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 high-dimensional multi-factor pricing model, with the number of assets much larger than the length of time series. Most of the existing tests are based on a quadratic form of estimated alphas. They suffer from low powers, however, due to the accumulation of errors in estimating high-dimensional parameters that overrides the signals of non-vanishing alphas. To resolve this issue, we develop a new class of tests, called "power enhancement" tests. It strengthens the power of existing tests in important sparse alternative hypotheses where market inefficiency is caused by a small portion of stocks with significant alphas. The power enhancement component is asymptotically negligible under the null hypothesis and hence does not distort much the size of the original test. Yet, it becomes large in a specific region of the alternative hypothesis and therefore significantly enhances the power. In particular, we design a screened Wald-test that enables us to detect and identify individual stocks with significant alphas. We also develop a feasible Wald statistic using a regularized high-dimensional covariance matrix. By combining those two, our proposed method achieves power enhancement while controlling the size, which is illustrated by extensive simulation studies and empirically applied to the components in the S&P 500 index. Our empirical study shows that market inefficiency is primarily caused by merely a few stocks with significant alphas, most of which are positive, instead of a large portion of slightly mis-priced assets.

*Address: Department of Operations Research and Financial Engineering, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA. Department of Mathematics, University of Maryland, College Park, MD 20742, USA. E-mail: *jqfan@princeton.edu*, *yuanliao@umd.edu*, *jiaweiy@princeton.edu*. The research was partially supported by DMS-1206464, NIH R01GM100474-01 and NIH R01-GM072611.

Keywords: high dimensionality, approximate factor model, test for mean-variance efficiency, power enhancement, thresholding, covariance matrix estimation, Wald-test, screening

JEL code: C12, C33, C58

1 Introduction

Factor pricing model is one of the most fundamental results in finance. It postulates how financial returns are related to market risks, and has many important practical applications, including portfolio selection, fund performance evaluation, and corporate budgeting. It also includes the Capital Asset Pricing Model (CAPM) as a specific case.

The multi-factor pricing model was derived by Ross (1976) using the arbitrage pricing theory and by Merton (1973) using the Intertemporal CAPM. Let y_{it} be the excess return of the i-th asset at time t and $\mathbf{f}_t = (f_{1t}, ..., f_{Kt})'$ be the excess returns of K market risk factors. Then, the excess return has the following decomposition:

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

where $\mathbf{b}_i = (b_{i1}, ..., b_{iK})'$ is a vector of factor loadings and u_{it} represents the idiosyncratic error. The key implication from the multi-factor pricing theory is that the intercept α_i should be zero, known as "mean-variance efficiency", for any asset i. An important question is then if such a pricing theory can be validated by empirical data, namely whether the null hypothesis

$$H_0 : \boldsymbol{\alpha} = 0, \tag{1.2}$$

is consistent with empirical data, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)'$ is the vector of intercepts for all N financial assets.

Most of the existing tests to the problem (1.2) 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. The Wald statistic, for instance, takes the form $a_T\widehat{\boldsymbol{\alpha}}'\widehat{\boldsymbol{\Sigma}}_u^{-1}\widehat{\boldsymbol{\alpha}}$, where $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ is the estimated inverse of the error covariance, and a_T is a positive number that depends on the factors $\mathbf{f}_t$ only. Other prominent examples are the test given by Gibbons, Ross and Shaken (1989, GRS test), the GMM test in MacKinlay and Richardson (1991), and the likelihood ratio test in Beaulieu, Dufour and Khalaf (2007, BDK test), all in quadratic forms. See Sentana (2009) for an overview.

The above tests are applicable only when the number of assets N is much smaller than the length of the time series T. When $N \geq T$, for instance, the sample covariance $\widehat{\boldsymbol{\Sigma}}_u$ becomes degenerate. However, one typically picks a testing period of $T = 60$ monthly data and does not increase the testing period any longer, because the factor pricing model is technically a one-period model whose factor loadings can be time-varying. As a result, it would often be

the case when the number of assets N is much larger than T. To overcome the difficulty, Pesaran and Yamagata (2012, PY test) proposed to ignore the correlations among assets and constructed a test statistic under working independence: choosing $\mathbf{V} = \text{diag}(\widehat{\mathbf{\Sigma}}_u)^{-1}$ to avoid the issue of invertibility. They obtained the asymptotic normality of the standardized quadratic form: $(W - EW)/\sqrt{\text{var}(W)}$. A closely related idea was employed by Chen and Qin (2010) and Chen, Zhang and Zhong (2010) to ameliorate high-dimensional Hotelling test. However, even such a simplified quadratic test suffers from a low power in a high-dimension-low-sample-size situation, as we now explain.

For simplicity, let us temporarily assume that $\{\mathbf{u}_t\}_{t=1}^T$ are i.i.d. Gaussian and $\mathbf{\Sigma}_u = \text{cov}(\mathbf{u}_t)$ is known, where $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})$. In this case, the Wald test statistic W equals $T\widehat{\boldsymbol{\alpha}}'\mathbf{\Sigma}_u\widehat{\boldsymbol{\alpha}}$ times a constant that only depends on the factors. Under H_0, it is χ^2_N distributed, with the critical value $\chi^2_{N,q}$, which is of order N, at significant level q. The test has no power at all when $T\boldsymbol{\alpha}'\mathbf{\Sigma}_u\boldsymbol{\alpha} = o(N)$ or $\|\boldsymbol{\alpha}\|^2 = o(N/T)$, assuming that $\mathbf{\Sigma}_u$ has bounded eigenvalues. This is not unusual for the high-dimension-low-sample-size situation we encounter, where there are thousands of assets to be tested over a relatively short time period (e.g. 60 monthly data). And it is especially the case when there are only a few significant alphas that arouse market inefficiency. By a similar argument, this problem can not be rescued by using the PY test, or any other genuine quadratic statistic, which are powerful only when a non-negligible fraction of assets are mispriced. Indeed, the factor N above reflects the noise accumulation in estimating N parameters of $\boldsymbol{\alpha}$.

We resolve the above problem by introducing a novel technique, called the "power enhancement" (PEM) technique. Let J_1 be a test statistic that has a correct asymptotic size (e.g., GRS, PY, BDK), which may suffer from small powers in all directions. Let us augment the test by adding a PEM component $J_0 \geq 0$ with the following two properties:

(a) Under H_0, J_0 has an order of magnitude smaller than J_1. If J_1 has been normalized, this requires that $J_0 \xrightarrow{P} 0$ under H_0.

(b) J_0 does not converge to zero and even diverges when the true parameters fall in a subset $\mathbf{\Theta}$ of the alternative hypothesis.

The constructed PEM test is then of the form

$$J = J_0 + J_1. \tag{1.3}$$

Property (a) shows that the size distortion is negligible, property (b) guarantees significant

power improvement on the set $\boldsymbol{\Theta}$, and the nonnegativity property of J_0 ensures that J is always at least as powerful as J_1.

As an example, we construct a screened Wald-test defined by

$$J_0 = \widehat{\boldsymbol{\alpha}}'_{\widehat{S}} \mathbf{V}_{\widehat{S}} \widehat{\boldsymbol{\alpha}}_{\widehat{S}} \tag{1.4}$$

where $\widehat{\boldsymbol{\alpha}}_{\widehat{S}} = (\widehat{\alpha}_j : |\widehat{\alpha}_j| > \delta_T)$ is a subvector of $\widehat{\boldsymbol{\alpha}}$ whose individual magnitude exceeds the threshold δ_T, and $\mathbf{V}_{\widehat{S}}$ is the corresponding submatrix of a weight matrix $\mathbf{V}$. 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 $\max_{j \le N} |\alpha_j| \gg \delta_T$. This enhances the power of J_1 in the sparse alternatives. The PEM component can be combined with either the PY test or the Wald test; the latter needs further development when $N > T$.

Another contribution of the paper is to develop an operational Wald statistic even when $N/T \to \infty$ and $\boldsymbol{\Sigma}_u$ is non-diagonal. The statistic is based on a regularized sparse estimator of $\boldsymbol{\Sigma}_u$. We show that as $N, T \to \infty$, with the covariance estimator $\widehat{\boldsymbol{\Sigma}}_u^{-1}$, the standardized Wald test

$$J_{sw} = \frac{a_T \widehat{\boldsymbol{\alpha}}' \widehat{\boldsymbol{\Sigma}}_u^{-1} \widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0,1)$$

under the null hypothesis for a given normalization factor a_T. This feasible 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,

$$a_T \widehat{\boldsymbol{\alpha}}' (\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}) \widehat{\boldsymbol{\alpha}} = o_P(\sqrt{N}). \tag{1.5}$$

Note that a simple inequality $|\widehat{\boldsymbol{\alpha}}'(\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1})\widehat{\boldsymbol{\alpha}}| \le \|\widehat{\boldsymbol{\alpha}}\|^2 \|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\|$ yields a too crude bound of order $O(N/T)$, due to the error accumulation in $\|\widehat{\boldsymbol{\alpha}}\|^2$ under high dimensions. Instead, we have developed a new technical strategy to prove (1.5), which would also be potentially useful in high-dimensional inference using GMM methods when one needs to estimate the optimal weight matrix of high dimensions. We further take $J_1 = J_{sw}$, and combine it with our PEM J_0 in (1.4) to propose a power enhancement test. It is much more powerful than existing tests, while maintaining the same asymptotic null distribution. We show that the power is enhanced uniformly over the true $\boldsymbol{\alpha}$.

Moreover, the screening step in the construction of J_0 also identifies the individual stocks with significant alphas. They provide us useful information on which stocks are mis-priced and contribute importantly to the market inefficiency. 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 market inefficiency is indeed primarily caused by a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market. Most of the significant alphas are positive, which result in extra returns. The market inefficiency is further evidenced by our newly created portfolio based on the PEM test, which outperforms the S&P500 index.

Importantly, the proposed PEM is widely applicable in many high-dimensional testing problems in applied econometrics. For instance, in panel data models

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \eta_i + u_{it},$$

we are often interested in testing slope homogeneity $H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}, \forall i \leq N$ (e.g., Phillips and Sul 2003, Breitung et al. 2013), and cross-sectional independence $H_0 : \text{cov}(u_{it}, u_{jt}) = 0, \forall i \neq j$ (Baltagi et al. 2012, Sarafidis et al. 2009, Pesaran et al. 2008). The proposed PEM principle carries over to these cases.

It is worth pointing out that in the statistical literature, there are many studies on testing the mean vector of a high-dimensional Gaussian vector (e.g., Hall and Jin 2010, Srivastava and Du 2008, etc.). These methods are not applicable here because they are mainly concerned with a special type of sparse alternatives, where the correlations among observations should decay fast enough. In contrast with our settings here, their estimated alphas can be strongly correlated.

The remainder of the paper is organized as follows. Section 2 sets up the preliminaries and discusses the limitations of traditional tests. Section 3 proposes the power enhancement method, derives the asymptotic behaviors of the screening statistic and analyzes its performances under different alternatives. Section 4 combines the PEM with the working-independent quadratic form. Section 5 studies a high-dimensional Wald-test. Section 6 discusses extended applications of PEM in panel data models and GMM. Simulation results are presented in Section 7, along with an empirical application to the stocks in the S&P 500 index in Section 8. Section 9 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 $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}^T\mathbf{A})$ 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, we denote $|S|_0$ as the number of elements in a set S.

2 Factor models and traditional tests

The multi-factor model (1.1) can be more compactly expressed as

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

where $\mathbf{y}_t = (y_{1t}, ..., y_{Nt})'$ is the excess return vector at time t; $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_N)'$ is a loading matrix, and $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})$ denotes the idiosyncratic component. Here both $\mathbf{y}_t$ and $\mathbf{f}_t$ are observable. Our goal is to test the hypothesis

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

namely the multi-factor pricing model is consistent with the data. As argued in the introduction, we are particularly interested in a high-dimensional situation where $N = \dim(\boldsymbol{\alpha})$ can be much larger than T, that is, $N/T \to \infty$.

We set up the model to have an approximate factor structure as in Chamberlain and Rothschild (1983), in which the covariance matrix $\boldsymbol{\Sigma}_u = \mathrm{cov}(\mathbf{u}_t)$ is sparse. Since the common factors have substantially mitigated the co-movement across the whole panel, a particular asset's idiosyncratic volatility is usually correlated significantly only with a couple 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 enables us to reliably estimate the error covariance later.

2.1 Wald-type tests

We can run OLS for regressions stock by stock to estimate $\boldsymbol{\alpha}$. Denote by $\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t$, $\mathbf{w} = (\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t')^{-1}\bar{\mathbf{f}}$, and $a_T = T(1-\bar{\mathbf{f}}'\mathbf{w})$. Then, the OLS estimator can be expressed as

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

Simple calculations yield

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

Assuming no serial correlation among $\{\mathbf{u}_t\}_{t=1}^{T}$, the conditional covariance of $\widehat{\boldsymbol{\alpha}}$ is $\boldsymbol{\Sigma}_u/a_T$, given the factors. If $\boldsymbol{\Sigma}_u$ is known, the classical Wald test statistic is

$$W = a_T\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}}. \tag{2.1}$$

When $N < T - K - 1$ (recall $K = \dim(\mathbf{f}_t)$), $\boldsymbol{\Sigma}_u$ can be estimated and replaced by the sample covariance matrix of the residual vector, and the resulting test statistic is then in line with the well-known test by Gibbons et al. (1989, GRS). Under the Gaussian assumption, GRS obtained the exact finite sample distribution of this test statistic. The asymptotic null distribution is χ_N^2-distribution when N is fixed. When N diverges, the traditional asymptotic theory for W does not apply directly. Instead, Pesaran and Yamagata (2012) developed an alternative asymptotic null distribution for the Wald statistic. They showed that under some regularity conditions,

$$J_1 = \frac{a_T\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0,1). \tag{2.2}$$

as $N \to \infty$. Hence at the significant level $q \in (0,1)$, $P(J_1 > z_q|H_0) \to q$ with critical value z_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 a Wald-type statistic, or quadratic test because W is a quadratic form of $\widehat{\boldsymbol{\alpha}}$.

2.2 Two main challenges of quadratic test

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 (2.2) only partially solves the problem. Specifically, there are two main challenges.

The first challenge arises from estimating $\mathbf{\Sigma}_u^{-1}$ in the presence of error cross-sectional correlations. It is well known that the sample residual covariance matrix becomes singular when $N > T - K - 1$. Even if $N < T - K - 1$, replacing $\mathbf{\Sigma}_u^{-1}$ 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 when the data deviates slightly from the normality.

Another challenge comes from the loss of power due to estimating the high-dimensional parameter vector $\boldsymbol{\alpha}$. Even when $\mathbf{\Sigma}_u^{-1}$ is known so that W is feasible, it still has very low powers against various alternative hypotheses when N is large. This again can be ascribed to the noise accumulation as illustrated in the following example.

Example 2.1. Suppose that $\mathbf{\Sigma}_u = \mathbf{I}_N$, the identity matrix, and the data are serially independent. Conditioning on the observed factors, W is distributed as $\chi_N^2(a_T\|\boldsymbol{\alpha}\|^2)$, a non-central χ^2-distribution with degree of freedom N and noncentrality parameter $a_T\|\boldsymbol{\alpha}\|^2$. Under the null hypothesis, $W \sim \chi_N^2$, which has mean N and standard deviation $\sqrt{2N}$. Now, consider the power of the test at the alternative, for some $r < N$,

$$H_a : \alpha_i = 0, \text{ for } i > r.$$

Then, the test statistic $W \sim \chi_N^2(a_T \sum_{i=1}^r \alpha_i^2)$ under H_a, which has mean $a_T \sum_{i=1}^r \alpha_i^2 + N$. Thus, if all α_i's are bounded, the test statistic has no power against H_a when $rT/\sqrt{N} = o(1)$ or $r = o(\sqrt{N}/T)$. ■

In the previous example, there are only a few non-vanishing alphas in the alternative, whose signals are dominated by the aggregated high-dimensional estimation errors: $T\sum_{i>r} \widehat{\alpha}_i^2$. On the other hand, the first r assets with non-vanishing alphas (fixed constants) are actually detectable (e.g., by confidence intervals) when $(\log N)/T$ converges, which should be considered when constructing the test. The above simple example illustrates how noise accumulation of the Wald test masks the power of the test. Such a result holds more generally as shown in the following Theorem.

Theorem 2.1. *Suppose that Assumption 3.2 below holds, and both* $\|\mathbf{\Sigma}_u\|_1$ *and* $\|\mathbf{\Sigma}_u^{-1}\|_1$ *are*

bounded. When $T = o(\sqrt{N})$, the test J_1 in (2.2) has low power at the following alternative:

$$H_a : \textit{there are at most } r = o(\frac{\sqrt{N}}{T}) \textit{ non-vanishing but bounded } \alpha_j\textit{'s.}$$

More precisely, $P(J_1 > z_q | H_a) \leq 2q + o(1)$ for any significant level $q \in (0, 0.5)$.

A more sensible approach is to focus on those alternatives that may have only a few nonzero alphas compared to N. This is particularly interesting when the market inefficiency is primarily caused by a minority of mis-priced stocks rather than systematic mis-pricing of the whole market. In what follows, we develop a new testing procedure that significantly improves the power of the Wald-type test against such sparse alternatives.

3 Power Enhancement

Traditional tests of factor pricing models are not powerful unless there are enough stocks that have non-vanishing alphas. Even if some individual assets are significantly mis-priced, their non-trivial contributions to the test statistic are insufficient to reject the null hypothesis. This problem can be resolved by introducing a power enhancement component (PEM) J_0 to the traditional test J_1 as in (1.3). The PEM J_0 is designed to detect sparse alternatives with significant individual alphas.

3.1 Screened Wald Statistic

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

$$\widehat{S} = \left\{ j : \frac{|\widehat{\alpha}_j|}{\widehat{\sigma}_j} > \delta_T, j = 1, ..., N \right\}, \tag{3.1}$$

where $\widehat{\alpha}_j$ is the OLS estimator and $\widehat{\sigma}_j^2 = \sum_{t=1}^T \widehat{u}_{jt}^2 / a_T$ is T times the estimated variance of $\widehat{\alpha}_j$, with $\widehat{u}_{jt}$ being the regression residuals. Denote a subvector of $\widehat{\boldsymbol{\alpha}}$ by

$$\widehat{\boldsymbol{\alpha}}_{\widehat{S}} = (\widehat{\alpha}_j : j \in \widehat{S}),$$

the screened-out alpha estimators, which can be interpreted as estimated alphas of mis-priced stocks. Let $\widehat{\boldsymbol{\Sigma}}_u$ be a consistent estimator of $\boldsymbol{\Sigma}_u$ to be defined later, and $\widehat{\boldsymbol{\Sigma}}_{\widehat{S}}$ the submatrix

of $\widehat{\mathbf{\Sigma}}_u$ formed by the rows and columns whose indices are in $\widehat{S}$. So $\widehat{\mathbf{\Sigma}}_{\widehat{S}}/a_T$ is an estimated conditional covariance matrix of $\widehat{\boldsymbol{\alpha}}_{\widehat{S}}$, given the common factors and $\widehat{S}$.

With the above notation, we define our screened Wald statistic as

$$J_0 = \sqrt{N} a_T \widehat{\boldsymbol{\alpha}}'_{\widehat{S}} \widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1} \widehat{\boldsymbol{\alpha}}_{\widehat{S}}, \tag{3.2}$$

where $a_T = T(1 - \bar{\mathbf{f}}'\mathbf{w})$. A similar idea of this kind of thresholding test appears in Fan (1996). 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

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

With this choice of δ_T, if we define, for $\sigma_j^2 = (\mathbf{\Sigma}_u)_{jj}/(1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t)$,

$$S = \left\{ j : \frac{|\alpha_j|}{\sigma_j} > 2\delta_T, j = 1, ..., N \right\}, \tag{3.3}$$

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

The screened Wald statistic depends on a nonsingular covariance matrix $\widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1}$ that consistently estimates $\mathbf{\Sigma}_S^{-1}$. The matrix $\mathbf{\Sigma}_S$ is a submatrix of $\mathbf{\Sigma}_u$ corresponding to the indices in S, and can be large when the set S is large. So it is difficult to estimate in general. To obtain an operational $\widehat{\mathbf{\Sigma}}_{\widehat{S}}^{-1}$, we assume $\mathbf{\Sigma}_u$ to be sparse and estimate it by thresholding (Bickel and Levina, 2008). For the sample covariance $s_{ij} = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt}$, let the thresholded sample covariance as

$$(\widehat{\mathbf{\Sigma}}_u)_{ij} = \begin{cases} s_{ij}, & \text{if } i = j, \\ \theta_{ij}(s_{ij}), & \text{if } i \neq j, \end{cases} \tag{3.4}$$

where $\theta_{ij}(\cdot)$ is a generalized thresholding function (Antoniadis and Fan, 2001, Rothman et al. 2009), with threshold value $h_{ij} = C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2}$ for some constant $C > 0$, designed to keep the sample correlation whose magnitude exceeds $C(\frac{\log N}{T})^{1/2}$. In particular, when the hard-thresholding function $\theta_{ij}(x) = x1\{|x| > h_{ij}\}$ is used, this is the estimator proposed by Fan et al. (2011). Many other thresholding functions such as soft-thresholding and SCAD can also be employed. In general, $\theta_{ij}(\cdot)$ should satisfy (Antoniadis and Fan, 2001):

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

(ii) $|\theta_{ij}(z) - z| \leq h_{ij}$;

(iii) there are constants $a > 0$ and $b > 1$ such that $|\theta_{ij}(z) - z| \leq a h_{ij}^2$ if $|z| > b h_{ij}$.

The thresholded error covariance matrix estimator sets most of the off-diagonal estimation noises in $(\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt})$ to zero, and produces a strictly positive definite matrix even when $N > T$. The constant C in the threshold can be selected in a data-driven way, which we refer to Fan et al. (2013) and Cai and Liu (2011). Moreover, it can be shown that $\widehat{\mathbf{\Sigma}}_u^{-1}$ consistently estimates $\mathbf{\Sigma}_u^{-1}$ under the operator norm.

Alternatively, we can use the working-independent version of the quadratic statistic, replacing $\widehat{\mathbf{\Sigma}}_{\widehat{S}}$ by $\widehat{\mathbf{D}}_{\widehat{S}} = \text{diag}\{s_{jj} : j \in \widehat{S}\}$. This is particularly convenient when $\mathbf{\Sigma}_u$ is close to diagonal, and corresponds to taking $h_{ij} = (s_{ii}s_{jj})^{1/2}$ in the thresholded covariance estimator with the hard-thresholding rule. The screening working-independent test statistic is then defined as

$$J_0 = \sqrt{N} a_T \widehat{\boldsymbol{\alpha}}_{\widehat{S}} \widehat{\mathbf{D}}_{\widehat{S}}^{-1} \widehat{\boldsymbol{\alpha}}_{\widehat{S}} = \sqrt{N} a_T \sum_{j \in \widehat{S}} \widehat{\alpha}_j^2 s_{jj}^{-1}. \tag{3.5}$$

3.2 Power enhancement test

The screened statistic J_0 is powerful in detecting stocks with large individual alphas. However, J_0 equals zero with probability approaching one under the null hypothesis. Hence, it can not live on its own to obtain the right size of the test, but can be combined with any other standard test statistics to achieve a proper size.

Suppose J_1 is a test statistic for $H_0 : \boldsymbol{\alpha} = 0$ and it has high power on a subset $\Omega_1 \subset \mathbb{R}^N \backslash \mathbf{0}$. Our power enhancement test is simply

$$J = J_0 + J_1.$$

Since $J_0 = 0$ with probability approaching one under the null hypothesis, the null distribution of J is asymptotically determined by that of J_1. Let F_q be the critical value for J_1 under the significant level q, that is, $P(J_1 > F_q | H_0) \leq q$. Our proposed PEM J-test then rejects H_0 when $J > F_q$. The size is not much distorted since

$$P(J > F_q | H_0) = P(J_1 > F_q | H_0) + o(1) \leq q + o(1).$$

Since $J \geq J_1$, the power of the enhanced test J is no smaller than that of J_1. On the other hand, if J_0 has high power on a subset $\Omega_0 \subset \mathbb{R}^N \setminus \mathbf{0}$ in the sense that $J_0 \xrightarrow{P} \infty$ for $\boldsymbol{\alpha} \in \Omega_0$, then J has power approaching one on the set Ω_0. In other words, we show that the PEM J-test has high power uniformly on the region $\Omega_0 \cup \Omega_1$. The power is enhanced whenever $\Omega_0 \cup \Omega_1$ is strictly larger than Ω_1. In high-dimensional testing problems, due to the error accumulations, Ω_0 and Ω_1 can be quite different.

To formally establish the aforementioned power enhancement properties, we need the sparsity of $\mathbf{\Sigma}_u$ in order for the thresholded covariance matrix estimator to work. This leads to defining a generalized sparsity measure

$$m_N = \max_{i \leq N} \sum_{j=1}^{N} |(\mathbf{\Sigma}_u)_{ij}|^k, \qquad \text{for some } k \in [0, 1). \tag{3.6}$$

Note that when $k = 0$, m_N is the maximum number of non-vanishing entries in each row (or column, with usual convention $0^0 = 0$). We assume $\log N = o(T)$ and that

Assumption 3.1. *There is $k \in [0, 1)$ such that*

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

Under the above generalized sparsity assumption, the thresholded covariance estimator $\widehat{\mathbf{\Sigma}}_u$ is positive definite, and consistently estimates $\mathbf{\Sigma}_u$ under the operator norm. When $\mathbf{\Sigma}_u$ is block diagonal with finite block sizes, Assumption 3.1 holds with bounded m_T for any $k \in [0, 1)$. Sparsity is one of the commonly used assumptions on high-dimensional covariance matrix estimation, which has been extensively studied recently. 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 α-mixing coefficient

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

Assumption 3.2. *(i)* Strict stationarity*: $\{\mathbf{f}_t, \mathbf{u}_t\}_{t \geq 1}$ is strictly stationary, $E\mathbf{u}_t = \mathbf{0}$ and $E\mathbf{u}_t\mathbf{f}_t = \mathbf{0}$. In addition, for $s \neq t$, $E\mathbf{u}_t\mathbf{u}_s' = \mathbf{0}$, and $E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t < 1$.*

(ii) There exist constants $c_1, c_2 > 0$ *such that* $\max_{i \le N} \|\mathbf{b}_i\| < c_2$,

$$c_1 < \lambda_{\min}(\mathbf{\Sigma}_u) \le \lambda_{\max}(\mathbf{\Sigma}_u) < c_2, \quad \textit{and} \quad c_1 < \lambda_{\min}(\mathrm{cov}(\mathbf{f}_t)) \le \lambda_{\max}(\mathrm{cov}(\mathbf{f}_t)) < c_2.$$

(iii) Exponential tail*: There exist* $r_1, r_2 > 0$, *and* $b_1, b_2 > 0$, *such that for any* $s > 0$,

$$\max_{i \le N} P(|u_{it}| > s) \le \exp(-(s/b_1)^{r_1}), \quad \max_{i \le K} P(|f_{it}| > s) \le \exp(-(s/b_2)^{r_2}).$$

(iv) Strong mixing*: There exists* $r_3 > 0$ *such that* $r_1^{-1} + r_2^{-1} + r_3^{-1} > 1$, *and* $C > 0$ *satisfying: for all* $T \in \mathbb{Z}^+$,

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

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 $\mathbf{\Sigma}_u / a_T$ given the factors. 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 with serial correlation. On the other hand, we allow the factors to be weakly dependent via the strong mixing condition, which holds when factors follow a VAR model. Also, it is always true that $E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t \le 1$. 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 screening Wald statistic J_0, and provides sufficient conditions for the selection consistency. Recall that $\widehat{S}$ and S are defined in (3.1) and (3.3) respectively. Define the "grey area set" as

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

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

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

In addition, 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 *empty grey area* condition.

Assumption 3.3. Empty grey area*: The alternative hypothesis H_a satisfies $\mathcal{G} = \emptyset$.*

The empty 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 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 screening, 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 $\mathcal{G}$ with probability approaching one.

Corollary 3.1. *Under the assumptions of Theorem 3.1 and Assumption 3.3, we have*

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

Note that the set S contains all individual stocks with large value of alphas and is the set of stocks that we wish to identify. Theorem 3.1 shows that they can be identified with no missing discoveries and Corollary 3.1 further asserts that the identification is consistent with no false discoveries either.

Define

$$\Omega_0 = \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j \le N} |\alpha_j| > 2\delta_T \max_{j \le N} \sigma_j\}.$$

Below we write $P(\cdot|\boldsymbol{\alpha})$ to denote the probability measure if the true alpha is $\boldsymbol{\alpha}$. A test is said to have high power uniformly on a set $A \subset \mathbb{R}^N \setminus \mathbf{0}$ if $\inf_{\boldsymbol{\alpha} \in A} P(\text{reject}|\boldsymbol{\alpha}) \to 1$.

We now present the asymptotic behavior of the PEM test.

Theorem 3.2. *Suppose* $\log N = o(T)$*, and Assumptions 3.1-3.3 hold. In addition, suppose there is a test J_1 such that (i) it has an asymptotic non-degenerate null distribution F; (ii) its critical region takes the form $J_1 > F_q$ for the level $q \in (0, 1)$, and (iii) it has high power uniformly on some set Ω_1. Then the PEM test $J = J_0 + J_1$ has the asymptotic null distribution F and high power uniformly on the set $\Omega_0 \cup \Omega_1$:*

$$\inf_{\boldsymbol{\alpha} \in \Omega_0 \cup \Omega_1} P(J > C|\boldsymbol{\alpha}) \to 1.$$

Theorem 3.2 shows that J_1 and J both have the critical regions $J > F_q$ and $J_1 > F_q$ respectively, where F_q is the qth quantile of F. It follows immediately from $J \ge J_1$ that

$P(J > F_q|\boldsymbol{\alpha}) \geq P(J_1 > F_q|\boldsymbol{\alpha})$, which means J is at least as powerful as J_1. In addition, J_0 plays exactly the role of power enhancement, in the sense that (i) it does not affect much the null distribution, and (ii) it substantially increases the power on Ω_0.

4 PEM for working-independent quadratic test

PEM can be combined with many tests. In this section, we consider the working-independent quadratic test recently developed by Pesaran and Yamagata (2012) in the current context. Without PEM, Theorem 2.1 shows that the working-independent test can suffer from low powers in high-dimensional testing problems.

4.1 Working-independent quadratic statistic

To avoid the degeneracy of the sample covariance $\widehat{\boldsymbol{\Sigma}}_u$ in normalizing the Wald statistic, one can simply ignore the cross-sectional correlations and utilize the diagonal matrix $\widehat{\mathbf{D}} = \text{diag}(\widehat{\boldsymbol{\Sigma}}_u)$ to normalize the quadratic form:

$$W = a_T \widehat{\boldsymbol{\alpha}}' \widehat{\mathbf{D}}^{-1} \widehat{\boldsymbol{\alpha}}, \qquad \widehat{\mathbf{D}} = \text{diag}\{s_{11}, ..., s_{NN}\}.$$

Note that W is a sum of N correlated individual squared t-statistics. One needs to calculate and estimate both $E(W)$ and $\text{var}(W)$ under the null hypothesis in order to establish the asymptotic normality. As $\widehat{\mathbf{D}}$ ignores the off-diagonal entries of $\boldsymbol{\Sigma}_u$, W is no longer χ^2 even under the normal assumption when $\boldsymbol{\Sigma}_u$ is indeed non-diagonal. Hence, efforts are needed to calculate the first two moments of W. This has been achieved by Pesaran and Yamagata (2012) and leads to a feasible version of the normalized test:

$$J_{wi} = \frac{W - N}{\sqrt{2N(1 + e_T)}} \tag{4.1}$$

where

$$e_T = N^{-1} \sum_{i \neq j} \hat{\rho}_{ij}^2 I(\hat{\rho}_{ij}^2 > c_T) \quad \text{with} \quad c_T = \frac{1}{T} \Phi^{-1}(1 - c/N)$$

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), it can be shown that J_{wi} is asymptotically standard normal, as $N, T \to \infty$. We call J_{wi} as the

working-independent quadratic statistic. Pesaran and Yamagata (2012) proposed a slightly different (but asymptotically equivalent) statistic that corrects the finite sample bias. For example, the expected value of squared t-statistic with degree of freedom ν is $\nu/(\nu-2)$. Therefore, the numerator of (4.1) should be centerized by $N(N-K-1)/(N-K-3)$ instead of N.

4.2 Asymptotic results

We now study the power of the PEM test $J = J_0 + J_{wi}$ where J_0 is given by either (3.2) or (3.5). It is summarized in the following theorem.

Theorem 4.1. *Suppose* $\log N = o(T)$ *and Assumptions 3.1-3.3 hold. Also,* $\sum_{i\neq j}(\boldsymbol{\Sigma}_u)_{ij}^2 = O(N)$. *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) PEM test J *has high power uniformly on the set*

$$\Omega_0 \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\} \equiv \Omega_0 \cup \Omega_1,$$

that is, for any $q \in (0,1)$, $\inf_{\boldsymbol{\alpha}\in\Omega_0\cup\Omega_1} P(J > z_q|\boldsymbol{\alpha}) \to 1$.

We see that including J_0 in the PEM test does not alter the asymptotic null distribution, but it significantly enhances the power of J_{wi} on the set Ω_0. The set $\Omega_1 = \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\}$ itself is the region where J_{wi} has a high power, but this is a restrictive region when $N > T$. For instance, it rules out the type of alternatives in which there are finitely many non-vanishing alphas, but such a sparse scenario is included in Ω_0. Hence, the PEM J will be able to detect it.

5 High-dimensional Wald test and its PEM

The working-independent quadratic test is a simple way of dealing with the singularity of the sample covariance matrix of $\mathbf{u}_t$ when $N > T$, and to avoid estimating the high-dimensional covariance matrix $\boldsymbol{\Sigma}_u$. On the other hand, when $\boldsymbol{\Sigma}_u$ were known, one typically prefers the Wald test or equivalently its normalized version

$$\frac{a_T \widehat{\boldsymbol{\alpha}}' \boldsymbol{\Sigma}_u^{-1} \widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}}, \tag{5.1}$$

which has an asymptotic null distribution $\mathcal{N}(0,1)$. A natural question is that when $\mathbf{\Sigma}_u$ admits a sparse structure, can it be estimated accurately enough for a substitution into (5.1)? The answer is affirmative if $N \log N = o(T^2)$, and still we can allow $N/T \to \infty$. However, such a simple question is far more technically involved than anticipated, as we now explain.

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

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

where $m_N = \max_{i \le N} \sum_{j=1}^N 1\{(\mathbf{\Sigma}_u)_{ij} \neq 0\}$. Using the lower bound derived by Cai and Zhou (2012), we infer that the convergence rate is minimax optimal for the sparse covariance estimation.

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

$$T\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}/\sqrt{N} = o_P(1).$$

Note that when $\boldsymbol{\alpha} = 0$, with careful analysis, $\|\widehat{\boldsymbol{\alpha}}\|^2 = O_P(N/T)$. Using this and (5.2), by the Cauchy-Schwartz inequality, we have

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

We see that even if m_N is bounded, it still requires $N \log N = o(T)$, which is basically a low-dimensional scenario.

The above simple derivation uses, however, a very crude Cauchy-Schwartz bound, which accumulates too many estimation errors in $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|^2$ (with $\boldsymbol{\alpha} = 0$ under H_0) 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 $\widehat{\boldsymbol{\alpha}}$ "average down" the accumulated estimation errors, and result in an improved rate of convergence. The formalization of this argument requires subtle technical details and further regularity conditions. These are formally presented in the following subsection.

5.2 Wald test and power enhancement

To simplify our discussion, let us focus on the serially independent Gaussian case.

Assumption 5.1. *Suppose:*
(i) $\mathbf{u}_t$ *is distributed as* $\mathcal{N}(0, \mathbf{\Sigma}_u)$*, where both* $\|\mathbf{\Sigma}_u\|_1$ *and* $\|\mathbf{\Sigma}_u^{-1}\|_1$ *are bounded;*
(ii) $\{\mathbf{u}_t, \mathbf{f}_t\}_{t \leq T}$ *is independent across* t*, and* $\mathbf{u}_t$ *and* $\mathbf{f}_t$ *are also independent.*
(iii) $\min_{(\mathbf{\Sigma}_u)_{ij} \neq 0} |(\mathbf{\Sigma}_u)_{ij}| \gg \sqrt{\frac{\log N}{T}}$.

Condition (iii) requires the minimal signal for the nonzero components be larger than the noise level, so that nonzero components are not thresholded off when estimating $\mathbf{\Sigma}_u$. Assumption 5.1 can be relaxed if u_{it} is i.i.d. across i and t. On the other hand, the imposed assumption allows investigations of non-diagonal covariance $\mathbf{\Sigma}_u$, whose off-diagonal structure can be unknown, and our goal is to develop the feasible Wald test in this case when $N/T \to \infty$. We are particularly interested in the sparse covariance. To formally quantify the level of sparsity, define

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

Here m_N represents the maximum number of nonzeros in each row, corresponding to $k = 0$ in (3.6), and D_N represents the total number of nonzero off-diagonal entries. We consider two kinds of sparse matrices, and develop our result 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 diverging 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 or banded matrices with finite number of bands. This is particularly useful when firms' individual shocks are correlated only within industries but not across industries. Formally, we assume:

Assumption 5.2. *One of the following cases holds:*
(i) $D_N = O(\sqrt{N})$*;*
(ii) $D_N = O(N)$*, and* $m_N = O(1)$.

With the help of the above two conditions, we show in the appendix (Proposition A.1) that

$$T\widehat{\boldsymbol{\alpha}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\alpha}}/\sqrt{N} = o_P(1).$$

As a result, the effect of replacing $\mathbf{\Sigma}_u^{-1}$ by its thresholded estimator is asymptotically negligible even if $N/T \to \infty$. With the thresholded estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$ defined in (3.4), we define the standardized Wald test in high-dimension as

$$J_{sw} = \frac{a_T \widehat{\boldsymbol{\alpha}}' \widehat{\mathbf{\Sigma}}_u^{-1} \widehat{\boldsymbol{\alpha}} - N}{\sqrt{2N}}.$$

Its power can be enhanced by

$$\tilde{J} = J_0 + J_{sw}.$$

We have the following properties.

Theorem 5.1. *Suppose $m_N^4 (\log N)^4 N = o(T^2)$, and Assumptions 3.1-3.3, 5.1, 5.2 hold. Then*

(i) under the null hypothesis $H_0 : \boldsymbol{\alpha} = \mathbf{0}$,

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

(ii) PEM $\tilde{J}$ has high power uniformly on the set

$$\Omega_0 \cup \{\boldsymbol{\alpha} \in \mathbb{R}^N : \|\boldsymbol{\alpha}\|^2 \gg (N \log N)/T\}.$$

6 Extended Applications of PEM

Though we have focused on the test of factor pricing models, the PEM principle is widely applicable to many high-dimensional problems in econometrics, where the conventional methods may suffer from low powers. As examples, we discuss two problems that have caused extensive attentions in the literature: testing the cross-sectional independence and testing the over-identifying condition.

6.1 Testing the cross-sectional independence

Consider a fixed effect panel data model

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

The regressor $\mathbf{x}_{it}$ could be correlated with the individual effect μ_i, but is uncorrelated with the idiosyncratic error u_{it}. Let $\mathbf{\Sigma}_u$ continue to denote the covariance matrix of $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})'$. The goal is to test the following hypothesis:

$$H_0 : (\mathbf{\Sigma}_u)_{ij} = 0, \text{ for all } i \neq j,$$

that is, whether the cross-sectional dependence is present. It is commonly known that cross-sectional dependence leads to efficiency loss for OLS, and sometimes it may even cause inconsistent estimations (e.g., Andrews 2005). Thus testing H_0 is an important problem in applied panel data models.

Most of the existing tests are based on the sum of squared correlations on the off-diagonal: $W = \sum_{i<j} T\hat{\rho}_{ij}^2$, where $\hat{\rho}_{ij}$ is the sample correlation between u_{it} and u_{jt}, estimated by the within-OLS (e.g., Baltagi 2008). Breusch and Pagan (1980) showed that when N is fixed, W is asymptotically $\chi^2_{N(N-1)/2}$. Pesaran et al. (2008) and Baltagi et al. (2012) studied the rescaled W, and showed that after a proper standardization, the rescaled W is asymptotically normal when both $N, T \to \infty$.

However, like the situation in the test of factor pricing model, the test based on W suffers from a low power against sparse alternatives, as long as N^2 is either comparable or larger than T. To see this, let ρ_{ij} denote the population correlation between u_{it} and u_{jt}. Using the χ^2-distribution with degree of freedom $N(N-1)/2$ as the null distribution, the power of the test is low when $T\sum_{i<j}\rho_{ij}^2 = o(N^2)$. This is not unusual when $\mathbf{\Sigma}_u$ is sparse and $N^2/T \to \infty$.

The proposed PEM method can resolve this problem by introducing a screened statistic

$$J_0 = \sum_{(i,j)\in\widehat{S}} NT\hat{\rho}_{ij}^2, \quad \widehat{S} = \{(i,j) : |\hat{\rho}_{ij}| > \delta_T, i < j \leq N\}$$

where $\delta_T = \log(\log T)\sqrt{\frac{\log N}{T}}$. The set $\widehat{S}$ screens off most of the estimation errors, and mimics $S = \{(i,j) : |\rho_{ij}| > 2\delta_T, i < j \leq N\}$. Let J_1 be a "standard" statistic. For instance, Baltagi et al. (2012) proposed to use a bias-corrected version of W, given by

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

They gave sufficient conditions under which $J_1 \to^d \mathcal{N}(0,1)$ under H_0. Then the PEM test can be constructed as $J = J_0 + J_1$. Because $P(J_0 = 0|H_0) \to 1$, the size is not distorted

much. Applying the PEM principle in Theorem 3.2, immediately we see that the power is substantially enhanced to cover the region

$$\Omega_0 = \{\mathbf{\Sigma}_u : \max_{i<j} |(\mathbf{\Sigma}_u)_{ij}| > 2\delta_T \max_{i,j} (\mathbf{\Sigma}_u)_{ii}^{1/2} (\mathbf{\Sigma}_u)_{jj}^{1/2}\}$$

in the alternative, in addition to the region detectable by J_1 itself.

As a by-product, PEM also identifies the entries where the idiosyncratic components are correlated through $\widehat{S}$. Empirically, by examining this set, researchers can understand better the underlying pattern of cross-sectional correlations.

6.2 Testing the over-identifying conditions

Consider testing a high-dimensional vector of moment conditions

$$H_0 : E\mathbf{g}(\mathbf{X}, \theta) = 0, \text{ for some } \theta \in \Theta$$

here $\dim(\mathbf{g}) = N$ can be much larger than the sample size T, and $\dim(\theta)$ is bounded. Such a high-dimensional moment condition arises typically when there are many instrumental variables (IV). For instance, when $\mathbf{g}(\mathbf{X}, \theta) = (y - \mathbf{x}^T\theta)\mathbf{z}$, where $\mathbf{z}$ is a high-dimensional vector of IV's, then we are testing the validity of the IV's. Recently Anatolyev and Gospodinov (2009) and Chao et al. (2012) studied the many IV problem, focusing on controlling the size. While many other methods have been proposed in the literature, we particularly discuss the power enhancement of the seminal GMM test in Hansen (1982).

We shall assume that there be a consistent estimator $\hat{\theta}$ for θ. Hansen's J-test is given by

$$J_{\text{GMM}} = T \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{g}(\mathbf{X}_t, \hat{\theta}) \right)' \mathbf{W}_T \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{g}(\mathbf{X}_t, \hat{\theta}) \right),$$

where $\mathbf{W}_T = [E\mathbf{g}(\mathbf{X}_T, \theta)\mathbf{g}(\mathbf{X}_T, \theta)']^{-1}$, and is often replaced with a consistent estimator $\widehat{\mathbf{W}}_T$. When N increases, Donald et al. (2003) showed that under the null, the normalized statistic satisfies

$$J_{\text{N-GMM}} = \frac{J_{\text{GMM}} - N}{\sqrt{2N}} \to^d \mathcal{N}(0, 1).$$

Besides the issue of estimating $\mathbf{W}_T$ under high dimensions, this test suffers from a low power if $N/T \to \infty$. Due to the accumulation of estimation errors, the critical value for J_{GMM} is of order N. Hence the test would not detect the sparse alternatives where a relatively small

number of moment conditions are violated.

With the help of PEM, the power of $J_{\text{N-GMM}}$ can be enhanced. Write $\mathbf{g} = (g_1, ..., g_N)'$. For each $j \leq N$, let $m_{Tj} = \frac{1}{T}\sum_{t=1}^{T} g_j(\mathbf{X}_t, \hat{\theta})$ and $v_{Tj} = \frac{1}{T}\sum_{t=1}^{T}[g_j(\mathbf{X}_t, \hat{\theta}) - m_{Tj}]^2$. Define

$$J_0 = \sqrt{N}T \sum_{j \in \widehat{S}} m_{Tj}^2 / v_{Tj}, \quad \widehat{S} = \{j \leq N : |m_{Tj}| > \delta_T \sqrt{v_{Tj}}\}$$

where $\delta_T = \log(\log T)\sqrt{\frac{\log N}{T}}$ is slightly larger than the noise level. By Theorem 3.2, immediately we see that the power is uniformly enhanced on the region

$$\Omega_0 = \{\mathbf{g} = (g_1, ..., g_N)' : \max_{j \leq N} |Eg_j(\mathbf{X}_t, \theta)| \gg \delta_T\}.$$

We can also use a non-diagonal version of PEM as

$$J_0' = \sqrt{N}T \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{g}_{\widehat{S}}(\mathbf{X}_t, \hat{\theta})) \right)' \widehat{\mathbf{W}}_{T,\widehat{S}} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{g}_{\widehat{S}}(\mathbf{X}_t, \hat{\theta} \right),$$

where $\mathbf{g}_{\widehat{S}}$ and $\widehat{\mathbf{W}}_{T,\widehat{S}}$ are subvector and submatrix of $\mathbf{g}$ and $\widehat{\mathbf{W}}_T$ respectively. Under special structures, the high-dimensional weight matrix $\mathbf{W}_T$ is consistently estimable. For instance, if $\mathbf{W}_T$ is either sparse (Bickel and Levina 2008) or conditionally sparse that admits a factor structure, we can apply thresholding and/or factor analysis along the line of Fan et al. (2013). Formal derivations are out of the scope of this paper, and is to be addressed separately.

7 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):

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

7.1 Simulation

We simulate $\{\mathbf{b}_i\}_{i=1}^N$, $\{\mathbf{f}_t\}_{t=1}^T$ and $\{\mathbf{u}_t\}_{t=1}^T$ independently from $\mathcal{N}_3(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, $\mathcal{N}_3(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, and $\mathcal{N}_N(0, \boldsymbol{\Sigma}_u)$ 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 con-

stituents, for the period from July 1st, 2008 to June 29th 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, whose correlation matrix has equi-off-diagonal entry ρ_j, generated from Uniform$[0, 0.5]$. The diagonal entries of $\mathbf{A}_j$ are generated as those of $\boldsymbol{\Sigma}_u^{(1)}$.

We evaluate the power of the test at two specific alternatives (we set $N > T$):

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

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

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.

7.2 Results

When $\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u^{(1)}$, we assume the diagonal structure to be known, and compare the performances of the working-independent quadratic test J_{wi} (considered by Pesaran and Yamagata 2012) with the power enhanced test $J_0 + J_{wi}$. When $\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u^{(2)}$, we do not

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_{wi}	PEM	$P(\widehat{S}=\emptyset)$	J_{wi}	PEM	$P(\widehat{S}=\emptyset)$	J_{wi}	PEM	$P(\widehat{S}=\emptyset)$
300	500	0.050	0.058	0.992	0.492	0.958	0.056	0.710	0.762	0.670
	800	0.048	0.056	0.992	0.542	0.984	0.016	0.692	0.762	0.654
	1000	0.050	0.052	0.998	0.500	0.978	0.030	0.758	0.826	0.580
	1200	0.056	0.058	0.998	0.688	0.992	0.012	0.742	0.810	0.620
500	500	0.058	0.062	0.994	0.390	0.986	0.018	0.696	0.720	0.778
	800	0.042	0.044	0.998	0.652	1.000	0.000	0.714	0.756	0.760
	1000	0.042	0.042	1.000	0.590	1.000	0.002	0.734	0.756	0.796
	1200	0.060	0.060	1.000	0.512	1.000	0.002	0.730	0.766	0.804

The frequencies of rejection and $\widehat{S}=\emptyset$ out of 500 replications are calculated. Here J_{wi} is the working independent test using diagonal weight matrix; PEM= $J_0 + J_{wi}$.

assume we know the block-diagonal structure. In this case, four tests are carried out and compared: (1) the working-independent quadratic test J_{wi}, (2) the proposed feasible Wald test J_{sw} based on estimated error covariance matrix, (3) the power enhanced test $J_0 + J_{wi}$, and (4) the power enhanced test $J_0 + J_{sw}$. For the thresholded covariance matrix, we use the soft-thresholding function and fix the threshold at $\sqrt{\log N/T}$, as suggested in Fan et al. (2013). For each test, we calculate the frequency of rejection under H_0, H_a^1 and H_a^2 based on 500 replications, with significance level $q = 0.05$. 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 $P(\widehat{S}=\emptyset)$ 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 hypothesis, which results in approximately the same size as the original test. 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 survive 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. But the PEM test still slightly improves the power of the quadratic test. For the non-diagonal covariance, similar patterns are observed. We additionally find that the power of Wald test is slightly larger than the working independent test.

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

		H_0				
T	N	J_{wi}	J_{sw}	PEM1	PEM2	$P(\widehat{S} = \emptyset)$
300	500	0.048	0.052	0.050	0.054	0.998
	800	0.056	0.054	0.064	0.062	0.992
	1000	0.040	0.040	0.046	0.046	0.990
	1200	0.052	0.050	0.056	0.054	0.996
500	500	0.058	0.058	0.060	0.060	0.994
	800	0.048	0.048	0.050	0.050	0.998
	1000	0.050	0.050	0.052	0.052	0.998
	1200	0.056	0.052	0.056	0.052	1.000

The frequencies of rejection and $\widehat{S} = \emptyset$ out of 500 replications are calculated. Here J_{wi} is the working independent test using diagonal weight matrix; J_{sw} is the Wald test using the thresholded $\widehat{\mathbf{\Sigma}}_u^{-1}$ as the weight matrix; PEM1= $J_0 + J_{wi}$; PEM2= $J_0 + J_{sw}$.

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

		H_a^1					H_a^2				
T	N	J_{wi}	J_{sw}	PEM1	PEM2	$P(\widehat{S} = \emptyset)$	J_{wi}	J_{sw}	PEM1	PEM2	$P(\widehat{S} = \emptyset)$
300	500	0.472	0.480	0.976	0.976	0.026	0.684	0.690	0.764	0.764	0.646
	800	0.588	0.600	0.990	0.990	0.012	0.690	0.696	0.772	0.778	0.622
	1000	0.546	0.546	0.984	0.984	0.026	0.708	0.716	0.756	0.764	0.632
	1200	0.636	0.642	0.992	0.992	0.008	0.726	0.726	0.790	0.790	0.636
500	500	0.334	0.338	0.992	0.992	0.008	0.720	0.734	0.772	0.772	0.778
	800	0.672	0.674	1.000	1.000	0.000	0.722	0.724	0.760	0.764	0.750
	1000	0.648	0.650	1.000	1.000	0.002	0.770	0.768	0.808	0.804	0.740
	1200	0.580	0.580	1.000	1.000	0.002	0.744	0.742	0.788	0.784	0.770

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

8 Empirical Study

We apply the proposed thresholded Wald test and the PEM to the securities in the S&P 500 index, by employing the Fama-French three-factor (FF-3) model to conduct our test. One of our empirical findings is that market inefficiency is primarily caused by a small portion of stocks with positive alphas, instead of a large portion of slightly mispriced assets. This provides empirical evidence of sparse alternatives. In addition, the market inefficiency is further evidenced by our newly created portfolio based on the PEM test, which outperforms the SP500 index.

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 stocks without missing observations 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 time-varying factor loadings. For testing months $\tau = 12/1984, ..., 12/2012$, we run the FF-3 regressions

$$r_{it}^{\tau} - r_{ft}^{\tau} = \alpha_i^{\tau} + \beta_{i,\mathrm{MKT}}^{\tau}(\mathrm{MKT}_t^{\tau} - r_{ft}^{\tau}) + \beta_{i,\mathrm{SMB}}^{\tau}\mathrm{SMB}_t^{\tau} + \beta_{i,\mathrm{HML}}^{\tau}\mathrm{HML}_t^{\tau} + u_{it}^{\tau}, \tag{8.1}$$

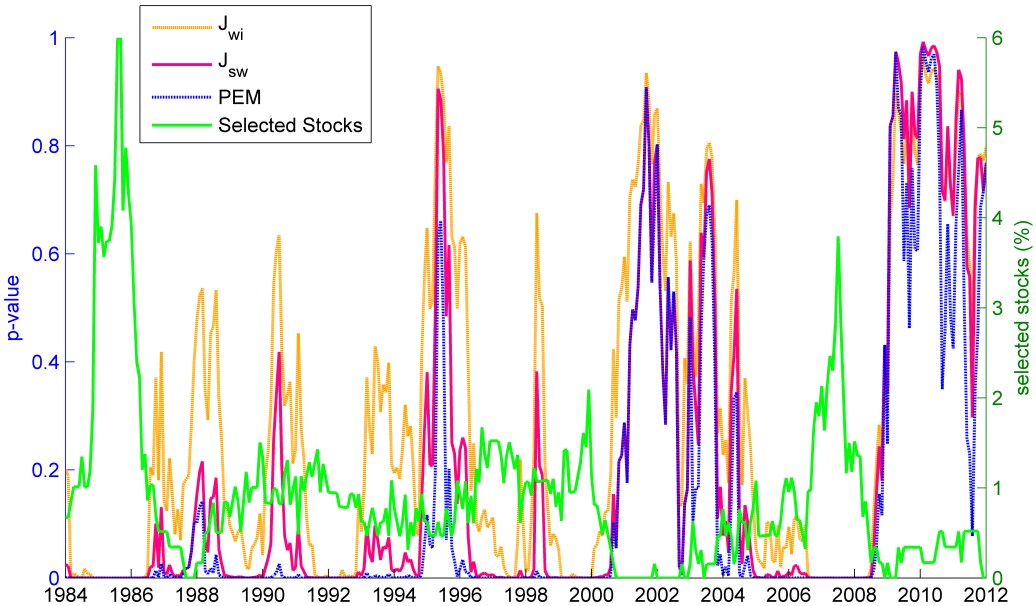
for $i = 1, ..., N_\tau$ and $t = \tau - 59, ..., \tau$, where r_{it} represents the return for stock i at month t, r_{ft} the risk free rate, and $\mathrm{MKT}, \mathrm{SMB}$ and HML constitute the FF-3 model's market, size and value factors. Our null hypothesis $\alpha_i^{\tau} = 0$ for all i implies that the market is mean-variance efficient.

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

Variables	Mean	Std dev.	Median	Min	Max
N_τ	617.70	26.31	621	574	665
$\lvert\widehat{S}\rvert_0$	5.49	5.48	4	0	37
$\overline{\lvert\widehat{\alpha}\rvert_i}^{\tau}(\%)$	0.9973	0.1630	0.9322	0.7899	1.3897
$\overline{\lvert\widehat{\alpha}\rvert_{i\in\widehat{S}}}^{\tau}(\%)$	4.3003	0.9274	4.1056	1.7303	8.1299
p-value of J_{wi}	0.2844	0.2998	0.1811	0	0.9946
p-value of J_{sw}	0.1861	0.2947	0.0150	0	0.9926
p-value of PEM	0.1256	0.2602	0.0003	0	0.9836

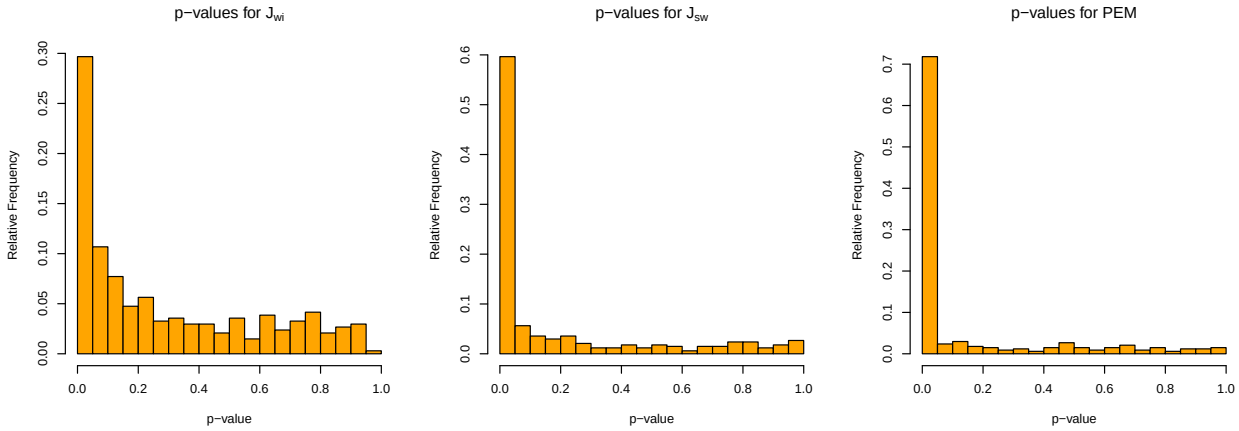
Table 5 summarizes descriptive statistics for different components and estimates in the model. On average, 618 stocks (which is more than 500 because we are recording stocks that have *ever* become the constituents of the index) enter the panel of the regression during each five-year estimation window, of which 5.5 stocks are selected by $\widehat{S}$. The threshold $\delta_T = \sqrt{\log N/T}\log(\log T)$ is about 0.45 on average, which changes as the panel size N changes for every window of estimation. The selected stocks have much larger alphas than other stocks do, as expected. As far as the signs of those alpha estimates are concerned, 61.84% of all the estimated alphas are positive, and 80.66% of all the selected alphas are positive. This indicates that market inefficiency is primarily contributed by stocks with extra returns, instead of a large portion of stocks with small alphas, demonstrating the sparse alternatives. In addition, we notice that the p-values of the thresholded Wald test J_{sw} are generally smaller than those of the working independent test J_{wi}.

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



Similar to Pesaran and Yamagata (2012), we plot the running p-values of J_{wi}, J_{sw} and the PEM test (augmented from J_{sw}) from December 1984 to December 2012. We also add the dynamics of the percentage of selected stocks ($|\widehat{S}|_0/N$) to the plot, as shown in Figure 1. There is a strong negative correlation between the stock selection percentage and the

Figure 2: Histograms of p-values for J_{wi}, J_{sw} and PEM



p-values 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 J_{sw} and J_{wi} tests as a result of enhanced power, and hence it captures several important market disruptions ignored by the latter two (e.g. Black Monday in 1987, collapse of Japanese bubble in late 1990, and the European sovereign debt crisis after 2010). Indeed, the null hypothesis of market efficiency is rejected by the PEM test at 5% level during almost all financial crisis, including major financial crisis such as Black Wednesday in 1992, Asian financial crisis in 1997, the financial crisis in 2008, which are also detected by J_{sw} and J_{wi} tests. For 30%, 60% and 72% of the study period, J_{wi}, J_{sw} 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 magnitude of alphas (more precisely, large t-test statistics). 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 196 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 18 consecutive months. As a result, 23 companies stand out, most of which have positive alphas. Table 6 lists these companies, their major periods when getting selected and the

associated alphas. We observe that companies such as Walmart, Dell and Apple exhibited large positive alphas during their periods of rapid growth, whereas some others like Massey Energy experienced very low alphas at their stressful times.

Table 6: Companies with longest selection period

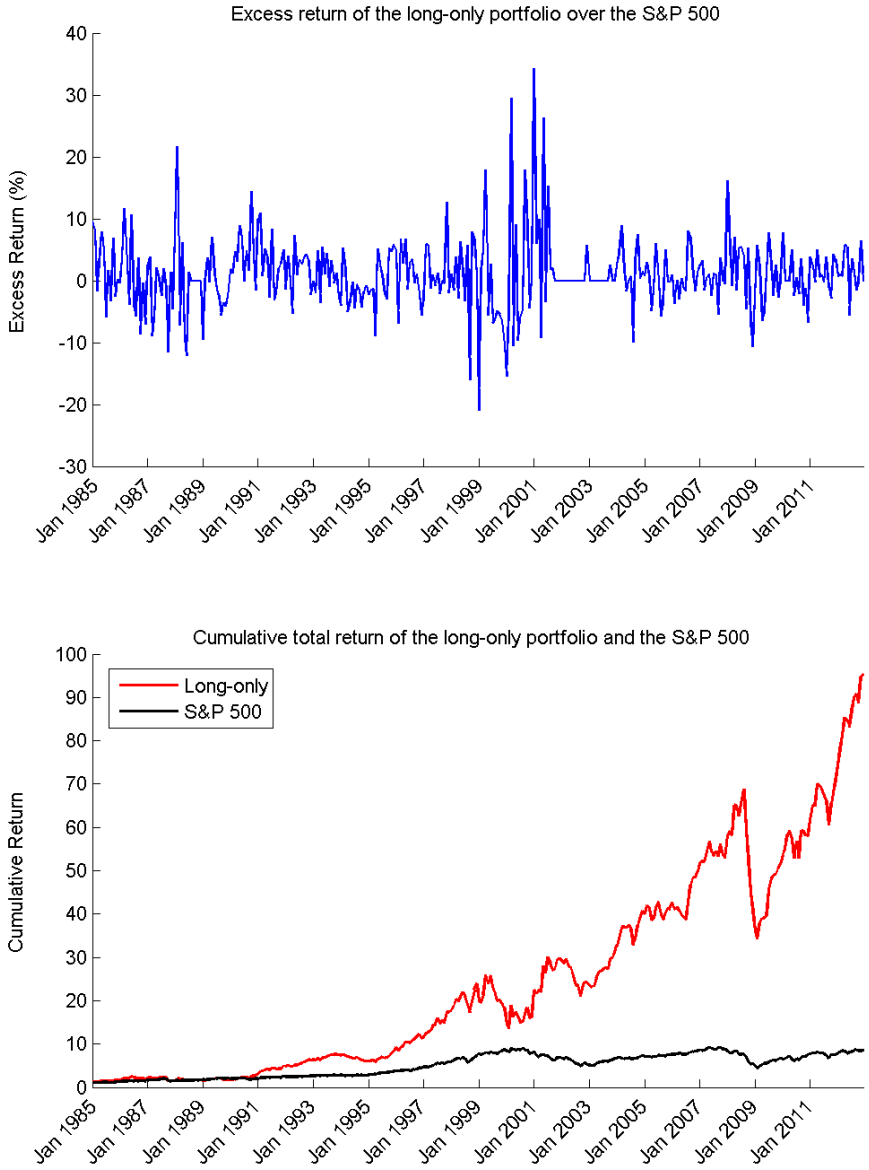
Company Name	Major period of selection	Average alpha (%)	Std. dev. (%)
T J X COMPANIES INC NEW	12/1984—08/1986	3.8230	0.2055
HASBRO INDUSTRIES INC *	12/1984—03/1987	5.7285	0.5362
WAL MART STORES INC *	12/1984—03/1987	2.7556	0.1672
MASSEY ENERGY CO	10/1985—03/1987	-3.7179	0.2778
DEERE & CO	11/1985—04/1987	-3.0385	0.3328
CIRCUIT CITY STORES INC	11/1985—09/1987	6.4062	0.7272
CLAIBORNE INC	06/1986—04/1988	4.1017	0.8491
ST JUDE MEDICAL INC *	10/1989—09/1991	4.1412	0.3794
HOME DEPOT INC ***	08/1990—01/1995	3.4988	0.4887
UNITED STATES SURGICAL CORP *	11/1990—03/1993	4.2471	0.4648
INTERNATIONAL GAME TECHNOLOGY	02/1992—11/1993	5.4924	0.4993
UNITED HEALTHCARE CORP **	07/1992—04/1995	4.9666	0.3526
H B O & CO ***	10/1995—09/1998	4.9303	0.6193
SAFEWAY INC	08/1997—04/1999	3.8200	0.3709
CISCO SYSTEMS INC ***	10/1997—12/2000	3.8962	0.3352
DELL COMPUTER CORP	04/1998—10/1999	7.5257	0.7335
SUN MICROSYSTEMS INC	01/1999—11/2000	4.6975	0.2630
J D S UNIPHASE CORP *	01/1999—05/2001	6.9504	0.5183
HANSEN NATURAL CORP *	11/2005—10/2007	8.1486	0.5192
CELGENE CORP	04/2006—11/2007	5.4475	0.4210
MONSANTO CO NEW	09/2007—05/2009	3.5641	0.3462
GANNETT INC	10/2007—03/2009	-2.7909	0.7952
APPLE COMPUTER INC	03/2008—10/2009	4.8959	0.3740

Major period of selection refers to the time interval of at least 18 months when those companies stay in the screening set. The average and standard deviations of the alphas are computed during the period of selection. We mark companies that stand out for at least 24 months by $*$ (30 months by $**$ and 36 months by $***$ respectively).

Our test also has a number of implications in terms of portfolio construction. In fact, a simple trading strategy can be designed based on the obtained screening set. At the beginning of month t, we run regressions (8.1) over a 60-month period from month $t-60$ to month $t-1$. From the estimated screening set $\widehat{S}$ we extract stocks that have positive alphas and construct a equal-weighted long-only portfolio. The portfolio is held for a one-month period and rebalanced at the beginning of month $t+1$. For diversification purposes, if $\widehat{S}$ contains less than 2 stocks, we invest all the money into the S&P 500 index. Transaction costs are ignored here.

Figure 3 compares our trading strategy with the S&P 500 index from January 1985 to December 2012. The top panel gives the excess return of the equal-weighted long-only

Figure 3: Monthly excess returns (top panel, returns of the new trading strategy subtracting S&P500 index) and cumulative total returns (bottom panel) of the long-only portfolio relative to the S&P 500



portfolio over the S&P 500 index during each month. The monthly excess return has a mean of 0.9%, and in particular, for 202 out of the 336 study months, our strategy generates a higher return than the S&P 500. The cumulative performance of the two portfolios is depicted in the bottom panel. Our long-only portfolio has a monthly mean return of 1.64% and a standard deviation of 7.48% during the period, whereas those for the S&P 500 are 0.74% and 4.48% respectively. The monthly Sharpe ratio for our strategy is 0.1759, and the Sharpe ratio for the S&P 500 is 0.0936. They are equivalent to an annual Sharpe ratio of 0.6093 and 0.3242 (multiplied by $\sqrt{12}$). Our strategy clearly outperforms the S&P 500 over the sample period with a higher Sharpe ratio. This is because the degree of market efficiency is dynamic and it often takes time for market participants to adapt to changing market conditions. Arbitrage opportunities as indicated by large positive alphas are not competed away immediately. By holding stocks with large positive alphas for a one-month period, we are likely to get higher returns without bearing as much risk.

9 Concluding remarks

The literature on testing mean-variance efficiency is predominated by low-dimensional tests based on constructed portfolios, which test only one part of the pricing theory and are subject to selection and unintentional biases. Recent efforts of extending them to high-dimensional tests are only able to detect market inefficiency in an average sense as measured by the weighted quadratic form $\widehat{\boldsymbol{\alpha}}'\mathbf{V}\widehat{\boldsymbol{\alpha}}$. However, when we deal with large panels, it is more appealing if we could identify individual departures from the factor pricing model, and handle the case when there are small portions of large alphas.

We propose a new concept for high dimensional statistical tests, namely, the power enhancement (PEM). The PEM test combines a PEM component and a Wald-type statistic. Under the null hypothesis, the PEM component equals zero with probability approaching one, whereas under the alternative hypothesis it is stochastically unbounded over some high-power regions. Hence while maintaining a good size asymptotically, the PEM test significantly enhances the power of Wald-type statistics. As a by-product, the selected subset $\widehat{S}$ also enables us to identify those significant alphas. Furthermore, the PEM technique is potentially widely applicable in many high-dimensional testing problems in panel data analysis, such as testing the cross-sectional independence and over-identifying constraints.

We also develop a high-dimensional Wald test when the covariance matrix of idiosyncratic noises is sparse, using the thresholded sample covariance matrix of the idiosyncratic

components. We 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 the market inefficiency is primarily caused by a small portion of significantly mis-priced stocks, instead of a large portion of slightly mis-priced stocks. In addition, most of the selected stocks have positive alphas. The market inefficiency is further evidenced by our newly created portfolio based on the PEM test, which outperforms the S&P500 index.

APPENDIX

A Proofs

We first cite a lemma that will be needed throughout the proofs. Write $\sigma_{ij} = (\mathbf{\Sigma}_u)_{ij}$ and $\widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt}$. Recall $\sigma_j^2 = (\mathbf{\Sigma}_u)_{jj}/(1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t)$, and $\widehat{\sigma}_j^2 = \widehat{\sigma}_{jj}/(1 - \bar{\mathbf{f}}'\mathbf{w})$.

Lemma A.1. *Under Assumption 3.2, there is $C > 0$,*
(i) $P(\max_{i,j\le N} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$
(ii) $P(\max_{i\le K, j\le N} |\frac{1}{T}\sum_{t=1}^{T} f_{it}u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$
(iii) $P(\max_{j\le N} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$

Proof. The proof follows from Lemmas A.3 and B.1 in Fan, Liao and Mincheva (2011). □

Lemma A.2. *When the distribution of $(\mathbf{u}_t, \mathbf{f}_t)$ is independent of $\boldsymbol{\alpha}$, under Assumption 3.2, there is $C > 0$,*
(i) $\sup_{\boldsymbol{\alpha}\in\mathbb{R}^N} P(\max_{j\le N} |\widehat{\alpha}_j - \alpha_j| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\alpha}) \to 0$
(ii) $\sup_{\boldsymbol{\alpha}\in\mathbb{R}^N} P(\max_{i,j\le N} |\widehat{\sigma}_{ij} - \sigma_{ij}| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\alpha}) \to 0,$
(iii) $\sup_{\boldsymbol{\alpha}\in\mathbb{R}^N} P(\max_{i\le N} |\widehat{\sigma}_i - \sigma_i| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\alpha}) \to 0.$

Proof. Note that $\widehat{\alpha}_j - \alpha_j = \frac{1}{\tau T}\sum_{t=1}^{T} u_{jt}(1 - \mathbf{f}_t'\mathbf{w})$. Here $\tau = 1 - \bar{\mathbf{f}}'\mathbf{w} \to^p 1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t > 0$, hence τ is bounded away from zero with probability approaching one. Thus by Lemma A.1, there is $C > 0$ independent of $\boldsymbol{\alpha}$, such that

$$\sup_{\boldsymbol{\alpha}\in\mathbb{R}^N} P(\max_{j\le N} |\widehat{\alpha}_j - \alpha_j| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\alpha}) = P(\max_{j} |\frac{1}{\tau T}\sum_{t=1}^{T} u_{jt}(1 - \mathbf{f}_t'\mathbf{w})| > C\sqrt{\frac{\log N}{T}}) \to 0$$

(ii) There is C independent of $\boldsymbol{\alpha}$, such that the event

$$A = \{\max_{i,j} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \sigma_{ij}| < C\sqrt{\frac{\log N}{T}}, \quad \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{f}_t\|^2 < C\}$$

has probability approaching one. Also, there is C_2 also independent of $\boldsymbol{\alpha}$ such that the event $B = \{\max_i \frac{1}{T}\sum_t u_{it}^2 < C_2\}$ occurs with probability approaching one. Then on the event $A \cap B$, by the triangular and Cauchy-Schwarz inequalities,

$$|\widehat{\sigma}_{ij} - \sigma_{ij}| \leq C\sqrt{\frac{\log N}{T}} + 2\max_i \sqrt{\frac{1}{T}\sum_t (\widehat{u}_{it} - u_{it})^2 C_2} + \max_i \frac{1}{T}\sum_t (u_{it} - \widehat{u}_{it})^2.$$

It can be shown that

$$\max_{i\leq N} \frac{1}{T}\sum_{t=1}^{T} (\widehat{u}_{it} - u_{it})^2 \leq \max_i (\|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|^2 + (\widehat{\alpha}_i - \alpha_i)^2)(\frac{1}{T}\sum_{t=1}^{T} \|\mathbf{f}_t\|^2 + 1).$$

Note that $\widehat{\mathbf{b}}_i - \mathbf{b}_i$ and $\widehat{\alpha}_i - \alpha_i$ only depend on $(\mathbf{f}_t, \mathbf{u}_t)$. By Lemma 3.1 of Fan et al. (2011), there is $C_3 > 0$ such that $\sup_{\mathbf{b},\boldsymbol{\alpha}} P(\max_{i\leq N} \|\widehat{\mathbf{b}}_i - \mathbf{b}_i\|^2 + (\widehat{\alpha}_i - \alpha_i)^2 > C_3\frac{\log N}{T}) = o(1)$. Combining the last two displayed inequalities yields, for $C_4 = (C+1)C_3$,

$$\sup_{\boldsymbol{\alpha}} P(\max_{i\leq N} \frac{1}{T}\sum_{t=1}^{T} (\widehat{u}_{it} - u_{it})^2 > C_4\frac{\log N}{T}|\boldsymbol{\alpha}) = o(1),$$

which yields the desired result.

(iii): Recall $\widehat{\sigma}_j^2 = \widehat{\sigma}_{jj}/\tau$, and $\sigma_j^2 = \sigma_{jj}/(1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t)$. Moreover, τ is independent of $\boldsymbol{\alpha}$. The result follows immediately from part (ii). □

Lemma A.3. *For any $\epsilon > 0$, $\sup_{\boldsymbol{\alpha}} P(\|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\| > \epsilon|\boldsymbol{\alpha}) = o(1)$.*

Proof. By Lemma A.2 (ii), $\sup_{\boldsymbol{\alpha}\in\mathbb{R}^N} P(\max_{i,j\leq N} |\widehat{\sigma}_{ij} - \sigma_{ij}| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\alpha}) \to 1$. By Theorem A.1 of Fan et al. (2013), on the event $\max_{i,j\leq N} |\widehat{\sigma}_{ij} - \sigma_{ij}| \leq C\sqrt{\frac{\log N}{T}}$, there is constant C' that is independent of $\boldsymbol{\alpha}$,

$$\|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\| \leq C' m_N (\frac{\log N}{T})^{(1-k)/2}.$$

Hence the result follows due to the sparse condition $m_N(\frac{\log N}{T})^{(1-k)/2} = o(1)$. □

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, $\mathbf{A} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}'(\boldsymbol{\Sigma}_2 - \boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}')^{-1}\boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}$. We also partition the estimator into $\widehat{\boldsymbol{\alpha}}' = (\widehat{\boldsymbol{\alpha}}_1', \widehat{\boldsymbol{\alpha}}_2')$. 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 $\boldsymbol{\xi} = \boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1$, whose i^{th} element is denoted by ξ_i. It follows from $\|\boldsymbol{\Sigma}_1^{-1}\|_1 < \infty$ that

$$\max_{i \le N-r} |\xi_i| = O_P(\max_{i \le N-r} |\widehat{\alpha}_{1i}|) = O_P(\max_{i \le N-r} |\widehat{\alpha}_{1i} - \alpha_{1i}|) = O_P(\sqrt{\frac{\log N}{T}}).$$

Also, $\max_{i \le r} \sum_{j=1}^{N-r} |\beta_{ij}| \le \|\boldsymbol{\Sigma}_u\|_1 = O(1)$, and $\lambda_{\max}((\boldsymbol{\Sigma}_2 - \boldsymbol{\beta}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}')^{-1}) = O(1)$. Hence

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

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

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

where we used the fact that $\max_{j \le r} |\widehat{\alpha}_{2j}| \le \max_j |\alpha_{2j}| + \max_j |\widehat{\alpha}_j - \alpha_j| = O_P(1)$. Also, $|\widehat{\boldsymbol{\alpha}}_2'\mathbf{C}\widehat{\boldsymbol{\alpha}}_2| \le \|\widehat{\boldsymbol{\alpha}}_2\|^2\|\mathbf{C}\| = O_P(r)$. It then yields, under H_a,

$$\kappa = \widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 = O_P(r).$$

It also follows from (2.2) that

$$Z \equiv \frac{a_T\widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1 - (N-r)}{\sqrt{2(N-r)}} \to^d \mathcal{N}(0,1).$$

As $a_T = T\tau = O_P(T)$, we have $(a_T\kappa - r/\sqrt{2})/\sqrt{N} = O_P(Tr/\sqrt{N})$. Since $Tr = o(\sqrt{N})$, for

any $\epsilon \in (0, z_q)$, for all large N, T, the following event holds with probability at least $1-\epsilon$:

$$A = \{|a_T\kappa - r/\sqrt{2}| < \sqrt{N}\epsilon\}.$$

Using $1-\Phi(z_q) = q$ and choosing ϵ small enough such that $1-\Phi(z_q-\epsilon)+\epsilon < 2q$, we have,

$$\begin{aligned} P(J_1 > z_q) &= P(\frac{a_T(\widehat{\boldsymbol{\alpha}}_1'\boldsymbol{\Sigma}_1^{-1}\widehat{\boldsymbol{\alpha}}_1+\kappa)-N}{\sqrt{2N}} > z_q) \\ &= P(Z\sqrt{\frac{N-r}{N}} + \frac{a_T\kappa - r/\sqrt{2}}{\sqrt{N}} > z_q) \\ &\leq P(Z\sqrt{\frac{N-r}{N}} + \epsilon > z_q) + P(A^c) \\ &\leq 1-\Phi(z_q-\epsilon)+\epsilon+o(1), \end{aligned}$$

which is bounded by $2q$. This implies the result.

A.2 Proof of Theorem 3.1 and Corollary 3.1

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

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

for some $C_2, C_3 > 0$. Lemma A.2 then implies that $\inf_{\boldsymbol{\alpha}} P(A_1 \cap A_2) \to 1$. Under $A_1 \cap A_2$,

$$\begin{aligned} \frac{|\widehat{\alpha}_j|}{\widehat{\sigma}_j} &\geq (|\alpha_j| - \max_j |\widehat{\alpha}_j - \alpha_j|)(\sigma_j^{-1} - \max_j |\widehat{\sigma}_j^{-1} - \sigma_j^{-1}|) \\ &\geq (|\alpha_j| - C_3\delta_T)(\sigma_j^{-1} - C_2) \geq \delta_T, \end{aligned}$$

where the last inequality holds for sufficiently small C_2, C_3, e.g., $C_3 < \min_j \frac{\sigma_j}{2}$ and $C_2 = \frac{1}{3}\min_j(\sigma_j^{-1})$. This implies that $j \in \widehat{S}$, hence $P(S \subset \widehat{S}) \to 1$. It can be readily seen that if $j \in \widehat{S}$, by similar arguments, we have $\frac{|\alpha_j|}{\sigma_j} > \frac{1}{2}\delta_T$ with probability tending to one. Consequently, $\widehat{S} \setminus S \subset \mathcal{G}$ with probability approaching one. In fact we have proved $\inf_{\boldsymbol{\alpha}} P(S \subset \widehat{S}) \to 1$ and $\inf_{\boldsymbol{\alpha}} P(\widehat{S} \setminus S \subset \mathcal{G}) \to 1$.

(ii) Suppose $\min_{j\leq N} \sigma_j > C_1$ for some $C_1 > 0$. For some constants $C_2 > 0$ and $C_3 <$

$(C_2 + C_1^{-1})^{-1}$, under the event $A_1 \cap A_2$ and H_0, we have

$$\begin{aligned}\max_{j\le N} \frac{|\widehat{\alpha}_j|}{\widehat{\sigma}_j} &\le \max|\widehat{\alpha}_j| \cdot \max_j(\widehat{\sigma}_j^{-1}) \le C_3\delta_T \cdot (\max|\widehat{\sigma}_j^{-1} - \sigma_j^{-1}| + \max \sigma_j^{-1}) \\ &\le C_3(C_2 + C_1^{-1})\delta_T \le \delta_T,\end{aligned}$$

where we note that under H_0, $\max_j |\widehat{\alpha}_j| = O_p(\sqrt{\frac{\log N}{T}})$. Hence $P(\max_{j\le N} \frac{|\widehat{\alpha}_j|}{\widehat{\sigma}_j} \le \delta_T) \to 1$, 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 $J_0 \ge F_q$. Since J_1 has high power uniformly on Ω_1, it follows from $J \ge J_1$

$$\inf_{\boldsymbol{\alpha}\in\Omega_1} P(J \ge F_q|\boldsymbol{\alpha}) \ge \inf_{\boldsymbol{\alpha}\in\Omega_1} P(J_1 \ge F_q|\boldsymbol{\alpha}) \to 1,$$

which implies that J also has power uniformly on Ω_1. We now show J also has high power uniformly on

$$\Omega_0 = \{\boldsymbol{\alpha} \in \mathbb{R}^N : \max_{j\le N} |\alpha_j| > 2\delta_T \max_{j\le N} \sigma_j\}.$$

First of all, $\Omega_0 \subset S$. It follows from $\inf_{\boldsymbol{\alpha}\in\Omega_0} P(S \subset \widehat{S}) \to 1$ that $\inf_{\boldsymbol{\alpha}\in\Omega_0} P(\widehat{S} \ne \emptyset) \to 1$. Lemma A.3 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 $\|\widehat{\boldsymbol{\alpha}}_{\widehat{S}}\| \ge \|\widehat{\boldsymbol{\alpha}}_S\|$. Hence there is a constant $C > 0$ so that with probability approaching one (uniformly in $\boldsymbol{\alpha}$),

$$J_0/\sqrt{N} \ge CT\|\widehat{\boldsymbol{\alpha}}_{\widehat{S}}\|^2 \ge CT\|\widehat{\boldsymbol{\alpha}}_S\|^2 \ge CT(\|\boldsymbol{\alpha}_S\| - \|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|)^2.$$

On one hand, on Ω_0, $\|\boldsymbol{\alpha}_S\|^2 \ge \min_{j\in S} \alpha_j^2 |S|_0 \ge |S|_0\delta_T^2 4 \min_j \sigma_j^2$. On the other hand, $\exists C_2 > 0$, by Lemma A.2(i),

$$\sup_{\boldsymbol{\alpha}} P\left(\|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|^2 > \frac{C_2 \log N}{T}|S|_0 \Big| \boldsymbol{\alpha}\right) = o(1).$$

Then for any $\boldsymbol{\alpha} \in \Omega_0$, because $\log(\log T) \min_j \sigma_j \to \infty$, as $\min_j \sigma_j$ is bounded away from

zero,

$$\|\boldsymbol{\alpha}_S\| - \sqrt{\frac{C_2 \log N |S|_0}{T}} \geq \sqrt{\frac{|S|_0 \log N}{T}} (\log(\log T) 2 \min_j \sigma_j - \sqrt{C_2}) \geq \sqrt{\frac{|S|_0 \log N}{T}} \log(\log T) \min_j \sigma_j.$$

Hence uniformly on $\boldsymbol{\alpha} \in \Omega_0$,

$$\begin{aligned}
&P(J_0/\sqrt{N} \geq CT(\|\boldsymbol{\alpha}_S\| - \|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|)^2 \big| \boldsymbol{\alpha}) \leq P(J_0/\sqrt{N} \geq CT(\|\boldsymbol{\alpha}_S\| - \sqrt{\frac{C_2 \log N |S|_0}{T}})^2 \big| \boldsymbol{\alpha}) \\
&+ \sup_{\boldsymbol{\alpha}} P(\|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|^2 > \frac{C_2 \log N}{T} |S|_0 \big| \boldsymbol{\alpha}) \\
&\leq P(J_0/\sqrt{N} \geq C|S|_0 \log N \log^2(\log T) \min_j \sigma_j^2 \big| \boldsymbol{\alpha}) + o(1),
\end{aligned}$$

where the term $o(1)$ is uniform in $\boldsymbol{\alpha} \in \Omega_0$.

Because $\inf_{\boldsymbol{\alpha} \in \Omega_0} P(J_0/\sqrt{N} \geq CT(\|\boldsymbol{\alpha}_S\| - \|\widehat{\boldsymbol{\alpha}}_S - \boldsymbol{\alpha}_S\|)^2 \big| \boldsymbol{\alpha}) \to 1$, and $|S|_0 \geq 1$ when there is $\boldsymbol{\alpha} \in \Omega_0$, for $g_T \equiv C \log N \log^2(\log T) \min_j \sigma_j^2 \to \infty$,

$$\inf_{\boldsymbol{\alpha} \in \Omega_0} P(J_0 \geq \sqrt{N} g_T \big| \boldsymbol{\alpha}) \to 1. \tag{A.1}$$

Note that J_1 is standardized such that $F_q = O(1)$ uniformly in $\boldsymbol{\alpha}$, and there is $c > 0$, $\inf_{\boldsymbol{\alpha} \in \Omega_0} P(J_1 \geq -c\sqrt{N} | \boldsymbol{\alpha}) \to 1$. Hence $g_T \to \infty$ implies

$$\inf_{\boldsymbol{\alpha} \in \Omega_0} P(J > F_q | \boldsymbol{\alpha}) \geq \inf_{\boldsymbol{\alpha} \in \Omega_0} P(\sqrt{N} g_T - c\sqrt{N} > F_q | \boldsymbol{\alpha}) - o(1) = 1 - o(1).$$

This proves the uniform powerfulness of J on Ω_0. Combining Ω_0 and Ω_1, we have

$$\inf_{\boldsymbol{\alpha} \in \Omega_1 \cup \Omega_0} P(J > F_q | \boldsymbol{\alpha}) \to 1.$$

A.4 Proof of Theorem 4.1

It follows from Pesaran and Yamagata (2012 Section 4.4) that $J_{wi} \to \mathcal{N}(0,1)$. For $\Omega_1 = \{\|\boldsymbol{\alpha}\|^2 \gg N \log N/T\}$, by Lemma A.2,

$$\inf_{\boldsymbol{\alpha} \in \Omega_1} P(\|\widehat{\boldsymbol{\alpha}}\|^2 \geq \frac{N \log N}{4T} \big| \boldsymbol{\alpha}) \to 1.$$

So $\inf_{\boldsymbol{\alpha}\in\Omega_1} P(J_{wi} > C\sqrt{N}\log N|\boldsymbol{\alpha}) \to 1$ for some $C > 0$. It then follows from $J \geq J_{wi}$ that

$$\inf_{\boldsymbol{\alpha}\in\Omega_1} P(J \geq C\sqrt{N}\log N|\boldsymbol{\alpha}) \geq \inf_{\boldsymbol{\alpha}\in\Omega_1} P(J_{wi} \geq C\sqrt{N}\log N|\boldsymbol{\alpha}) \to 1.$$

In addition, $J_{wi} \geq -\sqrt{N/2}$. By (A.1), we have

$$\begin{aligned}&\inf_{\boldsymbol{\alpha}\in\Omega_0} P(J \geq a_T\sqrt{N}/2|\boldsymbol{\alpha}) \geq \inf_{\boldsymbol{\alpha}\in\Omega_0} P(J_0 - \sqrt{N/2} \geq a_T\sqrt{N}/2|\boldsymbol{\alpha})\\ &\geq \inf_{\boldsymbol{\alpha}\in\Omega_0} P(J_0 \geq \sqrt{N}a_T|\boldsymbol{\alpha}) \to 1.\end{aligned}$$

This implies that both $\inf_{\boldsymbol{\alpha}\in\Omega_1} P(J > z_q)$ and $\inf_{\boldsymbol{\alpha}\in\Omega_0} P(J > z_q)$ converge to one, which yields $\inf_{\boldsymbol{\alpha}\in\Omega_1\cup\Omega_0} P(J > z_q) \to 1$.

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 $(a_T\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} - N)/\sqrt{2N} \to^d \mathcal{N}(0,1)$. So the theorem is proved by Proposition A.1 below.

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

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

Define $\mathbf{e}_t = \boldsymbol{\Sigma}_u^{-1}\mathbf{u}_t = (e_{1t}, ..., e_{Nt})'$, which is an N-dimensional vector with mean zero and covariance $\boldsymbol{\Sigma}_u^{-1}$, whose entries are stochastically bounded. Let $\bar{\mathbf{w}} = (E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t$. A key step of proving the above proposition is to establish the following two convergences:

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

$$\frac{1}{T}E|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1-\mathbf{f}_s'\bar{\mathbf{w}})][\frac{1}{\sqrt{T}}\sum_{k=1}^{T}e_{jk}(1-\mathbf{f}_k'\bar{\mathbf{w}})]|^2 = o(1), \tag{A.3}$$

where

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

The sparsity condition assumes that most of the off-diagonal entries of $\boldsymbol{\Sigma}_u$ are outside of S_U.

The above two convergences are weighted cross-sectional and serial double sums, where the weights satisfy $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} e_{it}(1-\mathbf{f}_t'\bar{\mathbf{w}}) = O_P(1)$ for each i. The proofs of (A.2) and (A.3) are given in Appendix A.6.

Proof of PropositionA.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 $a = o_P(1)$. 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,

$$\widehat{\sigma}_{ij} = (\widehat{\boldsymbol{\Sigma}}_u)_{ij} = \begin{cases} s_{ii}, & \text{if } i=j, \\ s_{ij}, & \text{if } i\neq j, |s_{ij}| > C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2} \\ 0, & \text{if } i\neq j, |s_{ij}| \leq C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2} \end{cases} \tag{A.4}$$

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

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

We first examine a_3. Note that

$$a_3 = \frac{T}{\sqrt{N}}\sum_{(i,j)\in S_U^c}(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_j\widehat{\sigma}_{ij} \leq \max_{i\leq N}|(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i|^2\frac{T}{\sqrt{N}}\sum_{(i,j)\in S_U^c}|\widehat{\sigma}_{ij}| \equiv a_{31}.$$

We have,

$$P(a_{31} > T^{-1}) \leq P(\max_{(i,j)\in S_U^c} |\widehat{\sigma}_{ij}| \neq 0) \leq P(\max_{(i,j)\in S_U^c} |s_{ij}| > C(s_{ii}s_{jj}\frac{\log N}{T})^{1/2}).$$

Because s_{ii} is uniformly (across i) bounded away from zero with probability approaching one, and $\max_{(i,j)\in S_U^c} |s_{ij}| = O_P(\sqrt{\frac{\log N}{T}})$. Hence for any $\epsilon > 0$, when C in the threshold is large enough, $P(a_{31} > T^{-1}) < \epsilon$, this implies $a_{31} = o_P(1)$, and thus $a_3 = o_P(1)$.

The proof is finished once we establish $a_i = o_P(1)$ for $i = 1, 2$, which are given respectively by the following lemmas.

□

Lemma A.4. *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 $(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i = (1 - \bar{\mathbf{f}}'\mathbf{w})^{-1} \frac{1}{T} \sum_{s=1}^T (1 - \mathbf{f}_s'\mathbf{w})(\mathbf{u}_s'\boldsymbol{\Sigma}_u^{-1})_i = c\frac{1}{T} \sum_{s=1}^T (1 - \mathbf{f}_s'\mathbf{w}) e_{is}$, where $c = (1 - \bar{\mathbf{f}}'\mathbf{w})^{-1} = O_P(1)$. Hence

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

By (A.2) (whose proof is given in Appendix A.6), $Ea_{12}^2 = o(1)$. 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 $\max_{i\leq N} \frac{1}{T} \sum_{t=1}^T (\widehat{u}_{it} - u_{it})^2 = O_P(\frac{\log N}{T})$ by Lemma 3.1 of Fan et al. (2011). Since $\|\widehat{\boldsymbol{\alpha}}\|^2 = O_P(\frac{N \log N}{T})$, $\|\boldsymbol{\Sigma}_u^{-1}\| = O(1)$ and $N(\log N)^3 = o(T^2)$,

$$a_{111} \leq O_P(\frac{\log N}{T}) \frac{T}{\sqrt{N}} \|\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1}\|^2 = O_P(\frac{(\log N)^2\sqrt{N}}{T}) = o_P(1),$$

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

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

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

Lemma A.5. *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

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

where

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

Under H_0, $\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\alpha}} = \frac{1}{T}(1 - \bar{\mathbf{f}}'\mathbf{w})^{-1}\sum_{t=1}^T \boldsymbol{\Sigma}_u^{-1}\mathbf{u}_t(1 - \mathbf{f}_t'\mathbf{w})$, and $\mathbf{e}_t = \boldsymbol{\Sigma}_u^{-1}\mathbf{u}_t$, we have

$$\begin{aligned} a_{22} &= \frac{T}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_j\frac{1}{T}\sum_{t=1}^T(u_{it}u_{jt} - Eu_{it}u_{jt}) \\ &= \frac{Tc}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}\frac{1}{T}\sum_{s=1}^T(1 - \mathbf{f}_s'\mathbf{w})e_{is}\frac{1}{T}\sum_{k=1}^T(1 - \mathbf{f}_k'\mathbf{w})e_{jk}\frac{1}{T}\sum_{t=1}^T(u_{it}u_{jt} - Eu_{it}u_{jt}). \end{aligned}$$

By (A.3) (whose proof is given in Appendix A.6), $Ea_{22}^2 = o(1)$.

On the other hand, $a_{21} = a_{211} + a_{212}$, where

$$a_{211} = \frac{T}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Sigma}_u^{-1})_j\frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it} - u_{it})(\widehat{u}_{jt} - u_{jt}),$$

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

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

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

Similar to the proof of term a_{112} in Lemma A.4, $\max_{ij} |\frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{u}_{jt}-u_{jt})| = O_P(\frac{\log N}{T})$.

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

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

A.6 Proof of (A.2) and (A.3)

For any index set A, we let $|A|_0$ denote its number of elements.

Lemma A.6. *Recall that* $\mathbf{e}_t = \boldsymbol{\Sigma}_u^{-1}\mathbf{u}_t$. e_{it} *and* u_{jt} *are independent if* $i \neq j$.

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

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

Then $\text{cov}(\mathbf{u}_t', \mathbf{e}_t') = \mathbf{A}\text{cov}(\mathbf{u}_t', \mathbf{u}_t')\mathbf{A}$, 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 (A.2)

Let $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^{T} e_{is}(1 - \mathbf{f}_s'\mathbf{w}))^2$. The goal is to show $EX^2 = o(T)$. We show respectively $\frac{1}{T}(EX)^2 = o(1)$ and $\frac{1}{T}\text{var}(X) = o(1)$. The proof of (A.2) is the same regardless of the type of sparsity in Assumption 5.2. For notational simplicity, let

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

Then $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi_{it}(\frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is})^2$. Because of the serial independence, ξ_{it} is independent of ζ_{js} if $t \neq s$, for any $i, j \leq N$, which implies $\text{cov}(\xi_{it}, \zeta_{is}\zeta_{ik}) = 0$ as long as either $s \neq t$ or $k \neq t$.

Expectation

For the expectation,

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

where the second last equality follows since $E\xi_{it} = E\zeta_{it} = 0$ and when $k \neq t$ $\text{cov}(\xi_{it}, \zeta_{it}\zeta_{ik}) = E\xi_{it}\zeta_{it}\zeta_{ik} = E\xi_{it}\zeta_{it}E\zeta_{ik} = 0$. 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,

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

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

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

Hence $B_1 = O(1)$.

We now show $\frac{1}{T}B_2 = o(1)$. Once this is done, it implies $\frac{1}{T}\text{var}(X) = o(1)$. The proof of (A.2) is then completed because $\frac{1}{T}EX^2 = \frac{1}{T}(EX)^2 + \frac{1}{T}\text{var}(X) = o(1)$.

For two variables X, Y, writing $X \perp Y$ if they are independent. Note that $E\xi_{it} = E\zeta_{is} = 0$, and when $t \neq s$, $\xi_{it} \perp \zeta_{js}$, $\xi_{it} \perp \xi_{js}$, $\zeta_{it} \perp \zeta_{js}$ 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 $\text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jl}\zeta_{jv}\zeta_{jp}) = 0$. 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

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

Hence

$$B_2 = \frac{1}{NT^3}\sum_{i\neq j}\sum_{(t,s,k,l,v,p)\in\Xi}\text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jl}\zeta_{jv}\zeta_{jp}).$$

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 $\frac{1}{NT^3}\sum_{i\neq j}\sum_{(t,s,k,l,v,p)\in\Xi_2}\text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jl}\zeta_{jv}\zeta_{jp}) = O(\frac{1}{NT^3}N^2T^2) = O(\frac{N}{T})$, which implies

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

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.

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

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

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

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

By Lemma A.6, u_{js} and e_{is} are independent for $i \neq j$. Also, u_{js} and $\mathbf{f}_s$ are independent, which implies ξ_{js} and ζ_{is} are independent. So $E\xi_{js}\zeta_{is} = 0$. It follows that $B_{22} = 0$.

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

By the definition $\mathbf{e}_s = \mathbf{\Sigma}_u^{-1}\mathbf{u}_s$, $\text{cov}(\mathbf{e}_s) = \mathbf{\Sigma}_u^{-1}$. Hence $Ee_{is}e_{js} = (\mathbf{\Sigma}_u^{-1})_{ij}$, which implies $B_{23} \leq O(\frac{N}{NT})\|\mathbf{\Sigma}_u^{-1}\|_1 = o(1)$.

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

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

Finally, $B_{25} = \frac{1}{NT^4}\sum_{i\neq j}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq s,t} E\xi_{it}\zeta_{jt}E\zeta_{is}\xi_{js}E\zeta_{il}\zeta_{jl} = 0$, because $E\zeta_{is}\xi_{js} = 0$ when $i \neq j$, following from Lemma A.6. Therefore, $\frac{1}{T}B_2 = o(1) + O(\frac{N}{T^2}) = o(1)$.

A.6.2 Proof of (A.3)

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

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

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

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

The goal is to show $\frac{1}{T}EZ^2 = o(1)$. We respectively show $\frac{1}{T}(EZ)^2 = o(1) = \frac{1}{T}\text{var}(Z)$.

Expectation

The proof for the expectation is the same regardless of the type of sparsity in Assumption 5.2, and is very similar to that of (A.2). In fact,

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

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

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

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

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

So $A_1 = O(1)$.

Note that $E\xi_{ijt} = E\zeta_{is} = 0$, and when $t \neq s$, $\xi_{ijt} \perp \zeta_{ms}$, $\xi_{ijt} \perp \xi_{mns}$, $\zeta_{it} \perp \zeta_{js}$ (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 $\text{cov}(\xi_{ijt}\zeta_{is}\zeta_{jk}, \xi_{mnl}\zeta_{mv}\zeta_{np}) = 0$. Hence for the same set Ξ defined as before, it satisfies: $|\Xi|_0 \leq CT^3$ for some $C > 1$, and

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

We proceed by studying the two cases of Assumption 5.2 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 $\frac{1}{T}\text{var}(Z) = o(1)$. The proof is then completed because $\frac{1}{T}EZ^2 = \frac{1}{T}(EZ)^2 + \frac{1}{T}\text{var}(Z) = o(1)$.

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

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

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

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

Similar to the proof of the first statement, for the same set Ξ_1 that contains exactly three distinct indices in each of its element, (recall $|H|_0 = O(N)$)

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

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

$$A(i) = \{j \leq N : \text{cov}(u_{it}, u_{jt}) \neq 0\} = \{j \leq N : (i,j) \in S_U\}.$$

Lemma A.7. *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 $\text{cov}(u_{it}u_{jt}, u_{kt}u_{pt}) \neq 0$.

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 definition, $k \notin A(i) \cup A(j)$, which implies $\text{cov}(u_{it}, u_{kt}) = \text{cov}(u_{jt}, u_{kt}) = 0$. By the Gaussianity, u_{kt} is independent of (u_{it}, u_{jt}). Hence if $p \in A(k)$ is such that $\text{cov}(u_{it}u_{jt}, u_{kt}u_{pt}) \neq 0$, then u_{pt} should be correlated with either u_{it} or u_{jt}. We thus must have $p \in A(i) \cup A(j)$. In other words, there is $p \in A(i) \cup A(j)$ such that $\text{cov}(u_{kt}, u_{pt}) \neq 0$,

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 $m_N = O(1)$, $\max_{i \le N} |A(i)|_0 = O(1)$, which implies $\max_{i,j} |M(i,j)|_0 = O(1)$, yielding the result. □

Now we define and bound each of A_{2h}. For any $(i,j) \in H = \{(i,j) : (\mathbf{\Sigma}_u)_{ij} \neq 0\}$, we must have $j \in A(i)$. So

$$\begin{aligned} A_{21} &= \frac{1}{NT^4} \sum_{(i,j) \in H,} \sum_{(m,n) \in H, (m,n) \neq (i,j),} \sum_{t=1}^{T} \sum_{s \neq t} \sum_{l \neq t,s} E\xi_{ijt}\xi_{mnt} E\zeta_{is}\zeta_{js} E\zeta_{ml}\zeta_{nl} \\ &\le O(\frac{1}{NT}) \sum_{(i,j) \in H,} \sum_{(m,n) \in H, (m,n) \neq (i,j)} |E\xi_{ijt}\xi_{mnt}| \\ &\le O(\frac{1}{NT}) \sum_{(i,j) \in H} (\sum_{m \in A(i) \cup A(j)} \sum_{n \in A(m)} + \sum_{m \notin A(i) \cup A(j)} \sum_{n \in A(m)}) |\mathrm{cov}(u_{it}u_{jt}, u_{mt}u_{nt})|. \end{aligned}$$

The first term is $O(\frac{1}{T})$ because $|H|_0 = O(N)$ and $|A(i)|_0$ is bounded uniformly by $m_N = O(1)$. So the number of summands in $\sum_{m \in A(i) \cup A(j)} \sum_{n \in A(m)}$ is bounded. 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.7. Hence $A_{21} = o(1)$.

Similarly, applying Lemma A.7,

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

which is proved in the same lines of those of A_{21}.

Also note three simple facts: (1) $\max_{j \le N} |A(j)|_0 = O(1)$, (2) $(m,n) \in H$ implies $n \in A(m)$, and (3) $\xi_{mms} = \xi_{nms}$. The term A_{23} is defined as

$$\begin{aligned} A_{23} &= \frac{1}{NT^4} \sum_{(i,j) \in H,} \sum_{(m,n) \in H, (m,n) \neq (i,j),} \sum_{t=1}^{T} \sum_{s \neq t} \sum_{l \neq t,s} E\xi_{ijt}\zeta_{it} E\zeta_{js}\xi_{mns} E\zeta_{ml}\zeta_{nl} \\ &\le O(\frac{1}{NT}) \sum_{j=1}^{N} \sum_{i \in A(j)} 1 \sum_{(m,n) \in H, (m,n) \neq (i,j)} |E\zeta_{js}\xi_{mns}| \end{aligned}$$

$$\leq \quad O(\frac{2}{NT})\sum_{j=1}^{N}\sum_{n\in A(j)}|E\zeta_{js}\xi_{jns}| + O(\frac{1}{NT})\sum_{j=1}^{N}\sum_{m\neq j, n\neq j}|E\zeta_{js}\xi_{mns}| = a + b.$$

Term $a = O(\frac{1}{T})$. For b, note that Lemma A.6 implies that when $m, n \neq j$, $u_{ms}u_{ns}$ and e_{js} are independent because of the Gaussianity. Also because $\mathbf{u}_s$ and $\mathbf{f}_s$ are independent, hence ζ_{js} and ξ_{mms} are independent, which implies that $b = 0$. Hence $A_{23} = o(1)$.

The same argument as of A_{23} also implies

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

Finally, because $\sum_{(i,j)\in H}1 \leq \sum_{i=1}^{N}\sum_{j\in A(i)}1 \leq m_N\sum_{i=1}^{N}1$, and $m_N = O(1)$, we have

$$\begin{aligned}
A_{25} &= \frac{1}{NT^4}\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j),}\sum_{t=1}^{T}\sum_{s\neq t}\sum_{l\neq t,s}E\xi_{ijt}\zeta_{it}E\zeta_{is}\zeta_{ms}E\xi_{mnl}\zeta_{nl} \\
&\leq O(\frac{1}{NT})\sum_{(i,j)\in H,}\sum_{(m,n)\in H,(m,n)\neq(i,j)}|E\xi_{ijt}\zeta_{it}E\zeta_{is}\zeta_{ms}E\xi_{mnl}\zeta_{nl}| \\
&\leq O(\frac{1}{NT})\sum_{i=1}^{N}\sum_{m=1}^{N}|E\zeta_{is}\zeta_{ms}| \leq O(\frac{1}{NT})\sum_{i=1}^{N}\sum_{m=1}^{N}|(\mathbf{\Sigma}_u^{-1})_{im}|E(1-\mathbf{f}_s'\mathbf{w})^2 \\
&\leq O(\frac{N}{NT})\|\mathbf{\Sigma}_u^{-1}\|_1 = o(1).
\end{aligned}$$

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

References

Anatolyev, S. and Gospodinov, N. (2009). Specification testing in models with many instruments. *Econometric Theory*. **27** 427-441

Andrews, D. (2005). Cross-sectional regression with common shocks. *Econometrica*. **73** 1551-1585.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Jour. Ameri. Statist. Assoc.*, **96**, 939-967.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*, 4 edition, Wiley.

Baltagi, B., Feng, Q. and Kao, C. (2012). A Lagrange Multiplier test for cross-sectional dependence in a fix effects panel data model. *Journal of Econometrics*. **170**, 164-177.

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.

Breitung, J., Roling, C. and Salish, N. (2013). LM-type tests for slope homogeneity in panel data models. *Manuscript*.

Breusch, T. and Pagan, A. (1980). The Lagrange multiplier test and its application to model specification in econometrics, *Review of Economic Studies*. **47**, 239-254.

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. *Ann. Statist.* **40**, 2389-2420.

Chao, J., Hausman, J., Newey, W., Swanson, N. and Woutersen, T. (2012). Testing overidentifying restrictions with many instruments and heteroskedasticity. *manuscript*.

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

Chen, S. X., and Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808-835.

Chen, S. X., Zhang, L. X., and Zhong, P. S. (2010). Tests for high-dimensional covariance matrices. *Jour. Ameri. Statist. Assoc.*, **105**, 810-819.

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

Donald, S., Imbens, G. and Newey, W. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econometrics*,**117** 55-93

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

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

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of American Statistical Association*, **91**, 674–688.

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

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

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statisitcs*. **38** 1686-1732

Hansen, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*. **50** 1029-1054.

Kelejian, 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.

Merton, R. C. (1973) Theory of rational option pricing. *Bell Journal of Economics and Management Science*, **4**, 141-183.

Pesaran, M. H., Ullah, A. and Yamagata, T. (2008). A bias-adjusted LM test of error cross section independence. *Econometrics Journal.* **11**, 105-127.

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

Phillips, P. and Sul, D. (2003). Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal.* **6**, 217-259

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.

Sarafidis, V., Yamagata, T. and Robertson, D. (2009) A test of cross section dependence for a linear dynamic panel model with regressors. *Journal of Econometrics.* **148**, 149-161.

Sentana, E.. (2009) The econometrics of mean-variance efficiency tests: a survey. *The Econometrics Journal.* **12**, C65-C101.

Srivastava, M. and Du, M. (2008) A test for the mean vector with fewer observations than the dimension. *Journal of multivariate analysis.* **99**, 386-402.