z}^+$,

$$\alpha(T) \le \exp(-CT^{r_2}).$$

iv. $\mathrm{cov}(\mathbf{f}_t)$ is positive definite, and $\max_{i\le N}\|\mathbf{b}_i\| < c_1$ for some $c_1 > 0$.

Some remarks are in order for the conditions in Assumption 4.1.

Remark 4.2—The above assumption, perhaps somewhat restrictive, substantially facilitates our technical analysis. Here $\mathbf{u}_t$ is required to be serially uncorrelated across t. Under this condition, the conditional covariance of $\hat{\boldsymbol{\theta}}$, given the factors, has a simple expression $\mathbf{\Sigma}_u/(Ta_{f,T})$. On the other hand, if serial correlations are present in $\mathbf{u}_t$, there would be additional autocovariance terms in the covariance matrix, which need to be further estimated via regularizations. Moreover, given that $\mathbf{\Sigma}_u$ is a sparse matrix, the Gaussianity ensures that most of the idiosyncratic errors are cross-sectionally independent so that $\mathrm{cov}(u_{it}^2, u_{jt}^l)=0$, l = 1, 2, for most of the pairs in $\{(i, j)$: $i \ne j\}$.

Note that we do allow the factors $\{\mathbf{f}_t\}_{t\le T}$ to be weakly correlated across t, but satisfy the strong mixing condition Assumption 4.1 (iii).

Remark 4.3—The conditional homoskedasticity $E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t)=E(\mathbf{u}_t\mathbf{u}_t')$ is assumed, granted by condition (ii). We admit that handling conditional heteroskedasticity, while important in empirical applications, is very technically challenging in our context. Allowing the high-dimensional covariance matrix $E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t)$ to be time-varying is possible with suitable *continuum of sparse* conditions on the time domain. In that case, one can require the sparsity condition to hold uniformly across t and continuously apply thresholding. However, unlike in the traditional case, technically, estimating the family of large inverse covariances $\{E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t)^{-1}:t=1,2,\ldots\}$ uniformly over t is highly challenging. As we shall see in the proof of Proposition 4.2, even in the homoskedastic case, proving the effect of estimating $\sum_u^{-1}$ to be first-order negligible when $N/T \to \infty$ requires delicate technical analysis.

To characterize the sparsity of $\mathbf{\Sigma}_u$ in our context, define

$$m_N = \max_{i \le N} \sum_{j=1}^{N} 1\{(\sum_u)_{ij} \neq 0\}, \quad D_N = \sum_{i \neq j} 1\{(\sum_u)_{ij} \neq 0\}.$$

Here m_N represents the maximum number of nonzeros in each row, and D_N represents the total number of nonzero off-diagonal entries. Formally, we assume:

Assumption 4.2—Suppose $N^{1/2}(\log N)^{\gamma} = o(T)$ for some $\gamma > 2$, and

i. $\min_{(\sum_u)_{ij} \neq 0} |(\sum_u)_{ij}| \gg \sqrt{(logN)/T}$;

ii. at least one of the following cases holds:

a. $D_N = O(N^{1/2})$, and $m_N^2 = O(\frac{T}{N^{1/2}(logN)^{\gamma}})$

b. $D_N = O(N)$, and $m_N^2 = O(1)$.

As regulated in Assumption 4.2, we consider two kinds of sparse matrices, and develop our results for both cases. In the first case (Assumption 4.2 (ii)(a)), $\mathbf{\Sigma}_u$ is required to have no more than $O(N^{1/2})$ off-diagonal nonzero entries, but allows a diverging m_N. One typical example of this case is that there are only a small portion (e.g., finitely many) of firms whose individual shocks (u_{it}) are correlated with many other firms'. In the second case (Assumption 4.2(ii)(b)), m_N should be bounded, but $\mathbf{\Sigma}_u$ can have $O(N)$ off-diagonal nonzero entries. This allows block-diagonal matrices with finite size of blocks or banded matrices with finite number of bands. This case typically arises when firms' individual shocks are correlated only within industries but not across industries.

Moreover, we require $N^{1/2}(\log N)^{\gamma} = o(T)$, which is the price to pay for estimating a large error covariance matrix. But still we allow $N/T \to \infty$. It is also required that the minimal signal for the nonzero components be larger than the noise level (Assumption 4.2 (i)), so that nonzero components are not thresholded off when estimating $\mathbf{\Sigma}_u$.

4.6 Asymptotic properties

The following result verifies the uniform convergence required in Assumption 3.1 over the entire parameter space that contains both the null and alternative hypotheses. Recall that the OLS estimator and its asymptotic standard error are defined in (4.2).

Proposition 4.1—Suppose the distribution of $(\mathbf{f}_t, \mathbf{u}_t)$ is independent of $\boldsymbol{\theta}$. Under Assumption 4.1, for $\delta_{N,T} = log(logT)\sqrt{logN}$, as $T, N \to \infty$,

$$\inf_{\boldsymbol{\theta} \in \Theta} P(\max_{j \le N} |\hat{\theta}_j - \theta_j| / \hat{v}_j^{1/2} < \delta_{N,T} | \boldsymbol{\theta}) \to 1.$$
$$\inf_{\boldsymbol{\theta} \in \Theta} P(4/9 < \hat{v}_j / v_j < 9/4, \forall_j = 1, \ldots, N | \boldsymbol{\theta}) \to 1.$$

Proposition 4.2—Under Assumptions 4.1, 4.2, and H_0,

$$J_{wald}=\frac{Ta_{f,T}\hat{\boldsymbol{\theta}}'\hat{\sum}_u^{-1}\hat{\boldsymbol{\theta}}-N}{\sqrt{2N}}\to^d\mathcal{N}(0,1).$$

As shown, the effect of replacing $\sum_u^{-1}$ by its thresholded estimator is asymptotically negligible and the size of the standard Wald statistic can be well controlled.

We are now ready to apply Theorem 3.3 to obtain the asymptotic properties of $J = J_0 + J_{wald}$ as follows. For $\delta_{N,T}=log(logT)\sqrt{logN}$, let

$$\Theta_s=\{\boldsymbol{\theta}\in\Theta:\max_{j\le N}\frac{T^{1/2}|\theta_j|}{\mathrm{var}^{1/2}(u_{jt})}>3a_f^{-1/2}\delta_{N,T}\},$$
$$\Theta(J_{wald})=\{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}\|^2>C\delta_{N,T}^2N/T\}.$$

Theorem 4.1—Suppose the assumptions of Propositions 4.1 and 4.2 hold.

i. Under the null hypothesis H_0: $\boldsymbol{\theta}=\mathbf{0}$, as $T, N\to\infty$,

$$P(J_0=0|H_0)\to 0,\quad J_{wald}\to^d\mathcal{N}(0,1),$$

and hence

$$J=J_0+J_{wald}\to^d\mathcal{N}(0,1).$$

ii. There is $C > 0$ so that for any $q\in(0,1)$, as $T, N\to\infty$,

$$\inf_{\boldsymbol{\theta}\in\Theta}P(J_0>\sqrt{N}|\boldsymbol{\theta})\to 1,\quad \inf_{\boldsymbol{\theta}\in\Theta(J_{wald})}P(J_{wald}>z_q|\boldsymbol{\theta})\to 1,$$

and hence

$$\inf_{\boldsymbol{\theta}\in\Theta_s\cup\Theta(J_{wald})}P(J>z_q|\boldsymbol{\theta})\to 1,$$

where z_q denotes the $1-q$ quantile of the standard normal distribution.

We see that the power is substantially enhanced after J_0 is added, as the region where the test has power is enlarged from $\Theta(J_{wald})$ to $\Theta_s\cup\Theta(J_{wald})$.

5 Application: Testing Cross-Sectional Independence

5.1 The model

Consider a mixed effect panel data model

$$y_{it}=\alpha+\mathbf{x}_{it}^{'}\boldsymbol{\beta}+\mu_i+u_{it},\ \ i\le n, t\le T,$$

where the idiosyncratic error u_{it} is assumed to be Gaussian. The regressor $\mathbf{x}_{it}$ could be correlated with the individual random effect μ_i, but is uncorrelated with u_{it}. Let ρ_{ij} denote the correlation between u_{it} and u_{jt}, assumed to be time invariant. The goal is to test the following hypothesis:

$$H_0{:}\rho_{ij}=0,\ \ \text{for all}\, i\neq j,$$

that is, whether the cross-sectional dependence is present. It is commonly known that the cross-sectional dependence leads to efficiency loss for OLS, and sometimes it may even cause inconsistency (Andrews, 2005). Thus testing H_0 is an important problem in applied panel data models. If we let $N = n(n - 1)/2$, and $\boldsymbol{\theta} = (\rho_{12}, \ldots, \rho_{1n}, \rho_{23}, \ldots, \rho_{2n}, \ldots, \rho_{n-1,n})'$ be an $N \times 1$ vector stacking all the mutual correlations, then the problem is equivalent to testing about a high-dimensional vector H_0: $\boldsymbol{\theta} = 0$. Note that often the cross-sectional dependences are weakly present. Hence the alternative hypothesis of interest is often a sparse vector $\boldsymbol{\theta}$, corresponding to a sparse covariance matrix $\boldsymbol{\Sigma}_u$ of u_{it}.

Most of the existing tests are based on the quadratic statistic $W=\sum_{i<j}T\hat{\rho}_{ij}^2=T\hat{\boldsymbol{\theta}}^{'}\hat{\boldsymbol{\theta}}$, where $\hat{\rho}_{ij}$ is the sample correlation between u_{it} and u_{jt}, estimated by the within-OLS (Baltagi, 2008), and $\hat{\boldsymbol{\theta}}=(\hat{\rho}_{12},\ldots,\hat{\rho}_{n-1,n})$. Pesaran et al. (2008) and Baltagi et al. (2012) studied the rescaled W, and showed that after a proper standardization, the rescaled W is asymptotically normal when both $n, T \to \infty$. However, the quadratic test suffers from a low power if $\boldsymbol{\Sigma}_u$ is a sparse matrix. In particular, as is shown in Theorem 3.4, when $n/T \to \infty$, the quadratic test cannot detect the sparse alternatives with $\Sigma_{i<j}1\{\rho_{ij} \neq 0\} = o(n/T)$, which can be restrictive. Such a sparse structure is present, for instance, when $\boldsymbol{\Sigma}_u$ is a block-diagonal sparse matrix with finite block sizes.

5.2 Power enhancement test

Following the conventional notation of panel data models, let $\hat{y}_{it}=y_{it}-\frac{1}{T}\sum_{t=1}^{T}y_{it}$, $\hat{\mathbf{x}}_{it}=\mathbf{x}_{it}-\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$, and $\tilde{u}_{it}=u_{it}-\frac{1}{T}\sum_{t=1}^{T}u_{it}$. Then $\tilde{y}_{it}=\tilde{\mathbf{x}}^{'}_{it}\boldsymbol{\beta}+\tilde{u}_{it}$. The within-OLS estimator $\hat{\boldsymbol{\beta}}$ is obtained by regressing $\tilde{y}_{it}$ on $\tilde{\mathbf{x}}_{it}$ for all i and t, which leads to the estimated residual $\hat{u}_{it}=\tilde{y}_{it}-\tilde{\mathbf{x}}^{'}_{it}\hat{\boldsymbol{\beta}}$. Then ρ_{ij} is estimated by

$$\hat{\rho}_{ij}=\frac{\hat{\sigma}_{ij}}{\hat{\sigma}_{ii}^{1/2}\hat{\sigma}_{jj}^{1/2}},\ \ \hat{\sigma}_{ij}=\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}.$$

For the within-OLS, the asymptotic variance of $\hat{\rho}_{ij}$ is given by $v_{ij}=(1-\rho_{ij}^2)^2/T$, and is estimated by $\hat{v}_{ij}=(1-\hat{\rho}_{ij}^2)^2/T$. Therefore the screening statistic for the power enhancement test is defined as

$$J_0=\sqrt{N}\sum_{(i,j)\in\hat{S}}\hat{\rho}_{ij}^2\hat{v}_{ij}^{-1},\quad \hat{S}=\{(i,j):|\hat{\rho}_{ij}|/\hat{v}_{ij}^{1/2}>\delta_{N,T},i<j\le n\}. \tag{5.1}$$

where $\delta_{N,T}=log(logT)\sqrt{logN}$ as before. The set $\hat{S}$ screens out most of the estimation errors.

To control the size, we employ Baltagi et al. (2012)'s bias-corrected quadratic statistic:

$$J_1=\sqrt{\frac{1}{n(n-1)}}\sum_{i<j}(T\hat{\rho}_{ij}^2-1)-\frac{n}{2(T-1)}. \tag{5.2}$$

Under regularity conditions (Assumptions 5.1, 5.2 below), $J_1\to^d\mathcal{N}(0,1)$ under H_0. Then the power enhancement test can be constructed as $J = J_0 + J_1$. The power is substantially enhanced to cover the region

$$\Theta_s=\{\boldsymbol{\theta}:\max_{i<j}\frac{\sqrt{T}|\rho_{ij}|}{1-\rho_{ij}^2}>3log(logT)\sqrt{logN}\}, \tag{5.3}$$

in addition to the region detectable by J_1 itself. As a byproduct, it also identifies pairs (i, j) for $\rho_{ij} \neq 0$ through $\hat{S}$. Empirically, this set helps us understand better the underlying pattern of cross-sectional correlations and subsequently the cause of the correlation.

5.3 Asymptotic properties

In order for the power to be uniformly enhanced, the parameter space of $\boldsymbol{\theta}=(\rho_{12},\ldots,\rho_{1n},\rho_{23},\ldots,\rho_{2n},\ldots,\rho_{n-1,n})'$ is required to be: $\boldsymbol{\theta}$ is element-wise bounded away from ±1: there is $\rho_{\max}\in(0,1)$,

$$\Theta=\{\boldsymbol{\theta}\in \mathrm{R}^{N}:\|\boldsymbol{\theta}\|_{\max}\le\rho_{\max}\}.$$

The following regularity conditions are imposed. They hold uniformly in $\boldsymbol{\theta}\in\Theta$.

Assumption 5.1—There are C_1, $C_2 > 0$, so that

i. $\sum_{i\neq j\le n}|E\tilde{\mathbf{x}}'_{it}\tilde{\mathbf{x}}_{it}E(u_{it}u_{jt})|<C_1n$,

ii. $\max_{j\le n}E(u_{jt}^4)<C_1,\min_{j\le n}E(u_{jt}^2)<C_2$,

Condition (i) is needed for the within-OLS to be $\sqrt{nT}$-consistent (see, e.g., Baltagi (2008)). It is usually satisfied by weak cross-sectional correlations (sparse alternatives) among the error terms, or weak dependence among the regressors. We require the second moment of u_{jt}

be bounded away from zero uniformly in $j \le n$ and $\boldsymbol{\theta} \in \Theta$, so that the cross-sectional correlations can be estimated stably.

The following conditions are assumed in Baltagi et al. (2012), which are needed for the asymptotic normality of J_1 under H_0.

Assumption 5.2

i. $\{\mathbf{u}_t\}_{t \le T}$ are i.i.d. N(0, $\boldsymbol{\Sigma}_u$), $E(\mathbf{u}_t | \{\mathbf{f}_t\}_{t \le T}, \boldsymbol{\theta}) = 0$ almost surely.

ii. With probability approaching one, all the eigenvalues of $\frac{1}{T}\sum_{t=1}^{T} \tilde{\mathbf{x}}_{jt} \tilde{\mathbf{x}}'_{jt}$ are bounded away from both zero and infinity uniformly for $j \le n$.

Proposition 5.1—Under Assumptions 5.1 and 5.2, for $\delta_{N,T} = log(logT)\sqrt{logN}$, and $N = n(n-1)/2$, as $T, N \to \infty$,

$$\inf_{\boldsymbol{\theta} \in \Theta} P(\max_{ij} |\hat{\rho}_{ij} - \rho_{ij}| / \hat{v}_{ij}^{1/2} < \delta_{N,T} | \boldsymbol{\theta}) \to 1$$
$$\inf_{\boldsymbol{\theta} \in \Theta} P(4/9 < \hat{v}_{ij}/v_{ij} < 9/4, \forall_i \ne j | \boldsymbol{\theta}) \to 1.$$

Define

$$\Theta(J_1) = \{\boldsymbol{\theta} \in \Theta : \sum_{i<j} \rho_{ij}^2 \ge C n^2 log n / T\}.$$

For J_1 defined in (5.2), let

$$J = J_0 + J_1. \quad (5.4)$$

The main result is presented as follows.

Theorem 5.1—Suppose Assumptions 5.1, 5.2 hold. As $T, N \to \infty$,

i. under the null hypothesis H_0: $\boldsymbol{\theta} = \mathbf{0}$,

$$P(J_0 = 0 | H_0) \to 0, \quad J_{wald} \to^d \mathcal{N}(0,1),$$

and hence

$$J = J_0 + J_1 \to^d \mathcal{N}(0,1);$$

ii. there is $C > 0$ in the definition of $\Theta(J_1)$ so that for any $q \in (0, 1)$,

$$\inf_{\boldsymbol{\theta} \in \Theta_s} P(J_0 > \sqrt{N} | \boldsymbol{\theta}) \to 1, \quad \inf_{\boldsymbol{\theta} \in \Theta(J_1)} P(J_1 > z_q | \boldsymbol{\theta}) \to 1,$$

and hence

$$\inf_{\boldsymbol{\theta}\in\Theta_s\cup\Theta(J_1)} P(J_1>z_q|\boldsymbol{\theta}) \to 1.$$

Note that the high power region is enhanced from $\Theta(J_1)$ to $\Theta_s \cup \Theta(J_1)$ uniformly over sparse alternatives. In particular, the required signal strength of Θ_s in (5.3) is mild: the maximum cross-sectional correlation is only required to exceed a magnitude of $log(logT)\sqrt{(logN)/T}$.

6 Numerical studies

In this section, Monte Carlo simulations are employed to examine the finite sample performance of the power enhancement tests. We also present empirical evidence of sparse alternatives in the factor pricing model using real data.

6.1 Testing factor pricing models

To mimic the real data application, we consider the Fama and French (1992) three-factor model:

$$y_{it}=\theta_i+\mathbf{b}_i'\mathbf{f}_t+u_{it}.$$

We simulate $\{\mathbf{b}_i\}_{t=1}^N$, $\{\mathbf{f}_t\}_{t=1}^T$ and $\{\mathbf{u}_t\}_{t=1}^T$ independently from $\mathcal{N}_3(\boldsymbol{\mu}_B, \sum_B)$, $\mathcal{N}_3(\boldsymbol{\mu}_f, \sum_f)$, and $\mathcal{N}_N(0, \sum_u)$ respectively. The parameters are set to be the same as those in the simulations of Fan et al. (2013), which are calibrated using daily returns of S&P 500's top 100 constituents, for the period from July 1st, 2008 to June 29th 2012. These parameters are listed in the following table.

Set $\boldsymbol{\Sigma}_u$ = diag{$\mathbf{A}_1$, …, $\mathbf{A}_{N/4}$} to be a block-diagonal correlation matrix. Each diagonal block $\mathbf{A}_j$ is a 4 × 4 positive definite matrix, whose correlation matrix has equi-off-diagonal entry ρ_j, generated from Uniform[0, 0.5].

We evaluate the powers of our tests under two specific alternatives (we set $N > T$):

$$\text{spares alternative } H_a^1{:}\theta_i=\begin{cases} 0.3, & i\le \frac{N}{T} \\ 0, & i>\frac{N}{T} \end{cases}$$
$$\text{weak theta } H_a^2{:}\theta_i=\begin{cases} \sqrt{\frac{logN}{T}}, & i\le N^{0.4} \\ 0, & i>N^{0.4} \end{cases}.$$

Under H_a^1, there are only a few nonzero θ's with a relatively large magnitude. Under H_a^2, there are many non-vanishing θ's, but their magnitudes are all relatively small. In our simulation setup, $\sqrt{logN/T}$ varies from 0.05 to 0.10. We therefore expect that under H_a^1, $P(\hat{S}=\varnothing)$ is close to zero, as most of the first N/T estimated θ's should survive from the screening step. These survived $\hat{\theta}$'s contribute importantly to the rejection of the null

hypothesis. In contrast, $P(\hat{S}=\varnothing)$ should be much larger under H_a^2 because the non-vanishing θ's are too weak to be detected.

Four testing methods are conducted and compared: the standardized Wald test J_{wald}, the thresholding test J_{thr} as in Fan (1996), and their power enhancement versions $J_0 + J_{wald}$ and $J_0 + J_{thr}$. In particular, the thresholding test J_{thr} is defined as, $\sigma_N^2 = \sqrt{2/\pi}a^{-1}t_N^3(1+3t_N^{-2})$ and $\mu_N = \sqrt{2/\pi}a^{-1}t_N(1+t_N^{-2})$,

$$J_{thr} = \sigma_N^{-1}\left(\sum_{j=1}^{N}\hat{\theta}_j^2\hat{v}_j^{-1}1\{|\hat{\theta}_j|\hat{v}_j^{-1/2} > t_N\} - \mu_N\right),$$

where $t_N = \sqrt{2log(Na)}$, $a = (\log N)^{-2}$. Here the threshold value t_N is chosen smaller than our $\delta_{N,T}$, and it results in a non-degenerate null distribution of J_{thr}. When Σ_u is diagonal, its asymptotic null distribution is $\mathcal{N}(0,1)$, but when Σ_u is non-diagonal, it can suffer from substantial size distortions (see Fan (1996) for detailed discussions). For each test, we calculate the relative frequency of rejection under H_0, H_a^1 and H_a^2 based on 2000 replications, with significance level $q = 0.05$. We also calculate the relative frequency of $\hat{S}$ being empty, which approximates $P(\hat{S}=\varnothing)$. We use the soft-thresholding to estimate the error covariance matrix.

Table II presents the empirical size and power of each testing method. Numerical findings are summarized as follows.

- **i.** The sizes of both J_{wald} and $J_0 + J_{wald}$ are close to the significance level. In contrast, the thresholding tests (J_{thr} and $J_{thr} + J_0$) have significant size distortions. Furthermore, adding J_0 results in just 0.1–0.2% increase of the size.
- **ii.** Under H_0, $P(\hat{S}=\varnothing)$ is close to one, indicating that the power enhancement component J_0 screens off most of the estimation errors. Under H_a^1, $P(\hat{S}=\varnothing)$ is less than 10% because the screening procedure manages to capture the big thetas. Under H_a^2, as the non-vanishing thetas are very weak, $\hat{S}$ has a large chance of being empty.
- **iii.** Under H_a^1, the power of the thresholding test is much higher than that of the Wald test, as the Wald test accumulates too many estimation errors. Besides, the power is significantly enhanced after J_0 is added.
- **iv.** Finally, under H_a^2, the power enhancement is not substantial as the nonzero thetas are very weak, and the thresholding test has higher power than J_0+J_{wald} does. The power of the thresholding test can be further enhanced after J_0 is added with little increase of false rejections. Note that in this case $\hat{S}$ still has more than 10% chance of being nonempty. Whenever it is non-empty, adding J_0 potentially enhances the power of the test.

6.2 Testing cross-sectional independence

We use the following data generating process in our experiments,

$$\begin{aligned} y_{it} &= \alpha + \beta x_{it} + \mu_i + u_{it}, \quad i \le n, t \le T, \\ x_{it} &= \xi x_{i,t-1} + \mu_i + \varepsilon_{it}. \end{aligned} \tag{6.1}$$

Note that we model $\{x_i\}$'s as AR(1) processes, so that x_{it} is possibly correlated with μ_i, but not with u_{it}, as was the case in Im et al. (1999). For each i, initialize $x_{it} = 0.5$ at $t = 1$. We specify the parameters as follows: μ_i is drawn from $\mathcal{N}(0, 0.25)$ for $i = 1, \ldots, n$. The parameters α and β are set −1 and 2 respectively. In regression (6.1), $\xi = 0.7$ and $\varepsilon_{it} \sim \mathcal{N}(0, 1)$.

We generate $\{\mathbf{u}_t\}_{t=1}^T$ from $\mathcal{N}_n(0, \sum_u)$. Under the null hypothesis, $\mathbf{\Sigma}_u$ is set to be a diagonal matrix $\sum_{u,0} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$. Following Baltagi et al. (2012), consider the heteroscedastic errors

$$\sigma_i^2 = \sigma^2 (1 + \kappa \bar{x}_i)^2 \tag{6.2}$$

with $\kappa = 0.5$, where $\bar{x}_i$ is the average of x_{it} across t. Here σ^2 is scaled to fix the average of σ_i^2's at one.

For alternative specifications, we use a spatial model for the errors u_{it}. Baltagi et al. (2012) considered a tri-diagonal error covariance matrix in this case. We extend it by allowing for higher order spatial autocorrelations, but require that not all the errors be spatially correlated with their immediate neighbors. Specifically, we start with $\mathbf{\Sigma}_{u,1} = \mathrm{diag}\{\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_{n/4}\}$ as a block-diagonal matrix with 4 × 4 blocks located along the main diagonal. Each $\mathbf{\Sigma}_i$ is assumed to be $\mathbf{I}_4$ initially. We then randomly choose $\lfloor n^{0.3} \rfloor$ blocks among them and make them non-diagonal by setting $\mathbf{\Sigma}_i(m, n) = \rho^{|m-n|}(m, n \le 4)$, with $\rho = 0.2$. To allow for error cross-sectional heteroscedasticity, we set $\sum_u = \sum_{u,0}^{1/2} \sum_{u,1} \sum_{u,0}^{1/2}$, where $\sum_{u,0} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$ as specified in (6.2).

The Monte Carlo experiments are conducted for different pairs of (n, T) with significance level $q = 0.05$ based on 2000 replications. The empirical size, power and the frequency of $\hat{S} = \emptyset$ as in (5.1) are recorded.

Table III gives the size and power of the bias-corrected quadratic test J_1 in (5.2) and those of the power enhanced test $J_0 + J_1$. The sizes of both tests are close to 5%. In particular, the power enhancement test has little distortion of the original size.

The bottom panel shows the power of the two tests under the alternative specification. The power enhancement test demonstrates almost full power under all combinations of (n, T). In contrast, the quadratic test J_1 only gains power when T gets large. As n increases, the proportion of nonzero off-diagonal elements in $\mathbf{\Sigma}_u$ gradually decreases. It becomes harder

for J_1 to effectively detect those deviations from the null hypothesis. This explains the low power exhibited by the quadratic test when facing a high sparsity level.

6.3 Empirical evidence of sparse alternatives

We present empirical evidence of sparse alternatives based on a real data example. Consider Carhart (1997)'s four-factor model on the S&P 500 index. We collect monthly excess returns on all the S&P 500 constituents from the CRSP database for the period January 1980 to December 2012, and construct the screening set $\hat{S}$ on a rolling window basis: for each month, we evaluate $\hat{S}$ using the preceding 60 months' returns ($T = 60$). The panel at each month consists of stocks without missing observations in the past five years, which yields a balanced panel with the cross-sectional dimension larger than the time-series dimension ($N > T$). In this manner we not only capture the up-to-date information in the market, but also mitigate the impact of time-varying factor loadings and sampling biases. In particular, for testing months $\tau = 1984.12, \ldots, 2012.12$, we run the regressions

$$r_{it}^{\tau} - r_{ft}^{\tau} = \theta_i^{\tau} + \beta_{i,\mathrm{MKT}}^{\tau}(\mathrm{MKT}_t^{\tau} - r_{ft}^{\tau}) + \beta_{i,\mathrm{SMB}}^{\tau}\mathrm{SMB}_t^{\tau} + \beta_{i,\mathrm{HML}}^{\tau}\mathrm{HML}_t^{\tau} + \beta_{i,\mathrm{MOM}}^{\tau}\mathrm{MOM}_t^{\tau} + u_{it}^{\tau}, \quad (6.3)$$

for $i = 1, \ldots, N_\tau$ and $t = \tau - 59, \ldots, \tau$, where r_{it} represents the return for stock i at month t, r_{ft} the risk free rate, and MKT, SMB, HML and MOM constitute market, size, value and momentum factors. The time series of factors are downloaded from Kenneth French's website. To make the notation consistent, we use θ_i^{τ} to represent the "alpha" of stock i.

Table IV summarizes descriptive statistics for different components and estimates in the model. On average, 618 stocks (which is more than 500 because we are recording stocks that have *ever* become the constituents of the index) enter the panel of the regression during each five-year estimation window. Of those, merely an average of 5.2 stocks are selected by the screening set $\hat{S}$ which directly implies the presence of sparse alternatives. The threshold $\delta_{N,T} = \sqrt{(\log N)}\log(\log T)$ varies as the panel size N changes at the end of each month, and is about 3.5 on average. The selected stocks have much larger alphas (θ) than other stocks do. Therefore empirically we find that there are only a few significant nonzero "alpha" components, corresponding to a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market.

The power enhancement procedure is particularly suited for the empirical setting where sparse alternatives are present. Note that finding only a few stocks with nonzero alphas is probably explained by the focus on a balanced panel of highly traded stocks with large capitalizations (cf. constituents of the S&P500). On the other hand, as in Gagliardini et al. (2011), the empirical finding would probably be different if we consider a much larger universe of stocks with possibly many more mis-pricing.

7 Discussions

We consider testing a high-dimensional vector H: $\theta = 0$ against sparse alternatives where the null hypothesis is violated only by a few components. We introduce a "power enhancement component" J_0 based on a screening technique, which is zero under the null, but diverges

quickly under sparse alternatives. We suggest constructing J_0 as described in the paper since the screening set $\hat{S}$ can reveal the sparse structure of θ, and a negative outcome of the test suggests a specific set of alternatives.

In the factor pricing model, the issue of missing a small number of factors is also important to consider. One one hand, when the unspecified factors are not "pervasive", only assets that are influenced by the missing factors are affected, which may lead to a sparse alternative. On the other hand, the unspecified pervasive factors may substantially affect the sparse structure of either θ or Σ_u, or both. In this case, we can extend the current model to allow for unobservable factors, which can be statistically inferred using principal components (PC) method as in Stock and Watson (2002) and Fan et al. (2013). Since the PC method is robust to over-specifying factors, the screening set $\hat{S}$ should be stable if the "working number of factors" $\hat{K}$ is no smaller than the true number of factors K. As a result, one can estimate θ and construct $\hat{S}$ using either a consistent estimator of K or a slightly overestimated $\hat{K}$. Once $\hat{S}$ is reasonably robust to the choice of $\hat{K}$, it indicates that no pervasive factors are omitted. We can then proceed to use the proposed J_0 to conduct the test.

In addition, this paper considers unconditional population moments of asset returns and focuses on the factor pricing model. Unconditional moments of financial returns, under a broad class of data generation processes, are time invariant and can thus be estimated from time series data. Though theoretical models often imply a conditional linear model with respect to investors' information sets, it is much more convenient to deduce testable implication that does not depend on this conditioning information (see Hansen and Richard (1987)). On the other hand, the use of conditioning information is also appealing and has been addressed by several authors (see,e.g., Gagliardini et al. (2011) and Ang and Kristensen (2012)). It will be an interesting direction to accommodate such conditioning information in deriving power enhancement tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

*We thank co-editor Lars Hansen and anonymous referees for many insightful comments and suggestions, which have greatly improved the paper. The authors are also grateful to the comments from Per Mykland and seminar and conference participants at UChicago, Princeton, Georgetown, George Washington, 2014 Econometric Society North America Summer Meeting, UCL workshop on High-dimensional Econometrics Models, The 2014 Annual meeting of Royal Economics Society, The 2014 Asian Meeting of the Econometric Society, 2014 International Conference on Financial Engineering and Risk Management, and 2014 Midwest Econometric Group meeting. The research was partially supported by National Science Foundation grants DMS-1206464 and DMS-1406266, and National Institute of Health grants R01GM100474-01 and R01-GM072611.

A Proofs for Section 3

Throughout the proofs, let C be a generic constant, which may differ at different places.

A.1 Proof of Theorem 3.1

Proof

Define events

$$A_1=\left\{\max_{j\le N}|\hat{\theta}_j-\theta_j|/\hat{v}_j^{1/2}<\delta_{N,T}\right\},\quad A_2=\left\{\frac{4}{9}\le \hat{v}_j/v_j\le\frac{9}{4},\forall_j=1,\ldots,N\right\}.$$

For any $j\in S(\boldsymbol{\theta})$, by the definition of $S(\boldsymbol{\theta})$, $|\theta_j|>3\delta_{N,T}v_j^{1/2}$. Under $A_1\cap A_2$,

$$\frac{|\hat{\theta}_j|}{\hat{v}_j^{1/2}}\ge\frac{|\theta_j|-|\hat{\theta}_j-\theta_j|}{\hat{v}_j^{1/2}}\ge\frac{2|\theta_j|}{3v_j^{1/2}}-\delta_{N,T}>\delta_{N,T}.$$

This implies that $j\in\hat{S}$, hence $S(\boldsymbol{\theta})\subset\hat{S}$. In fact, we have proved this statement on the event $A_1\cap A_2$ uniformly for $\boldsymbol{\theta}\in\Theta$:

$$\inf_{\boldsymbol{\theta}\in\Theta}P(S(\boldsymbol{\theta})\subset\hat{S}|\boldsymbol{\theta})\to 1.$$

Moreover, it is readily seen that, under H_0: $\boldsymbol{\theta}=0$, by Assumption 3.1,

$$P(J_0=0|H_0)=P(\hat{S}=\varnothing|H_0)=P(\max_{j\le N}\{|\hat{\theta}_j|/\hat{v}_j^{1/2}\}<\delta_{N,T}|H_0)\to 1.$$

In addition, by $\inf_{\boldsymbol{\theta}\in\Theta}P(S(\boldsymbol{\theta})\subset\hat{S}|\boldsymbol{\theta})\to 1$,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}P(J_0\le\sqrt{N}|S(\boldsymbol{\theta})\ne\varnothing)&\le\sup_{\boldsymbol{\theta}\in\Theta}P(J_0\le\sqrt{N},\hat{S}\ne\varnothing|S(\boldsymbol{\theta})\ne\varnothing)+\sup_{\boldsymbol{\theta}\in\Theta}P(\hat{S}=\varnothing|S(\boldsymbol{\theta})\ne\varnothing)\\&\le\sup_{\boldsymbol{\theta}\in\Theta}P(\sqrt{N}\sum_{j\in\hat{S}}\delta_{N,T}^2\le\sqrt{N},\hat{S}\ne\varnothing|S(\boldsymbol{\theta})\ne\varnothing)+o(1)\\&\le\sup_{\boldsymbol{\theta}\in\Theta}P(\sqrt{N}\delta_{N,T}^2\le\sqrt{N}|S(\boldsymbol{\theta})\ne\varnothing)+o(1)\to 0.\end{aligned}$$

Note that the last convergence holds uniformly in $\boldsymbol{\theta}\in\Theta$ because $\delta_{N,T}\to\infty$. Therefore, $\inf_{\boldsymbol{\theta}\in\Theta}P(J_0>\sqrt{N}|S(\boldsymbol{\theta})\ne\varnothing)\to 1$. This completes the proof.

A.2 Proof of Theorem 3.2

Proof

It follows immediately from $P(J_0=0|H_0)\to 1$ that $J\to^d F$, and hence the critical region $\{\mathbf{D}: J>F_q\}$ has size q. Moreover, by the power condition of J_1 and $J_0\ge 0$,

$$\inf_{\boldsymbol{\theta}\in\Theta(J_1)}P(J>F_q|\boldsymbol{\theta})\ge\inf_{\boldsymbol{\theta}\in\Theta(J_1)}P(J_1>F_q|\boldsymbol{\theta})\to 1.$$

This together with the fact

$$\inf_{\boldsymbol{\theta}\in\Theta_s\cup\Theta(J_1)} P(J>F_q|\boldsymbol{\theta}) \geq \min\{\inf_{\boldsymbol{\theta}\in\Theta_s} P(J>F_q|\boldsymbol{\theta}), \inf_{\boldsymbol{\theta}\in\Theta(J_1)} P(J>F_q|\boldsymbol{\theta})\},$$

establish the theorem, if we show $\inf_{\boldsymbol{\theta}\in\Theta_s} P(J>F_q|\boldsymbol{\theta}) \to 1$.

By the definition of $\hat{S}$ and J_0, we have $\{J_0<\sqrt{N}\delta_{N,T}^2\}=\{\hat{S}=\varnothing\}$. Since $\inf_{\theta\in\Theta} P(S(\boldsymbol{\theta}) \subset \hat{S}|\boldsymbol{\theta}) \to 1$ and $\Theta_s=\{\boldsymbol{\theta}\in\Theta: S(\boldsymbol{\theta})\neq\varnothing\}$, we have

$$\begin{aligned}\sup_{\Theta_s} P(J_0<\sqrt{N}\delta_{N,T}^2|\boldsymbol{\theta}) &= \sup_{\Theta_s} P(\hat{S}=\varnothing|\boldsymbol{\theta}) \\ &\leq \sup_{\{\boldsymbol{\theta}\in\Theta:S(\boldsymbol{\theta})\neq\varnothing\}} P(\hat{S}=\varnothing, S(\boldsymbol{\theta}) \subset \hat{S}|\boldsymbol{\theta}) + o(1),\end{aligned}$$

which converges to zero, since the first term is zero. This implies $\inf_{\Theta_s} P(J_0 \geq \sqrt{N}\delta_{N,T}^2|\boldsymbol{\theta}) \to 1$. Then by condition (ii), as $\delta_{N,T} \to \infty$,

$$\inf_{\boldsymbol{\theta}\in\Theta_s} P(J>F_q|\boldsymbol{\theta}) \geq \inf_{\boldsymbol{\theta}\in\Theta_s} P(\sqrt{N}\delta_{N,T}^2+J_1>F_q|\boldsymbol{\theta}) \geq \inf_{\theta\in\Theta_s} P(c\sqrt{N}+J_1>F_q|\boldsymbol{\theta}) \to 1.$$

This completes the proof.

A.3 Proof of Theorem 3.3

Proof

It suffices to verify conditions (i)–(iii) in Theorem 3.2 for $J_1 = J_Q$. Condition (i) follows from Assumption 3.2. Condition (iii) is fulfilled for $c > 2/\xi$, since

$$\inf_{\boldsymbol{\theta}\in\Theta_s} P(c\sqrt{N}+J_Q>F_q|\boldsymbol{\theta}) \geq \inf_{\boldsymbol{\theta}\in\Theta_s} P(c\sqrt{N} - \frac{N(1+\mu_{N,T})}{\xi_{N,T}\sqrt{N}}>F_q|\boldsymbol{\theta}) \to 1,$$

by using $F_q = O(1)$, $\xi_{N,T} \to \xi$, and $\mu_{N,T} \to 0$. We now verify condition (ii) for the $\Theta(J_Q)$ defined in the theorem. Let $\mathbf{D} = \text{diag}(v_1, \ldots, v_N)$. Then $\|\mathbf{D}\|_2 < C_3/T$ by Assumption 3.2(iv).

On the event $A=\{\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\mathbf{D}^{-1/2}\|^2<\delta_{N,T}^2 N/4\}$, we have

$$\begin{aligned}|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\mathbf{V}\boldsymbol{\theta}| &\leq \|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'\mathbf{D}^{-1/2}\|\|\mathbf{D}^{1/2}\mathbf{V}\boldsymbol{\theta}\| \\ &\leq \delta_{N,T}\sqrt{N}\|\mathbf{D}\|_2^{1/2}\|\mathbf{V}\|_2^{1/2}\left(\boldsymbol{\theta}'\mathbf{V}\boldsymbol{\theta}\right)^{1/2}/2 \\ &\leq \delta_{N,T}\sqrt{N}(C_3/T)^{1/2}\|\mathbf{V}\|_2^{1/2}\left(\boldsymbol{\theta}'\mathbf{V}\boldsymbol{\theta}\right)^{1/2}/2.\end{aligned}$$

For $\|\boldsymbol{\theta}\|^2>C\delta_{N,T}^2 N/T$ with $C = 4C_3\|\mathbf{V}\|_2/\lambda_{\min}(\mathbf{V})$, we can bound further that

$$|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{'}\mathbf{V}\boldsymbol{\theta}| \leq \boldsymbol{\theta}^{'}\mathbf{V}\boldsymbol{\theta}/4.$$

Hence, $\hat{\boldsymbol{\theta}}^{'}\mathbf{V}\hat{\boldsymbol{\theta}} \geq \boldsymbol{\theta}^{'}\mathbf{V}\boldsymbol{\theta} - 2(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{'}\mathbf{V}\boldsymbol{\theta} \geq \boldsymbol{\theta}^{'}\mathbf{V}\boldsymbol{\theta}/2$. Therefore,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta(J_Q)} P(J_Q \leq F_q|\boldsymbol{\theta}) &\leq \sup_{\Theta(J_Q)} P(\tfrac{T\boldsymbol{\theta}^{'}\mathbf{V}\boldsymbol{\theta}/2-2N}{\xi\sqrt{N}} \leq F_q|\boldsymbol{\theta}) + \sup_{\Theta(J_Q)} P(A^c|\boldsymbol{\theta}) \\ &\leq \sup_{\Theta(J_Q)} P(T\lambda_{\min}(\mathbf{V})\|\boldsymbol{\theta}\|^2 < 2F_q\xi\sqrt{N}+4N|\boldsymbol{\theta}) + o(1) \\ &\leq \sup_{\Theta(J_Q)} P(\lambda_{\min}(\mathbf{V})C\delta_{N,T}^2 N < 5N|\boldsymbol{\theta}) + o(1),\end{aligned}$$

which converges to zero since $\delta_{N,T}^2 \to \infty$. This implies $\inf_{\Theta(J_Q)} P(J_Q > F_q|\boldsymbol{\theta}) \to 1$ and finishes the proof.

A.4 Proof of Theorem 3.4

Proof

Through this proof, *C* is a generic constant, which can vary from one line to another. Without loss of generality, under the alternative, write

$$\boldsymbol{\theta}^{'} = (\boldsymbol{\theta}_1^{'}, \boldsymbol{\theta}_2^{'}) = (\mathbf{0}^{'}, \boldsymbol{\theta}_2^{'}),\ \ \hat{\boldsymbol{\theta}}^{'} = (\hat{\boldsymbol{\theta}}_1^{'}, \hat{\boldsymbol{\theta}}_2^{'}),$$

where $\dim(\boldsymbol{\theta}_1) = N - r_N$ and $\dim(\boldsymbol{\theta}_2) = r_N$. Corresponding to $(\theta^{'}{}_1, \theta^{'}{}_2)$, we partition $\mathbf{V}^{-1}$ and $\mathbf{V}$ into:

$$\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{M}_1 & \boldsymbol{\beta}^{'} \\ \boldsymbol{\beta} & \mathbf{M}_2 \end{pmatrix} \text{ and } \mathbf{V} = \begin{pmatrix} \mathbf{M}_1^{-1}+\mathbf{A} & \mathbf{G}^{'} \\ \mathbf{G} & \mathbf{C} \end{pmatrix},$$

where $\mathbf{M}_1$ and $\mathbf{A}$ are $(N - r_N) \times (N - r_N)$; $\boldsymbol{\beta}$ and $\mathbf{G}$ are $r_N \times (N - r_N)$; $\mathbf{M}_2$ and $\mathbf{C}$ are $r_N \times r_N$.

By the matrix inversion formula,

$$\mathbf{A} = \mathbf{M}_1^{-1}\boldsymbol{\beta}^{'}(\mathbf{M}_2 - \beta\mathbf{M}_1^{-1}\boldsymbol{\beta}^{'})^{-1}\boldsymbol{\beta}\mathbf{M}_1^{-1}.$$

Let $\Delta = T\hat{\theta}^{'}\mathbf{V}\hat{\theta} - T\hat{\theta}_1^{'}\mathrm{M}_1^{-1}\hat{\boldsymbol{\theta}}_1$. Note that

$$\Delta = T\hat{\theta}_1^{'}\mathrm{A}\widehat{\boldsymbol{\theta}_1} + 2T\hat{\theta}_2^{'}\mathbf{G}\widehat{\boldsymbol{\theta}_1} + T\hat{\theta}_2^{'}\mathbf{C}\widehat{\boldsymbol{\theta}_2}.$$

We first look at $T\hat{\boldsymbol{\theta}}_1^{'}\mathbf{A}\widehat{\boldsymbol{\theta}_1}$. Let $\lambda_{N,T}=T\lambda_{\max}((\mathbf{M}_2-\boldsymbol{\beta}\mathbf{M}_1^{-1}\boldsymbol{\beta}^{'})^{-1})$ and $\mathbf{D}_1=\mathrm{diag}(\frac{1}{T}\mathbf{M}_1)$. Note that the diagonal entries of $\frac{1}{T}\mathbf{V}^{-1}$ are given by $\mathrm{diag}(\frac{1}{T}\mathbf{V}^{-1})=\{v_j\}_{j\le N}$. Therefore $\mathbf{D}_1$ is a diagonal matrix with entries $\{v_j\}_{j\le N-r_N}$, and $\max_j v_j=O(T^{-1})$.

Since $\boldsymbol{\beta}$ is $r_N \times (N - r_N)$, using the expression of $\mathbf{A}$, we have

$$\begin{aligned} T\widehat{\boldsymbol{\theta}^{'}}_1\mathbf{A}\widehat{\boldsymbol{\theta}_1} &\le \lambda_{N,T}\|\boldsymbol{\beta}\mathbf{M}_1^{-1}\widehat{\boldsymbol{\theta}_1}\|_{\max}^2 \\ &\le \lambda_{N,Tr_N}\|\mathbf{M}_1^{-1}(\widehat{\boldsymbol{\theta}_1}-\boldsymbol{\theta}_1)\|_{\max}^2(\max_{i\le r_N}\sum_{j\le N-r}|\beta_{ij}|)^2 \\ &\le \lambda_{N,Tr_N}\|\mathbf{M}_1^{-1}\mathbf{D}_1^{1/2}\|_1^2\|\mathbf{D}_1^{-1/2}(\widehat{\boldsymbol{\theta}_1}-\boldsymbol{\theta}_1)\|_{\max}^2\|\mathbf{V}^{-1}\|_1^2, \end{aligned}$$

where we used $\boldsymbol{\theta}_1 = 0$ in the second inequality and the fact that $\max_i \le r_N\sum_{j\le N-r}|\beta_{ij}| \le \|\mathbf{V}^{-1}\|_1$. Note that $\|\mathbf{V}\|_1=O(1)=\|\mathbf{V}^{-1}\|_1$. Hence,

$$\|\mathbf{M}_1^{-1}\mathrm{D}_1^{1/2}\|_1^2=O(T^{-1}),\ \text{ and }\ \lambda_{N,T}=O(T).$$

Thus, there is $C > 0$, with probability approaching one,

$$T\hat{\boldsymbol{\theta}}_1^{'}\mathbf{A}\widehat{\boldsymbol{\theta}_1} \le Cr_N\|\mathbf{D}_1^{-1/2}(\widehat{\boldsymbol{\theta}_1}-\boldsymbol{\theta}_1)\|_{\max}^2 \le Cr_N\delta_{N,T}^2.$$

Note that the uniform convergence in Assumption 3.1 and boundness of $\|\theta\|_{\max}$ imply that $P(\|\hat{\boldsymbol{\theta}}\|_{\max} \le C) \to 1$ for a sufficient large constant C. For $\mathbf{G} = (g_{ij})$, note that $\max_{i\le r}\sum_{j=1}^{N-r}|g_{ij}| \le \|\mathbf{V}\|_1$ Hence, by using $\boldsymbol{\theta}_1 = 0$ again, with probability tending to one,

$$\begin{aligned} |T\hat{\boldsymbol{\theta}}_2^{'}\mathbf{G}\widehat{\boldsymbol{\theta}_1}| &= T|\hat{\boldsymbol{\theta}}_2^{'}\mathbf{G}\mathbf{D}_1^{1/2}\mathbf{D}_1^{-1/2}(\widehat{\boldsymbol{\theta}_1}-\boldsymbol{\theta}_1)| \\ &\le T\|\widehat{\theta_2}\|_{\max}\|\mathbf{D}_1^{-1/2}(\widehat{\boldsymbol{\theta}_1}-\boldsymbol{\theta}_1)\|_{\max}\sum_{i=1}^{r_N}\sum_{j=1}^{N-r}|g_{ij}|\sqrt{v_j} \\ &\le Cr_N\delta_{N,T}\sqrt{T}. \end{aligned}$$

Moreover, $T\hat{\theta}_2^{'}\mathbf{C}\widehat{\boldsymbol{\theta}_2} \le T\|\widehat{\boldsymbol{\theta}_2}\|^2\|\mathbf{C}\|_2=O_P(r_NT)$. Combining all the results above, it yields that for any $\boldsymbol{\theta}\in\Theta_b$,

$$\Delta=O_P(r_N\delta_{N,T}^2+r_NT).$$

We denote $\mathrm{var}(\hat{\boldsymbol{\theta}})$, $\mathrm{var}(\widehat{\boldsymbol{\theta}_1})$, $\mathrm{var}(\widehat{\boldsymbol{\theta}_2})$ to be the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}_1}$ and $\widehat{\boldsymbol{\theta}_2}$. Then $\frac{1}{T}\mathbf{V}^{-1}=\mathrm{var}(\hat{\boldsymbol{\theta}})$ and $\frac{1}{T}\mathbf{M}_1=\mathrm{var}(\widehat{\boldsymbol{\theta}_1})$. It then follows from (3.3) that

$$Z \equiv \frac{T\hat{\theta}_1^{'}\mathbf{M}_1^{-1}\widehat{\boldsymbol{\theta}_1} - (N - r_N)}{\sqrt{2(N - r_N)}} \to^d \mathcal{N}(0,1).$$

For any $0 < \varepsilon < F_q$, define the event $A=\{|\Delta - r_N| < \sqrt{2N} \in\}$. Hence, suppressing the dependence of $\boldsymbol{\theta}$,

$$\begin{aligned} P(J_Q > F_q) &= P\left(\frac{T\hat{\theta}_1^{'}\mathbf{M}_1^{-1}\widehat{\boldsymbol{\theta}_1} + \Delta - N}{\sqrt{2N}} > F_q\right) \\ &= P\left(Z\sqrt{\frac{N-r_N}{N}} + \frac{\Delta - r_N}{\sqrt{2N}} > F_q\right) \\ &\le P\left(Z\sqrt{\frac{N-r_N}{N}} + \in > F_q\right) + P(A^c), \end{aligned}$$

which is further bounded by $1-\Phi(F_q-\varepsilon)+P(A^c)+o(1)$. Since $1-\Phi(F_q) = q$, for small enough, ε, $1-\Phi(F_q-\varepsilon) = q+O(\varepsilon)$. By letting $\varepsilon \to 0$ slower than $O\left(Tr_N/\sqrt{N}\right)$, we have $P(A^c) = o(1)$, and $\limsup_{N\to\infty,T\to\infty} P(J_Q > F_q) \le q$. On the other hand, $P(J_Q > F_q) \ge P(J_1 > F_q)$, which converges to q. This proves the result.

B Proofs for Section 4

Lemma B.1

When $\mathrm{cov}(\mathbf{f}_t)$ is positive definite, $E\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t < 1$.

Proof

If $E\mathbf{f}_t = 0$, then $E\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t < 1$. If $E\mathbf{f}_t \neq 0$, because $\mathrm{cov}(\mathbf{f}_t)$ is positive definite, let $\mathbf{c} = (E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t$, then $\mathbf{c}^{'}(E\mathbf{f}_t\mathbf{f}_t^{'} - E\mathbf{f}_tE\mathbf{f}_t^{'})\mathbf{c} > 0$. Hence $\mathbf{c}^{'}E\mathbf{f}_tE\mathbf{f}_t^{'}\mathbf{c} < \mathbf{c}^{'}E\mathbf{f}_t\mathbf{f}_t^{'}\mathbf{c}$ implies $E\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t > (E\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t)^2$. This implies $E\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t < 1$.

B.1 Proof of Proposition 4.1

Recall that $v_j = \mathrm{var}(u_{jt})/(T - TE\mathbf{f}_t^{'}(E\mathbf{f}_t\mathbf{f}_t^{'})^{-1}E\mathbf{f}_t)$, and $\widehat{v_j} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{jt}^2/(Ta_{f,T})$. Write $\sigma_{ij} = (\sum_u)_{ij}$, $\widehat{\sigma_{ij}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{u_{it}}\widehat{u_{jt}}$, $\sigma_j^2 = Tv_j$, and $\widehat{\sigma_j^2} = T\widehat{v_j}$.

Simple calculations yield

$$\widehat{\theta_i} = \theta_i + a_{f,T}^{-1}\frac{1}{T}\sum_{t=1}^{T}u_{it}(1 - \mathbf{f}_t^{'}\mathbf{w}).$$

We first prove the second statement. Note that there is $\sigma_{\min} > 0$ (independent of $\boldsymbol{\theta}$) so that $\min_j \sigma_j > \sigma_{\min}$. By Lemma ??, there is $C > 0$, $\inf_\Theta P(\max_{j\le N}|\widehat{\sigma_j} - \sigma_j|) < C\sqrt{\frac{logN}{T}}|\boldsymbol{\theta}) \to 1$. On the event $\{\max_{j\le N}|\hat{\sigma}_j - \sigma_j| < C\sqrt{\frac{logN}{T}}\}$,

$$\max_{j\le N}\left|\frac{\hat{v}_j^{1/2}}{v_j^{1/2}} - 1\right| \le \max_{j\le N}\frac{|\hat{\sigma}_j - \sigma_j|}{\sigma_j} \le \frac{C\sqrt{logN}}{\sigma_{\min}\sqrt{T}}.$$

This proves the second statement. We can now use this to prove the first statement.

Note that v_j is independent of $\boldsymbol{\theta}$, so there is C_1 (independent of $\boldsymbol{\theta}$) so that $\max_{j\le N} v_j^{-1/2} < C_1\sqrt{T}$. On the event

$$\{\max_{j\le N} v_j^{1/2}/\hat{v}_j^{1/2} < 2\} \cap \{\max_{j\le N}|\hat{\theta}_j - \theta_j| < C\sqrt{\frac{logN}{T}}\},$$

$$\max_{j\le N}\frac{|\widehat{\theta_j} - \theta_j|}{\hat{v}_j^{1/2}} \le C\sqrt{\frac{logN}{T}}2\max_j v_j^{-1/2} \le 2CC_1\sqrt{logN} < \delta_{N,T}.$$

The constants C, C_1 appeared are independent of $\boldsymbol{\theta}$, and Lemma **??** holds uniformly in $\boldsymbol{\theta}$. Hence the desired result also holds uniformly in $\boldsymbol{\theta}$.

B.2 Proof of Proposition 4.2

By Theorem 1 of Pesaran and Yamagata (2012) (Theorem 1),

$$(Ta_{f,T}\widehat{\boldsymbol{\theta}'}\sum_u^{-1}\hat{\boldsymbol{\theta}} - N)/\sqrt{2N} \to^d \mathcal{N}(0,1).$$

Therefore, we only need to show

$$\frac{T\widehat{\boldsymbol{\theta}'}(\sum_u^{-1} - \hat{\sum}_u^{-1})\hat{\boldsymbol{\theta}}}{\sqrt{2N}} = o_P(1).$$

The left hand side is equal to

$$\frac{T\widehat{\boldsymbol{\theta}'}\sum_u^{-1}(\hat{\sum}_u - \sum_u)\sum_u^{-1}\widehat{\boldsymbol{\theta}'}}{\sqrt{N}} + \frac{T\widehat{\boldsymbol{\theta}'}(\hat{\sum}_u^{-1} - \sum_u^{-1})(\hat{\sum}_u - \sum_u)\sum_u^{-1}\widehat{\boldsymbol{\theta}'}}{\sqrt{N}} \equiv a + b.$$

It was shown by Fan et al. (2011) that

$\|\hat{\sum}_u - \sum_u\|_2 = O_P(m_N\sqrt{\frac{logN}{T}}) = \|\hat{\sum}_u^{-1} - \sum_u^{-1}\|_2$ In addition, under H_0,

$\|\hat{\boldsymbol{\theta}}\|^2 = O_P(NlogN/T)$. Hence $b = O_p(\frac{m_N^2\sqrt{N}(logN)^2}{T}) = o_P(1)$.

The challenging part is to prove $a = o_P(1)$ when $N > T$. As is described in the main text, simple inequalities like Cauchy-Schwarz accumulate estimation errors, and hence do not work. Define $e_t = \sum_u^{-1}\mathbf{u}_t = (e_{1t}, \ldots, e_{Nt})'$, which is an N-dimensional vector with mean zero and covariance $\sum_u^{-1}$, whose entries are stochastically bounded. Let $\overline{\mathbf{w}} = (E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t$. A key step of proving this proposition is to establish the following two convergences:

$$\frac{1}{T}E\Big|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2 - Eu_{it}^2)\Big(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1 - \mathbf{f}'_s\overline{\mathbf{w}})\Big)^2\Big|^2 = o(1), \quad \text{(B.1)}$$

$$\frac{1}{T}E\Big|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})\Big[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1 - \mathbf{f}'_s\overline{\mathbf{w}})\Big]\Big[\frac{1}{\sqrt{T}}\sum_{k=1}^{T}e_{jk}(1 - \mathbf{f}'_s\overline{\mathbf{w}})\Big]\Big|^2 = o(1), \quad \text{(B.2)}$$

where

$$S_U = \{(i,j):(\sum_u)_{ij} \neq 0\}.$$

The sparsity condition assumes that most of the off-diagonal entries of Σ_u are outside of S_U. The above two convergences are weighted cross-sectional and serial double sums, where the weights satisfy $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_{it}(1 - \mathbf{f}_t'\overline{\mathbf{w}}) = O_P(1)$ for each i. The proofs of (B.1) and (B.2) are given in the supplementary material in Appendix D.

We consider the hard-thresholding covariance estimator. The proof for the generalized sparsity case as in Rothman et al. (2009) is very similar. Let $s_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}$ and $\sigma_{ij} = (\Sigma_u)_{ij}$. Under hard-thresholding,

$$\hat{\sigma}_{ij} = (\hat{\sum}_u)_{ij} = \begin{cases} s_{ii}, & \text{if } i = j, \\ s_{ij}, & \text{if } i \neq j, |s_{ij}| > C(s_{ii}s_{jj}\frac{logN}{T})^{1/2} \\ 0, & \text{if } i \neq j, |s_{ij}| \leq C(s_{ii}s_{jj}\frac{logN}{T})^{1/2} \end{cases}$$

Write $(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_i$ to denote the ith element of $\hat{\boldsymbol{\theta}}'\sum_u^{-1}$, and $S_U^c = \{(i,j):(\sum_u)_{ij} = 0\}$. For $\sigma_{ij} \equiv (\sum_u)_{ij}$, and $\hat{\sigma}_{ij} = (\hat{\sum}_u)_{ij}$ we have

$$a=\frac{T}{\sqrt{N}}\sum_{i=1}^{N}(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_i^2(\hat{\sigma}_{ii}-\sigma_{ii})+\frac{T}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_i(\hat{\sigma}_{ij}-\sigma_{ij})$$
$$+\frac{T}{\sqrt{N}}\sum_{(i,j)\in S_U^c}(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_i(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_j(\hat{\sigma}_{ij}-\sigma_{ij})$$
$$=a_1+a_2+a_3$$

We first examine a_3. Note that

$$a_3=\frac{T}{\sqrt{N}}\sum_{(i,j)\in S_U^c}(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_i(\hat{\boldsymbol{\theta}}'\sum_u^{-1})_j\hat{\sigma}_{ij}.$$

Obviously,

$$P(a_3>T^{-1})\leq P(\max_{(i,j)\in S_U^c}|\hat{\sigma}_{ij}|\neq 0)\leq P(\max_{(i,j)\in S_U^c}|s_{ij}|>C(s_{ii}s_{jj}\frac{logN}{T})^{1/2}).$$

Because s_{ii} is uniformly (across i) bounded away from zero with probability approaching one, and $\max_{(i,j)\in S_U^c}|s_{ij}|=O_P(\sqrt{\frac{logN}{T}})$. Hence for any ε > 0, when C in the threshold is large enough, $P(a_3 > T^{-1}) < \varepsilon$, this implies $a_3 = o_P(1)$.

The proof is finished once we establish $a_i = o_P(1)$ for $i = 1, 2$, which are given in Lemmas ?? and ?? respectively in the supplementary material.

Proof of Theorem 4.1

Part (i) follows from Proposition 4.2 and that $P(J_0 = 0|H_0) \to 1$. Part (ii) follows immediately from Theorem 3.3.

C Proofs for Section 5

C.1 Proof of Proposition 5.1

Lemma C.1

Under Assumption 5.1, uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $P(\sqrt{nT}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|<\sqrt{logn})\to 1$.

Proof—Note that

$$\sqrt{nT}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|=\|(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}')^{-1}(\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it})\|.$$

Uniformly for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, due to serial independence, and $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}E\tilde{x}_{it}'\tilde{x}_{it}E\tilde{u}_{it}\tilde{u}_{it}\leq C_1$,

$$E\|\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\mathrm{x}}_{it}\tilde{u}_{it}\|^{2}=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T}E\tilde{\mathrm{x}}_{it}^{'}\tilde{\mathrm{x}}_{js}\tilde{u}_{it}\tilde{u}_{js}$$
$$=\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}E\tilde{\mathrm{x}}_{it}^{'}\tilde{\mathrm{x}}_{it}E\tilde{u}_{it}\tilde{u}_{it}+\frac{1}{nT}\sum_{i\neq j}\sum_{t=1}^{T}E\tilde{\mathrm{x}}_{it}^{'}\tilde{\mathrm{x}}_{jt}E\tilde{u}_{it}\tilde{u}_{jt}$$
$$\leq C_{1}+\frac{1}{n}\sum_{i\neq j}|E\tilde{\mathrm{x}}_{it}^{'}\tilde{\mathrm{x}}_{jt}||E\tilde{u}_{it}\tilde{u}_{jt}|\leq C.$$

Hence the result follows from the Chebyshev inequality and that

$\lambda_{\min}(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{\mathrm{x}}_{it}\tilde{\mathrm{x}}_{it}^{'})$ is bounded away from zero with probability approaching one, uniformly in $\boldsymbol{\theta}$.

Lemma C.2

Suppose $\max_{j\leq n}\|\frac{1}{T}\sum_{t}\tilde{\mathrm{x}}_{jt}\tilde{\mathrm{x}}_{it}^{'}\|_{2}<C^{'}$ with probability approaching one and $E(u_{jt}^{4})<C^{'}$. There is $C>0$, so that uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$,

i. $P(\max_{j\leq n}|\frac{1}{T}\sum_{t=1}^{T}u_{jt}|<C\sqrt{logn/T})\rightarrow 1$

ii. $P(\max_{i,j\leq n}|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-Eu_{it}u_{jt}|<C\sqrt{logn/T})\rightarrow 1$

iii. $P(\max_{j\leq n}\frac{1}{T}\sum_{t=1}^{T}(u_{jt}-\hat{u}_{jt})^{2}<Clogn/T)\rightarrow 1$

iv. $P(\max_{i,j\leq n}|\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}-Eu_{it}u_{jt}|<C\sqrt{logn/T})\rightarrow 1$

Proof

i. By the Bernstein inequality, for $C=(8\max_{j\leq n}E(u_{jt}^{2}))^{1/2}$, we have

$$P(\max_{j\leq n}|\frac{1}{T}\sum_{t=1}^{T}u_{jt}|\geq C\sqrt{\frac{logn}{T}})\leq n\max_{j\leq n}P(|\frac{1}{T}\sum_{t=1}^{T}u_{jt}|\geq C\sqrt{\frac{logn}{T}})$$
$$\leq\exp(logn-\frac{C^{2}logn}{4\max_{j\leq n}E(u_{jt}^{2})})=\frac{1}{n}.$$

Hence (i) is proved as $P(\max_{j\leq n}|\frac{1}{T}\sum_{t=1}^{T}u_{jt}|<C\sqrt{logn/T})\geq 1-\frac{1}{n}$.

ii. For $C=(12\max_{j\leq n}E(u_{jt}^{4}))^{1/2}$, we have, uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$,

$$P(\max_{i,j\le n}|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-Eu_{it}u_{jt}|\ge C\sqrt{\frac{logn}{T}})$$
$$\le n^2\max_{i,j\le n}P(|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-Eu_{it}u_{jt}|\ge C\sqrt{\frac{logn}{T}})$$
$$\le \exp(2logn-\frac{C^2logn}{4\max_{j\le n}E(u_{jt}^4)})=\frac{1}{n}.$$

iii. Note that $\hat{u}_{jt}-u_{jt}=-\frac{1}{T}\sum_{t=1}^{T}u_{jt}-\tilde{\mathrm{x}}'_{jt}(\hat{\beta}-\beta)$, and $\max_{j\le n}\|\frac{1}{T}\sum_t\tilde{\mathrm{x}}_{jt}\tilde{\mathrm{x}}'_{jt}\|_2<C$ with probability approaching one. The result then follows from part (i) and Lemma C.1.

iv. Observe that

$$|\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}-Eu_{it}u_{jt}|\le|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-Eu_{it}u_{jt}|+|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-\hat{u}_{it}\hat{u}_{jt}|$$
$$\le|\frac{1}{T}\sum_{t=1}^{T}u_{it}u_{jt}-Eu_{it}u_{jt}|+\frac{1}{T}\sum_{t=1}^{T}(\hat{u}_{jt}-u_{jt})^2+(\frac{2}{T}\sum_t u_{jt}^2)^{1/2}(\frac{2}{T}\sum_t(\hat{u}_{jt}-u_{jt})^2)^{1/2}$$

The first two terms and $(\frac{2}{T}\sum_t(\hat{u}_{jt}-u_{jt})^2)^{1/2}$ in the third term are bounded by results in (ii) and (iii). Therefore, it suffices to show that there is a constant $M > 0$ so that

$$P(\max_{j\le n}\frac{1}{T}\sum_t u_{jt}^2<M)\to 1.$$

Note that $\max_{j\le n}\frac{1}{T}\sum_t u_{jt}^2\le\max_{j\le n}|\frac{1}{T}\sum_t u_{jt}^2-Eu_{jt}^2|+\max_{j\le n}Eu_{jt}^2$. In addition, by (ii), there is $C > 0$ so that

$$P(\max_{j\le n}|\frac{1}{T}\sum_{t=1}^{T}u_{jt}^2-Eu_{jt}^2|<C\sqrt{logn/T})\to 1.$$

Hence we can pick up M so that $M-\max_{j\le n}E(u_{jt}^2)>C\sqrt{logn/T}$, and

$$P(\max_{j\le n}\frac{1}{T}\sum_t u_{jt}^2\ge M)\le P(\max_{j\le n}|\frac{1}{T}\sum_t u_{jt}^2-Eu_{jt}^2|\ge M-\max_{j\le n}Eu_{jt}^2)$$
$$\le P(\max_{j\le n}|\frac{1}{T}\sum_t u_{jt}^2-Eu_{jt}^2|\ge C\sqrt{\frac{logn}{T}})\to 0.$$

This proves the desired result.

Lemma C.3

Under Assumption 5.1, there is $C > 0$, uniformly in $\boldsymbol{\theta} \in \Theta$,

$$P(\max_{ij}|\hat{\rho}_{ij} - \rho_{ij}| < C\sqrt{logn/T}) \to 1.$$

Proof—By the definition $\hat{\rho}_{ij} = (\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}^2)^{-1/2}(\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{jt}^2)^{-1/2}\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}$. By the triangular inequality,

$$|\hat{\rho}_{ij} - \rho_{ij}| \leq \underbrace{\frac{|\frac{1}{T}\sum_t \hat{u}_{it}\hat{u}_{jt} - u_{it}u_{jt}|}{(\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}^2)^{1/2}(\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{jt}^2)^{1/2}}}_{X_1}$$
$$+\underbrace{|\frac{1}{T}\sum_t u_{it}u_{jt}||(\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}^2\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{jt}^2)^{-1/2} - (\frac{1}{T}\sum_{t=1}^{T}u_{it}^2\frac{1}{T}\sum_{t=1}^{T}u_{it}^2)^{-1/2}|}_{X_2}$$

By part (iv) of Lemma C.2, $P(\max_{i,j\leq n}|\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt} - Eu_{it}u_{jt}| < C\sqrt{logn/T}) \to 1$.

Hence for sufficiently large $M > 0$ such that $\min_j E(u_{jt}^2) - C/M > C\sqrt{logn/T}$,

$$P(\max_{ij}|X_1| > M\sqrt{\tfrac{logn}{T}}) \leq P(\min_j \tfrac{1}{T}\sum_t \hat{u}_{jt}^2 < C/M) + o(1)$$
$$\leq P(\max_j|\tfrac{1}{T}\sum_t \hat{u}_{jt}^2 - Eu_{jt}^2| > \min_j Eu_{jt}^2 - C/M) + o(1) = o(1).$$

By a similar argument, there is $M' > 0$ so that $P(\max_{ij}|X_2| > M'\sqrt{\frac{logn}{T}}) = o(1)$. The result then follows as, uniformly in $\boldsymbol{\theta} \in \Theta$,

$$P(\max_{ij}|\hat{\rho}_{ij} - \rho_{ij}| \geq 2(M+M')\sqrt{logn/T})$$
$$\leq P(\max_{ij}|X_1| \geq (M+M')\sqrt{logn/T}) + P(\max_{ij}|X_2| \geq (M+M')\sqrt{logn/T}) = o(1).$$

Proof of Proposition 5.1

As $1 - \rho_{ij}^2 > 1 - c$ uniformly for (i, j) and $\boldsymbol{\theta}$, the second convergence follows from Lemma C. 3. Also, with probability approaching one,

$$\frac{|\hat{\rho}_{ij} - \rho_{ij}|}{\hat{v}_{ij}^{1/2}} \leq \frac{3\sqrt{T}}{2(1-c)}C\sqrt{\frac{logn}{T}} < \delta_{N,T}.$$

C.2 Proof of Theorem 5.1

Lemma C.4

J_1 has power uniformly on $\Theta(J_1)=\{\sum_{i<j}\rho_{ij}^2 \geq Cn^2 logn/T\}$ for some C. Proof. By Lemma C.3, there is $C > 0$, $\inf_{\theta\in\Theta}P(\max_{ij}|\hat{\rho}_{ij}-\rho_{ij}|<C\sqrt{logn/T}|\boldsymbol{\theta}) \to 1$. Let

$$A=\{\sum_{i<j}(\hat{\rho}_{ij}-\rho_{ij})^2<C^2n^2(logn/T)\}.$$

Then $\inf_\Theta P(A|\theta) \to 1$. On the event A, we have, uniformly in $\boldsymbol{\theta}=\{\rho_{ij}\}$,

$$\sum_{i<j}(\hat{\rho}_{ij}-\rho_{ij})\rho_{ij} \leq (\sum_{i<j}(\hat{\rho}_{ij}-\rho_{ij})^2)^{1/2}(\sum_{i<j}\rho_{ij}^2)^{1/2} \leq \frac{Cn\sqrt{logn}}{\sqrt{T}}(\sum_{i<j}\rho_{ij}^2)^{1/2}.$$

Therefore, when $\sum_{i<j}\rho_{ij}^2 \geq 16C^2n^2logn/T$,

$$\sum_{i<j}\hat{\rho}_{ij}^2=\sum_{i<j}(\hat{\rho}_{ij}-\rho_{ij})^2+\rho_{ij}^2+2(\hat{\rho}_{ij}-\rho_{ij})\rho_{ij} \geq \sum_{i<j}\rho_{ij}^2-\frac{2Cn\sqrt{logn}}{\sqrt{T}}(\sum_{i<j}\rho_{ij}^2)^{1/2} \geq \frac{1}{2}\sum_{i<j}\rho_{ij}^2.$$

This entails that when $\sum_{i<j}\rho_{ij}^2 \geq 16Cn^2logn/T$, we have

$$\sup_{\Theta(J_1)}P(J_1<F_q|\boldsymbol{\theta}) \leq \sup_{\Theta(J_1)}P(\sum_{i<j}\hat{\rho}_{ij}^2<\frac{n(n-1)}{2T}+(F_q+\frac{n}{2(T-1)})\frac{\sqrt{n(n-1)}}{T}|\boldsymbol{\theta})$$
$$\leq \sup_{\Theta(J_1)}P(\frac{1}{2}\sum_{i<j}\rho_{ij}^2<\frac{n(n-1)}{2T}+(F_q+\frac{n}{2(T-1)})\frac{\sqrt{n(n-1)}}{T}|\boldsymbol{\theta})+\sup_{\Theta(J_1)}P(A^c|\boldsymbol{\theta}) \to 0.$$

Proof of Theorem 5.1

It suffices to verify conditions (i)–(iii) of Theorem 3.2. Condition (i) follows from Theorem 1 of Baltagi et al. (2012). As for condition (ii), note that $J_1 \geq -\frac{\sqrt{n(n-1)}}{2}-\frac{n}{2(T-1)}$ almost surely. Hence as n, $T \to \infty$,

$$\inf_{\boldsymbol{\theta}\in\Theta_s}P(c\sqrt{N}+J_1>z_q|\boldsymbol{\theta}) \geq \inf_{\boldsymbol{\theta}\in\Theta_s}P(c\sqrt{N}-\frac{\sqrt{n(n-1)}}{2}-\frac{n}{2(T-1)}>z_q|\boldsymbol{\theta})=1.$$

Finally, condition (iii) follows from Lemma C.4.

References

Andrews D. Hypothesis testing with a restricted parameter space. Journal of Econometrics. 1998; 84:155–199.

Andrews D. Cross-sectional regression with common shocks. Econometrica. 2005; 73:1551–1585.

Ang A, Kristensen D. Testing conditional factor models. Journal of Financial Economics. 2012; 106:132–156.

Antoniadis A, Fan J. Regularized wavelet approximations. Journal of the American Statistical Association. 2001; 96:939–967.

Bai ZD, Saranadasa H. Effect of high dimension: by an example of a two sample problem. Statistica Sinica. 1996; 6:311–329.

Baltagi, B. Econometric Analysis of Panel Data. fourth. Wiley; 2008. edition ed.

Baltagi B, Feng Q, Kao C. A lagrange multiplier test for cross-sectional dependence in a fix effects panel data model. Journal of Econometrics. 2012; 170:164–177.

Beaulieu M, Dufour J, Khalaf L. Multivariate tests of mean-variance efficiency with possibly non-gaussian errors: an exact simulation based approach. Journal of Business and Economic Statistics. 2007; 25:398–410.

Bickel P, Levina E. Covariance regularization by thresholding. Annals of Statistics. 2008; 36:2577–2604.

Breusch T, Pagan A. The lagrange multiplier test and its application to model specification in econometrics. Review of Economic Studies. 1980; 47:239–254.

Cai T, Liu W, Xia Y. Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. Journal of the American Statistical Association. 2013; 108:265–277.

Cai T, Zhang C, Zhou H. Optimal rates of convergence for covariance matrix estimation. Annals of Statistics. 2010; 38:2118–2144.

Cai TT, Yuan M. Adaptive covariance matrix estimation through block thresholding. The Annals of Statistics. 2012; 40:2014–2042.

Carhart MM. On persistence in mutual fund performance. The Journal of finance. 1997; 52:57–82.

Chamberlain G, Rothschild M. Arbitrage, factor structure and mean-variance analyssi in large asset markets. Econometrica. 1983; 51:1305–1324.

Chen SX, Qin YL. A two-sample test for high-dimensional data with applications to gene-set testing. The Annals of Statistics. 2010; 38:808–835.

Chernozhukov, V.; Chetverikov, D.; Kato, K. Tech rep. MIT; 2013. Testing many moment inequalities.

Connor G. A unified beta pricing theory. Journal of Economic Theory. 1984; 34:13–31.

Connor G, Korajczyk R. A test for the number of factors in an approximate factor model. Journal of Finance. 1993; 48:1263–1291.

Donald SG, Imbens GW, Newey WK. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics. 2003; 117:55–93.

Fama E, French K. The cross-section of expected stock returns. Journal of Finance. 1992; 47:427–465.

Fan J. Test of significance based on wavelet thresholding and neyman's truncation. Journal of the American Statistical Association. 1996; 91:674–688.

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

Fan J, Liao Y, Mincheva M. High dimensional covariance matrix estimation in approximate factor models. Annals of Statistics. 2011; 39:3320–3356. [PubMed: 22661790]

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

Fan, J.; Liao, Y.; Yao, J. Tech rep. Princeton University; 2014. Power enhancement in high dimensional cross-sectional tests.

Gagliardini, P.; Ossola, E.; Scaillet, O. Tech rep. Swiss Finance Institute; 2011. Time-varying risk premium in large cross-sectional equidity datasets.

Gibbons M, Ross S, Shanken J. A test of the efficiency of a given portfolio. Econometrica. 1989; 57:1121–1152.

Hall P, Jin J. Innovated higher criticism for detecting sparse signals in correlated noise. The Annals of Statistics. 2010; 38:1686–1732.

Hansen LP, Richard SF. The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. Econometrica. 1987; 55:587–613.

Hansen, P. Tech rep. CREATES; 2003. Asymptotic tests of composite hypotheses.

Hansen P. A test for superior predictive ability. Journal of Business and Economic Statistics. 2005; 23:365–380.

Im K, Ahn S, Schmidt P, Wooldridge J. Efficient estimation of panel data models with strictly exogenous explanatory variables. Journal of Econometrics. 1999; 93:177–201.

MacKinlay A, Richardson M. Using generalized method of moments to test mean-variance efficiency. Journal of Finance. 1991; 46:511–527.

Pesaran H, Ullah A, Yamagata T. A bias-adjusted lm test of error cross section independence. Econometrics Journal. 2008; 11:105–127.

Pesaran, H.; Yamagata, T. Tech rep. University of South California; 2012. Testing capm with a large number of assets.

Ross S. The arbitrage theory of capital asset pricing. Journal of Economic Theory. 1976; 13:341–360.

Rothman A, Levina E, Zhu J. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association. 2009; 104:177–186.

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

Zhong P, Chen S, Xu M. Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. Annals of Statistics. 2013; 41:2820–2851.

Table I

Means and covariances used to generate $\mathbf{b}_i$ and $\mathbf{f}_t$

μ_B		Σ_B		μ_f		Σ_f	
0.9833	0.0921	−0.0178	0.0436	0.0260	3.2351	0.1783	0.7783
−0.1233	−0.0178	0.0862	−0.0211	0.0211	0.1783	0.5069	0.0102
0.0839	0.0436	−0.0211	0.7624	−0.0043	0.7783	0.0102	0.6586

Table II

Size and Power (%) of tests for simulated Fama-French three-factor model

T	N	J_{wald}	J_{thr}	$J_0 + J_{wald}$	$J_0 + J_{thr}$	$P(\hat{S}=\varnothing)$
				H_0		
300	500	5.0%	7.2%	5.2%	7.3%	99.9%
	800	5.4%	7.7%	5.6%	7.9%	99.8%
	1000	4.7%	7.8%	4.9%	8.0%	99.7%
	1200	4.7%	6.6%	4.8%	6.7%	99.8%
500	500	5.2%	6.4%	5.3%	6.5%	99.9%
	800	5.0%	5.8%	5.1%	5.9%	99.9%
	1000	5.4%	6.6%	5.5%	6.6%	100.0%
	1200	4.5%	7.4%	4.6%	7.5%	99.8%
				H_a^1		
300	500	48.3%	72.2%	93.7%	94.5%	8.0%
	800	58.4%	87.9%	97.8%	98.8%	3.2%
	1000	53.6%	87.0%	96.4%	98.1%	6.3%
	1200	66.4%	94.3%	97.9%	98.8%	3.4%
500	500	37.9%	54.4%	96.3%	96.7%	4.0%
	800	68.3%	91.6%	99.9%	99.9%	0.1%
	1000	63.3%	89.8%	99.8%	99.8%	0.2%
	1200	55.1%	88.0%	99.7%	99.6%	0.6%
				H_a^2		
300	500	68.6%	83.2%	72.7%	84.9%	80.0%
	800	69.0%	86.4%	72.1%	87.5%	80.9%
	1000	74.5%	91.7%	77.7%	92.1%	78.9%
	1200	74.9%	93.6%	78.5%	94.3%	79.6%
500	500	70.6%	81.9%	72.8%	82.8%	89.0%
	800	71.4%	86.6%	73.4%	87.1%	88.3%

T	N	J_{wald}	J_{thr}	$J_0 + J_{wald}$	$J_0 + J_{thr}$	$P(\hat{S}=\varnothing)$
	1000	72.2%	89.7%	73.5%	90.0%	89.6%
	1200	75.9%	92.3%	77.5%	92.5%	87.8%

Note: This table reports the frequencies (in percentage) of rejection and $\hat{S}=\varnothing$ based on 2000 replications. These tests are conducted at 5% significance level.

Table III

Size and power (%) of tests for cross-sectional independence

T	***n* = 200** $J_1/\mathrm{PE}/P(\hat{S}=\varnothing)$	***n* = 400** $J_1/\mathrm{PE}/P(\hat{S}=\varnothing)$	***n* = 600** $J_1/\mathrm{PE}/P(\hat{S}=\varnothing)$	***n* = 800** $J_1/\mathrm{PE}/P(\hat{S}=\varnothing)$
		H_0		
100	4.7/5.5 /99.1	4.9/5.3 /99.6	5.5/5.7 /99.7	4.9/5.2 /99.7
200	5.3/5.3 /100.0	5.5/5.9 /99.6	4.7/5.1 /99.4	4.9/5.1 /99.8
300	5.2/5.2 /100.0	5.2/5.2 /100.0	4.6/4.6 /100.0	4.9/4.9 /100.0
500	4.7/4.7 /100.0	5.5/5.5 /100.0	5.0/5.0 /100.0	5.1/5.1 /100.0
		H_a		
100	26.4/95.5 /5.0	19.8/98.0 /2.3	13.5/98.2 /2.0	12.2/99.2 /0.9
200	54.6/98.8 /1.6	40.3/99.6 /0.5	24.8/99.6 /0.4	21/99.7 /0.3
300	78.9/99.2 /1.1	65.3/100.0 /0.1	41.7/99.9 /0.2	37.2/100.0 /0.1
500	93.5/99.8 /0.2	89.0/100.0 /0.0	69.1/100.0 /0.0	61.8/100.0 /0.0

Note: This table reports the frequencies of rejection by J_1 in (5.2) and PE = J_0 + J_1 in (5.4) under the null and alternative hypotheses, based on 2000 replications. The frequency of $\hat{S}$ being empty is also recorded. These tests are conducted at 5% significance level.

Table IV

Summary of descriptive statistics and testing results

Variables	Mean	Std dev.	Median	Min	Max
N_τ	617.70	26.31	621	574	665
$\lvert\hat{\theta}\rvert_0$	5.20	3.50	5	0	20
$\overline{\lvert\hat{\theta}\rvert_i}^{\tau}(\%)$	0.9767	0.1519	0.9308	0.7835	1.3816
$\overline{\lvert\hat{\theta}\rvert_{i\in\hat{S}}}^{\tau}(\%)$	4.5569	1.4305	4.1549	1.7839	10.8393