Large Panel Test of Factor Pricing Models

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

[*]Department of Operations Research and Financial Engineering, Princeton University

[†] Bendheim Center for Finance, Princeton University

[‡] Department of Mathematics, University of Maryland

Abstract

We consider testing the mean-variance efficiency in the context of a high-dimensional multi-factor model, with the number of assets much larger than the time-series dimension. Most of the existing tests are based on a quadratic form of estimated alphas. Under high dimensionality, however, they all suffer from low powers because the accumulation of a large amount of estimation errors overrides the signals of the true nonzero alphas. To resolve this issue, we propose a new test that deals with high-dimensional hypothesis testing problems, called "power enhancement". A screening statistic is introduced to screen off most of the estimation errors and consistently select stocks with significant alphas. We develop a feasible standardized Wald statistic using a consistent estimator of the high-dimensional weight matrix based on thresholding. In addition, by attaching the screening statistic to the traditional quadratic-form tests, our proposed test significantly enhances the power of the Wald-type tests under most of the alternatives, while keeping a correct asymptotic size. Finally, the proposed methods are applied to the securities in the S&P 500 index as an empirical application. The empirical study shows that market inefficiency is primarily caused by a small portion of mispriced stocks, instead of aggregated alphas. Moreover, most of the significant alphas are due to extra returns (underpriced).

Keywords: high dimensionality, approximate factor model, test for mean-variance efficiency, power enhancement

[*]Address: Department of Operations Research and Financial Engineering, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: *jqfan@princeton.edu*, *yuanliao@umd.edu*, *jiaweiy@princeton.edu*.
The research was partially supported by DMS-1206464 and NIH R01GM100474-01, NIH R01-GM072611.

1 Introduction

One of the fundamental assumptions of the Arbitrage Pricing Theory (APT) developed by Ross (1976) is that asset returns follow a factor model structure. Assume the excessive return of an asset over the risk-free rate satisfies

$$y_{it} = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{1.1}$$

where $\mathbf{f}_t = (f_{1t}, ..., f_{Kt})'$ are the excessive returns of K factors, $\mathbf{b}_i = (b_{i1}, ..., b_{iK})'$ are unknown factor loadings, and u_{it} represents the idiosyncratic error. The key implication from the asset pricing theory is that all the elements of the intercept vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$ should be zero, known as "mean-variance efficiency". Testing

$$H_0 : \boldsymbol{\alpha} = 0$$

in the multi-factor model (1.1) is of crucial importance in many practical applications, including portfolio selection and fund evaluation. It also includes the test of the Capital Asset Pricing Model (CAPM) as a special case.

Factor models have wide impacts on both economics and finance. In classical factor analysis, the cross-sectional dimension N is assumed fixed and the idiosyncratic components are cross-sectionally uncorrelated (that is, the idiosyncratic covariance matrix is diagonal). These assumptions are no longer suitable for modern financial applications when data are often widely available for a large number of assets over a short time span. In addition, a large panel often introduces correlations among idiosyncratic errors, making a strict factor model in the traditional sense very restrictive. In this paper, we consider the problem of testing the mean-variance efficiency when N is relatively large compared to the time-series dimension T. It is also desirable to allow for an approximate factor structure in the sense of Chamberlain and Rothschild (1983), which admits cross-sectional correlations among the idiosyncratic components.

Most of the existing tests are based on the quadratic statistic $W = \widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}}$, where $\widehat{\boldsymbol{\alpha}}$ is the OLS estimator for $\boldsymbol{\alpha}$, and $\mathbf{V}$ is some positive definite matrix whose eigenvalues are stochastically bounded. The Wald statistic, for instance, takes the form $T\tau\widehat{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\widehat{\boldsymbol{\alpha}}$, where $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ is the estimated inverse of the error covariance, and τ is a positive scalar that depends on the factors only. Another prominent example is the test given by Gibbons, Ross and Shaken (1989, GRS test), who showed that the exact distribution of an adjusted Wald test is F-distribution under normal assumptions. These tests are conducted when N does not grow with T, and therefore rule out the possibility of a large panel. When N diverges, Pesaran and Yamagata (2012, PY test) obtained the asymptotic null distribution of the standardized quadratic form: $(W - EW)/\sqrt{\mathrm{var}(W)}$. Noting that the sample residual covariance is no longer invertible when $N > T$, they chose $\mathbf{V} = \mathrm{diag}(\widehat{\mathbf{\Sigma}}_u)^{-1}$, and showed that the standardized W is asymptotically normal. In addition to these work, Beaulieu et al. (2007, BDK test) developed a likelihood ratio test, and MacKinlay and Richardson (1991) studied a GMM test; both are based on the quadratic form as well.

In practice, the number of assets under consideration can be of thousands. But to prevent possible structural changes on the factor loadings and risks, a relatively short time series is mostly appropriate, which contains only hundreds of daily observations, or tens of monthly data. Such a high-dimension-low-sample-size context causes the tests based on the quadratic statistic to have a very low power and inconsistent against many common alternatives. To see this, we note that the rejection region for the quadratic statistic at significant level q takes the form

$$\widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}} > c_q$$

for some critical value c_q, regardless of the choice of $\mathbf{V}$ or whether standardization is applied. When the dimension of $\boldsymbol{\alpha}$ is large, a large critical value has to be used in order to correctly control the size. For example, in the ideal case when $\{\mathbf{u}_t\}_{t=1}^T$ are i.i.d. normally distributed and $\mathbf{V} = \mathrm{cov}(\widehat{\boldsymbol{\alpha}})^{-1}$, the critical value is proportional to $\chi^2_{N,q}/T$, which is of order N/T, and diverges when $T = o(N)$. Consequently, such a quadratic test is only consistent when $\|\boldsymbol{\alpha}\|^2$ is large enough under the alternative, and will have a low detection power whenever $\|\boldsymbol{\alpha}\|^2$ is either bounded or growing at a slow rate. This is especially the case when it is only a few significant alphas that arouse market inefficiency. Therefore, one of the fundamental difficulties of the high-dimensional test arises from the quadratic form $\widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}}$: it accumulates a huge amount of estimation errors, and loses power against many alternatives under which $\|\boldsymbol{\alpha}\|$ does not grow so rapidly. Our empirical study on the consitituents of the S&P 500 index confirms this issue.

In this paper, we introduce a new concept for high-dimensional testing problems called "power enhancement" (PEM), and develop a PEM test as

$$J = J_0 + J_1.$$

Here $J_0 \geq 0$ is a "sure-screening" statistic that serves as a power enhancement part, and J_1 can be any quadratic form based test that has a correct asymptotic size (e.g., GRS, PY, BDK). The sure-screening statistic is defined to be

$$J_0 = \widehat{\boldsymbol{\alpha}}'_{\widehat{S}}\mathbf{V}_{\widehat{S}}\widehat{\boldsymbol{\alpha}}_{\widehat{S}}$$

where $\widehat{\boldsymbol{\alpha}}_{\widehat{S}} = (\widehat{\alpha}_j : |\widehat{\alpha}_j| > \delta_T)$ is a subvector of $\widehat{\boldsymbol{\alpha}}$ screened out by a threshold δ_T, and $\mathbf{V}_{\widehat{S}}$ is a corresponding submatrix of certain weight $\mathbf{V}$. The threshold δ_T is chosen such that under the null hypothesis,

$$P(J_0 = 0|H_0) \to 1,$$

and that J_0 diverges when $\max_{j\leq N} |\alpha_j| > O(\frac{\log N}{T})$. Hence the asymptotic null distribution of the PEM test is completely determined by that of J_1, while the sure-screening part significantly enhances the power of J_1 under many mild alternatives in which $\|\boldsymbol{\alpha}\|^2$ may not be large. This includes the "sparse alternative" as an example, where most of the α_j's are either zero or nearly so, with only a small (compared to N/T) portion of α_j's standing out. Since $J \geq J_1$, the rejection region of J strictly contains that of J_1.

We also develop an operational Wald statistic even when $N > T$ and $\boldsymbol{\Sigma}_u = \mathrm{cov}(\mathbf{u}_t)$ is not a diagonal matrix. The statistic is based on a consistent sparse estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ for the inverse error covariance. We show that as $N, T \to \infty$ and N is possibly much larger than T, for some scalar $\tau > 0$,

$$\widetilde{J}_{sq} = \frac{T\tau\widehat{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0,1)$$

under the null hypothesis. This test takes into account the cross-sectional dependence among the idiosyncratic errors. Technically, in order to show that the effect of replacing $\boldsymbol{\Sigma}_u^{-1}$ with the sparse estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ is negligible, we need to establish, under H_0,

$$\frac{T\tau\widehat{\boldsymbol{\alpha}}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\widehat{\boldsymbol{\alpha}}}{\sqrt{2N}} = o_p(1). \tag{1.2}$$

Note that a simple inequality $|\widehat{\boldsymbol{\alpha}}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\widehat{\boldsymbol{\alpha}}| \leq \|\widehat{\boldsymbol{\alpha}}\|^2\|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\|$ would not work when $N > T$ because the estimation errors in $\|\widehat{\boldsymbol{\alpha}}\|^2$ accumulate in high dimensions. Instead, we have developed a new technical strategy to prove (1.2), which would be also potentially useful in high-dimensional inference using GMM methods when one needs to estimate the optimal weight matrix. We further take $J_1 = \widetilde{J}_{sq}$, and combine it with our sure-screening statistic J_0 to propose a power enhancement test, which is much more powerful than using $\widetilde{J}_{sq}$ itself, while maintaining the same asymptotic null distribution.

As a by-product, the sure-screening step also identifies the individual significant alphas which we show to be useful to detect market inefficiency in a real application. In contrast, most of the existing tests do not possess this feature.

The proposed methods are applied to the securities in the S&P 500 index as an empirical application. The empirical study shows that indeed, market inefficiency is primarily caused by a small portion of mispriced stocks instead of aggregated alphas. In addition, most of the significant alphas are due to extra returns (that is, due to a large $\min_{\alpha_j>0}\alpha_j$ instead of a large $\|\boldsymbol{\alpha}\|^2$). This is captured by the proposed PEM test.

The remaining of the paper is organized as follows. Section 2 sets up the preliminaries and discuss the limitations of traditional tests. Section 3 proposes the power enhancement method, derives the asymptotic behaviours of the sure-screening statistic and analyzes its performances under different alternatives. Section 4 combines the PEM with the standardized quadratic form. An improved quadratic test based on thresholding is considered in Section 5. Simulation results are presented in Section 6, along with an empirical application to the securities in the S&P 500 index in Section 7. Section 8 concludes. All the proofs are given in the appendix.

Throughout the paper, for a square matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent its minimum and maximum eigenvalues. Let $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_1$ denote its operator norm and l_1 norm respectively, defined by $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A})$ and $\max_i \sum_j |\mathbf{A}_{ij}|$. 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, $a_T \asymp b_T$ if there are constants $C_1, C_2 > 0$ so that $C_1 b_T \leq a_T \leq C_2 b_T$ for all large T. Finally, for a finite set S, we denote $|S|_0$ as the number of elements in S.

2 Factor models and traditional tests

Consider the following linear factor pricing model in a matrix form

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, ..., T, \tag{2.1}$$

where $\mathbf{y}_t = (y_{1t}, ..., y_{Nt})'$ is an $N \times 1$ vector of observed asset returns at time t, $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_N)'$ with $\mathbf{b}_i$ being a $K \times 1$ vector of loadings, $\mathbf{f}_t = (f_{1t}, ..., f_{Kt})'$ is a vector of common factors, $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})$ denotes the idiosyncratic component, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$. We set up the model to have an approximate factor structure as in Chamberlain and Rothschild (1983) where the idiosyncratic components are cross-sectionally correlated over i. Here both $\mathbf{y}_t$ and the common factors $\mathbf{f}_t$ are observable.

Our goal is to test the hypothesis that those alphas are jointly zero across the panel:

$$H_0 : \boldsymbol{\alpha} = \mathbf{0}. \tag{2.2}$$

We are particularly interested in a high-dimensional situation where N can be much larger than T. In practice, the number of assets under consideration can reach as many as thousands, but the observation period might be much smaller, because a long observation window is likely to introduce structural breaks in factor loadings. In addition, endogeneity problems arise when using a testing strategy based on portfolios instead of individual securities. As a result, a reliable test of the above factor model often entails a large panel of high dimensions relative to T.

Moreover, a large panel naturally imposes a sparsity assumption on the error covariance matrix $\mathbf{\Sigma}_u = \text{cov}(\mathbf{u}_t)$, that is, many of the off-diagonal elements are either zeros or close to zero. Since the common factors have substantially mitigated the co-movement across the whole panel, a particular asset's idiosyncratic volatility is usually correlated with no more than a few number 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)]. Such a sparse assumption will be used to reliably estimate the error covariance later.

2.1 Wald-type tests

If we further denote $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$; $\mathbf{1}$ as an $N \times 1$ vector consisting of ones, $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_T)'$ and $\mathbf{u}_i = (u_{i1}, ..., u_{iT})$, then model (2.1) can be written as

$$\mathbf{y}_i = \alpha_i \mathbf{1} + \mathbf{F}\mathbf{b_i} + \mathbf{u_i}. \tag{2.3}$$

The model is a seemingly unrelated regression model with common factors. We therefore can run the regression stock by stock according to (2.3). Define $K \times 1$ vectors $\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$, $\mathbf{w} = (\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t')^{-1}\bar{\mathbf{f}}$, and a scalar $\tau = 1 - \bar{\mathbf{f}}'\mathbf{w}$. Then the ordinary least squares (OLS) estimator yields

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1, ..., \widehat{\alpha}_N)', \quad \widehat{\alpha}_i = \frac{1}{\tau T}\sum_{t=1}^{T} y_{it}(1 - \mathbf{f}_t'\mathbf{w}). \tag{2.4}$$

Further calculations yield

$$\widehat{\alpha}_i = \alpha_i + \frac{1}{\tau T}\sum_{t=1}^{T} u_{it}(1 - \mathbf{f}_t'\mathbf{w}). \tag{2.5}$$

Assuming no serial correlation among $\mathbf{u}_t$, the conditional covariance of $\widehat{\boldsymbol{\alpha}}$ is $\mathbf{\Sigma}_u/(T\tau)$, given the factors. If $\mathbf{\Sigma}_u$ is known, a classical way to build the test statistic would be

$$W = T\tau\widehat{\boldsymbol{\alpha}}'\mathbf{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}. \tag{2.6}$$

When N is fixed, $\mathbf{\Sigma}_u^{-1}$ can be estimated and replaced by the inverse of the sample residual covariance matrix, and the resulting test statistic is then in line with the well-known test by Gibbons, Ross and Shaken (1989, GRS test). With the normality assumption, GRS obtained the exact finite sample distribution of this test statistic. Asymptotically, the test statistic becomes the Wald statistic which follows a χ^2-distribution under the null hypothesis.

When N grows, the traditional asymptotics for the null distribution of W does not apply. Instead, Pesaran and Yamagata (2012, PY test) developed an alternative asymptotics for the Wald statistic, based on the standardized version of W. More specifically, under some regularity conditions, they showed

$$J_1 = \frac{T\tau\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0,1). \tag{2.7}$$

as $N \to \infty$. Hence at the significant level $q \in (0,1)$, $P(J_1 > z_q|H_0) \to q$. Regardless of the type of asymptotics, we shall refer to the test based on W (with $\boldsymbol{\Sigma}_u^{-1}$ possibly replaced with an estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$) as Wald-type statistic, or quadratic test because W is a quadratic form of $\widehat{\boldsymbol{\alpha}}$.

2.2 Two main challenges

In a data-rich environment, the panel size N can be much larger than the number of observations T. Such a high dimensionality brings new challenges to the test statistics based on W, and the alternative asymptotics introduced by Pesaran and Yamagata (2012) only partially solves the problem. Specifically, there are two main challenges.

The first challenge arises from estimating $\boldsymbol{\Sigma}_u^{-1}$. It is well known that the sample residual covariance matrix becomes singular when $N > T$. Even if $N < T$, replacing $\boldsymbol{\Sigma}_u^{-1}$ in W with the inverse sample covariance can still bring a huge amount of estimation errors when N^2 is close to T. This can distort the null distribution of the test statistic.

Another challenge comes from the concern of the power. Even when $\boldsymbol{\Sigma}_u^{-1}$ is known so that W is directly feasible, a test statistic based on W has very low powers against various alternative hypotheses when N is large, still due to the accumulation of estimation errors. This can be illustrated in the following example.

Example 2.1. Suppose $T = O(N)$, and we want to test H_0 against finitely many nonzero α's:

$$H_a : \alpha_i = c_i, i \le r, \quad \alpha_i = 0, r < i \le N,$$

where r is fixed, and all the $|c_i|$'s are bounded away from zero. Assume $\boldsymbol{\Sigma}_u = \mathbf{I}_N$ to be known, and $\widehat{\boldsymbol{\alpha}}$ is the OLS estimator with uniform accuracy: $\max_{i \le N} |\hat{\alpha}_i - \alpha_i| = O_p(\sqrt{\frac{1}{T}})$. Under H_0, $\widehat{\boldsymbol{\alpha}}$ is a pure estimation noise, and

$$W = T\tau\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}|_{H_0} = O_p(T)\sum_{i=1}^{N}\widehat{\alpha}_i^2 = O_p(N).$$

Under H_a, each of the first r components in $\widehat{\boldsymbol{\alpha}}$ is stochastically close to c_i, whereas the estimation noises constitute the remaining $N-r$ components. Hence

$$W|_{H_a} = O_p(T\sum_{i=1}^{r} c_i^2 + (N-r)) = O_p(N).$$

Apparently, the Wald-type test has the same order under both H_0 and H_a, making it hard to distinguish H_0 from H_a. Let us now derive the critical value. Suppose $\{\mathbf{u}_t\}_{t=1}^T$ are i.i.d. normal with a known covariance $\mathbf{I}_N$. Then conditional on $\{\mathbf{f}_t\}_{t=1}^T$, $\sqrt{T}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u/\tau)$. Therefore, W rejects H_0 if $W > c_q = \chi^2_{N,q}/(T\tau)$. In addition, as $N \to \infty$,

$$\frac{W-N}{\sqrt{2N}} \to^d \mathcal{N}(0,1).$$

Hence an alternative critical value is $\tilde{c}_q = (\sqrt{2N}z_q + N)/(T\tau)$. Either way, the critical value $c_q \asymp O(N/T) \asymp \tilde{c}_q$. But in this example, $\|\boldsymbol{\alpha}\|^2 = o(N/T)$ under the alternative. So a test based on W hardly has power against H_a. □

In the previous example, there are only a small portion of nonzero alphas in the alternative, whose signals are dominated by the aggregated high-dimensional estimation errors: $T\sum_{i>r}\widehat{\alpha}_i^2$. Note that when (2.7) holds, we can reject H_0 as long as $J_1 > z_q$ at the significant level $1-q \in (0,1)$, where z_q is the critical value for the standard normal distribution. However, it can be shown that when $T = o(\sqrt{N})$, and there are only $r = o(\frac{\sqrt{N}}{T})$ nonzero alphas in the alternative, this test is not consistent. We formally present this result in the following theorem. For simplicity, we assume both $\|\boldsymbol{\Sigma}_u\|_1$ and $\|\boldsymbol{\Sigma}_u^{-1}\|_1$ to be bounded.

Theorem 2.1. *Suppose $T = o(\sqrt{N})$. Consider J_1 that satisfies (2.7). In addition, suppose Assumption 3.2 below holds, and both $\|\boldsymbol{\Sigma}_u\|_1$ and $\|\boldsymbol{\Sigma}_u^{-1}\|_1$ are bounded away from infinity. When $T = O(\sqrt{N})$, consider the following alternative:*

$$H_a: \text{ there are at most } r = o(\frac{\sqrt{N}}{T}) \text{ nonzero } \alpha_j\text{'s, which are also bounded away from infinity.}$$

Then under H_a, $P(J_1 > z_q|H_a) \leq 2q + o(1)$ for any $q \in (0, 0.5)$. Therefore the test based on J_1 is inconsistent.

A more sensible approach is to focus on those alternatives that may have only a few nonzero alphas compared to N, which are also of interest in practice. In what follows, we develop a new testing procedure that significantly improves the power of the Wald-type test against such alternatives.

3 Power Enhancement

Traditional tests of factor pricing models are especially directed against alternatives regarding portfolios, rather than individual assets. The power of the GRS test relies on the squared Sharpe ratio for the tangency portfolio, regardless of how individual assets behave under the alternatives. It turns out that even if some individual asset are either significantly overpriced or underpriced, their trivial contribution to the whole portfolio is not enough for making inferences.

We aim to build a test statistic that not only takes care of alternatives involving portfolios, but also deals with sparse alternatives where only a few assets have nonzero alphas. By doing so, we keep track of features of both portfolios and individual assets. For such purposes, we propose a class of test statistics that consist of two parts:

$$J = J_0 + J_1, \tag{3.1}$$

where $J_0 \geq 0$ is a proposed sure-screening statistic, designed to detect both sparse alternatives and significant individual alphas, and J_1 is based on some existing Wald-type statistic that mainly controls the size of the test. We reconcile the two parts so that the asymptotic size of the test is that of J_1 under the null, and the asymptotic power of the test is mainly driven by J_0. Because $J \geq J_1$ always holds, the power of J_1 is enhanced as a result.

3.1 Sure-screening statistic

For a given factor pricing model with a large panel, we divide market inefficiency into two forms according to different configurations of alphas.

1. *Average case:* The average of alphas deviates from zero.
2. *Sparse case:* A few alphas are significantly away from zero, while all the others are close to zero.

Note that the above two cases are not mutually exclusive. A few alphas with large absolute values would drive the average up to some extent, implying market inefficiency in an average sense. In turn, a large enough average indicates at least a few alphas are as large. The separation, however, gives us an insight on how to select stocks by telling us whether we should focus on a small portion of them or all of them. And if it is a small portion, it compels our interest to identify those stocks that lead to the market inefficiency. As we shall see, our proposed test $J_0 + J_1$ automatically identifies significantly mispriced stocks.

To prevent the accumulation of estimation errors in a large panel, we propose a sure-screening statistic. For some predetermined threshold value $\delta_T > 0$, define a screening set

$$\widehat{S} = \left\{ j : \frac{|\widehat{\alpha}_j|}{\widehat{\sigma}_j} > \delta_T, j = 1, ..., N \right\}, \tag{3.2}$$

where $\widehat{\alpha}_j$ is the OLS estimator and $\widehat{\sigma}_j = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{jt}^2/\sqrt{\tau}$ is the sample estimator of the variance of $\sqrt{T}\widehat{\alpha}_j$. Denote by

$$\widehat{\boldsymbol{\alpha}}_{\widehat{S}} = (\widehat{\alpha}_j : j \in \widehat{S}) = \begin{pmatrix} \widehat{\alpha}_1 I_{1\in\widehat{S}} \\ \vdots \\ \widehat{\alpha}_N I_{N\in\widehat{S}} \end{pmatrix} \tag{3.3}$$

the screened-out alpha estimators, which can be interpreted as rejecting the efficiency of the corresponding assets individually. Let $\widehat{\boldsymbol{\Sigma}}_u$ be a nonsingular estimator of $\boldsymbol{\Sigma}_u$, to be defined later. Let $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}}$ denote the submatrix of $\widehat{\boldsymbol{\Sigma}}_u$ formed by the rows and columns in $\{j : j \in \widehat{S}\}$, so that $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}}/(T\tau)$ is the estimated conditional covariance matrix of $\widehat{\boldsymbol{\alpha}}_{\widehat{S}}$ given the common factors.

With the notations above, we define our sure-screening statistic as

$$J_0 = T\tau \widehat{\boldsymbol{\alpha}}'_{\widehat{S}} \widehat{\boldsymbol{\Sigma}}_{\widehat{S}}^{-1} \widehat{\boldsymbol{\alpha}}_{\widehat{S}}. \tag{3.4}$$

The choice of δ_T must suppress most of the noises, resulting in an empty set of $\widehat{S}$ under the null hypothesis. On the other hand, δ_T cannot be too large to filter out important signals of alphas under the alternative. For this purpose, noting that the maximum noise level is $O_p(\sqrt{\log N/T})$, we let

$$\delta_T = (\log\log T)\sqrt{\frac{\log N}{T}}. \tag{3.5}$$

With this choice of δ_T, if we define, for $\sigma_j = \Sigma_{u,jj}/\sqrt{\tau}$,

$$S = \left\{ j : \frac{|\alpha_j|}{\sigma_j} > 2\delta_T, j = 1, ..., N \right\}, \tag{3.6}$$

then under mild regularity conditions, $P(S = \widehat{S}) \to 1$, and $\widehat{\boldsymbol{\alpha}}_{\widehat{S}}$ mimics $\boldsymbol{\alpha}_S = (\alpha_j : j \in S)$.

Therefore, under H_0, $\widehat{S}$ gets rid of most of the estimation noises, and it follows that $P(J_0 = 0|H_0) \to 1$. Under most of the alternatives, $\widehat{S}$ preserves the important individual alphas by mimicking the oracle set S. When $\|\boldsymbol{\alpha}_S\|^2 \gg T^{-1}$, J_0 is stochastically unbounded. Hence the null and alternative hypotheses are well distinguished by the asymptotic behaviors of J_0.

The sure-screening statistic depends on a nonsingular covariance matrix $\widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1}$ that mimics $\mathbf{\Sigma}_S^{-1}$. Derived from an estimate of $\mathbf{\Sigma}_u^{-1}$, $\widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1}$ is used for standardization in order for J_0 to attain a proper scale. The covariance $\mathbf{\Sigma}_S$ can be very large when the true S is large, and is rather difficult to estimate. To obtain an operational $\widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1}$, we assume $\mathbf{\Sigma}_u$ to be a sparse covariance and estimate it by thresholding. Let $\widehat{u}_{jt}$ be the residual from the OLS estimator. For $s_{ij} = \frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}$, let

$$(\widehat{\mathbf{\Sigma}}_u)_{ij} = \begin{cases} s_{ij}, & \text{if } i = j, \\ th(s_{ij}), & \text{if } i \neq j, \end{cases} \tag{3.7}$$

where $th(\cdot)$ is a thresholding function, with threshold value $h_{ij} = C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2}$ for some constant $C > 0$. When the hard-thresholding function is used, this is the estimator proposed by Fan et al. (2011). Many other thresholdings also apply, e.g., soft thresholding and SCAD (Fan and Li 2001). In general, $th(\cdot)$ should satisfy:

(i) $th(z) = 0$ if $|z| < h_{ij}$;

(ii) $|th(z) - z| \leq h_{ij}$.

(iii) There are constants $a > 0$ and $b > 1$ such that $|th(z) - z| \leq ah_{ij}^2$ if $|z| > bh_{ij}$.

We can also replace $\widehat{\mathbf{\Sigma}}_{\widehat{S}}$ with $\widehat{\mathbf{D}}_{\widehat{S}} = \text{diag}\{s_{jj} : j \in \widehat{S}\}$. This is particularly useful when $\mathbf{\Sigma}_u$ is not sparse. The sure-screening statistic is then defined as

$$J_0 = T\tau\widehat{\boldsymbol{\alpha}}_{\widehat{S}}\widehat{\mathbf{D}}_{\widehat{S}}^{-1}\widehat{\boldsymbol{\alpha}}_{\widehat{S}} = T\tau\sum_{j\in\widehat{S}}\widehat{\alpha}_j^2 s_{jj}^{-1}.$$

3.2 Power enhancement test

The screening statistic J_0 is powerful in detecting significant alphas. However, under the null hypothesis, $J_0 = 0$ with probability approaching one. Hence J_0 by itself cannot be used directly. We combine it with other standard test statistics in order to control the size of the test.

Suppose J_1 is some test statistic for $H_0 : \boldsymbol{\alpha} = 0$, and assume that there is $\Omega \subset \mathbb{R}^N$ such that J_1 *has power against* $\boldsymbol{\alpha} \in \Omega$, in the sense that the test based on J_1 is consistent against the alternative hypothesis $H_a : \boldsymbol{\alpha} \in \Omega$. Our test statistic is formally defined as:

$$J = J_0 + J_1.$$

Because the sure-screening statistic J_0 is zero under the null with probability approaching one, the null distribution of J is asymptotically determined by that of J_1. Let c_q be the qth quantile of J_1 under the null, that is, $P(J_2 \leq c_q|H_0) \geq 1-q$. Our proposed J-test then reject H_0 if $J > c_q$ under the significant level $1-q$. Hence adding J_1 provides us a non-degenerate null distribution, which is needed to control the size of the test.

On the other hand, as demonstrated by Example 2.1, traditional tests often suffer from low powers due to the accumulation of estimation errors under the high dimensionality. Including the term J_0 enhances the power of J_1 because $J \geq J_1$ always holds. In fact, we will show below that the new test based on J has power against

$$\Omega \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j\leq N} |\alpha_j| > 2\delta_T \min_{j\leq N} \sigma_j\}.$$

As a result, adding J_0 significantly enhances the power of J_1. We shall thus address our test based on J to be *power enhancement test* (PEM test).

To formally see the fact of power enhancement, we impose the following assumptions. First of all, we assume $\boldsymbol{\Sigma}_u$ to be a sparse matrix so that one can apply the thresholding method to consistently estimate it. The notion of *generalized sparsity* in Bickel and Levina (2008) is used: for some $q \in [0, 1)$, define

$$m_N = \max_{i\leq N} \sum_{j=1}^{N} |(\boldsymbol{\Sigma}_u)_{ij}|^q. \tag{3.8}$$

Under the following sparsity assumption, the thresholded covariance estimator $\widehat{\boldsymbol{\Sigma}}_u$ is positive definite, and consistently estimates $\boldsymbol{\Sigma}_u$ under the operator norm.

Assumption 3.1. *There is $q \in [0, 1)$, so that, for $\log N = o(T)$,*

$$m_N = o\left((\frac{T}{\log N})^{(1-q)/2} \right).$$

A special case of Assumption 3.1 occurs when m_T is bounded from above, in the case of block-diagonal matrix with finite block sizes. Sparsity is one of the commonly used assumptions on high-dimensional covariance matrix estimations. There have been extensive studies recently in the statistical literature on estimating a large sparse covariance matrix. We refer to El Karoui (2008), Bickel and Levina (2008), Lam and Fan (2009), Cai and Liu (2011), and the references therein.

Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras generated by $\{(\mathbf{f}_t, \mathbf{u}_t) : -\infty \leq t \leq 0\}$ and $\{(\mathbf{f}_t, \mathbf{u}_t) : T \leq t \leq \infty\}$ respectively. In addition, define the mixing coefficient

$$\alpha(T) = \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_T^\infty} |P(A)P(B) - P(AB)|. \tag{3.9}$$

Assumption 3.2. *(i)* $\{\mathbf{f}_t, \mathbf{u}_t\}_{t\geq 1}$ *is strictly stationary, and* $E\mathbf{u}_t = 0$ *and* $E\mathbf{u}_t\mathbf{f}_t = 0$*. In addition, for* $s \neq t$*,* $E\mathbf{u}_t\mathbf{u}_s' = 0$*, and* $E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t < 1$*.*
(ii) There exit constants $c_1, c_2 > 0$ *such that* $\max_{i\leq N}\|\mathbf{b}_i\| < c_2$*,*
$c_1 < \lambda_{\min}(\mathbf{\Sigma}_u) \leq \lambda_{\max}(\mathbf{\Sigma}_u) < c_2$*, and* $c_1 < \lambda_{\min}(\mathrm{cov}(\mathbf{f}_t)) \leq \lambda_{\max}(\mathrm{cov}(\mathbf{f}_t)) < c_2$*.*
(iii) Exponential tail: There exist $r_1, r_2 > 0$*, and* $b_1, b_2 > 0$*, such that for any* $s > 0$*,* $\max_{i\leq N} P(|u_{it}| > s) \leq \exp(-(s/b_1)^{r_1})$*,* $\quad \max_{i\leq K} P(|f_{it}| > s) \leq \exp(-(s/b_2)^{r_2})$*.*
(iv) Strong mixing: There exists $r_3 > 0$ *and* $C > 0$ *satisfying: for all* $T \in \mathbb{Z}^+$*,*

$$\alpha(T) \leq \exp(-CT^{r_3}).$$

These conditions are standard in the time series literature. In Condition (i), we require the idiosyncratic error $\mathbf{u}_t$ be serially uncorrelated across t. Under this condition, the conditional covariance of $\widehat{\boldsymbol{\alpha}}$ is $\mathbf{\Sigma}_u/(T\tau)$. Estimating $\mathbf{\Sigma}_u$ when $N > T$ is already challenging. When the serial correlation is present, the autocovariance of $\mathbf{u}_t$ would also be involved in the covariance of the OLS estimator for alphas, and needs be estimated. We rule out these autocovariance terms to simplify the technicalities, and our method can be extended to the case of serial correlation, with further sparsity conditions. On the other hand, we allow the factors to be weakly dependent via the strong mixing condition. Also, it is always true that $E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t \leq 1$. We rule out the equality to guarantee that the asymptotic variance of $\sqrt{T}\widehat{\alpha}_j$ does not degenerate for each j.

The following theorem quantifies the asymptotic behavior of the sure-screening statistic J_0, and provides sufficient conditions for the set consistency in selecting significant alphas. Recall that $\widehat{S}$ and S are defined in (3.2) and (3.6) respectively. Define

$$\Delta = \{j : \alpha_j \asymp \delta_T, j = 1, ..., N\}.$$

Theorem 3.1. *Suppose* $\log N = o(T)$*, Assumption 3.2 hold. As* $T, N \to \infty$*,*
(i)

$$P(S \subset \widehat{S}) \to 1, \quad P(\widehat{S} \setminus S \subset \Delta) \to 1.$$

(ii) Under the null hypothesis, $P(\widehat{S} = \emptyset) \to 1$*. Hence*

$$P(J_0 = 0|H_0) \to 1,$$

We are particularly interested in a type of alternative hypothesis that satisfy the following *grey area* condition.

Assumption 3.3 (**Grey area**). *The alternative hypothesis H_a satisfies:*

$$\Delta = \emptyset.$$

The grey area represents a class of alternatives that have no nonzero α_j's on the boundary of the screening set S. This condition is very weak because the chance of falling exactly at the boundary is very low. Intuitively speaking, when an α_j is on the boundary of the threshold, it is hard to decide whether to eliminate it from the screening step or not. According to Theorem 3.1, the difference between the set estimator $\widehat{S}$ and the oracle set S is contained in the grey area Δ with probability approaching one. So the grey area condition suffices to achieve the screening consistency: $P(S = \widehat{S}) \to 1$.

Corollary 3.1. *Suppose Assumption 3.3 holds. Under the assumptions of Theorem 3.1,*

$$P(\widehat{S} = S) \to 1.$$

Note that S ranks the importance of individual alphas under the alternative, representing the true set of significant alphas we wish to identify. Hence the sure-screening consistency as in Corollary 3.1 enables us to identify these significant alphas.

We now formally present the asymptotic behavior of the PEM test. A test is said to have power against a set $A \subset \mathbb{R}^N$ if $\forall \boldsymbol{\alpha} \in A$, the probability of rejection converges to 1.

Theorem 3.2. *Suppose* $\log N = o(T)$*, and Assumptions 3.1-3.2 hold. In addition, suppose there is a test J_1 whose rejection region takes the form $J_1 > C$, and a non-degenerate distribution F, so that under the null hypothesis $J_1 \to^d F$, and J_1 has power against $\Omega \subset \mathbb{R}^N$. Then the PEM test $J = J_0 + J_1$ satisfies:*
(i) Under H_0,

$$J \to^d F,$$

(ii) under H_a such that $\boldsymbol{\alpha}$ satisfies:

$$\boldsymbol{\alpha} \in \Omega \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j \le N} |\alpha_j| > 2\delta_T \min_{j \le N} \sigma_j\} \equiv \bar{\Omega},$$

the PEM test has power against any subset of $\bar{\Omega}$.

Let F_q denote the qth quantile of F. Part (i) of Theorem 3.2 shows that J_1 and J reject the null if $J > F_q$ and $J_1 > F_q$ respectively. It follows immediately from $J \ge J_1$ that $P(J > F_q) \ge P(J_1 > F_q)$, which means J at least has power against Ω. In addition, J_0 indeed plays the role of power enhancement, in the sense that (i) it does not affect the null distribution, and (ii) it significantly enhances the power of the test by broadening the range of $\boldsymbol{\alpha}$ for which the test is consistent. Note that the introduced restriction for the alternative:

$$\{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j\le N} |\alpha_j| > 2\delta_T \min_{j\le N} \sigma_j\}$$

requires that S be nonempty. This is a very weak restriction for the alternative. For instance as long as there is an alpha standing out under H_a, the nonemptyness of S is satisfied. Under the null, J_1 dominates J_0, and gives a correct size, while under the alternative, J_0 dominates J_1, and is stochastically unbounded. In addition, the sure-screening statistic automatically identifies all the significant alphas that are in $\widehat{S}$.

4 PEM for the Standardized quadratic test

In principle, any existing consistent test can serve as J_1. In this section, we consider a standardized quadratic test (Wald-type) recently developed by Pesaran and Yamagata (2012). Theorem 2.1 shows that this test by itself suffers from low powers under the high dimensionality. We will see that PEM significantly enhances its power.

4.1 Standardized quadratic statistic

The Wald-type statistic $T\tau\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}$ depends on a high-dimensional inverse covariance $\boldsymbol{\Sigma}_u^{-1}$. As described in Section 2.2, if $\boldsymbol{\Sigma}_u$ is not diagonal, estimating $\boldsymbol{\Sigma}_u^{-1}$ is a challenging problem when $N > T$. Alternatively, one can use only the diagonal entries of $\boldsymbol{\Sigma}_u^{-1}$, and consider the following quadratic form, which takes into account the error cross-sectional heteroskedasticity only: for $s_{ii} = \frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}^2$,

$$W = T\tau\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\alpha}}, \quad \widehat{\mathbf{D}} = \mathrm{diag}\{s_{11}, ..., s_{NN}\}.$$

Note that W is a sum of individual squared t-statistics, the number of which grows in our large panel settings. A standardized version is given by:

$$\frac{W - E(W)}{\sqrt{\mathrm{var}(W)}}. \tag{4.1}$$

One needs to calculate and estimate both $E(W)$ and $\mathrm{var}(W)$ under the null. As $\widehat{\mathbf{D}}$ ignores the off diagonal entries of $\boldsymbol{\Sigma}_u$, when $\boldsymbol{\Sigma}_u$ is indeed non-diagonal, W is no longer χ^2 even under the normal assumption. Hence the calculation is not a trivial task. Pesaran and Yamagata

(2012) showed that a feasible version of (4.1) is

$$J_{sq} = \frac{W - N}{\sqrt{2N(1+\xi_T)}}$$

where

$$\xi_T = \frac{1}{N}\sum_{i\neq j}\hat{\rho}_{ij}^2 I_{\hat{\rho}_{ij}^2 > a_T}, \quad a_T = \frac{1}{T}\Phi^{-1}(1-c/N)$$

for some $c \in (0, 0.5)$. Here $\hat{\rho}_{ij}$ denotes the sample correlation between u_{it} and u_{jt} based on the residuals, and $\Phi^{-1}(\cdot)$ denotes the inverse standard normal cumulative distribution function. Applying a linear-quadratic form central limit theorem (Kelejian and Prucha 2001) yields that J_{sq} is standard normal as $N, T \to \infty$. We call J_{sq} to be the *standardized quadratic* statistic. Pesaran and Yamagata (2012) also proposed a slightly different (but asymptotically equivalent) statistic that corrects the finite sample bias.

4.2 Combining sure-screening with standardized quadratic test

We now formally investigate the power of J_{sq} combined with the sure-screening statistic J_0, namely, the PEM test:

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

The size of the test statistic is controlled by J_{sq}, as J_0 is zero with probability approaching one under the null hypothesis. To appreciate the power enhancement to J_{sq}, note that $J \geq J_{sq}$, so the rejection region of J_0 contains that of J_{sq}. In fact, the rejection region is significantly enlarged, as J_0 diverges under various interesting alternatives, including the sparse alternative case in the sense that there are only a few $\alpha_i \neq 0$. The PEM test J combines the rejection region of both two statistics, achieving a much enhanced power, without sacrificing the good size of J_{sq}.

Formally, we have the following results concerning the size and the power of the PEM test.

Theorem 4.1. *Suppose* $\log N = o(T)$ *and Assumptions 3.1-3.2 hold. Also,* $\sum_{i\neq j}(\mathbf{\Sigma}_u)_{ij}^2 = O(N)$. *As* $N, T \to \infty$ *and* $N/T^3 \to 0$, *we have:*
(i) under the null hypothesis $H_0 : \boldsymbol{\alpha} = \mathbf{0}$,

$$J \to^d \mathcal{N}(0, 1),$$

(ii) under the alternative H_a such that

$$\boldsymbol{\alpha} \in \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\} \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j \leq N} |\alpha_j| > 2\delta_T \min_{j \leq N} (\boldsymbol{\Sigma}_u)_{jj}\} \equiv \bar{B},$$

the PEM test has power against any subset of $\bar{B}$.

We see that adding J_0 in the PEM test does not lose anything under the null asymptotically, and it significantly enhances the power of J_{sq} under many important alternatives. The set $\{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\}$ itself is the region of $\boldsymbol{\alpha}$ in the alternative that J_{sq} has power against. This is a very restrictive region when $N > T$. For instance, it rules out the type of alternative in which there are finitely many nonzero alphas. In contrast, set $\bar{B}$ is much more enlarged, and contains many interesting alternatives. We investigate a few examples below that fall within this category. In these examples, we denote $H_a \subset A$ to mean that the alternative set for $\boldsymbol{\alpha}$ belongs to a given set A.

Example 4.1 (Minimum alpha)**.** Consider an alternative such that the minimum nonzero alpha is not too small:

$$H_a \subset A \equiv \{\min_j \{|\alpha_j| \neq 0\} \gg \delta_T\}.$$

In this case, the oracle set S is not empty, and meanwhile there exists no nonzero alphas that are at or below the level δ_T. This set is a subset of $\bar{B}$ in Theorem 4.1, and hence the PEM test J asymptotically has power against this alternative. Note that we allow $\min\{|\alpha_j|, j = 1, ..., N\}$ to vanish in the limit. A special case is the sparse alternative:

$$H_a : \min_{j \leq r} \{|\alpha_j|\} \gg \delta_T, \quad \alpha_j = 0, \text{ when } j > r$$

where r is fixed. In this case, there are only finitely many significant alphas. The PEM still has power against it (the power approaches one). However, J_{sq} itself has very low power because $\sum_i \alpha_i^2 = O(1)$ under the alternative. In fact, as long as r grows slowly compared to N, the power of any test based on $\widehat{\boldsymbol{\alpha}}' \mathbf{V} \widehat{\boldsymbol{\alpha}}$ with $\mathbf{V}$ being positive definite does not converge to one. □

Example 4.2 (Empty grey area)**.** Consider an alternative

$$H_a \subset \{j : |\alpha_j| \asymp \delta_T\} = \emptyset, \text{ and } \{j : |\alpha_j| \gg \delta_T\} \neq \emptyset.$$

Under this alternative, the grey area Δ is empty and there exists at least one significant alpha. We allow the existence of very small but nonzero alphas, that is, $\{|\alpha_j| \neq 0, |\alpha_j| = o(\delta_T)\} \neq \emptyset$, regardless of the number of them. The PEM test is still consistent by Theorem 4.1. In contrast, if the number of significant alphas is not large enough, J_{sq} itself still cannot detect H_a. $\square$

An important feature of the above alternatives is that we do not require $\|\boldsymbol{\alpha}\|^2$ under the alternative to be very large for the test consistency. In contrast, we only require a few alphas stand out. In addition, besides the enhanced power, the PEM test is able to identify all the significant alphas via $\widehat{S}$ if the grey area condition holds ($S = \widehat{S}$ with probability approaching one). Without screening, traditional tests such as GRS test and J_{sq} alone cannot detect these alternatives, especially when the panel is very large.

When there are no estimated alphas above the level δ_T under the alternative, $\widehat{S}$ can be empty, leading to $J_0 = 0$. However, due to the component J_{sq} in J, we may still achieve the consistency. This is illustrated in the following example.

Example 4.3. Consider the following alternative:

$$H_a \subset \max_{j \leq N} |\alpha_j| = O(\delta_T), \text{ and } \|\boldsymbol{\alpha}\|^2 \geq N(\frac{\log N}{T})^{1-c},$$

for $c \in (0, 1)$. Under such alternatives, a typical alpha lies in the interval $\sqrt{\frac{\log N}{T}}[1, \log \log T]$, and the number of nonzero alphas is not small. In this case, the probability that the sure-screening set $\widehat{S}$ is empty might be positive. We therefore expect $P(J_0 = 0) > 0$. However, since there are so many small but nonzero alphas, these aggregated alphas lead to a not-too-small $\sum_{j=1}^N \alpha_j^2$. As a result, J_{sq} stands out and gains power, making J still has power converging to one. Indeed, our simulated results demonstrate that the PEM test has a good power in this case (see Section 6). $\square$

Finally, when both the maximum magnitude of alphas and $\|\boldsymbol{\alpha}\|^2$ are small, the PEM test statistic has a low power. But the market can be considered approximately efficient in this situation.

5 An improved quadratic test based on thresholding

So far we have been using a diagonal weight matrix $\widehat{\mathbf{D}}^{-1}$ for the standardized quadratic test to combine with J_0, because the true $\boldsymbol{\Sigma}_u^{-1}$ is unknown and non-diagonal. On the other hand, this loses powers against certain local alternatives, because as long as $\boldsymbol{\Sigma}_u$ is not diagonal, using a diagonal weight matrix ignores the cross-sectional correlations among the error components. A better PEM should be:

$$J_0 + \frac{T\tau\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}}. \tag{5.1}$$

When $\mathbf{\Sigma}_u^{-1}$ is known, Pesaran and Yamagata (2012) showed that the above statistic without J_0 is asymptotically $\mathcal{N}(0,1)$ under the null. Hence, J_{sq} can be improved by replacing $\widehat{\mathbf{D}}^{-1}$ with a consistent estimator for $\mathbf{\Sigma}_u^{-1}$. In this section, we show that this indeed can be done by applying the thresholded covariance estimator under the sparsity condition of $\mathbf{\Sigma}_u$. We require $N \log N = o(T^2)$, but still allow N to be much larger than T.

5.1 A technical challenge

When $\mathbf{\Sigma}_u$ is a sparse matrix, Fan et al. (2011) obtained a thresholded estimator $\widehat{\mathbf{\Sigma}}_u$ as described in Section 3.1, which satisfies:

$$\|\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\| = O_p(m_N \sqrt{\frac{\log N}{T}}). \tag{5.2}$$

Here $m_N = \max_{i \leq N} \sum_{j=1}^{N} |(\mathbf{\Sigma}_u)_{ij}|$ is assumed to be either growing slowly with N or even bounded. For instance, m_N is bounded if $\mathbf{\Sigma}_u$ is a block-diagonal matrix with bounded block sizes, which is the case when the idiosyncratic noises are uncorrelated across industries. Note that the above convergence achieves the minimax optimal rate for sparse covariance estimation, as shown by Cai and Zhou (2012). However, it comes with a technical challenge when $N > T$, which we now explain.

When replacing $\mathbf{\Sigma}_u^{-1}$ in (5.1) with $\widehat{\mathbf{\Sigma}}_u^{-1}$, one needs to show that the effect of such a replacement is asymptotically negligible, namely, under H_0,

$$\frac{T\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}}{\sqrt{N}} = o_p(1). \tag{5.3}$$

Note that when $\boldsymbol{\alpha} = 0$, $\|\widehat{\boldsymbol{\alpha}}\|^2 = O_p(N(\log N)/T)$ (or $O_p(N/T)$ with a more careful analysis). A simple application of (5.2) yields

$$\frac{|T\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}|}{\sqrt{N}} \leq \frac{T\|\widehat{\boldsymbol{\alpha}}\|^2\|\widehat{\mathbf{\Sigma}}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\|}{\sqrt{N}} = O_p(m_N \log N \sqrt{\frac{N \log N}{T}}).$$

We see that even if m_N is bounded, it still requires $N \log N = o(T)$.

However, the above derivation uses a very crude bound $|\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}| \leq \|\widehat{\boldsymbol{\alpha}}\|^2\|\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\|$, which accumulates the estimation errors in $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2$ under a large N. In fact, $\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}$ is a weighted estimation error of $\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}$, where the weights $\widehat{\boldsymbol{\alpha}}$ help to reduce the curse of dimensionality, and should result in an improved rate of convergence. The formalization of this argument requires further regularity conditions, which we shall present in the following subsection.

5.2 Assumptions

In order to show that the effect of replacing $\mathbf{\Sigma}_u^{-1}$ with its consistent estimator is asymptotically negligible, especially when $N > T$, further assumptions are needed, but they are still quite reasonable.

First of all, we need to refine the sparsity condition on $\mathbf{\Sigma}_u$, which is similar to that of Lam and Fan (2009). Let S_L and S_U denote two disjoint sets and respectively include the indices of small and large entries of $\mathbf{\Sigma}_u$, and

$$\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U. \tag{5.4}$$

We assume that most of the indices (i,j) belong to S_L when $i \neq j$. For the banded matrix as an example, $(\mathbf{\Sigma}_u)_{ij} \neq 0$ if $|i-j| \leq k$; $(\mathbf{\Sigma}_u)_{ij} = 0$ if $|i-j| > k$ for some fixed $k \geq 1$. Then $S_L = \{(i,j) : |i-j| > k\}$ and $S_U = \{(i,j) : |i-j| \leq k\}$.

Formally, we assume:

Assumption 5.1. *There is a partition* $\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U$ *such that* $\sum_{i\neq j,(i,j)\in S_U} 1 = O(N)$ *and* $\sum_{(i,j)\in S_L} |(\mathbf{\Sigma}_u)_{ij}| = O(1)$. *In addition,*

$$\max_{(i,j)\in S_L} |(\mathbf{\Sigma}_u)_{ij}| \ll \sqrt{\frac{\log N}{T}} \ll \min_{(i,j)\in S_U} |(\mathbf{\Sigma}_u)_{ij}|.$$

We require the elements in S_L and S_U be well-separable. For example, if $\mathbf{\Sigma}_u$ is a block covariance matrix with finite block sizes, this assumption is naturally satisfied as long as the signal is not too-weak (that is, $\sqrt{\frac{\log N}{T}} = o(\min_{(i,j)\in S_U} |(\mathbf{\Sigma}_u)_{ij}|)$). The partition $\{(i,j) : i \leq N, j \leq N\} = S_L \cup S_U$ may not be unique. Most importantly, we do not need to know either S_L or S_U; hence the block size, the banding length, or the locations of the zero entries can be completely unknown. Our analysis suffices as long as such a partition exists.

To introduce our next assumption, define $\boldsymbol{\xi}_t = \mathbf{\Sigma}_u^{-1}\mathbf{u}_t = (\xi_{1t}, ..., \xi_{Nt})'$, which is an N-dimensional vector with mean zero and covariance $\mathbf{\Sigma}_u^{-1}$, whose entries are stochastically bounded. Since $E\mathbf{u}_t\mathbf{f}_t' = 0$ and $E\mathbf{u}_t = 0$, we have $\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_{it} = O_p(1)$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_{it}\mathbf{f}_t = O_p(1)$ for each $i \leq N$. We assume the following:

Assumption 5.2. *Let* $\bar{\mathbf{w}} = (E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t$, *then*

(i)

$$\frac{1}{T}E\left|\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T(u_{it}^2 - Eu_{it}^2)\left(\frac{1}{\sqrt{T}}\sum_{s=1}^T \xi_{is}(1-\mathbf{f}_s'\mathbf{w})\right)^2\right|^2 = o(1)$$

(ii)

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

Note that in the literature of high-dimensional panels and factor analysis (e.g., Stock and Watson 2002, Bai 2003, 2009), it is usually assumed that the cross-sectional and serial double sum is bounded: $E|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2-Eu_{it}^2)|^2=O(1)$, which is usually guaranteed by the central limit theorem across both i and t. Here, Condition (i) of Assumption 5.2 is with respect to the weighted double sums, where the weight $(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\xi_{is}(1-\mathbf{f}_s'\bar{\mathbf{w}}))^2$ is stochastically bounded because both $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{it}$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{it}\mathbf{f}_t$ are asymptotically normal. Condition (ii) is new in the literature because it is related to the sparsity condition. It is however very similar to (i) in that the index set of the cross-sectional sum is changed from $\{(i,j):i=j\}$ to $\{(i,j):i\neq j,(i,j)\in S_U\}$, where S_U is defined as the set of "big entry" indices. Recall that Assumption 5.1 assumes $\sum_{i\neq j,(i,j)\in S_U}1=O(N)$. So this condition is still reasonable for sparse enough covariances.

Primitive conditions for Assumption 5.2 will be provided in Section 5.4.

5.3 PEM test with improved standardized quadratic form

With the help of Assumptions 5.1 and 5.2, we show in the appendix (Proposition A.1) that (5.3) indeed holds. As a result, the effect of replacing $\mathbf{\Sigma}_u^{-1}$ with its consistent thresholded estimator is asymptotically negligible even if $N>T$. Now define a new PEM test with an improved standardized quadratic form:

$$\widetilde{J}=J_0+\widetilde{J}_{sq}$$

where, with $\widehat{\mathbf{\Sigma}}_u^{-1}$ defined in (3.7) as in Fan et al. (2011),

$$\widetilde{J}_{sq}=\frac{T\tau\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\boldsymbol{\alpha}}-N}{\sqrt{2N}}.$$

We have the following theorem.

Theorem 5.1. *Suppose $m_N^4(\log N)^4N=o(T^2)$, and Assumptions 3.1, 3.2, 5.1, 5.2 hold. Then*
(i) under the null hypothesis $H_0:\boldsymbol{\alpha}=\mathbf{0}$,

$$\widetilde{J}\to^d\mathcal{N}(0,1),\quad \widetilde{J}_{sq}\to^d\mathcal{N}(0,1),$$

(ii) under the alternative H_a such that

$$\boldsymbol{\alpha} \in \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\} \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j \le N} |\alpha_j| > 2\delta_T \min_{j \le N} (\boldsymbol{\Sigma}_u)_{jj}\} \equiv \bar{B},$$

the PEM test $\widetilde{J}$ has power against any subset of $\bar{B}$.

Note that when m_N is bounded, we then only require $N(\log N)^4 = o(T^2)$. Therefore N is allowed to be much larger than T.

5.4 Sufficient conditions for Assumption 5.2

As a simple example, Assumption 5.2 is satisfied if u_{it} is i.i.d. across both i and t.

Example 5.1. As an example, Assumption 5.2 can be simply verified if u_{it} is i.i.d. across both i and t and $N = o(T^2)$, but still, N can be larger than T. Write $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T \xi_{is}(1 - \mathbf{f}_s'\mathbf{w}))^2$. Let $e_{is} = \xi_{is}(1 - \mathbf{f}_s'\mathbf{w})$. When u_{it} is i.i.d. across both i and t,

$$\mathrm{var}(X) = \mathrm{var}[\frac{1}{\sqrt{T}} \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is})^2] \le E[\frac{1}{\sqrt{T}} \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is})^2]^2.$$

It is bounded by $\{E[\frac{1}{\sqrt{T}} \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)]^4\}^{1/2} \{E[\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is}]^8\}^{1/2}$, by the Cauchy-Schwarz inequality, which is $O(1)$ by the central limit theorem. On the other hand,

$$\begin{aligned} EX &= \sqrt{N} E[\frac{1}{\sqrt{T}} \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is})^2] = \sqrt{N} \mathrm{cov}(\frac{1}{\sqrt{T}} \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2), (\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is})^2) \\ &= \frac{\sqrt{N}}{T\sqrt{T}} (\sum_{t=1}^T \mathrm{cov}(u_{it}^2, e_{it}^2) + 2 \sum_{t=1}^T \sum_{s \ne t} \mathrm{cov}(u_{it}^2 - Eu_{it}^2, e_{is}e_{it}) + \sum_{t=1}^T \sum_{s,k \ne t} \mathrm{cov}(u_{it}^2, e_{is}e_{ik})). \end{aligned}$$

The second term on the right is zero because when $s \ne t$, $\mathrm{cov}(u_{it}^2 - Eu_{it}^2, e_{is}e_{it}) = E[(u_{it}^2 - Eu_{it}^2)e_{is}e_{it}] = E[(u_{it}^2 - Eu_{it}^2)e_{it}]E(e_{is}) = 0$. The third term is also zero because $\mathbf{u}_t$ is serially independent. Thus $EX = \frac{\sqrt{N}}{\sqrt{T}} \mathrm{cov}(u_{it}^2, e_{it}^2)$. Together, we have $\frac{1}{T} EX^2 = \frac{1}{T} \mathrm{var}(X) + \frac{1}{T}(EX)^2 = o(1) + o(\frac{N}{T^2}) = o(1)$ as long as $N = o(T^2)$. Condition (ii) is naturally satisfied because $\{(i,j) : i \ne j\} = \emptyset$ so the sums are zero. $\square$

When cross-sectional heteroskedasticity and correlations are present, Assumption 5.2 can still be verified when $\boldsymbol{\Sigma}_u$ is sparse enough. For simplicity, we assume $\boldsymbol{\Sigma}_u$ to be strictly sparse, in the sense that all its small entries are zero: $(\boldsymbol{\Sigma}_u)_{ij} = 0$ for $(i,j) \in S_L$. In addition, it is assumed that both $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_u^{-1}$ have bounded row sums. These are stated in the assumption as follows.

Assumption 5.3. *(i)* $(\mathbf{\Sigma}_u)_{ij} = 0$ *for all* $(i, j) \in S_L$*, where* S_L *is defined in Assumption 5.1. (ii) There is a constant* $C > 0$ *such that* $\|\mathbf{\Sigma}_u\|_1 + \|\mathbf{\Sigma}_u^{-1}\|_1 < C$*.*

Let

$$m_N = \max_{i\leq N} \sum_{j=1}^N I_{(\mathbf{\Sigma}_u)_{ij}\neq 0}, \quad D_N = \sum_{i,j\leq N} I_{i\neq j,(\mathbf{\Sigma}_u)_{ij}\neq 0} = \sum_{i\neq j,(i,j)\in S_U} 1.$$

Here m_N represents the maximum number of nonzeros in each row, corresponding to $q = 0$ in (3.8), and D_N represents the number of nonzero off-diagonal entries. We consider two kinds of sparse matrices, and verify Assumption 5.2 in both cases. In the first case, $\mathbf{\Sigma}_u$ is required to have no more than $O(\sqrt{N})$ off-diagonal nonzero entries, but allows a growing m_N; in the second case, m_N should be bounded, but $\mathbf{\Sigma}_u$ can have $O(N)$ off-diagonal nonzero entries. The latter allows block-diagonal matrices with finite size of blocks. This is particularly useful when firms' individual shocks are correlated only within industries but not across industries. Formally, we assume:

Assumption 5.4. *One of the following cases holds:*
(i) $D_N = O(\sqrt{N})$*;*
(ii) $D_N = O(N)$*, and* $m_N = O(1)$*.*

The following lemma shows that Assumption 5.2 can be verified under the required sparsity assumptions when $N > T$. For simplicity, we focus on the Gaussian errors and serially independent time series.

Lemma 5.1. *Suppose* $\mathbf{u}_t \sim \mathcal{N}(0, \mathbf{\Sigma}_u)$*, where* $\mathbf{\Sigma}_u$ *satisfies Assumptions 5.3 and 5.4. In addition, the sequence* $\{\mathbf{u}_t, \mathbf{f}_t\}_{t\leq T}$ *is independent across* t*, and satisfy Assumption 3.2 (i)-(iii);* $\mathbf{u}_t$ *and* $\mathbf{f}_t$ *are also independent. Then as* $N, T \to \infty$*,* $N \log N = o(T^2)$*, Assumption 5.2 is satisfied.*

6 Monte Carlo Experiments

We examine the power enhancement via several numerical examples. Excess returns are assumed to follow the three-factor model by Fama and French (1992):

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

6.1 Simulation

We simulate $\{\mathbf{b}_i\}_{i=1}^N$, $\{\mathbf{f}_t\}_{t=1}^T$ and $\{\mathbf{u}_t\}_{t=1}^T$ independently from $\mathcal{N}_3(\boldsymbol{\mu}_B, \mathbf{\Sigma}_B)$, $\mathcal{N}_3(\boldsymbol{\mu}_f, \mathbf{\Sigma}_f)$, and $\mathcal{N}_N(0, \mathbf{\Sigma}_u)$ respectively. The same parameters as in the simulations of Fan et al. (2013)

are used, which are calibrated using the data on daily returns of S&P 500's top 100 constituents, for the period from July 1^{st}, 2008 to June 29^{th} 2012. These parameters are listed in the following table.

$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_B$			$\boldsymbol{\mu}_f$	$\boldsymbol{\Sigma}_f$		
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 1: Means and covariances used to generate $\mathbf{b}_i$ and $\mathbf{f}_t$

Two types of $\boldsymbol{\Sigma}_u$ are considered.

Diagonal $\boldsymbol{\Sigma}_u^{(1)}$ is a diagonal matrix with diagonal entries $(\boldsymbol{\Sigma}_u)_{ii} = 1 + \|\mathbf{v}_i\|^2$, where $\mathbf{v}_i$ are generated independently from $\mathcal{N}_3(0, 0.01\mathbf{I}_3)$. In this case no cross-sectional correlations are present.

Block-diagonal $\boldsymbol{\Sigma}_u^{(2)} = \text{diag}\{\mathbf{A}_1, ..., \mathbf{A}_{N/5}\}$ is a block-diagonal covariance, where each diagonal block $\mathbf{A}_j$ is a 5×5 positive definite matrix, generated from a cross-sectional MA(3) process as follows: for each j, generate $\{a_i, b_i, c_i\}_{i \le 5}$ independently from $\mathcal{N}(0, 0.01)$. Let $\{e_i\}_{i \le 5}$ be i.i.d. $\mathcal{N}(0,1)$, $v_1 = e_1$, $v_2 = e_2 + a_1 e_1$, $v_3 = e_3 + a_2 e_2 + b_1 e_1$, and for $i = 3, 4$,

$$v_{i+1} = e_{i+1} + a_i e_i + b_{i-1} e_{i-1} + c_{i-2} e_{i-2}.$$

Set $\mathbf{A}_j$ be the covariance matrix of $(v_i : i \le 5)$, which only depends on $\{a_i, b_i, c_i\}_{i \le 5}$. In the numerical study, we would not assume we know the block-diagonal structure though, and apply soft-thresholdings to estimate $\boldsymbol{\Sigma}_u^{(2)}$.

We study two types of alternatives (we set $N > T$):

$$\textbf{sparse alternative } H_a^1: \quad \alpha_i = \begin{cases} 0.3, & i \le \frac{N}{T} \\ 0, & i > \frac{N}{T} \end{cases}$$

$$\textbf{weak alpha } H_a^2: \quad \alpha_i = \begin{cases} \sqrt{\frac{\log N}{T}}, & i \le N^{0.4} \\ 0, & i > N^{0.4} \end{cases}.$$

Under H_a^1, many components of $\boldsymbol{\alpha}$ are zero but the nonzero alphas are not weak. Under H_a^2, the nonzero alphas are all very weak. In our simulation setup, $\sqrt{\log N/T}$ varies from 0.05 to 0.10. We therefore expect that under H_a^1, $P(\widehat{S} = \emptyset)$ is close to zero because most of the first N/T estimated alphas should survive from the screening step. In contrast, $P(\widehat{S} = \emptyset)$ should be much larger under H_a^2 because the nonzero alphas are too week.

Table 2: Size and Power comparison when $\mathbf{\Sigma}_u = \mathbf{\Sigma}_u^{(1)}$ is diagonal

		H_0			H_a^1			H_a^2		
T	N	J_{sq}	PEM	$P(\widehat{S}=\emptyset)$	J_{sq}	PEM	$P(\widehat{S}=\emptyset)$	J_{sq}	PEM	$P(\widehat{S}=\emptyset)$
300	500	0.076	0.090	0.986	0.536	0.950	0.082	0.730	0.790	0.642
	800	0.096	0.116	0.984	0.714	0.982	0.026	0.782	0.822	0.634
	1000	0.088	0.100	0.982	0.648	0.976	0.032	0.810	0.852	0.630
	1200	0.098	0.110	0.980	0.770	0.992	0.012	0.808	0.860	0.602
500	500	0.070	0.070	0.994	0.436	0.980	0.024	0.788	0.808	0.796
	800	0.076	0.078	0.996	0.718	1	0	0.788	0.810	0.796
	1000	0.068	0.076	0.994	0.670	0.996	0.004	0.796	0.828	0.754
	1200	0.056	0.058	0.998	0.624	1	0	0.786	0.812	0.788

The frequencies of rejection and $\widehat{S}=\emptyset$ out of 500 replications are calculated. Here J_{sq} is the standardized quadratic test using diagonal weight matrix as in Pesaran and Yamagata (2012); PEM represents the power enhancement test $J_0 + J_{sq}$.

6.2 Results

When $\mathbf{\Sigma}_u = \mathbf{\Sigma}_u^{(1)}$, we assume the diagonal structure to be known, and compare the performances of the standardized quadratic test J_{sq} (considered by Pesaran and Yamagata 2012) with the power enhanced test $J_0 + J_{sq}$. When $\mathbf{\Sigma}_u = \mathbf{\Sigma}_u^{(2)}$, we do not assume we know the block-diagonal structure. In this case, four tests are carried out and compared: (1) the standardized quadratic test with diagonal weight J_{sq} based on $\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\alpha}}$, (2) the improved standardized quadratic test $\widetilde{J}_{sq}$ with thresholded covariance weight matrix, based on $\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\boldsymbol{\alpha}}$, (3) the power enhanced test $J_0 + J_{sq}$, and (4) the power enhanced test $J_0 + \widetilde{J}_{sq}$. For the thresholded covariance matrix, we use the soft-thresholding function and fix the threshold at $1.5\sqrt{\log N/T}$.

For each test, we calculate the frequency of rejection under H_0, H_a^1 and H_a^2 based on 500 replications, with the 0.05 significance level. We also calculate the frequency of $\widehat{S}$ being empty, which approximates $P(\widehat{S}=\emptyset)$. Results are summarized in Tables 2-4.

When $\mathbf{\Sigma}_u$ is diagonal, we see that under the null the PEM test has slightly larger rejection probabilities, and $P(\widehat{S}=\emptyset)$ is close to one, which demonstrates that the screening statistic J_0 indeed manages to screen out most of the estimation errors under the null. On the other hand, under H_a^1, the PEM test significantly improves the power of the standardized quadratic test. In this case, $P(\widehat{S}=\emptyset)$ is nearly zero because the estimated nonzero alphas still stay after screening. Under H_a^2, however, the nonzero alphas are very week, which leads to a large probability that $\widehat{S}$ is an empty set. As a result, the PEM test only slightly improves the power of the quadratic test.

Table 3: Size comparison when $\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u^{(2)}$ is block-diagonal

T	N	J_{sq}	$\widetilde{J}_{sq}$	PEM1	PEM2	$P(\widehat{S} = \emptyset)$
		H_0				
300	500	0.074	0.102	0.090	0.116	0.982
	800	0.078	0.102	0.088	0.112	0.984
	1000	0.102	0.136	0.108	0.142	0.992
	1200	0.088	0.136	0.104	0.152	0.980
500	500	0.068	0.074	0.070	0.076	0.996
	800	0.096	0.104	0.104	0.112	0.992
	1000	0.064	0.080	0.066	0.082	0.996
	1200	0.104	0.118	0.104	0.118	1.000

The frequencies of rejection and $\widehat{S} = \emptyset$ out of 500 replications are calculated. Here J_{sq} is the standardized quadratic test using diagonal weight matrix; $\widetilde{J}_{sq}$ uses the thresholded $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ as the weight matrix; PEM1 represents the test $J_0 + J_{sq}$; PEM2 represents the test $J_0 + \widetilde{J}_{sq}$. $\widetilde{J}_{sq}$ does not assume the block-diagonal structure to be known, and uses soft-thresholding.

Table 4: Power comparison when $\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u^{(2)}$ is block-diagonal

T	N	H_a^1 J_{sq}	H_a^1 $\widetilde{J}_{sq}$	H_a^1 PEM1	H_a^1 PEM2	H_a^1 $P(\widehat{S} = \emptyset)$	H_a^2 J_{sq}	H_a^2 $\widetilde{J}_{sq}$	H_a^2 PEM1	H_a^2 PEM2	H_a^2 $P(\widehat{S} = \emptyset)$
300	500	0.568	0.602	0.966	0.966	0.050	0.768	0.816	0.808	0.846	0.624
	800	0.710	0.756	0.992	0.994	0.012	0.780	0.832	0.836	0.874	0.570
	1000	0.690	0.734	0.990	0.992	0.018	0.830	0.858	0.874	0.896	0.592
	1200	0.762	0.806	0.994	0.994	0.016	0.836	0.874	0.872	0.900	0.612
500	500	0.438	0.468	0.986	0.986	0.016	0.762	0.790	0.796	0.812	0.784
	800	0.750	0.768	0.998	0.998	0.002	0.748	0.770	0.778	0.794	0.758
	1000	0.670	0.706	1	1	0	0.786	0.816	0.814	0.840	0.760
	1200	0.628	0.652	1	1	0	0.784	0.812	0.820	0.844	0.752

The frequencies of rejection and $\widehat{S} = \emptyset$ out of 500 replications are calculated.

For the nondiagonal covariance, similar patterns are observed. We additionally find interesting behaviors of $\widetilde{J}_{sq}$ when a thresholded covariance matrix is used. Under the alternatives, $\widetilde{J}_{sq}$ has larger powers than J_{sq} does because it takes into account the cross-sectional correlations, and the power is further significantly improved by the PEM tests. In addition, ignoring the cross-sectional correlation structure, J_{sq} yields more stable test statistics than the thresholded test $\widetilde{J}_{sq}$, so the sizes of J_{sq} can be more accurately determined under the null.

7 Real data analysis

We apply the test statistics J_{sq}, $\widetilde{J}_{sq}$ and PEM ($= J_0 + \widetilde{J}_{sq}$) to the securities in the S&P 500 index. Composed of large cap U.S. stocks, the S&P 500 index has diverse constituency and is therefore a good market representation. We collect monthly returns on all the S&P 500 constituents from the CRSP database for the period January 1980 to December 2012, during which a total of 1170 stocks have entered the index for our study. Testing of market efficiency is performed on a rolling window basis: for each month from December 1984 to December 2012, we evaluate our test statistics using the preceding 60 months' returns ($T = 60$). The panel at each testing month consists of those stocks without missing data in the past five years, which yields a cross-sectional dimension much 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 possible structural breaks in the factor loadings.

We employ the Fama-French three-factor (FF-3) model to conduct our test. For testing months $\tau = 12/1984, ..., 12/2012$, we estimate the FF-3 regressions

$$r_{it}^{\tau} - r_{ft}^{\tau} = \alpha_i^{\tau} + b_{MKT}^{\tau}(MKT_t^{\tau} - r_{ft}^{\tau}) + b_{SMB}^{\tau} SMB_t^{\tau} + b_{HML}^{\tau} HML^{\tau} + u_{it}^{\tau},$$

for $i = 1, ..., N_\tau$ and $t = \tau - 59, ..., \tau$, where N_τ is the panel size for testing window $[\tau - 59, \tau]$, r_{it}^{τ} represents the monthly return of security i, r_{ft}^{τ} is the corresponding risk-free rate, MKT, SMB and HML constitute the FF-3 model's market, size and value factors. Our null hypothesis $\alpha_i^{\tau} = 0$ for all i implies that the market is mean-variance efficient.

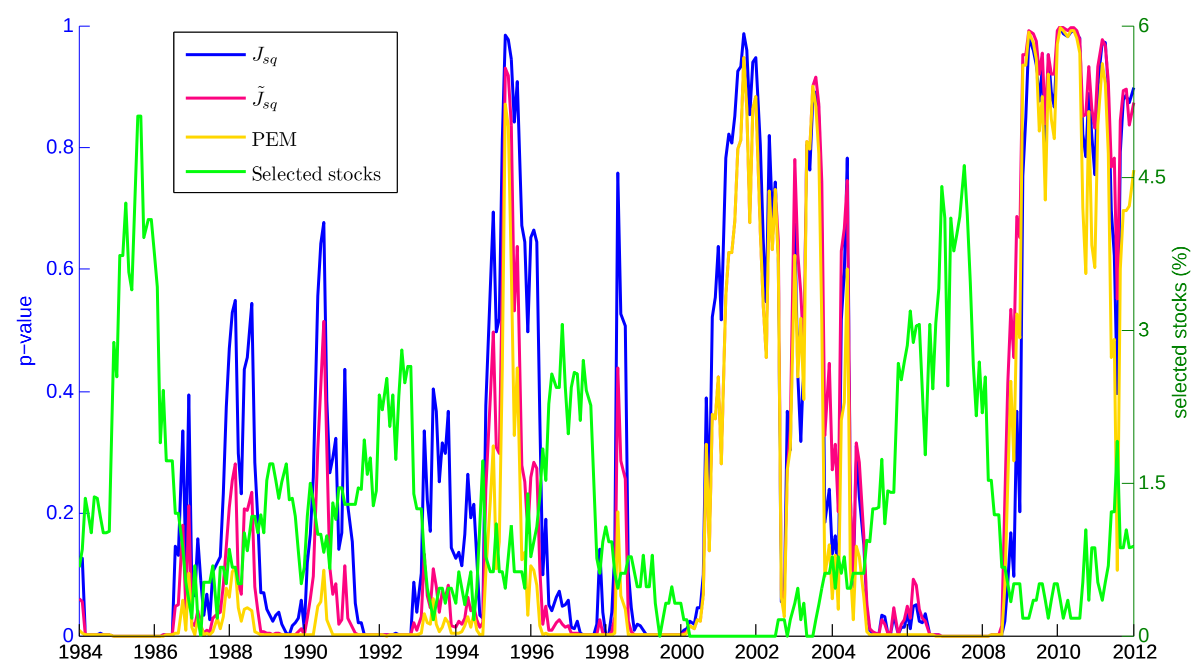
Table 5 summarizes descriptive statistics for different components and estimates in the model. On average, 618 stocks enter the panel of the regression during each five-year estimation window, of which 7.3 stocks are selected by $\widehat{S}$. The selected stocks have much larger alphas than other stocks do, as we could expect. As far as the signs of those alpha estimates are concerned, 61.84% of all the alpha estimates during the entire study period are positive. In contrast, 81.05% of the selected alphas are positive. It is therefore more likely for stocks

Table 5: Variable descriptive statistics for the FF-3 model

Variables	Mean	Std dev.	Median	Min	Max
N_τ	617.70	26.31	621	574	665
$\lvert\widehat{S}\rvert_0$	7.30	6.73	5	0	30
$\overline{\widehat{\alpha}}_i^\tau(\%)$	0.3729	0.1990	0.3338	-0.1735	0.9763
$\overline{\widehat{\alpha}}_{i\in\widehat{S}}^\tau(\%)$	1.1272	0.7881	1.2351	-2.6182	3.3757
p-value of J_{sq}	0.2789	0.3385	0.0992	0	0.9969
p-value of $\widetilde{J}_{sq}$	0.2350	0.3339	0.0291	0	0.9984
p-value of PEM	0.1846	0.3101	0.0031	0	0.9960

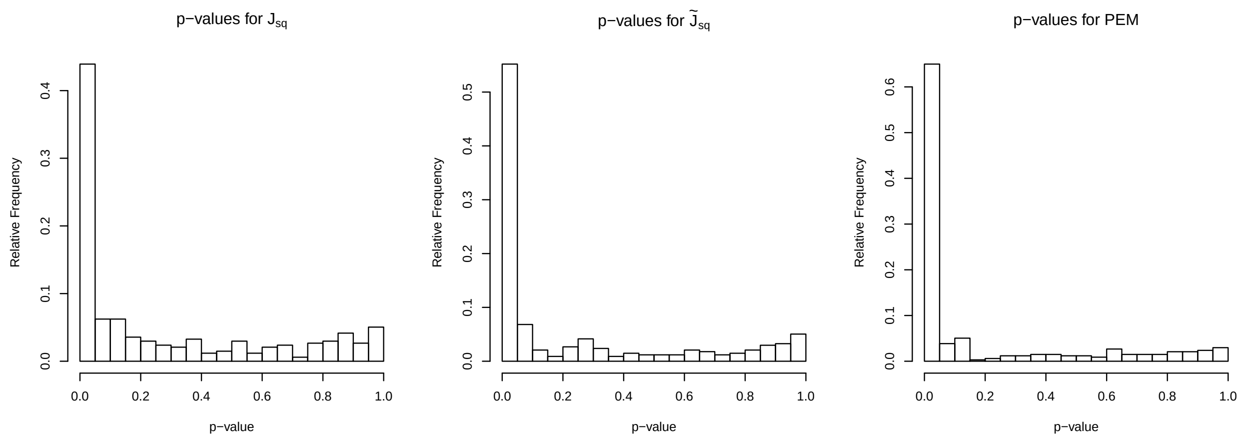
with large positive alphas to enter $\widehat{S}$, and less so for those with large negative alphas. This indicates that market inefficiency is primarily caused by stocks with extra returns, instead of an average (or aggregated) sense (that is, more likely due to a large $\min_{\alpha_j>0}\alpha_j$ other than a large $\|\boldsymbol{\alpha}\|$). In addition, we notice that the p-values of $\widetilde{J}_{sq}$ test are generally smaller than those of J_{sq}, because $\widetilde{J}_{sq}$ has fully exploited the cross-sectional correlations. Due to an enhanced power, PEM exhibits even lower p-values than the previous two.

Figure 1: Evolutions of p-values and selected stocks (%)



Similar to Pesaran and Yamagata (2012), we plot the running p-values of J_{sq}, $\widetilde{J}_{sq}$ and the PEM test from December 1984 to December 2012. We also add the evolution of the percentage of selected stocks ($|\widehat{S}|_0/N$) to the plot, as shown in Figure 1. There is a strikingly

Figure 2: Histograms of p-values for J_{sq}, $\widetilde{J}_{sq}$ and PEM



negative correlation between stock selection percentage and the p-value of these tests. This shows that the degree of market efficiency is influenced not only by the aggregation of alphas, but also by those extreme ones. We also observe that the p-value line of the PEM test lies beneath those of $\widetilde{J}_{sq}$ and J_{sq} test as a result of enhanced power, and it captures some important market disruptions ignored by the latter two. The null hypothesis of market efficiency is rejected at 5% level during major financial crisis, including Black Wednesday in 1992, Asian financial crisis in 1997, the recent subprime crisis in 2008 and the European sovereign debt crisis in 2010. For 44%, 55% and 65% of the study period, J_{sq}, $\widetilde{J}_{sq}$ and the PEM test conclude that the market is inefficient respectively. The histograms of the p-values of the three test statistics are displayed in Figure 2.

We now take a closer look at the screening set $\widehat{S}$, which consists of stocks with large positive or negative alphas. By definition, the selected stocks have statistically significant alphas for the given window of estimation, suggesting that their returns are not commensurate with their risks. In practice, such stocks could often contribute to the construction of a market-neutral high-alpha portfolio. During the entire study period, we record 223 different stocks that have entered $\widehat{S}$ at least once. We extract those who have persistent performance—in particular, those who stay in the screening set for at least 24 consecutive months. As a result, 7 companies stand out. Table 6 lists these companies, their major periods when getting selected and the associated alphas. Figure 3 specifies the selection periods. Interestingly, these companies span different industries and they all have positive alphas when picked by $\widehat{S}$.

Table 6: Companies with longest selection period

Company Name	Major period of selection	Average alpha (%)	Std. dev. (%)
COCA COLA CO	09/1989—08/1992	1.3714	0.0921
U S T INC	02/1986—10/1988	1.9997	0.2154
MCDONALDS CORP	03/2003—12/2012	1.5502	0.2013
WAL MART STORES INC	08/1985—12/1988	1.4927	0.1425
HOME DEPOT INC	05/1986—02/1990	3.4931	0.5243
KIMCO REALTY CORP	09/1999—06/2002	1.6195	0.1833
PRAXAIR INC	07/2001—11/2004	1.3430	0.2764

Major period of selection refers to the time interval when those companies stay in the screening set for at least 24 months. Average alpha and Std. dev. are computed during the period of selection.

8 Concluding remarks

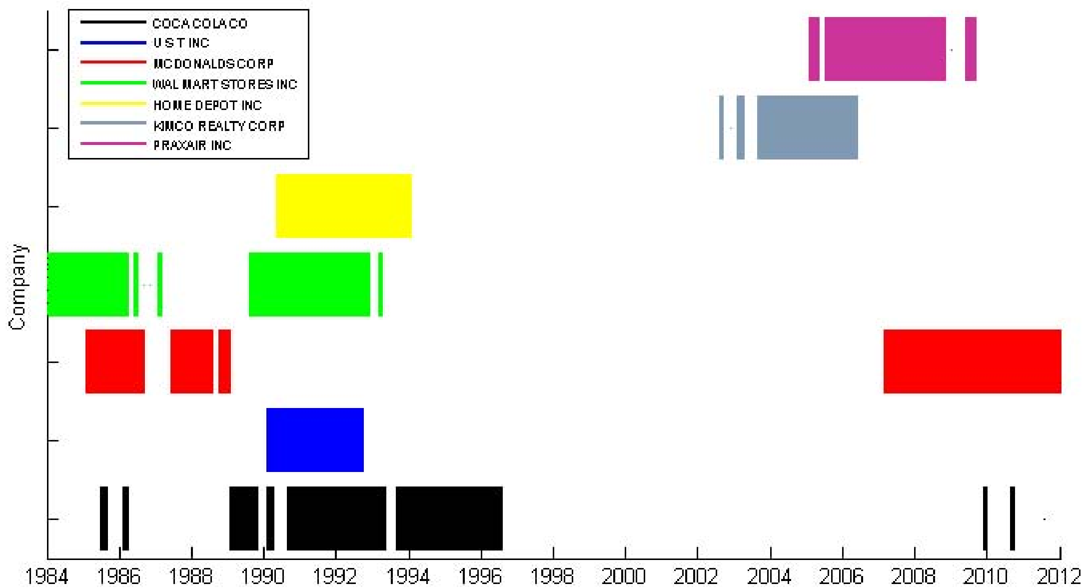
The literature on testing mean-variance efficiency is only able to detect market inefficiency in an average sense as measured by the quadratic form $\boldsymbol{\alpha}'\boldsymbol{\alpha}$. However, when we deal with large panels, it is more appealing if we could identify individual departures from the factor pricing model, and deal with the case when there are small portions of significant alphas.

We propose a new concept for high dimensional statistical tests, namely, the power enhancement (PEM). The PEM test combines a sure-screening statistic and a Wald-type statistic. Under the null, the sure-screening part equals zero with probability approaching one, and under alternatives where there are significant alphas, it is stochastically unbounded. Hence while maintaining a good size asymptotically, the PEM test significantly enhances the power of Wald-type statistics. As a by-product, the screening-set $\widehat{S}$ also enables us to identify those significant alphas.

We also improve the standardized quadratic test by taking into account the off-diagonal structure of the error covariance matrix. Assuming this matrix to be sparse, we estimate it using the thresholding technique, and develop new techniques to prove that the effect of estimating the inverse covariance matrix is asymptotically negligible. Therefore, the aggregation of estimation errors is successfully avoided. This technique is potentially useful in other high-dimensional econometric applications, where an optimal weight matrix needs to be estimated, such as GMM and GLS.

Our empirical study shows that indeed, market inefficiency is primarily caused by a small portion of significantly mispriced stocks, instead of aggregated alphas. In addition, most of the selected stocks are with extra returns. The PEM test serves as an appropriate test of factor pricing models.

Figure 3: Selected companies and their periods of selection



A Proofs

We first present a lemma that will be needed throughout the proofs. Suppose $\sigma_{ij} = (\mathbf{\Sigma}_u)_{ij}$ and $\widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt}$.

Lemma A.1. *(i)* $\max_{i,j\leq N} |\frac{1}{T}\sum_{t=1}^T u_{it}u_{jt} - Eu_{it}u_{jt}| = O_p(\sqrt{(\log N)/T})$.
(ii) $\max_{i\leq K, j\leq N} |\frac{1}{T}\sum_{t=1}^T f_{it}u_{jt}| = O_p(\sqrt{(\log N)/T})$.
(iii) $\max_{i,j\leq N} |\widehat{\sigma}_{ij} - \sigma_{ij}| = O_p(\sqrt{\log N/T})$
(iv) $\max_{j\leq N} |\frac{1}{T}\sum_{t=1}^T u_{jt}| = O_p(\sqrt{(\log N)/T})$.
(v) $\max_{j\leq N} |\widehat{\alpha}_j - \alpha_j| = O_p(\sqrt{\log N/T})$.

Proof. Parts (i)-(iii) follow from Lemmas A.3 and B.1 in Fan, Liao and Mincheva (2011). Part (iv) follows from the same proof of that of Lemma A.3 of Fan et al. (2011). For part (v), first note that $\widehat{\alpha}_j - \alpha_j = \frac{1}{\tau T}\sum_{t=1}^T u_{jt}(1 - \mathbf{f}_t'\mathbf{w})$. Here $\tau \to^p 1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t > 0$, hence τ is bounded away from zero with probability arbitrarily close to one. Part (v) then follows from (ii) and (iv).

□

Lemma A.2. $\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| = o_p(1)$.

Proof. See Theorem 3.1 of Fan et al. (2011). Note that Fan et al. (2011) showed this result for the hard-thresholding. In fact, it holds for more general thresholdings, implied by Theorem A.1 of Fan et al. (2013). □

A.1 Proof of Theorem 2.1

Without loss of generality, under the alternative, let $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2') = (\mathbf{0}', \boldsymbol{\alpha}_2')$, where $\dim(\boldsymbol{\alpha}_1) = N - r$ and $\dim(\boldsymbol{\alpha}_2) = r$. Correponding to $(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$, we partition $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_u^{-1}$ into:

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\beta}' \\ \boldsymbol{\beta} & \boldsymbol{\Sigma}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_u^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_1^{-1} + \mathbf{A} & \mathbf{G}' \\ \mathbf{G} & \mathbf{C} \end{pmatrix}.$$

By the matrix inversion formula, we know that $\mathbf{A} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}'(\boldsymbol{\Sigma}_2 - \boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}')^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}$. In addition, we partition the estimator into $\widehat{\boldsymbol{\alpha}}' = (\widehat{\boldsymbol{\alpha}}_1', \widehat{\boldsymbol{\alpha}}_2')$ where $\dim(\widehat{\boldsymbol{\alpha}}_1) = N - r$ and $\dim(\widehat{\boldsymbol{\alpha}}_2) = r$. Note that $\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{\alpha}}_1'\mathbf{A}\widehat{\boldsymbol{\alpha}}_1 + 2\widehat{\boldsymbol{\alpha}}_2'\mathbf{G}\widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{\alpha}}_2'\mathbf{C}\widehat{\boldsymbol{\alpha}}_2$.

We first look at $\widehat{\boldsymbol{\alpha}}_1'\mathbf{A}\widehat{\boldsymbol{\alpha}}_1$. Write $\boldsymbol{\xi} = \boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1$. It follows from $\|\boldsymbol{\Sigma}_1^{-1}\|_1 < \infty$ that $\max_{i\leq N-r}|\xi_i| = O_p(\max_{i\leq N-r}|\widehat{\alpha}_{1i}|) = O_p(\max_{i\leq N-r}|\widehat{\alpha}_{1i} - \alpha_{1i}|) = O_p(\sqrt{\frac{\log N}{T}})$. Also, $\max_{i\leq r}\sum_{j=1}^{N-r}|\beta_{ij}| \leq \|\boldsymbol{\Sigma}_u\|_1 = O(1)$, and $\lambda_{\max}((\boldsymbol{\Sigma}_2 - \boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}')^{-1}) = O(1)$. Hence

$$|\widehat{\boldsymbol{\alpha}}_1'\mathbf{A}\widehat{\boldsymbol{\alpha}}_1| = O(1)\|\boldsymbol{\beta}\boldsymbol{\xi}\|^2 \leq O(1)\max_j |\xi_j|^2 \sum_{i=1}^{r}(\sum_{j=1}^{N-r}|\beta_{ij}|)^2 = O_p(\frac{r\log N}{T}).$$

For $\mathbf{G} = (g_{ij})$, note that $\max_{i\leq r}\sum_{j=1}^{N-r}|g_{ij}| \leq \|\boldsymbol{\Sigma}_u^{-1}\|_1 = O(1)$. Hence

$$|\widehat{\boldsymbol{\alpha}}_2'\mathbf{G}\widehat{\boldsymbol{\alpha}}_1| \leq \max_{j\leq N-r}|\widehat{\alpha}_{1j}|\max_{j\leq r}|\widehat{\alpha}_{2j}|\sum_{i=1}^{r}\sum_{j=1}^{N-r}|g_{ij}| \leq O_p(r\sqrt{\frac{\log N}{T}})$$

where we used the fact that $\max_{j\leq r}|\widehat{\alpha}_{2j}| \leq \max_j|\alpha_{2j}| + \max_j|\widehat{\alpha}_j - \alpha_j| = O_p(1)$. Also, $|\widehat{\boldsymbol{\alpha}}_2'\mathbf{C}\widehat{\boldsymbol{\alpha}}_2| \leq \|\widehat{\boldsymbol{\alpha}}_2\|^2\|\mathbf{C}\| = O_p(r)$. It then yields $\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} \leq \widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 + O_p(r)$ under H_a. It also follows from (2.7) that

$$Z \equiv \frac{T\tau\widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 - (N-r)}{\sqrt{2(N-r)}} \to^d \mathcal{N}(0,1).$$

Hence under H_a, since $Tr = o(\sqrt{N})$, for any $\epsilon \in (0, z_q)$, we have

$$P(J_1 > z_q) \quad \leq \quad P(\frac{T\tau\widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 - N}{\sqrt{2N}} + O_p(\frac{Tr}{\sqrt{N}}) > z_q)$$

$$
\begin{aligned}
&\leq \; P(Z(1+o(1)) + O_p(\frac{Tr}{\sqrt{N}}) > z_q) \\
&\leq \; P(Z + o_p(1) > z_q) \leq 1 - \Phi(z_q - \epsilon) + o(1).
\end{aligned}
$$

Choose ϵ such that $\Phi(z_q - \epsilon) \geq 1 - 2q$. Then $P(J_1 > z_q) \leq 2q + o(1)$.

A.2 Proof of Theorem 3.1 and Corollary 3.1

(i) For any $j \in S$, by the definition of S, $\frac{|\alpha_j|}{\sqrt{\sigma_{jj}}} > 2\delta_T$. Define events

$$
A_1 = \left\{ \max_{j \leq N} |\widehat{\sigma}_{jj}^{-1} - \sigma_{jj}^{-1}| \leq C_2 \right\}, \quad A_2 = \left\{ \max_{j \leq N} |\widehat{\alpha}_j - \alpha_j| \leq C_3 \delta_T \right\}
$$

for some $C_2, C_3 > 0$. Lemma A.1 then implies that $P(A_1 \cap A_2) \to 1$. Under $A_1 \cap A_2$,

$$
\begin{aligned}
\frac{|\widehat{\alpha}_j|}{\sqrt{\widehat{\sigma}_{jj}}} &\geq \; (|\alpha_j| - \max_j |\widehat{\alpha}_j - \alpha_j|)(\sigma_{jj}^{-1} - \max_j |\widehat{\sigma}_{jj}^{-1} - \sigma_{jj}^{-1}|)^{1/2} \\
&\geq \; (|\alpha_j| - C_3 \delta_T)(\sigma_{jj}^{-1} - C_2)^{1/2} \geq \delta_T,
\end{aligned}
$$

where the last inequality holds for sufficiently small C_2, C_3, e.g., $C_3 < \min_j \sqrt{\sigma_{jj}}(2 - \sqrt{3})$ and $C_2 = \frac{2}{3}\min_j(\sigma_{jj}^{-1})$. This implies that $j \in \widehat{S}$. Now $P(S \subset \widehat{S}) \to 1$ follows from the fact that $P(A_1 \cap A_3) \geq 1 - o(1)$. It can be readily seen that if $j \in \widehat{S}$, by similar arguments, we have $\frac{|\alpha_j|}{\sqrt{\sigma_{jj}}} > \frac{1}{2}\delta_T$ with probability tending to one. Consequently, $\widehat{S} \setminus S \subset \Delta$ with probability approaching one.

(ii) Suppose $\min_{j \leq N} \sigma_{jj} > C_1$ for some $C_1 > 0$. For some constants $C_2 > 0$ and $C_3 < (C_2 + C_1^{-1})^{1/2}$ in the event $A_1 \cap A_2$, under the event $A_1 \cap A_2$ and H_0, we have

$$
\begin{aligned}
\max_{j \leq N} \frac{|\widehat{\alpha}_j|}{\sqrt{\widehat{\sigma}_{jj}}} &\leq \; \max |\widehat{\alpha}_j| \cdot (\max \widehat{\sigma}_{jj}^{-1})^{1/2} \leq C_3 \delta_T \cdot (\max |\widehat{\sigma}_{jj}^{-1} - \sigma_{jj}^{-1}| + \max \sigma_{jj}^{-1})^{1/2} \\
&\leq \; C_3 (C_2 + C_1^{-1})^{1/2} \delta_T \leq \delta_T,
\end{aligned}
$$

where we note that under H_0, $\alpha_j = 0$. Hence $P(\max_{j \leq N} \frac{|\widehat{\alpha}_j|}{\sqrt{\widehat{\sigma}_{jj}}} \leq \delta_T) \to 1$, which implies $P(\widehat{S} = \emptyset) \to 1$. This immediately implies $P(J_0 = 0) \to 1$.

Corollary 3.1 is implied by part (i).

A.3 Proof of Theorem 3.2

Part (i) follows immediately from that $P(J_0 = 0 | H_0) \to 1$.

(ii) Let F_q be the qth quantile of F, then under the level $(1 - q)$, then both J and J_1

reject H_0 if $J_0 \geq F_q$. Below we write $P(\cdot|\boldsymbol{\alpha})$ to denote the probability measure if the true alpha is $\boldsymbol{\alpha}$. For any $\boldsymbol{\alpha} \in \bar{A}$, if $\boldsymbol{\alpha} \in A$ under the alternative, because J_1 has power against A and $J \geq J_1$, so

$$P(J \geq F_q|\boldsymbol{\alpha}) \geq P(J_1 \geq F_q|\boldsymbol{\alpha}) \to 1,$$

which implies that J also has power against A. We now consider the case when

$$\boldsymbol{\alpha} \in \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j \leq N} |\alpha_j| > 2\delta_T \min_{j \leq N} \sigma_j\}.$$

First of all, since $\boldsymbol{\alpha} \in \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j\leq N} |\alpha_j| > 2\delta_T \min_{j\leq N} \sigma_j\}$, $S \neq \emptyset$. It follows from $P(S \subset \widehat{S}) \to 1$ that $P(\widehat{S} \neq \emptyset) \to 1$. Lemma A.2 then implies $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\widehat{S}}^{-1})$ is bounded away from zero with probability approaching one. Under the event $S \subset \widehat{S}$, we have $\|\widehat{\boldsymbol{\alpha}}_{\widehat{S}}\| \geq \|\widehat{\boldsymbol{\alpha}}_S\|$. Hence there is a constant $C > 0$ so that with probability approaching one,

$$J_0 \geq CT\|\widehat{\boldsymbol{\alpha}}_{\widehat{S}}\|^2 \geq CT\|\widehat{\boldsymbol{\alpha}}_S\|^2 \geq CT(\|\boldsymbol{\alpha}_S\| - \|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|)^2.$$

On one hand, $\|\boldsymbol{\alpha}_S\|^2 \geq \min_{j\in S} \alpha_j^2 |S|_0 \geq |S|_0 \delta_T^2 4 \min_j \sigma_j^2$. On the other hand,

$$\|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|^2 = O_p(\frac{\log N}{T}|S|_0).$$

Since $\log N/T = o(\delta_T^2)$, with probability approaching one, $J_0 \geq \min_j \sigma_j^2 CT\delta_T^2$, which is $C'(\log\log T)(\log N)$ for some constant $C' > 0$. Note that J_{sq} is standardized such that $F_q = O(1)$. For example, if a quadratic statistic $\widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}}$ is used, then $J_{sq} = T(\widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}} - N)/\sqrt{2N}$ is asymptotically normal, and $F_q = z_q$. Hence $P(J > F_q) \to 1$ as $J \geq J_0$, which is stochastically unbounded.

A.4 Proof of Theorem 4.1

It follows from Pesaran and Yamagata (2012 Section 4.4) that $J_{sq} \to \mathcal{N}(0, 1)$. According to Theorem 3.2, it suffices to show that J_{sq} has power against $\{\|\boldsymbol{\alpha}\|^2 \gg N\log N/T\}$. In fact, $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| = O_p(\sqrt{N(\log N)/T})$. When $\|\boldsymbol{\alpha}\|^2 \gg N\log N/T$, $\|\widehat{\boldsymbol{\alpha}}\|^2 \geq (\|\boldsymbol{\alpha}\| - \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|)^2 \geq \frac{N\log N}{4T}$ with probability approaching one. So $P(W \geq CN\log N) \to 1$ for some $C > 0$, which implies $P(J_{sq} > C'\sqrt{N}\log N) \to 1$ for some $C' > 0$. Thus the test is consistent since $J \geq J_{sq}$.

A.5 Proof of Theorem 5.1

The proof of part (ii) is the same as that of Theorem 3.2. Moreover, it follows from Pesaran and Yamagata (2012, Theorem 1) that $(T\tau\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - N)/\sqrt{2N} \to^d \mathcal{N}(0, 1)$. So the theorem is proved by Proposition A.1 below.

Proposition A.1. *Under the assumptions of Theorem 5.1, and under* H_0,

$$\frac{T\widehat{\boldsymbol{\alpha}}'(\boldsymbol{\Sigma}_u^{-1}-\widehat{\boldsymbol{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}}{\sqrt{N}}=o_p(1)$$

Proof. The left hand side is equal to

$$\frac{T\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}(\widehat{\boldsymbol{\Sigma}}_u-\boldsymbol{\Sigma}_u)\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}'}{\sqrt{N}}+\frac{T\widehat{\boldsymbol{\alpha}}'(\widehat{\boldsymbol{\Sigma}}_u^{-1}-\boldsymbol{\Sigma}_u^{-1})(\widehat{\boldsymbol{\Sigma}}_u-\boldsymbol{\Sigma}_u)\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}'}{\sqrt{N}}\equiv a+b.$$

It was shown by Fan et al. (2011) that $\|\widehat{\boldsymbol{\Sigma}}_u-\boldsymbol{\Sigma}_u\|=O_p(m_N\sqrt{\frac{\log N}{T}})=\|\widehat{\boldsymbol{\Sigma}}_u^{-1}-\boldsymbol{\Sigma}_u^{-1}\|$. In addition, $\|\widehat{\boldsymbol{\alpha}}\|^2=O_p(N\log N/T)$. Hence $b=O_p(\frac{m_N^2\sqrt{N}(\log N)^2}{T})=o_p(1)$. It suffices to show $a=o_p(1)$. For simplicity, we consider the same setup as in Fan et al. (2011) in the sense that $S_L=\{(i,j):(\boldsymbol{\Sigma}_u)_{ij}=0\}$, and $\min_{(i,j)\in S_U}|(\boldsymbol{\Sigma}_u)_{ij}|\gg\sqrt{\frac{\log N}{T}}$, which corresponds to the strictly sparse case. In addition, we consider the hard-thresholding covariance estimator. The proof for the generalized sparsity case as in Rothman et al. (2009) is very similar.

Let $s_{ij}=\frac{1}{T}\sum_{t=1}^T\widehat{u}_{it}\widehat{u}_{jt}$ and $\sigma_{ij}=(\boldsymbol{\Sigma}_u)_{ij}$. Under hard-thresholding,

$$\widehat{\sigma}_{ij}=(\widehat{\boldsymbol{\Sigma}}_u)_{ij}=\begin{cases}s_{ii}, & \text{if } i=j,\\ s_{ij}, & \text{if } i\neq j, |s_{ij}|>C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2}\\ 0, & \text{if } i\neq j, |s_{ij}|\leq C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2}\end{cases}\tag{A.1}$$

Write $(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i$ to denote the ith element of $\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}$. We have,

$$\begin{aligned}a &= \frac{T}{\sqrt{N}}\sum_{i=1}^N(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2(\widehat{\sigma}_{ii}-\sigma_{ii})+\frac{T}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_j(\widehat{\sigma}_{ij}-\sigma_{ij})\\ &\quad+\frac{T}{\sqrt{N}}\sum_{(i,j)\in S_L}(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_j\widehat{\sigma}_{ij}=a_1+a_2+a_3.\end{aligned}$$

Let us investigate $a_i, i=1,2,3$ separately.

First, we look at a_3. Note that

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

Because s_{ii} is uniformly (across i) bounded away from zero with probability approaching one, and $\max_{(i,j)\in S_L}|s_{ij}|=O_p(\sqrt{\frac{\log N}{T}})$. Hence for any $\epsilon>0$, when C in the threshold (A.1) is large enough, $P(a_3 > T^{-1}) < \epsilon$, this implies $a_3 = o_p(1)$.

The proofs for $a_1, a_2 = o_p(1)$ are given by the following lemmas. □

Lemma A.3. *Under H_0, $a_1 = o_p(1)$.*

Proof. We have $a_1 = \frac{T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}^2 - Eu_{it}^2)$, which is

$$\frac{T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}^2 - u_{it}^2) + \frac{T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T(u_{it}^2 - Eu_{it}^2) = a_{11} + a_{12}.$$

For a_{12}, note that $(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i = \frac{1}{T}\sum_{s=1}^T (1 - \mathbf{f}_s'\mathbf{w})(\mathbf{u}_s'\boldsymbol{\Sigma}_u^{-1})_i = \frac{1}{T}\sum_{s=1}^T(1-\mathbf{f}_s'\mathbf{w})\xi_{is}$. Hence

$$a_{12} = \frac{T}{\sqrt{N}}\sum_{i=1}^N (\frac{1}{T}\sum_{s=1}^T (1-\mathbf{f}_s'\mathbf{w})\xi_{is})^2 \frac{1}{T}\sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)$$

By Assumption 5.2 (i), $Ea_{12}^2 = o(1)$. On the other hand,

$$a_{11} = \frac{T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T (\widehat{u}_{it} - u_{it})^2 + \frac{2T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{u}_{it} - u_{it}) = a_{111} + a_{112}.$$

Note that $\max_{i\le N}\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})^2 = O_p(\sqrt{\frac{\log N}{T}})$ by Lemma 3.1 of Fan et al. (2011). Hence

$$a_{111} \le O_p(\sqrt{\frac{\log N}{T}})\frac{T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 = O_p(\sqrt{\frac{\log N}{N}})\|\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\|^2 = o_p(1),$$

since $\|\widehat{\boldsymbol{\alpha}}\|^2 = O_p(\frac{N\log N}{T})$, $\|\boldsymbol{\Sigma}_u^{-1}\| = O(1)$ and $N(\log N)^3 = o(T^2)$. To bound a_{112}, note that

$$\widehat{u}_{it} - u_{it} = \widehat{\alpha}_i - \alpha_i + (\widehat{\mathbf{b}}_i - \mathbf{b}_i)'\mathbf{f}_t, \quad \max_i \|\widehat{\alpha}_i - \alpha_i\| = O_p(\sqrt{\frac{\log N}{T}}) = \max_i \|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|.$$

Also, $\max_i |\frac{1}{T}\sum_{t=1}^T u_{it}| = O_p(\sqrt{\frac{\log N}{T}}) = \max_i \|\frac{1}{T}\sum_{t=1}^T u_{it}\mathbf{f}_t\|$. Hence

$$\begin{aligned} a_{112} &= \frac{2T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 \frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{\alpha}_i - \alpha_i) + \frac{2T}{\sqrt{N}}\sum_{i=1}^N (\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i^2 (\widehat{\mathbf{b}}_i - \mathbf{b}_i)'\frac{1}{T}\sum_{t=1}^T \mathbf{f}_t u_{it} \\ &\le O_p(\frac{\log N}{\sqrt{N}})\|\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\|^2 = O_p(\frac{\sqrt{N}\log N}{T}) = o_p(1). \end{aligned}$$

In summary, $a_1 = a_{12} + a_{111} + a_{112} = o_p(1)$. □

Lemma A.4. *Under H_0, $a_2 = o_p(1)$.*

Proof. We have $a_2 = \frac{T}{\sqrt{N}} \sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt})$, which is

$$\frac{T}{\sqrt{N}} \sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j \left(\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt} - u_{it}u_{jt}) + \frac{1}{T}\sum_{t=1}^T(u_{it}u_{jt} - Eu_{it}u_{jt})\right) = a_{21}+a_{22}.$$

$$a_{22} = \frac{T}{\sqrt{N}} \sum_{i\neq j,(i,j)\in S_U} \frac{1}{T}\sum_{s=1}^T(1-\mathbf{f}_s'\mathbf{w})\xi_{is}\frac{1}{T}\sum_{k=1}^T(1-\mathbf{f}_k'\mathbf{w})\xi_{jk}\frac{1}{T}\sum_{t=1}^T(u_{it}u_{jt} - Eu_{it}u_{jt}).$$

By Assumption 5.2 (ii), $Ea_{22}^2 = o(1)$. On the other hand, $a_{21} = a_{211} + a_{212}$, where

$$\begin{aligned} a_{211} &= \frac{T}{\sqrt{N}} \sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})(\widehat{u}_{jt} - u_{jt}), \\ a_{212} &= \frac{2T}{\sqrt{N}} \sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j \frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{u}_{jt} - u_{jt}). \end{aligned}$$

By the Cauchy-Schwarz inequality, $\max_{ij}|\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})(\widehat{u}_{jt} - u_{jt})| = O_p(\frac{\log N}{T})$. Hence

$$\begin{aligned} |a_{211}| &\leq O_p(\frac{\log N}{\sqrt{N}}) \sum_{i\neq j,(i,j)\in S_U} |(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i||(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j| \\ &\leq O_p(\frac{\log N}{\sqrt{N}}) \left(\sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i^2\right)^{1/2} \left(\sum_{i\neq j,(i,j)\in S_U} (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j^2\right)^{1/2} \\ &= O_p(\frac{\log N}{\sqrt{N}}) \sum_{i=1}^N (\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i^2 \sum_{j:\sigma_{ij}\neq 0} 1 \leq O_p(\frac{\log N}{\sqrt{N}})\|\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1}\|^2 m_N = O_p(\frac{m_N\sqrt{N}(\log N)^2}{T}) \end{aligned}$$

which is $o_p(1)$. Similar to the proof of Lemma A.3, $\max_{ij}|\frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{u}_{jt} - u_{jt})| = O_p(\frac{\log N}{T})$.

$$|a_{212}| \leq O_p(\frac{\log N}{\sqrt{N}}) \sum_{i\neq j,(i,j)\in S_U} |(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_i||(\widehat{\boldsymbol\alpha}'\boldsymbol\Sigma_u^{-1})_j| = O_p(\frac{m_N\sqrt{N}(\log N)^2}{T}) = o_p(1).$$

In summary, $a_2 = a_{22} + a_{211} + a_{212} = o_p(1)$. □

A.6 Proof of Lemma 5.1

Recall that $\boldsymbol\xi_t = \boldsymbol\Sigma_u^{-1}\mathbf{u}_t$. The key observation is the following lemma.

Lemma A.5. *ξ_{it} and u_{jt} are independent if $i \neq j$.*

Proof. Because $\mathbf{u}_t$ is Gaussian, it suffices to show that $\mathrm{cov}(\xi_{it}, u_{jt}) = 0$ when $i \neq j$. Consider the vector $(\mathbf{u}_t', \boldsymbol{\xi}_t')' = \mathbf{A}(\mathbf{u}_t', \mathbf{u}_t')'$, where

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_N & 0 \\ 0 & \boldsymbol{\Sigma}_u^{-1} \end{pmatrix}.$$

Then $\mathrm{cov}(\mathbf{u}_t', \boldsymbol{\xi}_t') = \mathbf{A}\mathrm{cov}(\mathbf{u}_t', \mathbf{u}_t')\mathbf{A}$, which is

$$\begin{pmatrix} \mathbf{I}_N & 0 \\ 0 & \boldsymbol{\Sigma}_u^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_u \\ \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_u \end{pmatrix} \begin{pmatrix} \mathbf{I}_N & 0 \\ 0 & \boldsymbol{\Sigma}_u^{-1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_u & \mathbf{I}_N \\ \mathbf{I}_N & \boldsymbol{\Sigma}_u^{-1} \end{pmatrix}.$$

This completes the proof. □

A.6.1 Proof of Lemma 5.1, first statement

The proof for the first statement (Condition (i) of Assumption 5.2) is the same regardless of the type of $\boldsymbol{\Sigma}_u$ in Assumption 5.4.

Let $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T \xi_{is}(1 - \mathbf{f}_s'\mathbf{w}))^2$. For notational simplicity, let

$$e_{it} = u_{it}^2 - Eu_{it}^2, \quad \zeta_{is} = \xi_{is}(1 - \mathbf{f}_s'\mathbf{w}).$$

Then $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T e_{it}(\frac{1}{\sqrt{T}} \sum_{s=1}^T \zeta_{is})^2$. In addition, $E\xi_{is} = 0$. We show respectively $\frac{1}{T}(EX)^2 = o(1)$ and $\frac{1}{T}\mathrm{var}(X) = o(1)$.

Note that $\mathbf{u}_t$ is Gaussian and independent across t and also independent of $\mathbf{f}_t$, hence e_{it} is independent of ζ_{js} if $t \neq s$, for any $i, j \leq N$, which implies $\mathrm{cov}(e_{it}, \zeta_{is}\zeta_{ik}) = 0$ as long as $s, k \neq t$.

Expectation

For the expectation,

$$\begin{aligned} EX &= \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \mathrm{cov}(e_{it}, (\frac{1}{\sqrt{T}} \sum_{s=1}^T \zeta_{is})^2) = \frac{1}{T\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \sum_{k=1}^T \mathrm{cov}(e_{it}, \zeta_{is}\zeta_{ik}) \\ &= \frac{1}{T\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (\mathrm{cov}(e_{it}, \zeta_{it}^2) + 2\sum_{k \neq t} \mathrm{cov}(e_{it}, \zeta_{it}\zeta_{ik})) \\ &= \frac{1}{T\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \mathrm{cov}(e_{it}, \zeta_{it}^2) = O(\sqrt{\frac{N}{T}}), \end{aligned}$$

where the second last equality follows since $Ee_{it} = E\zeta_{it} = 0$ and when $k \neq t$ $\mathrm{cov}(e_{it}, \zeta_{it}\zeta_{ik}) = Ee_{it}\zeta_{it}\zeta_{ik} = Ee_{it}\zeta_{it}E\zeta_{ik} = 0$. It then follows that $\frac{1}{T}(EX)^2 = O(\frac{N}{T^2}) = o(1)$, given $N = o(T^2)$.

Variance

Consider the variance. We have,

$$\begin{aligned}\mathrm{var}(X) &= \frac{1}{N}\sum_{i=1}^{N}\mathrm{var}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_{it}(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\zeta_{is})^2) \\ &\quad +\frac{1}{NT^3}\sum_{i\neq j}\sum_{t,s,k,l,v,p\leq T}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}) = B_1 + B_2.\end{aligned}$$

B_1 can be bounded by the Cauchy-Schwarz inequality. Note that $Ee_{it} = E\zeta_{js} = 0$,

$$B_1 \leq \frac{1}{N}\sum_{i=1}^{N}E(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_{it}(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\zeta_{is})^2)^2 \leq \frac{1}{N}\sum_{i=1}^{N}[E(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_{it})^4]^{1/2}[E(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\zeta_{is})^8]^{1/2}.$$

Hence $B_1 = O(1)$.

We now show $\frac{1}{T}B_2 = o(1)$. Once this is done, it implies $\frac{1}{T}\mathrm{var}(X) = o(1)$. The proof of the lemma's first statement is then completed because $\frac{1}{T}EX^2 = \frac{1}{T}(EX)^2 + \frac{1}{T}\mathrm{var}(X) = o(1)$.

Note that $Ee_{it} = E\zeta_{is} = 0$, and when $t \neq s$, $e_{it} \perp \zeta_{js}$, $e_{it} \perp e_{js}$, $\zeta_{it} \perp \zeta_{js}$ (independent) for any $i, j \leq N$. Therefore, it is straightforward to verify that if the set $\{t, s, k, l, v, p\}$ contains more than three distinct elements, then $\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}) = 0$. Hence if we denote Ξ as the set of (t, s, k, l, v, p) such that $\{t, s, k, l, v, p\}$ contains no more than three distinct elements, then its cardinality satisfies: $|\Xi|_0 \leq CT^3$ for some $C > 1$, and

$$\sum_{t,s,k,l,v,p\leq T}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}) = \sum_{(t,s,k,l,v,p)\in\Xi}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}).$$

Hence $B_2 = \frac{1}{NT^3}\sum_{i\neq j}\sum_{(t,s,k,l,v,p)\in\Xi}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp})$. Let us partition Ξ into $\Xi_1 \cup \Xi_2$ where each element (t, s, k, l, v, p) in Ξ_1 contains exactly three distinct indices, while each element in Ξ_2 contains less than three distinct indices. We know that $\frac{1}{NT^3}\sum_{i\neq j}\sum_{(t,s,k,l,v,p)\in\Xi_2}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}) = O(\frac{1}{NT^3}N^2T^2) = O(\frac{N}{T})$, which implies

$$\frac{1}{T}B_2 = \frac{1}{NT^4}\sum_{i\neq j}\sum_{(t,s,k,l,v,p)\in\Xi_1}\mathrm{cov}(e_{it}\zeta_{is}\zeta_{ik}, e_{jl}\zeta_{jv}\zeta_{jp}) + O_p(\frac{N}{T^2}).$$

The first term on the right hand side can be written as $\sum_{h=1}^{5}B_{2h}$. Each of these five terms is defined and analyzed separately as below.

$$B_{21} = \frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t}Ee_{it}e_{jt}E\zeta_{is}^2E\zeta_{jl}^2 \leq O(\frac{1}{NT})\sum_{i\neq j}|Ee_{it}e_{jt}|.$$

Note that if $(\mathbf{\Sigma}_u)_{ij} = 0$, u_{it} and u_{jt} are independent, and hence $Ee_{it}e_{jt} = 0$. This implies

$\sum_{i\neq j}|Ee_{it}e_{jt}|\leq O(1)\sum_{i\neq j,(i,j)\in S_U}1=O(N)$. Hence $B_{21}=o(1)$.

$$B_{22}=\frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t}Ee_{it}\zeta_{it}E\zeta_{is}e_{js}E\zeta_{jl}^2.$$

By Lemma A.5, u_{js} and ξ_{is} are independent for $i\neq j$. Also, u_{js} and $\mathbf{f}_s$ are independent, which implies e_{js} and ζ_{is} are independent. So $Ee_{js}\zeta_{is}=0$. It follows that $B_{22}=0$.

$$\begin{aligned}B_{23} &= \frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t}Ee_{it}\zeta_{it}E\zeta_{is}\zeta_{js}Ee_{jl}\zeta_{jl}=O(\frac{1}{NT})\sum_{i\neq j}|E\zeta_{is}\zeta_{js}|\\ &=O(\frac{1}{NT})\sum_{i\neq j}|E\xi_{is}\xi_{js}E(1-\mathbf{f}_s'\mathbf{w})^2|=O(\frac{1}{NT})\sum_{i\neq j}|E\xi_{is}\xi_{js}|.\end{aligned}$$

By the definition $\boldsymbol{\xi}_s=\boldsymbol{\Sigma}_u^{-1}\mathbf{u}_s$, $\mathrm{cov}(\boldsymbol{\xi}_s)=\boldsymbol{\Sigma}_u^{-1}$. Hence $E\xi_{is}\xi_{js}=(\boldsymbol{\Sigma}_u^{-1})_{ij}$, which implies $B_{23}\leq O(\frac{N}{NT})\|\boldsymbol{\Sigma}_u^{-1}\|=o(1)$.

$$B_{24}=\frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t}Ee_{it}e_{jt}E\zeta_{is}\zeta_{js}E\zeta_{il}\zeta_{jl}=O(\frac{1}{T}),$$

which is analyzed in the same way as B_{21}.

Finally, $B_{25}=\frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t}Ee_{it}\zeta_{jt}E\zeta_{is}e_{js}E\zeta_{il}\zeta_{jl}=0$, because $E\zeta_{is}e_{js}=0$ when $i\neq j$, following from Lemma A.5. Therefore, $\frac{1}{T}B_2=o(1)+O(\frac{N}{T^2})=o(1)$.

A.6.2 Proof of Lemma 5.1, second statement

For notational simplicity, let $e_{ijt}=u_{it}u_{jt}-Eu_{it}u_{jt}$. Because of the serial independence and the Gaussianity, $\mathrm{cov}(e_{ijt},\zeta_{ls}\zeta_{nk})=0$ when $s,k\neq t$, for any $i,j,l,n\leq N$. In addition, define a set

$$H=\{(i,j)\in S_L: i\neq j\}.$$

Then by the sparsity assumption, $\sum_{(i,j)\in H}1=D_N=O(N)$. Now let

$$\begin{aligned}Z &= \frac{1}{\sqrt{NT}}\sum_{(i,j)\in H}\sum_{t=1}^{T}(u_{it}u_{jt}-Eu_{it}u_{jt})[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\xi_{is}(1-\mathbf{f}_s'\mathbf{w})][\frac{1}{\sqrt{T}}\sum_{k=1}^{T}\xi_{jk}(1-\mathbf{f}_k'\mathbf{w})]\\ &= \frac{1}{\sqrt{NT}}\sum_{(i,j)\in H}\sum_{t=1}^{T}e_{ijt}[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\zeta_{is}][\frac{1}{\sqrt{T}}\sum_{k=1}^{T}\zeta_{jk}]=\frac{1}{T\sqrt{NT}}\sum_{(i,j)\in H}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{k=1}^{T}e_{ijt}\zeta_{is}\zeta_{jk}.\end{aligned}$$

We respectively show $\frac{1}{T}(EZ)^2=o(1)=\frac{1}{T}\mathrm{var}(Z)$.

Expectation

The proof for the expectation is the same regardless of the type of $\mathbf{\Sigma}_u$ in Assumption 5.4, and is very similar to that of the first statement of Lemma 5.1. In fact,

$$EZ = \frac{1}{T\sqrt{NT}}\sum_{(i,j)\in H}\sum_{t=1}^T\sum_{s=1}^T\sum_{k=1}^T \mathrm{cov}(e_{ijt},\zeta_{is}\zeta_{jk}) = \frac{1}{T\sqrt{NT}}\sum_{(i,j)\in H}\sum_{t=1}^T \mathrm{cov}(e_{ijt},\zeta_{it}^2).$$

Because $\sum_{(i,j)\in H} 1 = O(N)$, $EZ = O(\sqrt{\frac{N}{T}})$. Thus $\frac{1}{T}(EZ)^2 = o(1)$.

Variance

For the variance, we have

$$\begin{aligned}\mathrm{var}(Z) &= \frac{1}{T^3N}\sum_{(i,j)\in H}\mathrm{var}(\sum_{t=1}^T\sum_{s=1}^T\sum_{k=1}^T e_{ijt}\zeta_{is}\zeta_{jk}) \\ &\quad + \frac{1}{T^3N}\sum_{(i,j)\in H,}\ \sum_{(m,n)\in H,(m,n)\neq(i,j),}\ \sum_{t,s,k,l,v,p\leq T}\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}) \\ &= A_1 + A_2.\end{aligned}$$

By the Cauchy-Schwarz inequality and the serial independence of e_{ijt},

$$\begin{aligned}A_1 &\leq \frac{1}{N}\sum_{(i,j)\in H} E[\frac{1}{\sqrt{T}}\sum_{t=1}^T e_{ijt}\frac{1}{\sqrt{T}}\sum_{s=1}^T\zeta_{is}\frac{1}{\sqrt{T}}\sum_{k=1}^T\zeta_{jk}]^2 \\ &\leq \frac{1}{N}\sum_{(i,j)\in H}[E(\frac{1}{\sqrt{T}}\sum_{t=1}^T e_{ijt})^4]^{1/2}[E(\frac{1}{\sqrt{T}}\sum_{s=1}^T\zeta_{is})^8]^{1/4}[E(\frac{1}{\sqrt{T}}\sum_{k=1}^T\zeta_{jk})^8]^{1/4}.\end{aligned}$$

So $A_1 = O(1)$.

Note that $Ee_{ijt} = E\zeta_{is} = 0$, and when $t\neq s$, $e_{ijt}\perp\zeta_{ms}$, $e_{ijt}\perp e_{mns}$, $\zeta_{it}\perp\zeta_{js}$ (independent) for any $i,j,m,n\leq N$. Therefore, it is straightforward to verify that if the set $\{t,s,k,l,v,p\}$ contains more than three distinct elements, then $\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}) = 0$. Hence for the same set Ξ defined as before, it satisfies: $|\Xi|_0 \leq CT^3$ for some $C>1$, and

$$\sum_{t,s,k,l,v,p\leq T}\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}) = \sum_{(t,s,k,l,v,p)\in\Xi}\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}).$$

We proceed by studying the two cases of Assumption 5.4 separately, and show that in both cases $\frac{1}{T}A_2 = o(1)$. Once this is done, because we have just shown $A_1 = O(1)$, then $\frac{1}{T}\mathrm{var}(Z) = o(1)$. The proof is then completed because $\frac{1}{T}EZ^2 = \frac{1}{T}(EZ)^2 + \frac{1}{T}\mathrm{var}(Z) = o(1)$.

When $D_N = O(\sqrt{N})$

Because $|\Gamma|_0 \leq CT^3$ and $|H|_0 = D_N = O(\sqrt{N})$, and $|\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np})|$ is bounded uniformly in $i, j, m, n \leq N$, we have

$$\frac{1}{T}A_2 = \frac{1}{T^4 N}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t,s,k,l,v,p\in\Xi}\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}) = O(\frac{1}{T}).$$

When $D_n = O(N)$, and $m_N = O(1)$

Similar to the proof of the first statement, for the same set Ξ that contains exactly three distinct indices in each of its element, (recall $|H|_0 = O(N)$)

$$\frac{1}{T}A_2 = \frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t,s,k,l,v,p\in\Xi_1}\mathrm{cov}(e_{ijt}\zeta_{is}\zeta_{jk}, e_{mnl}\zeta_{mv}\zeta_{np}) + O(\frac{N}{T^2}).$$

The first term on the right hand side can be written as $\sum_{h=1}^5 A_{2h}$. Each of these five terms is defined and analyzed separately as below. Before that, let us introduce a useful lemma.

The following lemma is needed when $\mathbf{\Sigma}_u$ has bounded number of nonzero entries in each row ($m_N = O(1)$). Let $|S|_0$ denote the number of elements in a set S if S is countable. For any $i \leq N$, let $A(i) = \{j \leq N : \mathrm{cov}(u_{it}, u_{jt}) \neq 0\} = \{j \leq N : (i,j) \in S_U\}$.

Lemma A.6. *Suppose $m_N = O(1)$. For any $i, j \leq N$, let $B(i,j)$ be a set of $k \in \{1, ..., N\}$ such that:*
(i) $k \notin A(i) \cup A(j)$
(ii) there is $p \in A(k)$ such that $\mathrm{cov}(u_{it}u_{jt}, u_{kt}u_{pt}) \neq 0$. Then $\max_{i,j\leq N}|B(i,j)|_0 = O(1)$.

Proof. First we note that if $B(i,j) = \emptyset$, then $|B(i,j)|_0 = 0$. If it is not empty, for any $k \in B(i,j)$, by the Gaussianity, u_{kt} is independent of (u_{it}, u_{jt}). Hence if $p \in A(k)$ is such that $\mathrm{cov}(u_{it}u_{jt}, u_{kt}u_{pt}) \neq 0$, we must have $p \in A(i)\cup A(j)$. In other words, there is $p \in A(i)\cup A(j)$ such that $\mathrm{cov}(u_{kt}, u_{pt}) \neq 0$, which implies $k \in A(p)$. Hence,

$$k \in \bigcup_{p\in A(i)\cup A(j)} A(p) \equiv M(i,j),$$

and thus $B(i,j) \subset M(i,j)$. Because $m_N = O(1)$, $\max_{i\leq N}|A(i)|_0 = O(1)$, which implies $\max_{i,j}|M(i,j)|_0 = O(1)$, yielding the result. □

Now we define and bound each of A_{2h}.

$$\begin{aligned} A_{21} &= \frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^T\sum_{s\neq t}\sum_{l\neq t,s} Ee_{ijt}e_{mnt}E\zeta_{is}\zeta_{js}E\zeta_{ml}\zeta_{nl} \\ &\leq O(\frac{1}{NT})\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j)} |Ee_{ijt}e_{mnt}| \end{aligned}$$

$$\leq O(\frac{1}{NT})\sum_{(i,j)\in H}(\sum_{m\in A(i)\cup A(j)}\sum_{n\in A(m)}+\sum_{m\notin A(i)\cup A(j)}\sum_{n\in A(m)})|\mathrm{cov}(u_{it}u_{jt},u_{mt}u_{nt})|.$$

The first term is $O(\frac{1}{T})$ because $|H|_0=O(N)$ and $|A(i)|_0$ is bounded uniformly in $i\leq N$. For the second term, if $m\notin A(i)\cup A(j)$, $n\in A(m)$ and $\mathrm{cov}(u_{it}u_{jt},u_{mt}u_{nt})\neq 0$, then $m\in B(i,j)$. Hence the second term is bounded by $O(\frac{1}{NT})\sum_{(i,j)\in H}\sum_{m\in B(i,j)}\sum_{n\in A(m)}|\mathrm{cov}(u_{it}u_{jt},u_{mt}u_{nt})|$, which is also $O(\frac{1}{T})$ by Lemma A.6. Hence $A_{21}=o(1)$.

Similarly, applying Lemma A.6 as for A_{21}, we can show

$$A_{22}=\frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq t,s}Ee_{ijt}e_{mnt}E\zeta_{is}\zeta_{ms}E\zeta_{jl}\zeta_{nl}=o(1).$$

The term A_{23} is defined as

$$\begin{aligned}A_{23} &= \frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq t,s}Ee_{ijt}\zeta_{it}E\zeta_{js}e_{mns}E\zeta_{ml}\zeta_{nl}\\ &\leq O(\frac{1}{NT})\sum_{j=1}^{N}\sum_{i\in A(j)}1\sum_{(m,n)\in H,(m,n)\neq(i,j)}|E\zeta_{js}e_{mns}|\\ &\leq O(\frac{2}{NT})\sum_{j=1}^{N}\sum_{n\in A(j)}|E\zeta_{js}e_{jns}|+O(\frac{1}{NT})\sum_{j=1}^{N}\sum_{m,n\neq j}|E\zeta_{js}e_{mns}|=a+b.\end{aligned}$$

Term $a=O(\frac{1}{T})$. For b, note that Lemma A.5 implies that when $m,n\neq j$, $u_{ms}u_{ns}$ and ξ_{js} are independent because of the Gaussianity. Also because $\mathbf{u}_s$ and $\mathbf{f}_s$ are independent, hence ζ_{js} and e_{mms} are independent, which implies that $b=0$. Hence $A_{23}=o(1)$.

The same argument also implies

$$A_{24}=\frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq t,s}Ee_{ijt}\zeta_{mt}E\zeta_{is}e_{mns}E\zeta_{il}\zeta_{nl}=o(1)$$

Finally,

$$\begin{aligned}A_{25} &= \frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq t,s}Ee_{ijt}\zeta_{it}E\zeta_{is}\zeta_{ms}Ee_{mnl}\zeta_{nl}\\ &\leq O(\frac{1}{NT})\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j)}|Ee_{ijt}\zeta_{it}E\zeta_{is}\zeta_{ms}Ee_{mnl}\zeta_{nl}|\\ &\leq O(\frac{1}{NT})\sum_{i=1}^{N}\sum_{m=1}^{N}|E\zeta_{is}\zeta_{ms}|\leq O(\frac{1}{NT})\sum_{i=1}^{N}\sum_{m=1}^{N}|(\mathbf{\Sigma}_u^{-1})_{im}|E(1-\mathbf{f}_s'\mathbf{w})^2\end{aligned}$$

$$\leq \quad O(\frac{N}{NT})\|\mathbf{\Sigma}_u^{-1}\|_1 = o(1).$$

In summary, $\frac{1}{T}A_2 = o(1) + O(\frac{N}{T^2}) = o(1)$. This completes the proof.

References

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

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

Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577-2604.

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

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672-684.

Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. Forthcoming in *Ann. Statist.*

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica.* **51** 1305-1324.

Connor, G. and Korajczyk, R. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance.* **48**, 1263-1291

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance.* **47** 427-465.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39**, 3320-3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Ser. B.* **75**, 603-680

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757-2790.

Gibbons, M., Ross, S. and Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica*, **57**, 1121-1152.

Keleijan, H. and Prucha, I. (2001). On the asymptotic distribution of the Moran I test statistic with applications *Journal of Econometrics*. **101**, 219-257

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254-4278.

MacKinlay, A.C. and Richardson, M. (1991). Using generalized method of moments to test mean-variance efficiency. *Journal of Finance*, **46**, 511-527.

Pesaran, M. H. and Yamagata, T. (2012). Testing CAPM with a large number of assets. *Manuscript*.

Ross, S.A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, **13**, 341-360.

Rothman, A., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177-186.

Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**, 1167-1179.