Factor Models: Testing and Forecasting

Jiawei Yao

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Operations Research and Financial Engineering

Adviser: Professor Jianqing Fan

January 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Abstract

This dissertation focuses on two aspects of factor models, testing and forecasting. For testing, we investigate a more general high-dimensional testing problem, with an emphasis on panel data models. Specifically, we propose a novel technique to boost the power of testing a high-dimensional vector $H : \boldsymbol{\theta} = 0$ against sparse alternatives. Existing tests based on quadratic forms such as the Wald statistic often suffer from low powers, whereas more powerful tests such as thresholding and extreme-value tests require either stringent conditions or bootstrap to derive the null distribution, and often suffer from size distortions. Based on a screening technique, we introduce a "power enhancement component", which is zero under the null hypothesis with high probability, but diverges quickly under sparse alternatives. The proposed test statistic combines the power enhancement component with an asymptotically pivotal statistic, and strengthens the power under sparse alternatives. As a byproduct, the power enhancement component also consistently identifies the elements that violate the null hypothesis.

Next, we consider forecasting a single time series using many predictors when nonliearity is present. We develop a new methodology called sufficient forecasting, by connecting sliced inverse regression with factor models. The sufficient forecasting correctly estimates projections of the underlying factors and provides multiple predictive indices for further investigation. We derive asymptotic results for the estimate of the central space spanned by these projection directions. Our method allows the number of predictors larger than the sample size, and therefore extends the applicability of inverse regression. Numerical experiments demonstrate that the proposed method improves upon a linear forecasting model. Our results are further illustrated in an empirical study of macroeconomic variables, where sufficient forecasting is found to deliver additional predictive power over conventional methods.

Acknowledgements

I would like to express my sincerest gratitude to my adviser, Professor Jianqing Fan, for his great mentorship, constant support and unreserved encouragement during the past four years. Not only has he guided me into the wondrous beauty of academic research, but he also helped me build my character in life. I benefited immensely from his unsurpassed knowledge, his genuine care for students and his appetite for adventure over love of ease. Without him this dissertation would never be possible.

It is my great honor to have Professor Marc Hallin and Professor Samory Kpotufe sitting on my committee. In particular, I thank Professor Marc Hallin for teaching and inspiring me through many courses and discussions we had. I am also truly grateful to Professor Han Liu for the precious time he spent on reading my thesis. His vast knowledge and thought-provoking comments have greatly improved my work, which I would always cherish. In addition, I am deeply indebted to my collaborators as well as sincere friends, Professor Yuan Liao and Professor Lingzhou Xue, whose help and devotion have greatly enriched my research experience.

My warmest thanks to Professors Yacine Ait-Sahalia, Erhan Çinlar, Philippe Rigollet, Sébastien Bubeck, Ramon van Handel and Patrick Cheridito, who offered generous help along my graduate studies as well as research. I also thank the lovely members of Statlab and all my friends; it is their encouragement that helps me achieve this far. They have left me unforgettable memories, caressed by the smiles of yesterday.

Finally, I owe deep gratitude to my beloved parents, Jieding Yao and Yumei Hui, for bringing me to the joy and inspiration of life. I wish to thank my dearest brother, Jiaxiong Yao, whose constant support gives me wisdom and strength. Most importantly, I thank my wife, Bo Feng, for always giving me strongest faith, limitless love and unfailing hopes in our adventurous journey. I simply can't imagine a life without them.

To my family

Contents

Abstract iii

Acknowledgements iv

List of Tables ix

List of Figures x

1 Power Enhancement in High Dimensional Cross-Sectional Tests 1

1.1 Introduction 1

1.2 Power Enhancement in high dimensions 6

1.2.1 Power enhancement 7

1.2.2 Construction of power enhancement component 9

1.2.3 Comparisons with thresholding and extreme-value tests 11

1.3 Asymptotic properties 13

1.3.1 Main results 13

1.3.2 Power enhancement for quadratic tests 17

1.3.3 Low power of quadratic statistics under sparse alternatives 18

1.4 Application: Testing Factor Pricing Models 19

1.4.1 The model 19

1.4.2 Power enhancement component 21

1.4.3 Feasible Wald test in high dimensions 22

1.4.4 Does the thresholded covariance estimator affect the size? 24

1.4.5 Regularity conditions 25

1.4.6 Asymptotic properties . . . 27
1.5 Application: Testing Cross-Sectional Independence . . . 29
1.5.1 The model . . . 29
1.5.2 Power enhancement test . . . 30
1.5.3 Asymptotic properties . . . 31

2 Sufficient Forecasting Using Factor Models 34
2.1 Introduction . . . 34
2.2 Methodology . . . 37
2.2.1 Factor models and forecasting . . . 37
2.2.2 Sliced inverse regression . . . 39
2.2.3 Sufficient forecasting . . . 41
2.2.4 Determining the number of factors . . . 43
2.3 Asymptotic properties . . . 44
2.3.1 Assumptions . . . 44
2.3.2 Convergence of $\widehat{\mathbf{\Sigma}}_{f|y}$ . . . 45
2.3.3 Connection to linear estimators . . . 46
2.4 Applications . . . 48
2.5 Future work . . . 49

3 Numerical studies 51
3.1 Numerical studies for power enhancement test . . . 51
3.1.1 Testing factor pricing models . . . 51
3.1.2 Testing cross-sectional independence . . . 54
3.1.3 Empirical Study . . . 56
3.2 Numerical studies for sufficient forecasting . . . 59
3.2.1 Linear forecast . . . 59
3.2.2 Factor Interaction . . . 61

3.2.3 An Empirical Example 62

4 Concluding remarks **65**

5 Technical Proofs **67**

5.1 Proofs for Chapter 1 67

5.1.1 Proofs for Section 1.3 67

5.1.2 Proofs for Section 1.4 72

5.1.3 Proofs for Section 1.5 76

5.1.4 Supplementary Material 81

5.2 Proofs for Chapter 2 93

5.2.1 Proof of Proposition 2.2.1 94

5.2.2 Proof of Theorem 2.3.1 94

5.2.3 Proof of Proposition 2.3.1 98

Bibliography **100**

List of Tables

3.1 Means and covariances used to generate $\mathbf{b}_i$ and $\mathbf{f}_t$ 52

3.2 Size and power (%) of tests for simulated Fama-French three-factor model . . 53

3.3 Size and power (%) of tests for cross-sectional independence 55

3.4 Summary of descriptive statistics and testing results 57

3.5 Simulated Forecast Performance (Linear Model) 60

3.6 Simulated Forecast Performance (Factor Interaction) 62

3.7 Out-of-sample Macroeconomic Forecasting 63

List of Figures

3.1 Dynamics of p-values and percents of selected stocks 58

3.2 Histograms of p-values for J_{wald} and PE. 58

3.3 Forecasting results for CCINRV (consumer credit outstanding) 64

Chapter 1

Power Enhancement in High Dimensional Cross-Sectional Tests

1.1 Introduction

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

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

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

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

tradable market risk factors. Then, the excess return has the following decomposition:

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

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 θ_i should be zero, known as the "mean-variance efficiency" pricing, for any asset i. An important question is then if such a pricing theory can be validated by empirical data, namely we wish to test the null hypothesis $H_0 : \boldsymbol{\theta} = 0$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)'$ is the vector of intercepts for all N financial assets. As the factor pricing model is derived from theories of financial economics (Merton, 1973; Ross, 1976), one would expect that inefficient pricing by the market should only occur to a small fractions of exceptional assets. Indeed, our empirical study of the constituents in the S&P 500 index indicates that there are only a couple of significant nonzero-alpha stocks, corresponding to a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market. Therefore, it is important to construct tests that have high power when $\boldsymbol{\theta}$ is sparse.

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

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

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

critical values that can dominate the signals in the sparse alternatives. A formal proof of this will be given in Section 1.3.

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

Power Enhancement Properties:

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

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

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

Our constructed power enhancement test takes the form

$$J = J_0 + J_1.$$

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

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

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

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

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

As for specific applications, we study the tests of the aforementioned factor pricing model, and of cross-sectional independence in mixed effect panel data models:

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

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

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

that is, under the null hypothesis, the $n \times n$ covariance matrix $\mathbf{\Sigma}_u$ of $\{u_{it}\}_{i \leq n}$ is diagonal. In empirical applications, weak cross-sectional correlations are often present, which results in a sparse covariance $\mathbf{\Sigma}_u$ with just a few nonzero off-diagonal elements. This leads to a sparse vector $\boldsymbol{\theta} = (\rho_{12}, \rho_{13}, ..., \rho_{n-1,n})$. The dimensionality $N = n(n-1)/2$ can be much larger than the number of observations. Therefore, the power enhancement in sparse alternatives is very important to the testing problem.

There has been a large literature on high-dimensional cross-sectional tests. For instance, the literature on testing the factor pricing model is found in Gibbons et al. (1989), MacKinlay and Richardson (1991), Beaulieu et al. (2007) and Pesaran and Yamagata (2012), all in quadratic forms. Moreover, for the mixed effect panel data model, most of the existing statistics in the literature are based on the sum of squared residual correlations, which also accumulates many off-diagonal estimation errors in the covariance matrix of $(u_{1t}, ..., u_{nt})$. The literature includes Breusch and Pagan (1980), Pesaran et al. (2008), Baltagi et al. (2012), etc. In addition, our problem is also related to the test with a restricted parameter space, previously considered by Andrews (1998), who improves the power by directing towards the "relevant" alternatives (also see Hansen (2003) for a related idea). Recently, Chernozhukov et al. (2013) proposed a high-dimensional inequality test, and employed an extreme value statistic, whose critical value is determined through applying the moderate deviation theory on an upper bound of the rejection probability. In contrast, the asymptotic distribution of our proposed power enhancement statistic is determined through the pivotal statistic J_1, and the power is improved via screening off most of the noises under sparse alternatives.

In a related recent paper by Gagliardini et al. (2011), they studied estimating and testing about the risk premia in a CAPM model. While we also study a large panel of stock returns as a specific example and double asymptotics (as $N, T \to \infty$), the problems and approaches being considered are very different. This chapter addresses a general problem of enhancing powers under high-dimensional sparse alternatives.

The remainder of this chapter is organized as follows. Section 1.2 sets up the preliminaries and highlights the major differences from existing tests. Section 1.3 presents the main result of power enhancement test. As applications to specific cases, Section 1.4 and Section 1.5 respectively study the factor pricing model and test of cross-sectional independence. We defer simulation results in Section 3.1, Chapter 3, along with an empirical application to the stocks in the S&P 500 index. All the proofs are given in Section 5.1, Chapter 5.

Throughout this dissertation, for a symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent its minimum and maximum eigenvalues. Let $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_1$ denote its operator norm and l_1-norm respectively, defined by $\|\mathbf{A}\|_2 = \lambda_{\max}^{1/2}(\mathbf{A}'\mathbf{A})$ and $\max_i \sum_j |\mathbf{A}_{ij}|$. For a vector $\boldsymbol{\theta}$, define $\|\boldsymbol{\theta}\| = (\sum_j \theta_j^2)^{1/2}$ and $\|\boldsymbol{\theta}\|_{\max} = \max_j |\theta_j|$. For two deterministic sequences a_T and b_T, we write $a_T \ll b_T$ (or equivalently $b_T \gg a_T$) if $a_T = o(b_T)$. Also, $a_T \asymp b_T$ 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.

1.2 Power Enhancement in high dimensions

This section introduces power enhancement techniques and provides heuristics to justify the techniques. Their differences with related ideas in the literature are also highlighted.

1.2.1 Power enhancement

Consider a testing problem:

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

where $\Theta_a \subset \mathbb{R}^N \backslash \{\mathbf{0}\}$ is an alternative set. A typical example is $\Theta_a = \{\boldsymbol{\theta} : \boldsymbol{\theta} \neq 0\}$. Suppose we observe a stationary process $\mathbf{D} = \{\mathbf{D}_t\}_{t=1}^T$ of size T. Let $J_1(\mathbf{D})$ be a certain test statistic, and for notational simplicity, we write $J_1 = J_1(\mathbf{D})$. Often J_1 is constructed such that under H_0, it has a non-degenerate limiting distribution F: As $T, N \to \infty$,

$$J_1 | H_0 \to^d F. \tag{1.1}$$

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

$$\limsup_{T,N\to\infty} P(J_1 > F_q | H_0) = q. \tag{1.2}$$

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

$$\liminf_{T,N\to\infty} \inf_{\boldsymbol{\theta}\in\Theta(J_1)} P(J_1 > F_q | \boldsymbol{\theta}) \to 1. \tag{1.3}$$

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

In a data-rich environment, econometric models often involve high-dimensional parameters in which $\dim(\boldsymbol{\theta}) = N$ can grow fast with the sample size T. We are particularly interested in *sparse alternatives* $\Theta_s \subset \Theta_a$ under which H_0 is violated only on a couple of exceptional components of $\boldsymbol{\theta}$. Specifically, when $\boldsymbol{\theta} \in \Theta_s$, the number of non-vanishing com-

ponents is much less than N. As a result, its l_2-norm is relatively small. Therefore, under sparse alternative Θ_s, the omnibus test J_1 typically has lower power, due to the accumulation of high-dimensional estimation errors. Detailed explanations are given in Section 1.3.3 below.

We introduce a *power enhancement principle* for high-dimensional sparse testing, by bringing in a data-dependent component J_0 that satisfies the **Power Enhancement Properties** as defined in Section 1.1. The introduced component J_0 does not serve as a test statistic on its own, but is added to a classical statistic J_1 that is often pivotal (e.g., Wald-statistic), so the proposed test statistic is defined by

$$J = J_0 + J_1.$$

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

1. The critical region of J is defined by

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

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

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

3. Under mild conditions, $P(J_0 = 0|H_0) \to 0$. Hence when (1.1) is satisfied, we have

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

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

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

1.2.2 Construction of power enhancement component

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

$$\widehat{S} = \{j : |\widehat{\theta}_j| > \widehat{v}_j^{1/2}\delta_{N,T}, j = 1, ..., N\}, \tag{1.4}$$

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

$$\inf_{\boldsymbol{\theta}\in\Theta_a\cup\{\mathbf{0}\}} P(\max_{j\le N} |\widehat{\theta}_j - \theta_j|/\widehat{v}_j^{1/2} < \delta_{N,T}/2|\boldsymbol{\theta}) \to 1 \tag{1.5}$$

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

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

By (1.4) and (1.5), under $H_0 : \boldsymbol{\theta} = 0$,

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

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

Let $\{v_j\}_{j\leq N}$ be the population counterpart of $\{\widehat{v}_j\}_{j\leq N}$. For instance, one can take v_j as the asymptotic variance of $\widehat{\theta}_j$, and $\widehat{v}_j$ as its estimator. To satisfy the power-enhancement property, note that the screening set mimics

$$S(\boldsymbol{\theta}) = \left\{ j : |\theta_j| > 2v_j^{1/2}\delta_{N,T}, j = 1, ..., N \right\}, \tag{1.6}$$

and in particular $S(\mathbf{0}) = \emptyset$. We shall show in Theorem 1.3.1 below that $P(\widehat{S} = S(\boldsymbol{\theta})|\boldsymbol{\theta}) \to 1$, for all $\boldsymbol{\theta} \in \Theta_a \cup \{\mathbf{0}\}$. Thus, the subvector $\widehat{\boldsymbol{\theta}}_{\widehat{S}} = (\widehat{\theta}_j : j \in \widehat{S})$ behaves like $\boldsymbol{\theta}_S = (\theta_j : j \in S(\boldsymbol{\theta}))$, which can be interpreted as estimated significant signals. If $S(\boldsymbol{\theta}) \neq \emptyset$, then by the definition of $\widehat{S}$ and $\delta_{N,T} \to \infty$, we have

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

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

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

As a byproduct, the screening set consistently identifies the elements of $\boldsymbol{\theta}$ that violate the null hypothesis.

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

$$J = J_0 + J_1.$$

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

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

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

In sparse alternatives where $\|\boldsymbol{\theta}\|$ may not grow fast with N but $\boldsymbol{\theta} \in \Theta_s$, the combined test $J_0 + J_1$ can be very powerful. In contrast, we will formally show in Theorem 1.3.4 below that the conventional Wald test J_1 can have very low power on its own. On the other hand, when the alternative is "dense" in the sense that $\|\boldsymbol{\theta}\|$ grows fast with N, the conventional test J_1 itself is consistent. In this case, J is still as powerful as J_1. Therefore, if we denote $\Theta(J_1) \subset \mathbb{R}^N / \{\mathbf{0}\}$ as the set of alternative $\boldsymbol{\theta}$'s against which the classical J_1 test has power converging to one, then the combined $J = J_0 + J_1$ test has power converging to one against $\boldsymbol{\theta}$ on

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

We shall show in Section 1.3 that the power is enhanced uniformly over $\boldsymbol{\theta} \in \Theta_s \cup \Theta(J_1)$.

1.2.3 Comparisons with thresholding and extreme-value tests

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

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

In the high-dimensional testing literature, there are mainly two types of statistics with good power under sparse alternatives: extreme value test and thresholding test respectively. The test based on extreme values studies the maximum deviation from the null hypothesis across the components of $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, ..., \widehat{\theta}_N)$, and forms the statistic based on $\max_{j \leq N} |\frac{\widehat{\theta}_j}{w_j}|^\delta$ for some $\delta > 0$ and a weight w_j (e.g., Cai et al. (2013), Chernozhukov et al. (2013)). Such a test statistic typically converges slowly to its asymptotic counterpart. An alternative test is based on thresholding: for some $\delta > 0$ and pre-determined threshold level t_T,

$$R = \sqrt{T} \sum_{j=1}^{N} |\frac{\widehat{\theta}_j}{w_j}|^\delta 1\{|\widehat{\theta}_j| > t_T w_j\} \tag{1.7}$$

The accumulation of estimation errors is prevented due to the threshold $1\{|\widehat{\theta}_j| > t_T w_j\}$ (see, e.g., Fan (1996) and Zhong et al. (2013)) for sufficiently large t_T. In a low-dimensional setting, Hansen (2005) suggested using a threshold to enhance the power in a similar way.

Although (1.7) looks similar to J_0, the ideas behind are very different. Both extreme value test and thresholding test require regularity conditions that may be restrictive in econometric applications. For instance, it can be difficult to employ the central limit theorem directly on (1.7), as it requires the covariance between $\widehat{\theta}_j$ and $\widehat{\theta}_{j+k}$ decay fast enough as $k \to \infty$ (Zhong et al., 2013). In cross-sectional testing problems, this essentially requires an explicit ordering among the cross-sectional units which is, however, often unavailable in panel data applications. In addition, as (1.7) involves effectively limited terms of summations due to thresholding, the asymptotic theory does not provide adequate approximations,

resulting size-distortion in applications. For example, when t_T is taken slightly less than $\max_{j\leq N}|\widehat{\theta}_j|/w_j$, R becomes the extreme statistic. When t_T is small (e.g. 0), R becomes a traditional test, which is not powerful in detecting sparse alternatives, though it can have good size properties.

1.3 Asymptotic properties

1.3.1 Main results

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

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

Assumption 1.3.1. *As $T, N \to \infty$, the sequence $\delta_{N,T} \to \infty$, and the estimators $\{\widehat{\theta}_j, \widehat{v}_j\}_{j\leq N}$ are such that*

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

(ii) $\inf_{\boldsymbol{\theta}\in\Theta} P(4/9 < \widehat{v}_j/v_j < 16/9, \forall j = 1, ..., N|\boldsymbol{\theta}) \to 1$.

The normalizing constant v_j is often taken as the asymptotic variance of $\widehat{\theta}_j$, with $\widehat{v}_j$ being its consistent estimator. The constants 4/9 and 16/9 in condition (ii) are not optimally chosen, as this condition only requires $\{\widehat{v}_j\}_{j\leq N}$ be *not-too-bad* estimators of their population counterparts.

In many high-dimensional problems with strictly stationary data that satisfy strong mixing conditions, following from the large-deviation theory, typically, $\max_{j\leq N}|\widehat{\theta}_j - \theta_j|/\widehat{v}_j^{1/2} =$

$O_P(\sqrt{\log N})$. Therefore, we shall fix

$$\delta_{N,T} = \log(\log T)\sqrt{\log N}, \tag{1.8}$$

which is a high criticism that slightly dominates the standardized noise level. We shall provide primitive conditions for this choice of $\delta_{N,T}$ in the subsequent sections, so that Assumption 1.3.1 holds.

Recall that $\widehat{S}$ and $S(\boldsymbol{\theta})$ are defined by (1.4) and (1.6) respectively for a given $\boldsymbol{\theta} \in \Theta$ and its consistent estimator $\widehat{\boldsymbol{\theta}}$. In particular, $S(\boldsymbol{\theta}) = \left\{ j : |\theta_j| > 2v_j^{1/2}\delta_{N,T}, j = 1, ..., N \right\}$, so under $H_0 : \boldsymbol{\theta} = 0$, $S(\boldsymbol{\theta}) = \emptyset$. Note that Θ denotes the parameter space containing both the null and alternative hypotheses. The following theorem characterizes the asymptotic behavior of $J_0 = \sqrt{N}\sum_{j\in\widehat{S}} \widehat{\theta}_j^2 \widehat{v_j}^{-1}$ under both the null and alternative hypotheses.

Define the "grey area set" as

$$\mathcal{G}(\boldsymbol{\theta}) = \{ j : |\theta_j|/v_j^{1/2} \asymp \delta_{N,T}, j = 1, ..., N \}.$$

Theorem 1.3.1. *Let Assumption 1.3.1 hold. As $T, N \to \infty$, we have, under $H_0 : \boldsymbol{\theta} = 0$, $P(\widehat{S} = \emptyset | H_0) \to 1$. Hence*

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

In addition,

$$\inf_{\boldsymbol{\theta}\in\Theta} P(S(\boldsymbol{\theta}) \subset \widehat{S}|\boldsymbol{\theta}) \to 1 \quad \textit{and} \quad \inf_{\boldsymbol{\theta}\in\Theta} P(\widehat{S} \setminus S(\boldsymbol{\theta}) \subset \mathcal{G}(\boldsymbol{\theta})|\boldsymbol{\theta}) \to 1.$$

Besides the asymptotic behavior of J_0, Theorem 1.3.1 also provides a "sure screening" property of $\widehat{S}$. Sometimes we wish to find out the identities of the elements in $S(\boldsymbol{\theta})$, which

represent the components of $\boldsymbol{\theta}$ that deviate from zero. Therefore, we are particularly interested in a type of alternative hypothesis that satisfies the following *empty grey area* condition.

Assumption 1.3.2 (Empty grey area)**.** *For any* $\boldsymbol{\theta} \in \Theta$, $\mathcal{G}(\boldsymbol{\theta}) = \emptyset$.

Theorem 1.3.1 shows that the "large" θ_j's can be selected with no missing discoveries and Corollary 1.3.1 below further asserts that the selection is consistent with no false discoveries either, under both the null and alternative hypotheses.

Corollary 1.3.1. *Under Assumptions 1.3.1, 1.3.2, as* $T, N \to \infty$,

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

Proof. Corollary 1.3.1 follows immediately from Theorem 1.3.1 and Assumption 1.3.2:

$$\inf_{\boldsymbol{\theta} \in \Theta} P(\widehat{S} \setminus S(\boldsymbol{\theta}) = \emptyset | \boldsymbol{\theta}) \geq \inf_{\boldsymbol{\theta} \in \Theta} P(\widehat{S} \setminus S(\boldsymbol{\theta}) \subset \mathcal{G}(\boldsymbol{\theta}) | \boldsymbol{\theta}) \to 1.$$

□

Remark 1.3.1. Corollary 1.3.1 and its required assumptions (Assumptions 1.3.1 and 1.3.2) are stated uniformly over $\boldsymbol{\theta} \in \Theta$. The empty grey area condition (Assumption 1.3.2) rules out $\boldsymbol{\theta}$'s that have components on the boundary of the screening set. Intuitively, when a component θ_j is on the boundary of the screening, it is hard to decide whether or not to eliminate it from the screening step. Note that the boundary of the screening depends on (N, T), which is similar in spirit to the local alternatives in classical testing problems, and is also a common practice for asymptotic analysis of high-dimensional tests (e.g., Cai et al. (2010); Chernozhukov et al. (2013)).

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

$$\Theta_s = \{\boldsymbol{\theta} \in \Theta : \max_{j \leq N} \frac{|\theta_j|}{v_j^{1/2}} > 2\delta_{N,T}\}. \tag{1.9}$$

In particular, if $\widehat{\theta}_j$ is $\sqrt{T}$-consistent, and $v_j^{1/2}$ is the asymptotic standard deviation of $\widehat{\theta}_j$, then $\sigma_j = \sqrt{Tv_j}$ is bounded away from both zero and infinity. Using (1.8), we have

$$\Theta_s = \{\boldsymbol{\theta} \in \Theta : \max_{j \leq N} |\theta_j|/\sigma_j > 2\log(\log T)\sqrt{\frac{\log N}{T}}\}.$$

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

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

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

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

Theorem 1.3.2. *Let Assumptions 1.3.1-1.3.2 hold. Suppose there is a test J_1 such that*

(i) it has an asymptotic non-degenerate null distribution F, and the critical region takes the form $\{\mathbf{D} : J_1 > F_q\}$ for the significance level $q \in (0,1)$,

(ii) it has high power uniformly on some set $\Theta(J_1) \subset \Theta$,

(iii) there is $c > 0$ so that $\inf_{\boldsymbol{\theta} \in \Theta_s} P(c\sqrt{N} + J_1 > F_q | \boldsymbol{\theta}) \to 1$, as $T, N \to \infty$.

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

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

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

Theorem 1.3.2 also shows that J_1 and J have the critical regions $\{\mathbf{D} : J_1 > F_q\}$ and $\{\mathbf{D} : J > F_q\}$ respectively, but the power is enhanced from $\Theta(J_1)$ to $\Theta_s \cup \Theta(J_1)$. In high-dimensional testing problems with a fast-growing dimension, $\Theta_s \cup \Theta(J_1)$ can be much larger than $\Theta(J_1)$. As a result, the power of J_1 can be substantially enhanced by adding J_0.

1.3.2 Power enhancement for quadratic tests

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

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

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

Assumption 1.3.3. *(i) There is a non-degenerate distribution F so that under H_0, $J_Q \to^d F$*
(ii) The critical value $F_q = O(1)$ and the critical region of J_Q is $\{\mathbf{D} : J_Q > F_q\}$,
(iii) $\mathbf{V}$ is positive definite, and there exist two positive constants C_1 and C_2 such that $C_1 \leq \lambda_{\min}(\mathbf{V}) \leq \lambda_{\max}(\mathbf{V}) \leq C_2$.
(iv) $C_3 \leq Tv_j \leq C_4, j = 1, ..., N$ for positive constants C_3 and C_4.

Analyzing the power properties of J_Q and applying Theorem 1.3.2, we obtain the following theorem. Recall that $\delta_{N,T}$ and Θ_s are defined by (1.8) and (1.9).

Theorem 1.3.3. *Under Assumptions 1.3.1-1.3.3, the power enhancement test $J = J_0 + J_Q$ satisfies: as $T, N \to \infty$,*

(i) under the null hypothesis $H_0 : \boldsymbol{\theta} = \mathbf{0}$, $J \to^d F$,

(ii) there is $C > 0$ so that J has high power uniformly on the set

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

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

1.3.3 Low power of quadratic statistics under sparse alternatives

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

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

$$\frac{T\widehat{\boldsymbol{\theta}}'\mathbf{V}\widehat{\boldsymbol{\theta}} - N}{\sqrt{2N}} | H_0 \to^d \mathcal{N}(0, 1). \tag{1.10}$$

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

Theorem 1.3.4. *Suppose that (1.10) holds with $\|\mathbf{V}\|_1 < C$ and $\|\mathbf{V}^{-1}\|_1 < C$ for some $C > 0$. Under Assumptions 1.3.1- 1.3.3, $T = o(\sqrt{N})$ and $\log N = o(T^{1-\gamma})$ for some $0 < \gamma < 1$, the*

quadratic test J_Q has low power at the sparse alternative Θ_b given by

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

In other words, $\forall \boldsymbol{\theta} \in \Theta_b$, for any significance level q,

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

where z_q is the qth quantile of standard normal distribution.

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

1.4 Application: Testing Factor Pricing Models

1.4.1 The model

The multi-factor pricing model, derived by Ross (1976) and Merton (1973), is one of the most fundamental results in finance. It postulates how financial returns are related to market risks, and has many important practical applications. Let y_{it} be the excess return of the i-th asset at time t and $\mathbf{f}_t = (f_{1t}, ..., f_{Kt})'$ be the observable excess returns of K market risk factors. Then, the excess return has the following decomposition:

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

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

The key implication from the multi-factor pricing theory for tradable factors is that under no-arbitrage restrictions, the intercept θ_i should be zero for any asset i (Ross, 1976; Merton, 1973; Chamberlain and Rothschild, 1983). An important question is then testing the null hypothesis

$$H_0 : \boldsymbol{\theta} = 0, \tag{1.11}$$

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

Most existing tests to the problem (1.11) are based on the quadratic statistic $W = T\widehat{\boldsymbol{\theta}}'\mathbf{V}\widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is the OLS estimator for $\boldsymbol{\theta}$, and $\mathbf{V}$ is some positive definite matrix. Prominent examples are given by Gibbons et al. (1989), MacKinlay and Richardson (1991) and Beaulieu et al. (2007). When N is possibly much larger than T, Pesaran and Yamagata (2012) showed that, under regularity conditions (Assumption 1.4.1 below),

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

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

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

1.4.2 Power enhancement component

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

Denote by $\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$ and $\mathbf{w} = (\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t\mathbf{f}_t')^{-1}\bar{\mathbf{f}}$. Also define

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

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

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, ..., \widehat{\theta}_N)', \quad \widehat{\theta}_j = \frac{1}{Ta_{f,T}}\sum_{t=1}^{T} y_{jt}(1 - \mathbf{f}_t'\mathbf{w}). \tag{1.12}$$

When $\text{cov}(\mathbf{f}_t)$ is positive definite, under mild regularity conditions (Assumption 1.4.1 below), $a_{f,T}$ consistently estimates a_f, and $a_f > 0$. In addition, without serial correlations, the

conditional variance of $\widehat{\theta}_j$ (given $\{\mathbf{f}_t\}$) converges in probability to

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

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

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

We show in Proposition 1.4.1 below that $\max_{j\leq N} |\widehat{\theta}_j - \theta_j|/\widehat{v}_j^{1/2} = O_P(\sqrt{\log N})$. Therefore, $\delta_{N,T} = \log(\log T)\sqrt{\log N}$ slightly dominates the maximum estimation noise. The screening set and the power enhancement component are defined as

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

and

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

1.4.3 Feasible Wald test in high dimensions

Assuming no serial correlations among $\{\mathbf{u}_t\}_{t=1}^T$ and conditional homoskedasticity (Assumption 1.4.1 below), given the observed factors, the conditional covariance of $\widehat{\boldsymbol{\theta}}$ is $\boldsymbol{\Sigma}_u/(Ta_{f,T})$. If the covariance matrix $\boldsymbol{\Sigma}_u$ of $\mathbf{u}_t$ were known, the standardized Wald test statistic is

$$\frac{Ta_{f,T}\widehat{\boldsymbol{\theta}}'\boldsymbol{\Sigma}_u^{-1}\widehat{\boldsymbol{\theta}} - N}{\sqrt{2N}}. \tag{1.13}$$

Under $H_0 : \boldsymbol{\theta} = 0$, it converges in distribution to $\mathcal{N}(0, 1)$. Note that the idiosyncratic errors $(u_{1t}, ..., u_{Nt})$ are often cross-sectionally correlated, which leads to a non-diagonal inverse covariance matrix $\boldsymbol{\Sigma}_u^{-1}$. When $N/T \to \infty$, it is practically difficult to estimate $\boldsymbol{\Sigma}_u^{-1}$, as there are $O(N^2)$ free off-diagonal parameters.

To consistently estimate $\mathbf{\Sigma}_u^{-1}$ when $N/T \to \infty$, without parametrizing the off-diagonal elements, we assume $\mathbf{\Sigma}_u = \text{cov}(\mathbf{u}_t)$ be a sparse matrix. This assumption is natural for large covariance estimations for factor models, and was previously considered by Fan et al. (2011). Since the common factors dictate preliminarily the co-movement across the whole panel, a particular asset's idiosyncratic shock is usually correlated significantly only with a few of other assets. For example, some shocks only exert influences on a particular industry, but are not pervasive for the whole economy (Connor and Korajczyk, 1993).

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

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

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

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

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

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

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

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

$$J_{wald} = \frac{Ta_{f,T}\widehat{\boldsymbol{\theta}}'\widehat{\mathbf{\Sigma}}_u^{-1}\widehat{\boldsymbol{\theta}} - N}{\sqrt{2N}}, \tag{1.14}$$

whose power can be enhanced under sparse alternatives by:

$$J = J_0 + J_{wald}. \tag{1.15}$$

1.4.4 Does the thresholded covariance estimator affect the size?

A natural but technical question to address is that when $\mathbf{\Sigma}_u$ indeed admits a sparse structure, is the thresholded estimator $\widehat{\mathbf{\Sigma}}_u^{-1}$ accurate enough so that the feasible J_{wald} is still asymptotically normal? The answer is affirmative if $N(\log N)^4 = o(T^2)$, and still we can allow $N/T \to \infty$. However, such a simple question is far more technically involved than anticipated, as we now explain.

When $\mathbf{\Sigma}_u$ is a sparse matrix, under regularity conditions (Assumption 1.4.2 below), Fan et al. (2011) showed that

$$\|\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1}\|_2 = O_P(\sqrt{\frac{\log N}{T}}). \tag{1.16}$$

By the lower bound derived by Cai et al. (2010), the convergence rate is minimax optimal for the sparse covariance estimation. On the other hand, when replacing $\mathbf{\Sigma}_u^{-1}$ in (1.13) by $\widehat{\mathbf{\Sigma}}_u^{-1}$, one needs to show that the effect of such a replacement is asymptotically negligible, namely, under H_0,

$$T\widehat{\boldsymbol{\theta}}'(\mathbf{\Sigma}_u^{-1} - \widehat{\mathbf{\Sigma}}_u^{-1})\widehat{\boldsymbol{\theta}}/\sqrt{N} = o_P(1). \tag{1.17}$$

However, when $\boldsymbol{\theta} = 0$, with careful analysis, $\|\widehat{\boldsymbol{\theta}}\|^2 = O_P(N/T)$. Using this and (1.16), by the Cauchy-Schwartz inequality, we have

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

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

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

1.4.5 Regularity conditions

We are now ready to present the regularity conditions. These conditions are imposed for three technical purposes: (i) Achieving the uniform convergence for $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ as required in Assumption 1.3.1, (ii) defining the sparsity of $\boldsymbol{\Sigma}_u$ so that $\widehat{\boldsymbol{\Sigma}}_u^{-1}$ is consistent, and (iii) showing (1.17), so that the errors from estimating $\boldsymbol{\Sigma}_u^{-1}$ do not affect the size of the test.

Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_T^\infty$ denote the σ-algebras generated by $\{\mathbf{f}_t : -\infty \leq t \leq 0\}$ and $\{\mathbf{f}_t : T \leq t \leq \infty\}$ 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 1.4.1. *(i)* $\{\mathbf{u}_t\}_{t \leq T}$ *is i.i.d.* $\mathcal{N}(0, \boldsymbol{\Sigma}_u)$*, where both* $\|\boldsymbol{\Sigma}_u\|_1$ *and* $\|\boldsymbol{\Sigma}_u^{-1}\|_1$ *are bounded;*

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

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

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

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

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

Some remarks are in order for the conditions in Assumption 1.4.1.

Remark 1.4.1. Condition (i), perhaps somewhat restrictive, substantially facilitates our technical analysis. Here $\mathbf{u}_t$ is required to be serially uncorrelated across t. Under this condition, the conditional covariance of $\widehat{\boldsymbol{\theta}}$, given the factors, has a simple expression $\boldsymbol{\Sigma}_u/(Ta_{f,T})$. On the other hand, if serial correlations are present in $\mathbf{u}_t$, there would be additional autocovariance terms in the covariance matrix, which need to be further estimated via regularizations. Moreover, given that $\boldsymbol{\Sigma}_u$ is a sparse matrix, the Gaussianity ensures that most of the idiosyncratic errors are cross-sectionally independent so that $\text{cov}(u_{it}^2, u_{jt}^l) = 0$, $l = 1, 2$, for most of the pairs in $\{(i, j) : i \neq j\}$.

Note that we do allow the factors $\{\mathbf{f}_t\}_{t\leq T}$ to be weakly correlated across t, but satisfy the strong mixing condition Assumption 1.4.1 (iii).

Remark 1.4.2. The conditional homoskedasticity $E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t) = E(\mathbf{u}_t\mathbf{u}_t')$ is assumed, granted by condition (ii). We admit that handling conditional heteroskedasticity, while important in empirical applications, is very technically challenging in our context. Allowing the high-dimensional covariance matrix $E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t)$ to be time-varying is possible with suitable *continuum of sparse* conditions on the time domain. In that case, one can require the sparsity condition to hold uniformly across t and continuously apply thresholding. However, unlike in the traditional case, technically, estimating the family of large inverse covariances $\{E(\mathbf{u}_t\mathbf{u}_t'|\mathbf{f}_t)^{-1} : t = 1, 2, ...\}$ uniformly over t is highly challenging. As we shall see in the proof of Proposition 1.4.2, even in the homoskedastic case, proving the effect of estimating $\boldsymbol{\Sigma}_u^{-1}$ to be first-order negligible when $N/T \to \infty$ requires delicate technical analysis.

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

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

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

Assumption 1.4.2. *Suppose* $N^{1/2}(\log N)^\gamma = o(T)$ *for some* $\gamma > 2$, *and*
(i) $\min_{(\mathbf{\Sigma}_u)_{ij}\neq 0} |(\mathbf{\Sigma}_u)_{ij}| \gg \sqrt{(\log N)/T}$;
(ii) at least one of the following cases holds:
(a) $D_N = O(N^{1/2})$, *and* $m_N^2 = O(\frac{T}{N^{1/2}(\log N)^\gamma})$
(b) $D_N = O(N)$, *and* $m_N^2 = O(1)$.

As regulated in Assumption 1.4.2, we consider two kinds of sparse matrices, and develop our results for both cases. In the first case (Assumption 1.4.2 (ii)(a)), $\mathbf{\Sigma}_u$ is required to have no more than $O(N^{1/2})$ off-diagonal nonzero entries, but allows a diverging m_N. One typical example of this case is that there are only a small portion (e.g., finitely many) of firms whose individual shocks (u_{it}) are correlated with many other firms'. In the second case (Assumption 1.4.2(ii)(b)), m_N should be bounded, but $\mathbf{\Sigma}_u$ can have $O(N)$ off-diagonal nonzero entries. This allows block-diagonal matrices with finite size of blocks or banded matrices with finite number of bands. This case typically arises when firms' individual shocks are correlated only within industries but not across industries.

Moreover, we require $N^{1/2}(\log N)^\gamma = o(T)$, which is the price to pay for estimating a large error covariance matrix. But still we allow $N/T \to \infty$. It is also required that the minimal signal for the nonzero components be larger than the noise level (Assumption 1.4.2 (i)), so that nonzero components are not thresholded off when estimating $\mathbf{\Sigma}_u$.

1.4.6 Asymptotic properties

The following result verifies the uniform convergence required in Assumption 1.3.1 over the entire parameter space that contains both the null and alternative hypotheses. Recall that the OLS estimator and its asymptotic standard error are defined in (1.12).

Proposition 1.4.1. *Suppose the distribution of* $(\mathbf{f}_t, \mathbf{u}_t)$ *is independent of* $\boldsymbol{\theta}$*. Under Assumption 1.4.1, for* $\delta_{N,T} = \log(\log T)\sqrt{\log N}$*, as* $T, N \to \infty$*,*

$$\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq N} |\widehat{\theta}_j - \theta_j|/\widehat{v}_j^{1/2} < \delta_{N,T}/2|\boldsymbol{\theta}) \to 1.$$
$$\inf_{\boldsymbol{\theta}\in\Theta} P(4/9 < \widehat{v}_j/v_j < 16/9, \forall j = 1, ..., N|\boldsymbol{\theta}) \to 1.$$

Proposition 1.4.2. *Under Assumptions 1.3.2, 1.4.1, 1.4.2, and* H_0*,*

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

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

We are now ready to apply Theorem 1.3.3 to obtain the asymptotic properties of $J = J_0 + J_{wald}$ as follows. For $\delta_{N,T} = \log(\log T)\sqrt{\log N}$, let

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

Theorem 1.4.1. *Suppose the assumptions of Propositions 1.4.1 and 1.4.2 hold.*

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

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

and hence

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

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

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

and hence

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

where z_q denotes the qth quantile of the standard normal distribution.

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

1.5 Application: Testing Cross-Sectional Independence

1.5.1 The model

Consider a mixed effect panel data model

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

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

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

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

hypothesis of interest is often a sparse vector $\boldsymbol{\theta}$, corresponding to a sparse covariance matrix $\boldsymbol{\Sigma}_u$ of u_{it}.

Most of the existing tests are based on the quadratic statistic $W = \sum_{i<j} T\widehat{\rho}_{ij}^2 = T\widehat{\boldsymbol{\theta}}'\widehat{\boldsymbol{\theta}}$, where $\widehat{\rho}_{ij}$ is the sample correlation between u_{it} and u_{jt}, estimated by the within-OLS (Baltagi, 2008), and $\widehat{\boldsymbol{\theta}} = (\widehat{\rho}_{12}, ..., \widehat{\rho}_{n-1,n})$. Pesaran et al. (2008) and Baltagi et al. (2012) studied the rescaled W, and showed that after a proper standardization, the rescaled W is asymptotically normal when both $n, T \to \infty$. However, the quadratic test suffers from a low power if $\boldsymbol{\Sigma}_u$ is a sparse matrix under the alternative. In particular, as is shown in Theorem 1.3.4, when $n/T \to \infty$, the quadratic test cannot detect the sparse alternatives with $\sum_{i<j} 1\{\rho_{ij} \neq 0\} = o(n/T)$, which is very restrictive. Such a sparse structure is present, for instance, when $\boldsymbol{\Sigma}_u$ is a block-diagonal sparse matrix with finitely many blocks and finite block sizes.

1.5.2 Power enhancement test

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

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

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

$$J_0 = \sqrt{N} \sum_{(i,j)\in\widehat{S}} \widehat{\rho}_{ij}^2 \widehat{v}_{ij}^{-1}, \qquad \widehat{S} = \{(i,j) : |\widehat{\rho}_{ij}|/\widehat{v}_{ij}^{1/2} > \delta_{N,T}, i < j \leq n\}. \tag{1.18}$$

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

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

$$J_1 = \sqrt{\frac{1}{n(n-1)}} \sum_{i<j} (T\widehat{\rho}_{ij}^2 - 1) - \frac{n}{2(T-1)}. \tag{1.19}$$

Under regularity conditions (Assumptions 1.5.1, 1.5.2 below), $J_1 \to^d \mathcal{N}(0,1)$ under H_0. Then the power enhancement test can be constructed as $J = J_0 + J_1$. The power is substantially enhanced to cover the region

$$\Theta_s = \{\boldsymbol{\theta} : \max_{i<j} \frac{\sqrt{T}|\rho_{ij}|}{1-\rho_{ij}^2} > 2\log(\log T)\sqrt{\log N}\}, \tag{1.20}$$

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

1.5.3 Asymptotic properties

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

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

We denote $E(u_{it}^r|\boldsymbol{\theta})$ as the rth moment of u_{it} when the correlation vector of the underlying data generating process is $\boldsymbol{\theta}$. The following regularity conditions are imposed.

Assumption 1.5.1. *There are $C_1, C_2 > 0$, so that*
(i) $\sup_{\boldsymbol{\theta}\in\Theta} \sum_{i\neq j\leq n} |E\widetilde{\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{jt} E(u_{it}u_{jt}|\boldsymbol{\theta})| < C_1 n$,
(ii) $\sup_{\boldsymbol{\theta}\in\Theta} \max_{j\leq n} E(u_{jt}^4|\boldsymbol{\theta}) < C_1$, $\inf_{\boldsymbol{\theta}\in\Theta} \min_{j\leq n} E(u_{jt}^2|\boldsymbol{\theta}) > C_2$,

Condition (i) is needed for the within-OLS to be $\sqrt{nT}$-consistent (see, e.g., Baltagi (2008)). It is usually satisfied by weak cross-sectional correlations (sparse alternatives)

among the error terms, or weak dependence among the regressors. We require the second moment of u_{jt} be bounded away from zero uniformly in $j \leq n$ and $\boldsymbol{\theta} \in \Theta$, so that the cross-sectional correlations can be estimated stably.

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

Assumption 1.5.2. *(i)* $\{\mathbf{u}_t\}_{t\leq T}$ *are i.i.d.* $N(0, \boldsymbol{\Sigma}_u)$, $E(\mathbf{u}_t|\{\mathbf{f}_t\}_{t\leq T}, \boldsymbol{\theta}) = 0$ *almost surely.* *(ii) With probability approaching one, all the eigenvalues of* $\frac{1}{T}\sum_{t=1}^{T} \widetilde{\mathbf{x}}_{jt}\widetilde{\mathbf{x}}'_{jt}$ *are bounded away from both zero and infinity uniformly in* $j \leq n$.

Proposition 1.5.1. *Under Assumptions 1.5.1 and 1.5.2, for* $\delta_{N,T} = \log(\log T)\sqrt{\log N}$, *and* $N = n(n-1)/2$, *as* $T, N \to \infty$,

$$\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{ij} |\widehat{\rho}_{ij} - \rho_{ij}|/\widehat{v}_{ij}^{1/2} < \delta_{N,T}/2|\boldsymbol{\theta}) \to 1$$
$$\inf_{\boldsymbol{\theta}\in\Theta} P(4/9 < \widehat{v}_{ij}/v_{ij} < 16/9, \forall i \neq j|\boldsymbol{\theta}) \to 1.$$

Define

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

For J_1 defined in (1.19), let

$$J = J_0 + J_1. \tag{1.21}$$

The main result is presented as follows.

Theorem 1.5.1. *Suppose Assumptions 1.3.2, 1.5.1, 1.5.2 hold. As* $T, N \to \infty$,

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

$$P(J_0 = 0|H_0) \to 0, \quad J_1 \to^d \mathcal{N}(0, 1),$$

and hence

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

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

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

and hence

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

Therefore the power is enhanced from $\Theta(J_1)$ to $\Theta_s \cup \Theta(J_1)$ uniformly over sparse alternatives. In particular, the required signal strength of Θ_s in (1.20) is mild: the maximum cross-sectional correlation is only required to exceed a magnitude of $\log(\log T)\sqrt{(\log N)/T}$.

Chapter 2

Sufficient Forecasting Using Factor Models

2.1 Introduction

Forecasting using a data-rich environment has been an important research topic in statistics, economics and finance. Typical examples include forecasts of a macroeconomic output using a large number of employment and production variables (Stock and Watson, 1989; Bernanke et al., 2005), and forecasts of the values of market prices and dividends using cross-sectional asset returns (Sharpe, 1964; Lintner, 1965). The predominant framework to harness vast predictive information is via the *factor model*, which proves effective in simultaneously modeling the commonality and cross-sectional dependence of the observed data. Turning the curse of dimensionality into blessing, factor models have been widely demonstrated in many applications, such as portfolio management (Fama and French, 1992; Carhart, 1997), large-scale multiple testing (Leek and Storey, 2008; Fan et al., 2012), high-dimensional covariance matrix estimation (Fan et al., 2008, 2013), and in particular, linear forecasting using many predictors (Stock and Watson, 2002*a*,*b*).

With little knowledge of the relationship between the forecast target and the latent factors, most research focuses on a linear model and its refinement. Motivated by the classic principal component regression (Kendall, 1957; Hotelling, 1957), Stock and Watson (2002*a*,*b*) employed a similar idea to forecast a single time series from a large number of predictors: first used the traditional principal component analysis (PCA) to estimate the underlying common factors, followed by a linear regression of the target on the estimated factors. The key insight here is to condense information from many cross-sectional predictors into several predictive indices. As an improvement to this procedure, Paul et al. (2006) used a correlation screening to remove the irrelevant predictors before performing PCA. In a similar fashion, Bai and Ng (2008) employed thresholding rules to select "targeted predictors", and Stock and Watson (2012) used shrinkage methods to downweight the unrelated principal components. Recently, Kelly and Pruitt (2014) took into account the covariance with the forecast target, and proposed a three-pass regression filter that generalizes partial least squares to forecast a single time series.

However, a linear principle components framework only reveals one dimension of the predictive power of the underlying factors. When the link function of the target and the factors is nonlinear, an thorough exploration of the factor space often leads to additional gains. In order to address this issue, we introduce an alternative method called *sufficient forecasting*. Our procedure springs from the idea of sufficient dimension reduction, which were first introduced as the sliced inverse regression in the seminal work of Li (1991). We are interested in constructing the sufficient predictive indices, given which the forecast target is independent of the common factors. Put it another way, the forecast target relates to the common factors only through these sufficient predictive indices. Such a goal is closely related to the estimation of the *central space* in dimension reduction literature (Cook, 2009). In a linear forecasting model, the central space consists of only one dimension. By contrast, when a nonlinear link function is present, the central space goes beyond one dimension and our proposed method can effectively estimate all the sufficient predictive indices. This

procedure therefore greatly enlarges the scope of factor forecasting. As demonstrated in our numerical studies, the sufficient forecasting has improved performance over benchmark methods, especially under a nonlinear forecasting equation.

To our knowledge, relatively few work has been done in forecasting a nonlinear time series using factor models, partially because of the linear tradition. Bai and Ng (2008) discussed the use of squared factors (i.e., volatility of the factors) in augmenting forecasting equation. Ludvigson and Ng (2007) found that the square of the first factor estimated from a set of financial factors is significant in the regression model for the mean excess returns. This, however, naturally leads to the question of which factor, or more precisely, which direction of factor space to include for higher moments. The sufficient predictive indices provide guidelines for these directions, leaving questions such as how to model nonlinearity for further investigation.

In summary, the contribution of this work is at least twofold. On one hand, our work advances existing forecasting methods, and fills the important gap between incorporating target information and dealing with nonlinear forecasting. We also provide a rigorous theoretical guarantee for the sufficient forecasting without requiring the i.i.d assumption. On the other hand, our work actually presents a promising dimension reduction technique through factor models. It is well-known that existing dimension reduction methods are limited to either a fixed dimension or a diverging dimension that is smaller than the sample size (Zhu et al., 2006). With the aid of factor models, our work alleviates what plagues sufficient dimension reduction in high-dimensional regimes, where the dimension might be much higher than the sample size.

The rest of this chapter is organized as follows. Section 2.2 presents the complete methodological details of the sufficient forecasting, and Section 2.3 establishes the asymptotic properties. We give a few applications of the sufficient forecasting in Section 2.4 and put a few remarks on future directions in Section 2.5. The numerical performance are demonstrated in Section 3.2, Chapter 3. Proofs are given in Section 5.2, Chapter 5.

2.2 Methodology

2.2.1 Factor models and forecasting

Consider the following factor model with a target variable y_t which we wish to forecast:

$$y_{t+1} = h(\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t, \epsilon_{t+1}), \tag{2.1}$$

$$x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it}, \quad 1 \leq i \leq p,\ 1 \leq t \leq T, \tag{2.2}$$

where x_{it} is the i-th predictor observed at time t, $\mathbf{b}_i$ is a vector of factor loadings, $\mathbf{f}_t$ is a $K \times 1$ vector of common factors driving the predictors and u_{it} is the error term, or the idiosyncratic component. The target variable y_{t+1} depends on the factors $\mathbf{f}_t = (f_{1t}, ...f_{Kt})'$ through $L(\leq K)$ linear combinations $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L$, which are orthogonal unit $K \times 1$ vectors. In (2.1), we assume that $h(\cdot)$ is an unknown link function and ϵ_{t+1} is some stochastic error independent of $\mathbf{f}_t$ and u_{it}. Note that the unknown function $h(\cdot)$ poses a significant challenge in forecasting y_{t+1}. As a special case, when the target is linearly related to the underlying factors, we simply have $L = 1$, and (2.1) reduces to

$$y_{t+1} = \boldsymbol{\phi}_1'\mathbf{f}_t + \epsilon_{t+1}.$$

Such linear forecasting problems using many predictors have been addressed extensively in the literature, for example, Stock and Watson (2002a), Stock and Watson (2002b), Bai and Ng (2008), Stock and Watson (2012) and Kelly and Pruitt (2014), among others.

In order to forecast y_{t+1}, we seek to find out certain projections of $\mathbf{f}_t$ that is target-relevant, i.e., $\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t$. We call these projections *sufficient predictive indices* throughout this work. They can be seen as a way of weighing common factors and reducing dimensions. Traditional factor analysis, such as PCA, will yield factors whose loadings account for the variation of predictors. However, those factors in PCA do not necessarily contain information about the forecasting target. In particular, if the target-relevant factors contribute only a

small fraction of the total variability in the predictors, principal component regression will involve many irrelevant factors. To make things worse, the possible non-linearity will only deteriorate the situation. This is seen through the following example.

Example 2.2.1. Suppose we have the following factor structure

$$y_{t+1} = f_{(K-1)t} + f_{(K-1)t} f_{Kt} + \epsilon_{t+1},$$
$$x_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}, \quad 1 \le i \le p,\ 1 \le t \le T,$$

where $\{\boldsymbol{\lambda}_k = (b_{1k}, ..., b_{pk})'\}_{k=1}^K$ are orthogonal unit vectors and $\{f_{kt}\}_{k=1}^K$ are uncorrelated (i.e. $\text{cov}(\mathbf{f}_t)$ is a diagonal matrix). We further assume that only the first factor f_{1t} is dominant, that is $\text{var}(f_{kt}) = o(\text{var}(f_{1t}))$ for $k \ge 2$. The covariance matrix of $\mathbf{x}_t = (x_{1t}, ..., x_{pt})'$ can be decomposed as

$$\text{cov}(\mathbf{x}_t) = \sum_{k=1}^{K} \text{var}(f_{kt}) \boldsymbol{\lambda}_k \boldsymbol{\lambda}_k' + \text{cov}(\mathbf{u}_t).$$

When $\text{cov}(\mathbf{u}_t)$ is very small, a direct application of PCA will deliver f_{1t} as the first principal component. However, since $\text{cov}(f_1, f_{(K-1)t} f_{Kt})$ is not necessarily zero, PCR will include f_{1t} in the regression function and possibly other f_{it}'s. (As an example, let X, Y, Z be independent standard normal variables and $f_{1t} = |X|\text{sign}(YZ), f_{(K-1)t} = Y, f_{Kt} = Z$. It's easy to verify that f_{it} are pairwise uncorrelated but $\text{cov}(f_1, f_{(K-1)t} f_{Kt}) > 0$.)

One way to tackle this issue is to use statistical methods to select relevant factors. Bai and Ng (2009) applied boosting method in the screen of factors. However, such an approach is more tailored to handle overfitting issues, and is often limited to linear forecasting, even if we augment the factor set.

Traditional analysis of factor models focuses on the covariance of these predictors, which we denote by a $p \times p$ matrix $\boldsymbol{\Sigma}_x$. Writing $\mathbf{x}_t = (x_{1t}, ..., x_{pt})', \mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_p)'$ and $\mathbf{u}_t = (u_{1t}, ..., u_{pt})'$, we have

$$\boldsymbol{\Sigma}_x = \mathbf{B}\text{cov}(\mathbf{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u, \tag{2.3}$$

where $\mathbf{\Sigma}_u$ is the covariance matrix of $\mathbf{u}_t$ or the error covariance matrix. However, the covariance within the predictors is often not enough to construct an optimal linear forecast since it does not incorporate the target information. Kelly and Pruitt (2014) resort to the covariance with the target to produce a better linear forecast. In the presence of a possibly nonlinear forecast target, the factor models (2.1)-(2.2) are more challenging than the linear forecast of Kelly and Pruitt (2014). We adopt a different perspective, by considering the covariance matrix of conditional expectation given the forecast target. This allows us to fully utilize the target information without knowing the nonlinear dependence, a feature we shall demonstrate below.

2.2.2 Sliced inverse regression

Suppose the factor model (2.2) has the following canonical normalization

$$\text{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.4}$$

which serves as an indentifiability condition because $\mathbf{B}\mathbf{f}_t = \mathbf{B}\mathbf{\Omega}\mathbf{\Omega}^{-1}\mathbf{f}_t$ holds for any nonsingular matrix $\mathbf{\Omega}$. Also assume for simplicity that x_{it}'s and $\mathbf{f}_t$'s in (2.2) have already been de-meaned. If the common factors $\mathbf{f}_t$ are *observed*, we can rely on the semi-parametric index model (2.1) itself to forecast y_{t+1}. The so-called sufficient dimension reduction (SDR) direction $\boldsymbol{\phi}_i$'s in the link function form the *central subspace* $S_{y|\mathbf{f}}$ (Cook, 2009), given which y_{t+1} is independent of $\mathbf{f}_t$. Li (1991) developed a sliced inverse regression (SIR) method to effectively estimate these SDR directions. Under model (2.1), Li (1991) showed that if $E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t)$ is a linear function of $\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t$ for any $\mathbf{b} \in \mathbb{R}^p$, $E(\mathbf{f}_t|y_{t+1})$ is contained in $S_{y|\mathbf{f}}$. Thus, SDR directions can be obtained by the eigenvectors corresponding to the L largest eigenvalues of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$, which we denote by $\mathbf{\Sigma}_{f|y}$.

Since the factors $\mathbf{f}_t$ are *unobserved* in practice, the SIR can not be directly pursued by looking at the conditional information of these underlying factors. A natural solution is to

use estimated factors to approximate $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$, which leads to

$$\boldsymbol{\Sigma}^1_{f|y} := \text{cov}(E(\widehat{\mathbf{f}}_t|y_{t+1})), \tag{2.5}$$

where $\widehat{\mathbf{f}}_t$ is some consistent estimator for $\mathbf{f}_t$.

Alternatively, we can start with the observed predictors x_{it}. By conditioning on the target y_{t+1}, we obtain

$$\text{cov}(E(\mathbf{x}_t|y_{t+1})) = \mathbf{B}\text{cov}(E(\mathbf{f}_t|y_{t+1}))\mathbf{B}'.$$

Here, the error covariance matrix $\boldsymbol{\Sigma}_u$ in (2.3) disappears since $\mathbf{u}_t$ and y_{t+1} are independent. This allows a flexible structure on $\boldsymbol{\Sigma}_u$, and includes the *approximate factor model* (Chamberlain and Rothschild, 1983) as a special case. As is well-known that principal component analysis is not scale-invariant, we cannot directly deal with $\text{cov}(E(\mathbf{x}_t|y_{t+1}))$. Letting $\boldsymbol{\Lambda}_b = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$ be a $K \times p$ matrix, the following linear transformation connects $\mathbf{x}_t$ to the factors $\mathbf{f}_t$,

$$\text{cov}(E(\boldsymbol{\Lambda}_b\mathbf{x}_t|y_{t+1})) = \text{cov}(E(\mathbf{f}_t|y_{t+1})).$$

Note that $\boldsymbol{\Lambda}_b$ also needs to be estimated, which only involves factor loadings. With a consistent estimator $\widehat{\boldsymbol{\Lambda}}_b$, we immediately obtain a second estimator for $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$,

$$\boldsymbol{\Sigma}^2_{f|y} := \widehat{\boldsymbol{\Lambda}}_b\text{cov}(E(\mathbf{x}_t|y_{t+1}))\widehat{\boldsymbol{\Lambda}}_b'. \tag{2.6}$$

Remark 2.2.1. The SIR (Li, 1991) takes into account the target information through covariance of the inverse regression curve $E(\mathbf{f}_t|y_{t+1})$. As pointed out in Chen and Li (1998), the largest eigenvalue of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$ corresponds to the largest R-squared value among all transformations of y_{t+1}, i.e.

$$\max_{\mathbf{b},T} \text{Corr}^2(T(y_{t+1}), \mathbf{b}'\mathbf{f}_t),$$

where the maximum is taken over any transformation $T(\cdot)$ and $\mathbf{b} \in \mathbb{R}^p$. This further justifies the need in considering $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$ and the corresponding SDR directions $\boldsymbol{\phi}_j$'s, especially in the presence of a nonlinear relationship between the target and the common factors.

Remark 2.2.2. Under the factor model (2.1)-(2.2), the distribution of the forecast target relates to the common factors only through the sufficient predictive indices. Thus, this is the problem of estimating the *central space* (Cook, 2009). Since the seminal work of Li (1991), various methods have been developed for identifying the central space, for example, the sliced average variance estimation (Cook and Weisberg, 1991), the directional regression (Li and Wang, 2007), and so on. This problem is also closely related to the problem of estimating the *central mean space* (Cook and Li, 2002) where the conditional mean $E(y_{t+1}|\mathbf{f}_t)$ relates to the common factors only through several predictive indices. Several other dimension reduction techniques are developed to recover the central mean space, such as the ordinary least squares (Li and Duan, 1989), the method of principal Hessian directions (Li, 1992), etc. One should generally distinguish the two different goals when applying corresponding techniques.

2.2.3 Sufficient forecasting

To make forecast, we first elucidate how factors and factor loadings are estimated. We temporarily assume that the number of underlying factors K is known to us. Consider the following constrained least squares problem

$$\arg\min_{\mathbf{B},\mathbf{F}} ||\mathbf{X} - \mathbf{B}\mathbf{F}'||_F^2, \tag{2.7}$$

$$\text{subject to} \quad T^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_K, \quad \mathbf{B}'\mathbf{B} \text{ is diagonal}, \tag{2.8}$$

where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ and $\mathbf{F}' = (\mathbf{f}_1, ..., \mathbf{f}_T)$. This is a classical principal components problem, and has been used by many researchers to extract underlying factors (Stock and Watson, 2002*a*; Fan et al., 2013). The constraints (2.8) correspond to the normalization (2.4). The

minimizers $\widehat{\mathbf{F}}_K, \widehat{\mathbf{B}}_K$ are such that the columns of $\widehat{\mathbf{F}}_K/\sqrt{T}$ are the eigenvectors corresponding to the K largest eigenvalues of the $T \times T$ matrix $\mathbf{X}'\mathbf{X}$ and $\widehat{\mathbf{B}}_K = T^{-1}\mathbf{X}\widehat{\mathbf{F}}_K$.

To fully estimate $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$, we follow the sliced inverse regression scheme in Li (1991), replacing the expectation and covariance by their sample counterparts. Denote the order statistics of $\{(y_{t+1}, \widehat{\mathbf{f}}_t)\}_{t=1}^{T-1}$ by $\{(y_{(t+1)}, \widehat{\mathbf{f}}_{(t)})\}_{t=1}^{T-1}$ according to the values of y, where $y_{(2)} \leq \cdots \leq y_{(T)}$ and we only use information up to time T. We divide the range of y into H slices, each of which contains an even number of observations $c > 0$. By introducing a double script (h, j) in which h refers to the slice number and j refers to the order number of an observation in the given slice, we write the data as

$$y_{(h,j)} = y_{(c(h-1)+j)+1}, \quad \widehat{\mathbf{f}}_{(h,j)} = \widehat{\mathbf{f}}_{(c(h-1)+j)}.$$

The estimate $\widehat{\mathbf{\Sigma}}^1_{f|y}$ of $\mathbf{\Sigma}_{f|y} = \text{cov}(E(\mathbf{f}_t|y_{t+1}))$ has the form

$$\widehat{\mathbf{\Sigma}}^1_{f|y} = \frac{1}{H}\sum_{h=1}^{H}[\frac{1}{c}\sum_{l=1}^{c}\widehat{\mathbf{f}}_{(h,l)}][\frac{1}{c}\sum_{l=1}^{c}\widehat{\mathbf{f}}_{(h,l)}]'. \tag{2.9}$$

Since H is typically fixed in practice, the fact that the last slice may have less than c observations exerts little influence on SIR asymptotically. Analogously, for $\mathbf{\Sigma}^2_{f|y}$, we have

$$\widehat{\mathbf{\Sigma}}^2_{f|y} = \widehat{\mathbf{\Lambda}}_b(\frac{1}{H}\sum_{h=1}^{H}[\frac{1}{c}\sum_{l=1}^{c}\mathbf{x}_{(h,l)}][\frac{1}{c}\sum_{l=1}^{c}\mathbf{x}_{(h,l)}]')\widehat{\mathbf{\Lambda}}_b'. \tag{2.10}$$

The following proposition shows that the estimates of $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$ based on either the estimated factors or the estimated factor loadings are equivalent.

Proposition 2.2.1. *Suppose we have predictors x_{it} that follow a factor structure (2.1), along with a target y_{t+1}. Let $\widehat{\mathbf{f}}_t$ and $\widehat{\mathbf{B}}$ be estimated from the method of principal components. $\widehat{\mathbf{\Lambda}}_b$ is obtained by substitution. Then, the two estimators (2.9) and (2.10) for* $\text{cov}(E(\mathbf{f}_t|y_{t+1}))$ *are equivalent, i.e.*

$$\widehat{\mathbf{\Sigma}}^1_{f|y} = \widehat{\mathbf{\Sigma}}^2_{f|y}.$$

Remark 2.2.3. There are alternative ways for estimating factors and loadings. For example, Forni et al. (2000) studied factor estimation based on projection. Connor et al. (2012) applied a weighted additive nonparametric estimation procedure to estimate characteristic-based factor models. These methods do not necessarily lead to the identity above.

We denote the two equivalent terms by $\widehat{\mathbf{\Sigma}}_{f|y}$. We shall show that under mild conditions, $\widehat{\mathbf{\Sigma}}_{f|y}$ consistently estimate $\mathbf{\Sigma}_{f|y}$ as $p, T \to \infty$. As a result, the eigenvectors of $\widehat{\mathbf{\Sigma}}_{f|y}$, denoted as $\widehat{\boldsymbol{\phi}}_j (j = 1, ..., K)$, converge to the corresponding eigenvectors of $\mathbf{\Sigma}_{f|y}$, which span the central space discussed before. This will yield consistent estimates of sufficient predictive indices $\boldsymbol{\phi}_i' \mathbf{f}_t$, and provides baselines for further investigation.

2.2.4 Determining the number of factors

In practice, the number of factors K might be unknown to us. There are many existing approaches to determining K in the literature, e.g., Bai and Ng (2002), Hallin and Liška (2007), Alessi et al. (2010). Recently, Lam et al. (2012) and Ahn and Horenstein (2013) proposed a ratio-based estimator by maximizing the ratio of two adjacent eigenvalues of $\mathbf{X}'\mathbf{X}$ arranged in descending order, i.e.

$$\hat{K} = \underset{1 \le i \le kmax}{\arg\max}\, \hat{\lambda}_i / \hat{\lambda}_{i+1},$$

where $\hat{\lambda}_1 \ge ... \ge \hat{\lambda}_T$ are the eigenvalues. The estimator enjoys good finite-sample performances and was motivated by the following observation: the K largest eigenvalues of $\mathbf{X}'\mathbf{X}$ grow unboundedly as p increases, while the others remain bounded.

We note here that once a consistent estimator of K is found, the asymptotic results in this paper hold true for the unknown K case by a conditioning argument. Unless otherwise specified, we shall assume a known K in the sequel.

2.3 Asymptotic properties

2.3.1 Assumptions

We first detail the modeling assumptions on model (2.1) and (2.2), in which $\{(\mathbf{x}_t, y_t)\}_{t=1}^T$ are observable.

Assumption 2.3.1 (Factors and Loadings)**.** *(1)* $||\mathbf{b}_i|| \leq M$ *for some* $M > 0$ *(*$i = 1, ..., p$*). And as* $p \to \infty$*, there exists positive constants* c_1 *and* c_2 *such that*

$$c_1 < \lambda_{\min}(\frac{1}{p}\mathbf{B}'\mathbf{B}) < \lambda_{\max}(\frac{1}{p}\mathbf{B}'\mathbf{B}) < c_2.$$

(2) $E||\mathbf{f}_t||^4 < \infty$*, and* $T^{-1}\mathbf{F}'\mathbf{F} \to \text{cov}(\mathbf{f}_t) = \mathbf{I}$ *as* $T \to \infty$*.*
(3) Linearity: $E(\mathbf{b}'\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t)$ *is a linear function of* $\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t$ *for any* $\mathbf{b} \in \mathbb{R}^p$*, where* $\boldsymbol{\phi}_i$*'s come from model (2.1).*

Condition (1) is often known as the *pervasive* condition (Bai and Ng, 2002; Fan et al., 2013) in that the factors impact a non-vanishing portion of the predictors. Condition (2) is also standard for factor models. Condition (3) ensures that the (centered) inverse regression curve $E(\mathbf{f}_t|y_{t+1})$ is contained in the central space, and is satisfied when the distribution of $\mathbf{f}_t$ is elliptically symmetric (Hall and Li, 1993). If the distribution of $\mathbf{f}_t$ is non-elliptically distributed, we can follow Li and Dong (2009) to greatly relax the linearity condition in Assumption 2.3.1, and assume that $E(\mathbf{f}_t|\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t)$ is a polynomial function of $\boldsymbol{\phi}_1'\mathbf{f}_t, ..., \boldsymbol{\phi}_L'\mathbf{f}_t$, where $\boldsymbol{\phi}_i$'s come from model (2.1).

We impose the strong mixing condition on the data generating process. Let $\mathcal{F}_\infty^0$ and $\mathcal{F}_T^\infty$ denote the $\sigma-$algebras generated by $\{(\mathbf{f}_t, y_{t+1}) : t \leq 0\}$ and $\{(\mathbf{f}_t, y_{t+1}) : t \geq T\}$ respectively. Define the mixing coefficient

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

Assumption 2.3.2 (Data generating process)**.** $\{\mathbf{f}_t, \mathbf{u}_t, \epsilon_{t+1}\}_{t\geq 1}$ *is strictly stationary,* $E(||\mathbf{f}_t||^2|y_{t+1}) < \infty$ *and for* $T \in \mathbb{Z}^+$ *and some* $\rho \in (0,1)$, $\alpha(T) < c\rho^T$.

The assumption above ensures that the sample average of $\mathbf{f}_t|y_{t+1}$ for any range of y_{t+1} is root T consistent. In addition, we impose the following assumption on the residuals and dependence of the factor model. Conditions (1)-(4) in Assumption 2.3.3 are similar assumptions as those in Bai (2003), which are needed to consistently estimate the common factors as well as the factor loadings.

Assumption 2.3.3 (Residuals and Dependence)**.** *For some* $M > 0$,

(1) $E(\mathbf{u}_t) = \mathbf{0}$, *and* $E|u_{it}|^8 \leq M$.

(2) $||\mathbf{\Sigma}_u||_1 \leq M$, *and for every* $i, j, t, s > 0$, $(pT)^{-1}\sum_{i,j,t,s} \text{cov}(u_{it}, u_{js}) \leq M$

(3) For every (t, s), $E|p^{-1/2}\sum_{i=1}^{p}(\mathbf{u}_s'\mathbf{u}_t - E(\mathbf{u}_s\mathbf{u}_t)|^4 \leq M$.

(4) Weak dependence between factors and idiosyncratic errors

$$E(\frac{1}{p}\sum_{i=1}^{p}||\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{f}_t u_{it}||^2) \leq M$$

2.3.2 Convergence of $\widehat{\mathbf{\Sigma}}_{f|y}$

As with many other applications, the factor structure (2.2) often involves estimation of the unknown factors. Note that the factor loadings $\mathbf{B}$ and the common factors $\mathbf{f}_t$ are not separably identifiable. Let $\mathbf{V}$ denote the $K \times K$ diagonal matrix of the first K largest eigenvalues of the sample covariance matrix $T^{-1}\mathbf{X}'\mathbf{X}$ in descending order. Define a $K \times K$ matrix $\mathbf{H} = (1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$, where $\mathbf{F}' = (\mathbf{f}_1, ..., \mathbf{f}_T)$. Since $\mathbf{H}\mathbf{f}_t = (1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'(\mathbf{B}\mathbf{F}')'\mathbf{B}\mathbf{f}_t$ depends only on an identifiable part $(\mathbf{B}\mathbf{F}')'\mathbf{B}\mathbf{f}_t$ and the data $\mathbf{V}^{-1}\widehat{\mathbf{F}}'$, $\mathbf{H}$ eliminates identifiability issues when estimating $(\mathbf{B}, \mathbf{f}_t)$ simultaneously.

The following theorem gives the rate of convergence of the estimated covariance of inverse regression curve $\widehat{\mathbf{\Sigma}}_{f|y}$.

Theorem 2.3.1. *Suppose that assumptions 2.3.1-2.3.3 hold and let* $\omega_{p,T} = p^{-1/2} + T^{-1/2}$. *Then under model (2.1) and (2.2), we have*

$$||\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{H}\mathbf{\Sigma}_{f|y}\mathbf{H}'|| = O_p(\omega_{p,T}). \tag{2.11}$$

If the eigenvalues of $\mathbf{\Sigma}_{f|y} = \text{cov}(E(\mathbf{f}_t|y_{t+1}))$ *are positive and distinct, then the eigenvectors,* $\widehat{\boldsymbol{\phi}}_j (j = 1, .., L)$ *associated with* L *largest eigenvalues of* $\widehat{\mathbf{\Sigma}}_{x|y}$ *give consistent estimate of SDR directions up to rotation* $\mathbf{H}$*, i.e.*

$$||\widehat{\boldsymbol{\phi}}_j - \mathbf{H}\boldsymbol{\phi}_j|| = O_p(\omega_{p,T}) \tag{2.12}$$

for any $j \leq L$.

The proof of this theorem relies on the fact that $\widehat{\mathbf{\Lambda}}_b$ can be consistently estimated, and it is straightforward given existing details in the literature. We render it in Section 5.2, Chapter 5 . As a consequence of theorem 2.3.1, we have $\widehat{\boldsymbol{\phi}}_j'\widehat{\mathbf{f}}_t \to^p \boldsymbol{\phi}_j\mathbf{f}_t$ for any j. The sufficient predictive indices can therefore be consistently estimated. Define $\widehat{\boldsymbol{\xi}}_j := \widehat{\mathbf{\Lambda}}_b'\widehat{\boldsymbol{\phi}}_j$, then similarly $\widehat{\boldsymbol{\xi}}_j'\mathbf{x}_t \to^p \boldsymbol{\phi}_j\mathbf{f}_t$. This supplies linear combinations of observed data $\mathbf{x}_t$ with powerful forecast performance. Traditional sliced inverse regression can not handle the case when the number of predictors p is larger than the number of observations T. By condensing the cross-sectional information, a factor structure effectively reduces the dimension of predictors and extends the applicability of inverse regression.

2.3.3 Connection to linear estimators

What are the consequences of forcing a linear forecast? Forecasting problems within the framework of factor model typically focuses on a linear target, which shares the same factor representation as the predictors. When the underlying relationship between the target and the driving factors is nonlinear, directly applying linear forecasts would violate the link func-

tion $h(\cdot)$. Despite the validity issues of such forecast, linear estimates are easy to construct and usually provide benchmarks for our analysis. We shall see that linear forecast actually averages the sufficient predictive indices.

With a large number of predictors, the underlying factors are first estimated via PCA. The target is then regressed on the extracted factors to form the predictive model. Note that the normalization (2.8) serves as an orthogonal design for the estimated factors. One may then employ the following linear estimate of the target's loadings on $\mathbf{f}_t$

$$\widehat{\boldsymbol{\phi}} = \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}\widehat{\mathbf{f}}_t, \tag{2.13}$$

where $\widehat{\mathbf{f}}_t$'s are estimated via the optimization (2.7) and (2.8). To examine the behavior of this projection direction, we shall assume normality of the underlying factors. The following proposition shows that, regardless of the specification of the link function $h(\cdot)$, $\widehat{\boldsymbol{\phi}}$ falls into the central space spanned by $\phi_1, ..., \phi_L$ as $p, T \to \infty$.

Proposition 2.3.1. *Consider model (2.1) and (2.2) under assumptions of theorem 2.3.1. Suppose* $\{\mathbf{f}_t, \epsilon_{t+1}\}_{t\geq 1}$ *is i.i.d., the factors* $\mathbf{f}_t$ *are normally distributed and that* $E(y_t^2) < \infty$. *Then,*

$$||\widehat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}|| = O_p(\omega_{p,T}), \tag{2.14}$$

where $\bar{\boldsymbol{\phi}} = \sum_{i=1}^{L} E((\boldsymbol{\phi}_i'\mathbf{f}_t)y_{t+1})\boldsymbol{\phi}_i$.

It is interesting to see that when $L = 1$, the coefficient $\widehat{\boldsymbol{\phi}}$ delivers asymptotically efficient estimate of the projection direction of factors. The nonlinearity does not significantly decrease one's ability to estimate such direction. When $L \geq 2$, this is no longer the case. The estimated coefficient belongs to the linear subspace spanned by $\boldsymbol{\phi}_i$'s, and its coordinates depend on the correlation between the target and the sufficient predictive indices. This subspace, however, is entirely contained in $\boldsymbol{\Sigma}_{f|y}$. By estimating $\boldsymbol{\Sigma}_{f|y}$ directly, sliced inverse

regression tries to recover all the effective directions and would therefore capture most of the driving forces.

2.4 Applications

We give two examples to which the preceding results can be readily applied. Although detailed pursuits are beyond our scope, we demonstrate the corresponding numerical results in the next Chapter.

Example 2.4.1. (Linear forecast)

When we have *a priori* knowledge that the link function $h(\cdot)$ in (2.1) is in fact linear, only a single index needs to be estimated, i.e. $L = 1$. Prominent examples include asset return predictability, where we use the cross section of book-to-market ratios to forecast aggregate market returns (Campbell and Shiller, 1988; Polk et al., 2006; Kelly and Pruitt, 2013). The Arbitrage Pricing Theory (APT) by Ross (1976) states that the excessive return of a financial asset can be explained by a linear combination of risk factors, which justifies linear forecast. In such cases, the target admits the following linear factor structure

$$y_{t+1} = \boldsymbol{\phi}_1' \mathbf{f}_t + \epsilon_{t+1}, \quad t \leq T.$$

By Theorem 2.3.1 and Proposition 2.3.1, the eigenvector corresponding to the largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}_{f|y}$ provides estimation of target factor loadings equivalent to linear regression (2.13), up to a scale factor. However, the motivations are different. The linear regression is predicated on the assumption that the same set of factors drive both the target and the cross section of predictors. By contrast, sliced inverse regression finds projections of factors most relevant to the target. This incorporates the case when the target is a linear function of a strict subset of the latent factors.

Example 2.4.2. (Interaction effect)

Consider model (2.1) with an interaction effect

$$y_{t+1} = (\boldsymbol{\phi}_1' \mathbf{f}_t)(\boldsymbol{\phi}_2' \mathbf{f}_t) + \epsilon_{t+1},$$

where the interaction terms are formed by the two directions $\boldsymbol{\phi}_{1,2}$, which we are interested in determining. Note that since the underlying factors are extracted with the normalization (2.9), direct interaction terms such as $f_i f_j$ may not make much sense. Interaction models have been considered by many researchers in both economics and statistics. For example, in an influential paper, Rajan and Zingales (1998) examined the interaction between financial dependence and economic growth. Recently, Jiang and Liu (2014) studied variable selection with interaction detection via inverse modeling.

Including all the interaction terms $f_i f_j$ would require $K(K-1)/2$ parameters, and can deteriorate prediction significantly. We can successfully solve this problem by applying theorem 2.3.1. The eigenvectors corresponding to the largest two eigenvalues are consistent estimators of $\boldsymbol{\phi}_{1,2}$. Regression models can be subsequently built to account for such interaction effects.

2.5 Future work

We identify two avenues for future research. One is on the selection of the number of sufficient predictive indices L. There are some existing methods to tackle this problem, for example, Li (1991) and Schott (1994), whose approaches tend to be based on probabilistic assumptions on the underlying factors. An alternative way is a cross-validation approach which penalizes on the complexity of the forecasting model. Although heuristic methods such as eigenvalue ratio test (as used in picking the number of underlying factors) can be used in practice, a consistent estimate L is no doubt helpful.

A more fundamental direction is to remove the linearity condition (Li, 1991) or the polynomial condition (Li and Dong, 2009) as in Assumption 2.3.1. Such conditions are for

technical convenience and often difficult to check in practice. Recent advances in dimension-reduction literature have solved this problem at the cost of performing additional nonparametric regression. This could enrich the applicability of sufficient forecasting.

Chapter 3

Numerical studies

We first present Monte Carlo experiments for the power enhancement test proposed in Chapter 1, which is also applied to the components of S&P 500 as an empirical study. We next examine numerical performance on sufficient forecasting with the use of factor models in Section 3.2

3.1 Numerical studies for power enhancement test

In this section, Monte Carlo simulations are employed to examine the finite sample performance of the power enhancement tests. We respectively study the factor pricing model and the cross-sectional independence test. The proposed test is then applied to S&P 500 components to examine the market efficiency between 1985-2012.

3.1.1 Testing factor pricing models

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

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

We simulate $\{\mathbf{b}_i\}_{i=1}^N$, $\{\mathbf{f}_t\}_{t=1}^T$ 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 parameters are set to be the same as those in the simulations of Fan et al. (2013), which are calibrated using daily returns of S&P 500's top 100 constituents, for the period from July 1^{st}, 2008 to June 29^{th} 2012. These parameters are listed in the following table.

Table 3.1: Means and covariances used to generate $\mathbf{b}_i$ and $\mathbf{f}_t$

$\boldsymbol{\mu}_B$	$\boldsymbol{\Sigma}_B$			$\boldsymbol{\mu}_f$	$\boldsymbol{\Sigma}_f$		
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

Set $\boldsymbol{\Sigma}_u = \text{diag}\{\mathbf{A}_1, ..., \mathbf{A}_{N/4}\}$ to be a block-diagonal covariance matrix. Each diagonal block $\mathbf{A}_j$ is a 4×4 positive definite matrix, whose correlation matrix has equi-off-diagonal entry ρ_j, generated from Uniform$[0, 0.5]$. The diagonal entries of $\mathbf{A}_j$ are obtained via $(\boldsymbol{\Sigma}_u)_{ii} = 1 + \|\mathbf{v}_i\|^2$, where $\mathbf{v}_i$ is generated independently from $\mathcal{N}_3(0, 0.01\mathbf{I}_3)$.

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

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

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

Under H_a^1, there are only a few nonzero θ's with a relative large magnitude. Under H_a^2, there are many non-vanishing θ's, but their magnitudes are all relatively small. In our simulation setup, $\sqrt{\log N/T}$ 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 θ's should survive from the screening step. These survived $\hat{\theta}$'s contribute importantly to the rejection of the null hypothesis. In contrast, $P(\widehat{S} = \emptyset)$ should be much larger under H_a^2 because the non-vanishing θ's are too weak to be detected.

For each test, we calculate the relative frequency of rejection under H_0, H_a^1 and H_a^2 based on 2000 replications, with significance level $q = 0.05$. We also calculate the relative frequency of $\widehat{S}$ being empty, which approximates $P(\widehat{S} = \emptyset)$. We use the soft-thresholding to estimate the error covariance matrix.

Table 3.2: Size and power (%) of tests for simulated Fama-French three-factor model

		H_0			H_a^1			H_a^2		
T	N	J_{wald}	PE	$P(\widehat{S} = \emptyset)$	J_{wald}	PE	$P(\widehat{S} = \emptyset)$	J_{wald}	PE	$P(\widehat{S} = \emptyset)$
300	500	5.2	5.4	99.8	48.0	97.6	2.6	69.0	76.4	64.6
	800	4.9	5.1	99.8	60.0	99.0	1.2	69.2	76.2	62.2
	1000	4.6	4.7	99.8	54.6	98.4	2.6	75.8	82.6	63.2
	1200	5.0	5.4	99.6	64.2	99.2	0.8	74.2	81.0	63.6
500	500	5.2	5.3	99.8	33.8	99.2	0.8	73.4	77.2	77.8
	800	4.8	5.0	99.8	67.4	100.0	0.0	72.4	76.4	75.0
	1000	5.0	5.2	99.8	65.0	100.0	0.2	76.8	80.4	74.0
	1200	5.2	5.2	100.0	58.0	100.0	0.2	74.2	78.4	77.0

Notes: This table reports the frequencies of rejection and $\widehat{S} = \emptyset$ based on 2000 replications. Here J_{wald} is the standardized Wald test, and PE the power enhanced test. These tests are conducted at 5% significance level.

Table 3.2 presents the empirical size and power of the feasible standardized Wald test J_{wald} as well as those of the power enhanced test $J = J_0 + J_{wald}$. First of all, the size of J_{wald} is close to the significance level. Under H_0, $P(\widehat{S} = \emptyset)$ is close to one, implying that the power enhancement component J_0 screens off most of the estimation errors. The power enhanced test (PE) has approximately the same size as the original test J_{Wald}. Under H_a^1, the PE test significantly improves the power of the standardized Wald-test. In this case, $P(\widehat{S} = \emptyset)$ is nearly zero because the screening set manages to capture the big thetas. Under H_a^2, as the non-vanishing thetas are very week, it follows that $\widehat{S}$ has a large probability of

being empty. But, whenever $\hat{S}$ is non-empty, it contributes to the power of the test. The PE test still slightly improves the power of the quadratic test.

3.1.2 Testing cross-sectional independence

We use the following data generating process in our experiments,

$$y_{it} = \alpha + \beta x_{it} + \mu_i + u_{it}, \quad i \leq n, t \leq T, \tag{3.1}$$

$$x_{it} = \xi x_{i,t-1} + \mu_i + \varepsilon_{it}. \tag{3.2}$$

Note that we model $\{x_i\}$'s as AR(1) processes, so that x_{it} is possibly correlated with μ_i, but not with u_{it}, as was the case in Im et al. (1999). For each i, initialize $x_{it} = 0.5$ at $t = 1$. We specify the parameters as follows: μ_i is drawn from $\mathcal{N}(0, 0.25)$ for $i = 1, ..., n$. The parameters α and β are set -1 and 2 respectively. In regression (3.2), $\xi = 0.7$ and $\varepsilon_{it} \sim \mathcal{N}(0, 1)$.

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

$$\sigma_i^2 = \sigma^2(1 + \kappa \bar{x}_i)^2 \tag{3.3}$$

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

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

cross-sectional heteroscedasticity, we set $\mathbf{\Sigma}_u = \mathbf{\Sigma}_{u,0}^{1/2}\mathbf{\Sigma}_{u,1}\mathbf{\Sigma}_{u,0}^{1/2}$, where $\mathbf{\Sigma}_{u,0} = \text{diag}\{\sigma_1^2, ..., \sigma_n^2\}$ as specified in (3.3).

The Monte Carlo experiments are conducted for different pairs of (n, T) with significance level $q = 0.05$ based on 2000 replications. The empirical size, power and the frequency of $\widehat{S} = \emptyset$ as in (1.18) are recorded.

Table 3.3: Size and power (%) of tests for cross-sectional independence

H_0	T	$n = 200$	$n = 400$	$n = 600$	$n = 800$
		J_1/PE /$P(\widehat{S} = \emptyset)$	J_1/PE /$P(\widehat{S} = \emptyset)$	J_1/PE /$P(\widehat{S} = \emptyset)$	J_1/PE /$P(\widehat{S} = \emptyset)$
	100	4.7/5.5 /99.1	4.9/5.3 /99.6	5.5/5.7 /99.7	4.9/5.2 /99.7
	200	5.3/5.3 /100.0	5.5/5.9 /99.6	4.7/5.1 /99.4	4.9/5.1 /99.8
	300	5.2/5.2 /100.0	5.2/5.2 /100.0	4.6/4.6 /100.0	4.9/4.9 /100.0
	500	4.7/4.7 /100.0	5.5/5.5 /100.0	5.0/5.0 /100.0	5.1/5.1 /100.0
H_a	T	$n = 200$	$n = 400$	$n = 600$	$n = 800$
	100	26.4/95.5 /5.0	19.8/98.0 /2.3	13.5/98.2 /2.0	12.2/99.2 /0.9
	200	54.6/98.8 /1.6	40.3/99.6 /0.5	24.8/99.6 /0.4	21/99.7 /0.3
	300	78.9/99.25 /1.1	65.3/100.0 /0.1	41.7/99.9 /0.2	37.2/100.0 /0.1
	500	93.5/99.85 /0.2	89.0/100.0 /0.0	69.1/100.0 /0.0	61.8/100.0 /0.0

Notes: This table reports the frequencies of rejection by J_1 in (1.19) and PE in (1.21) under the null and alternative hypotheses, based on 2000 replications. The frequency of $\widehat{S}$ being empty is also recorded. These tests are conducted at 5% significance level.

Table 3.3 gives the size and power of the bias-corrected quadratic test J_1 in (1.19) and those of the power enhanced test J in (1.21). The sizes of both tests are close to 5%. In particular, the power enhancement test has little distortion of the original size.

The bottom panel shows the power of the two tests under the alternative specification. The PE test demonstrates almost full power under all combinations of (n, T). In contrast, the quadratic test J_1 as in (1.19) only gains power when T gets large. As n increases, the

proportion of nonzero off-diagonal elements in $\mathbf{\Sigma}_u$ gradually decreases. It becomes harder for J_1 to effectively detect those deviations from the null hypothesis. This explains the low power exhibited by the quadratic test when facing a high sparsity level.

3.1.3 Empirical Study

As an empirical application, we consider a test of Carhart (1997)'s four-factor model on the S&P 500 index. Our empirical findings show that there are only a few significant nonzero "alpha" components, corresponding to a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market.

We collect monthly excess returns on all the S&P 500 constituents from the CRSP database for the period January 1980 to December 2012. We test whether $\boldsymbol{\theta} = 0$ (all alpha's are zero) in the factor-pricing model on a rolling window basis: for each month, we evaluate our test statistics J_{wald} and J (as in (1.14) and (1.15) respectively) 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 balanced panel with the cross-sectional dimension larger than the time-series dimension ($N > T$). In this manner we not only capture the up-to-date information in the market, but also mitigate the impact of time-varying factor loadings and sampling biases. In particular, for testing months $\tau = 1984.12, ..., 2012.12$, we run the regressions

$$r_{it}^{\tau} - r_{ft}^{\tau} = \theta_i^{\tau} + \beta_{i,\mathrm{MKT}}^{\tau}(\mathrm{MKT}_t^{\tau} - r_{ft}^{\tau}) + \beta_{i,\mathrm{SMB}}^{\tau}\mathrm{SMB}_t^{\tau} + \beta_{i,\mathrm{HML}}^{\tau}\mathrm{HML}_t^{\tau} + \beta_{i,\mathrm{MOM}}^{\tau}\mathrm{MOM}_t^{\tau} + u_{it}^{\tau}, \quad (3.4)$$

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 $\mathrm{MKT}, \mathrm{SMB}$, HML and MOM constitute market, size, value and momentum factors. The time series of factors are downloaded from Kenneth French's website. To make the notation consistent, we use θ_i^{τ} to represent the "alpha" of stock i.

Table 3.4: Summary of descriptive statistics and testing results

Variables	Mean	Std dev.	Median	Min	Max
N_τ	617.70	26.31	621	574	665
$\lvert\widehat{S}\rvert_0$	5.20	3.50	5	0	20
$\overline{\lvert\widehat{\theta}\rvert_i}^\tau(\%)$	0.9767	0.1519	0.9308	0.7835	1.3816
$\overline{\lvert\widehat{\theta}\rvert_{i\in\widehat{S}}}^\tau(\%)$	4.5569	1.4305	4.1549	1.7839	10.8393
p-value of J_{wald}	0.2351	0.2907	0.0853	0	0.9992
p-value of J (PE)	0.1148	0.2164	0.0050	0	0.9982

Table 3.4 summarizes descriptive statistics for different components and estimates in the model. On average, 618 stocks (which is more than 500 because we are recording stocks that have *ever* become the constituents of the index) enter the panel of the regression during each five-year estimation window. Of those, merely 5.2 stocks are selected by the screening set $\widehat{S}$, which directly implies the presence of sparse alternatives. The threshold $\delta_{N,T} = \sqrt{(\log N)}\log(\log T)$ varies as the panel size N changes at the end of each month, and is about 3.5 on average, a high-criticism thresholding. The selected stocks have much larger alphas (θ) than other stocks do. In addition, 64.05% of all the estimated alphas are positive, whereas 87.33% of the selected alphas in $\widehat{S}$ are positive. This indicates that the power enhancement component in our test is primarily contributed by stocks with extra returns. We also notice that the p-values of the Wald test J_{wald} are generally smaller than those of the power enhanced test J.

Similar to Pesaran and Yamagata (2012), we plot the running p-values of J_{wald} and the PE test 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 3.1. There is a strong negative correlation between the stock selection percentage and the p-values of these tests. In other words, the months at which the null hypothesis is rejected typically correspond to a few stocks with alphas exceeding the threshold. Such evidence of sparse alternatives has originally motivated our study. We also observe that the p-values of the PE test lie beneath those of J_{wald} test as a result of enhanced power, and hence it captures several important market disruptions ignored by the latter (e.g. collapse of Japanese bubble in 1990). Indeed,

Figure 3.1: Dynamics of p-values and percents of selected stocks

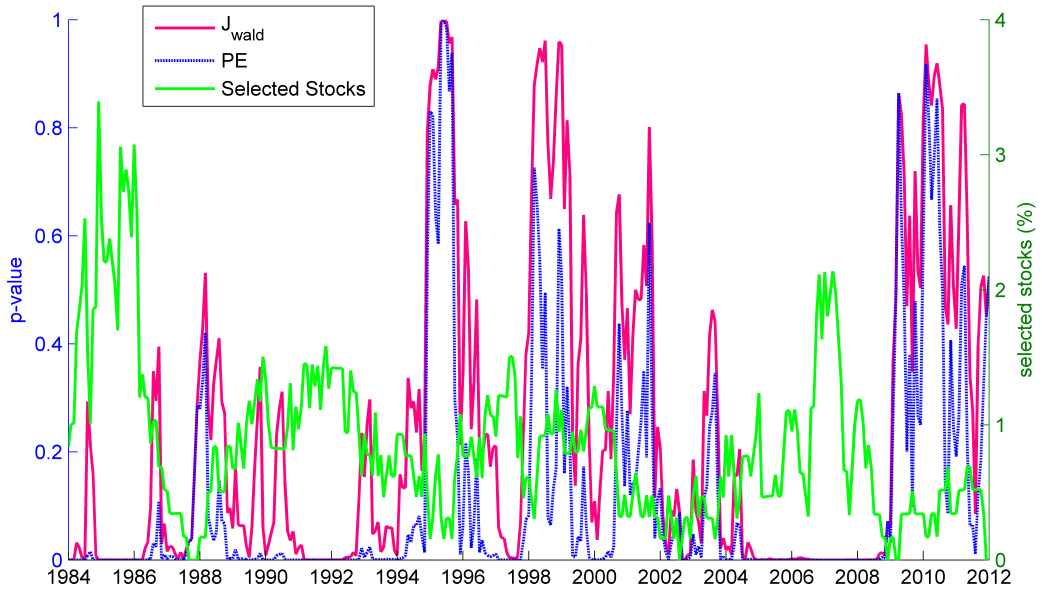
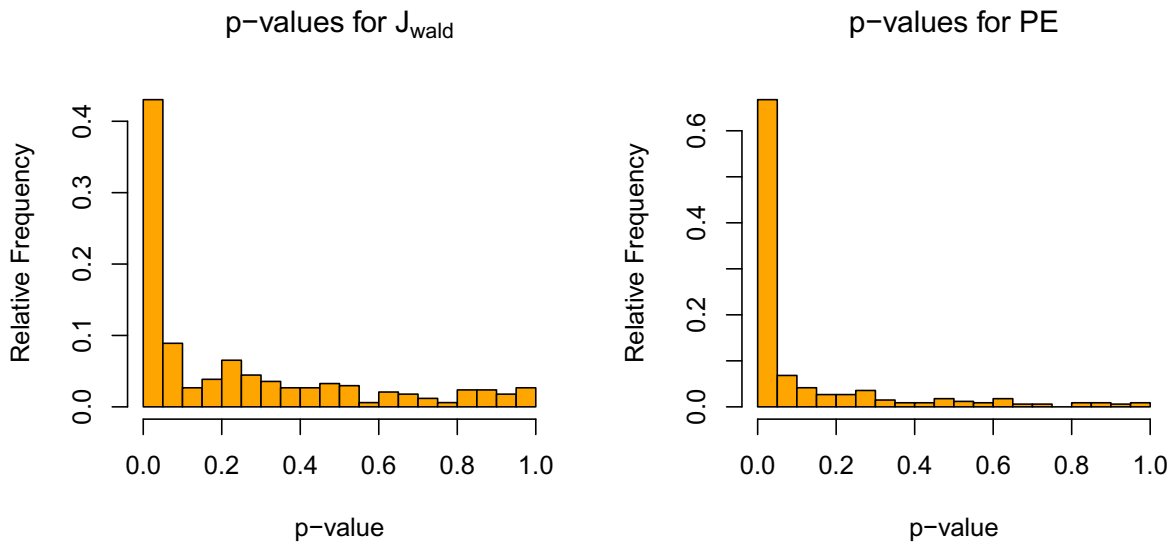


Figure 3.2: Histograms of p-values for J_{wald} and PE.



the null hypothesis of $\boldsymbol{\theta} = 0$ is rejected by the PE test at 5% level for almost all months during financial crisis, including major financial crisis such as Black Wednesday in 1992, Asian financial crisis in 1997, the financial crisis in 2008, which is also partially detected by J_{wald} tests. The histograms of the p-values of the two test statistics are displayed in Figure 3.2. By inspection, we see that of 43.03% and 66.07% of the study months, J_{wald} and the PE test reject the null hypothesis respectively. Again, the test results indicate the existence of sparse alternatives when faced with high crosse-sectional dimension.

3.2 Numerical studies for sufficient forecasting

In this section, we conduct Monte Carlo experiments to evaluate the numerical performance of sufficient forecast using factor models. The empirical results on forecasting macroeconomic variables are presented subsequently, which provides substantial evidence for the predictive power of sufficient forecasting.

3.2.1 Linear forecast

We first consider the case when the target is a linear function of a subset of the latent factors plus some noise. To this end, we specify our data generating process as

$$y_{t+1} = \boldsymbol{\phi}'\mathbf{f}_t + \sigma_y \epsilon_{t+1},$$
$$x_{it} = \mathbf{b}_i'\mathbf{f}_t + u_{it},$$

where we let $K = 5$ and $\boldsymbol{\phi} = (0.8, 0.5, 0.3, 0, 0)'$. Factor loadings are drawn from standard normal distribution. To account for serial correlation, we set f_{jt} and u_{it} as AR(1) processes

$$f_{jt} = \alpha_j f_{jt-1} + e_{jt}, \quad u_{it} = \rho_i u_{it-1} + \nu_{it}.$$

We draw α_j, ρ_i from $\sim U[0.2, 0.8]$ and fix them during simulations, while the shocks e_{jt}, ν_{it} and ϵ_{t+1} are standard normal respectively. σ_y is adjusted to equal the variance of the factors, so that the infeasible best forecast when knowing $\boldsymbol{\phi}'\mathbf{f}_t$ has an R^2 of 50%.

Table 3.5 reports in-sample and out-of-sample comparisons between principal component regression and sufficient forecasting. The out-of-sample R^2 is defined as root-mean-squared forecast error (RMSE) relative to the variance of y, and is computed from recursive out-of-sample forecast begun at the middle of the time series. In all cases, the single effective factor yields comparable results as PCR, which employs all the factors and therefore slightly outperforms the former. In contrast, using first principal component alone has very poor performance in general, as it may not be relevant to the forecasting target.

Table 3.5: Simulated Forecast Performance (Linear Model)

		In-sample			Out-of-sample		
p	T	SIR	PCR	PC1	SIR	PCR	PC1
50	100	46.9	47.7	7.8	35.1	39.5	2.4
50	200	46.3	46.5	6.6	42.3	41.7	4.4
100	100	49.3	50.1	8.9	37.6	40.3	3.0
100	500	47.8	47.8	5.5	43.6	43.5	1.1
500	100	48.5	48.8	7.9	40.0	43.1	4.7
500	500	48.2	48.3	7.2	48.0	47.9	6.0

Notes: In-sample and out-of-sample median R^2, recorded in percentage, over 2000 simulations. SIR denotes the sufficient forecast using sliced inverse regression, PCR denotes principal component regression, and PC1 uses only the first principal component.

3.2.2 Factor Interaction

We next turn to the case when the interaction between factors is present. Consider the model

$$y_{t+1} = f_{1t}(f_{2t} + f_{3t} + 1) + \epsilon_{t+1},$$

where ϵ_{t+1} is standard normal. The data generating process for the predictors x_{it} is set to be the same as that in the previous section, but we let $K = 10$. To measure the distance between the estimated directions $\widehat{\boldsymbol{\phi}}_{1,2}$ and the central subspace $S_{f|y}$ spanned by $\boldsymbol{\phi}_1 = (1, 0, ..., 0)$ and $\boldsymbol{\phi}_2 = (0, 1, 1, ..., 0)$, we first rotate $\widehat{\boldsymbol{\phi}}_i$ by left multiplying $\mathbf{H}'$ as in theorem 2.3.1 to obtain consistent estimators of $\boldsymbol{\phi}_{1,2}$. Following Li (1991), we use the squared multiple correlation coefficient $R^2(\widehat{\boldsymbol{\phi}}_j)$ as an affine invariant criterion for each j, where

$$R^2(\widehat{\boldsymbol{\phi}}_j) = \max_{\boldsymbol{\phi} \in S_{f|y}} \frac{[(\mathbf{H}'\widehat{\boldsymbol{\phi}}_j)'\boldsymbol{\Sigma}_f \boldsymbol{\phi}]^2}{(\mathbf{H}'\widehat{\boldsymbol{\phi}}_j)'\boldsymbol{\Sigma}_f(\mathbf{H}'\widehat{\boldsymbol{\phi}}_j) \cdot (\boldsymbol{\phi}'\boldsymbol{\Sigma}_f\boldsymbol{\phi})}.$$

Note that we have the convenience $\boldsymbol{\Sigma}_f = \mathbf{I}$ corresponding to the normalization (2.4).

The simulation results are summarized in Table 3.2 based on 1000 replicates. We observe that the time-series dimension T has a major effect in $R^2(\widehat{\boldsymbol{\phi}}_j)$. When T gets larger from 100 to 500, so are the $R^2(\widehat{\boldsymbol{\phi}}_j)$'s for both $j = 1, 2$. A large cross-sectional dimension p helps ensure the convergence of estimated factors, but only has slight influence on $R^2(\widehat{\boldsymbol{\phi}}_j)$. As theory shows, sliced inverse regression successfully picks up the effective dimension in the simulation.

A practical question is how to use the two effective factors to make forecast. We adopt a simple approach, by including $\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_1$, $\widehat{\boldsymbol{\phi}}_2'\widehat{\mathbf{f}}_2$ and $(\widehat{\boldsymbol{\phi}}_1'\widehat{\mathbf{f}}_1) \cdot (\widehat{\boldsymbol{\phi}}_2'\widehat{\mathbf{f}}_2)$ in the regression of y_{t+1}, which takes into account constant terms in the factor interaction. For comparison purposes, we report results from linear forecasts (PCR). In addition, we add the interaction between the first two principal components to PCR. As can be seen from Table 3.6, the in-sample R^2's of the linear forecast hover around 35%, and its out-of-sample R^2's are relatively low. Including

interaction between the first two PCs does not help much. SIR picks up the correct form of interaction and exhibit better performance, especially when T gets reasonably large.

Table 3.6: Simulated Forecast Performance (Factor Interaction)

		Squared multiple correlation		In-sample			Out-of-sample		
p	T	$R^2(\widehat{\phi}_1)$	$R^2(\widehat{\phi}_2)$	SIR	PCR	PCRi	SIR	PCR	PCRi
100	100	66.2 (19.1)	46.9 (25.0)	46.2	38.5	42.4	20.8	12.7	13.5
100	200	80.4 (14.4)	68.4 (21.1)	57.7	35.1	38.6	41.6	24.0	24.7
100	500	91.0 (9.4)	87.6 (10.6)	77.0	31.9	34.9	69.7	29.1	31.5
200	100	68.2 (18.2)	45.0 (24.0)	48.2	39.0	44.1	26.1	17.9	19.1
500	200	80.3 (14.1)	69.4 (20.7)	58.9	34.7	39.0	40.2	22.2	24.0
500	500	91.5 (9.0)	88.4 (10.4)	79.8	32.5	35.6	72.3	26.9	28.2

Notes: Squared multiple correlation coefficients, in-sample and out-of-sample median R^2 recorded in percentage over 1000 replications. The values in parentheses are the standard deviations. SIR uses first two predictive indices and includes their interaction effect; PCR uses all principal components; PCRi extends PCR by including an extra interaction term built on the first two principal components.

3.2.3 An Empirical Example

As an empirical investigation, we apply factor models and inverse regression to forecast several macroeconomic variables. Our dataset is taken from Stock and Watson (2012), which consists of quarterly observations on 108 U.S. low-level disaggregated macroeconomic time series from 1959:I through 2008:IV. Similar datasets have been employed to forecast other time series in the literature (Bai and Ng, 2008; Ludvigson and Ng, 2009). We study out-of-sample performance of each time series with all the others forming the predictor set. The procedure involves fully recursive factor estimation and parameter estimation starting half-way of the sample, using data only through quarter t for forecasting in quarter $t+1$.

Table 3.7: Out-of-sample Macroeconomic Forecasting

Category	Label	SIR	SIR(2)	PCR-ER	PC1
GDP components	GDP261	7.4	8.6	2.6	-2.5
IP	IPS13	21.2	31.8	24.4	13.3
Employment	CES033	38.8	27.5	38.6	40.4
Unemployment rate	LHU15	28.7	33.0	30.1	23.7
Housing	HSNE	33.6	28.8	31.8	30.5
Inventories	PMDEL	27.4	20.1	16.8	17.1
Prices	GDP276_3	8.3	10.9	8.2	4.9
Wages	CES278.R	22.0	15.4	20.3	19.6
Interest rates	FYFF	8.0	13.9	10.4	10.6
Money	CCINRV	3.1	11.8	0.5	-0.1
Exchange rates	EXRCAN	1.4	3.5	1.3	-1.8
Stock prices	FSPCOM	19.0	15.5	16.0	15.0
Consumer expectations	HHSNTN	6.1	8.2	8.2	7.6

Notes: Out-of-sample R^2 for one quarter ahead forecasts. SIR uses single predictive index built on 8 estimated factors to forecast, SIR(2) is fit by local linear regression using the first two predictive indices, PCR-ER uses as many principal components as determined by eigenvalue ratio test and PC1 uses the first principal component.

Table 3.7 presents the forecasting results for a few representatives in each macroeconomic category. The time series are chosen such that the second eigenvalue of $\widehat{\mathbf{\Sigma}}_{f|y}$ exceeds 60% of the first eigenvalue in the training sample, so we could consider the effect of second predictive index. In terms of linear forecast, sliced inverse regression yields comparable performance as PCR. There are cases where SIR exhibits more predictability than PCR, e.g., GDP components, Inventories and Wages. This is due to the fact that the first predictive index obtained from our procedure gives a parsimonious representation of linear predictors, and it is therefore less prone to over-fit. SIR(2) is fit by local linear regression using an

additional predictive index, which improves predictability in a few cases. Taking CCINRV (consumer credit outstanding) for example, Figure 3.2.3 plots the eigenvalues of its corresponding $\widehat{\mathbf{\Sigma}}_{f|y}$, the estimated regression surface and the running out-of-sample R^2's. As can been seen from the plot, there is a non-linear effect of the two underlying macro factors on the target. By taking such effect into account, SIR(2) consistently outperforms the other methods.

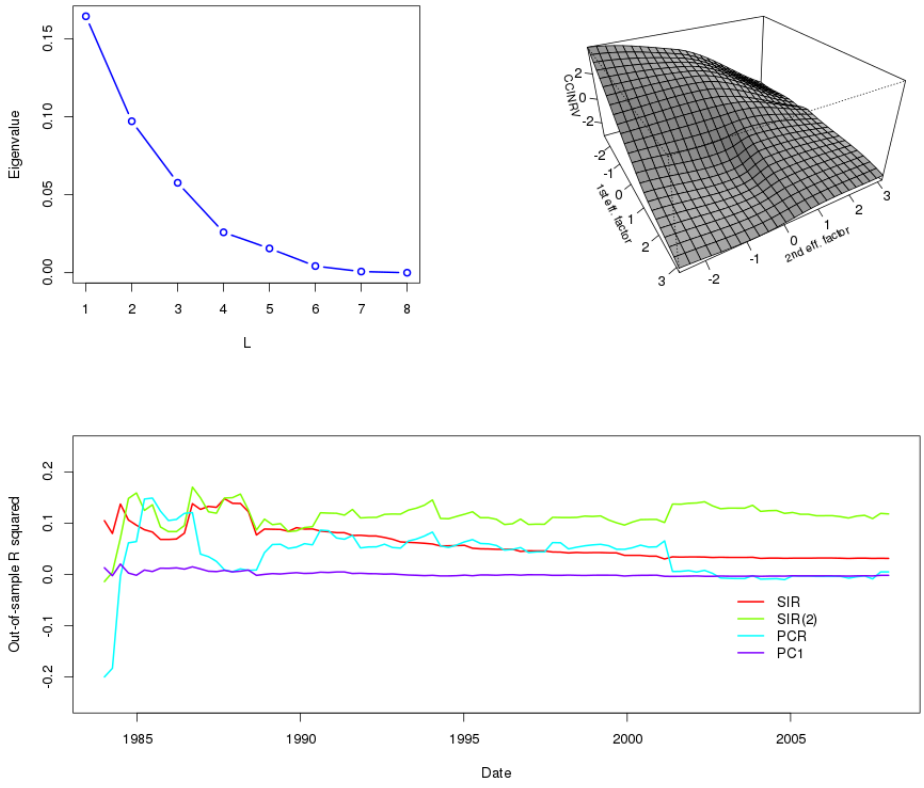


Figure 3.3: Forecasting results for CCINRV (consumer credit outstanding). The top left panel shows the eigenvalues of $\widehat{\mathbf{\Sigma}}_{f|y}$. The top right panel gives a 3-d plot of the estimated regression surface. The lower panel displays the running out-of-sample R^2's for the four methods described in Table 3.7.

Chapter 4

Concluding remarks

In this dissertation, we first consider testing a high-dimensional vector $H : \boldsymbol{\theta} = 0$ against sparse alternatives where the null hypothesis is violated only by a few components. Existing tests based on quadratic forms such as the Wald statistic often suffer from low powers due to the accumulation of errors in estimating high-dimensional parameters. We introduce a "power enhancement component" based on a screening technique, which is zero under the null, but diverges quickly under sparse alternatives. The proposed test statistic combines the power enhancement component with a classical statistic that is often asymptotically pivotal, and strengthens the power under sparse alternatives. On the other hand, the null distribution does not require stringent regularity conditions, and is completely determined by that of the pivotal statistic. As a byproduct, the screening statistic also consistently identifies the elements that violate the null hypothesis. As specific applications, the proposed methods are applied to testing the mean-variance efficiency in factor pricing models and testing the cross-sectional independence in panel data models. Our empirical study on the S&P500 index shows that there are only a few significant nonzero components, corresponding to a small portion of mis-priced stocks instead of systematic mis-pricing of the whole market. This provides empirical evidence of sparse alternatives.

We then address how to employ factor models for nonlinear forecasting. We introduce sufficient forecasting in a many-predictor environment to predict a single time series. By connecting factor models and inverse regression, the proposed method enlarges the scope of traditional factor forecasting. The key feature of the sufficient forecasting is its ability in extracting multiple predictive indices when the target is a nonlinear function of underlying factors. We have demonstrated its efficacy through Monte Carlo experiments. Our empirical results on macroeconomic forecasting also suggest that such procedure can contribute to substantial improvement beyond conventional linear models.

Chapter 5

Technical Proofs

5.1 Proofs for Chapter 1

We detail the proofs for the theories of power enhancement test. Throughout the proofs, let C denote a generic constant, which may differ at difference places.

5.1.1 Proofs for Section 1.3

Proof of Theorem 1.3.1

Proof. Define events

$$A_1 = \left\{ \max_{j \leq N} |\widehat{\theta}_j - \theta_j| / \widehat{v}_j^{1/2} < \delta_{N,T} \right\}, \quad A_2 = \left\{ \frac{4}{9} < \widehat{v}_j / v_j < \frac{16}{9}, \forall j = 1, ..., N \right\}.$$

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

$$\frac{|\widehat{\theta}_j|}{\widehat{v}_j^{1/2}} \geq \frac{|\theta_j| - |\widehat{\theta}_j - \theta_j|}{\widehat{v}_j^{1/2}} \geq \frac{3|\theta_j|}{4 v_j^{1/2}} - \frac{\delta_{N,T}}{2} > \delta_{N,T}.$$

This implies that $j \in \widehat{S}$, hence $S(\boldsymbol{\theta}) \subset \widehat{S}$. If $j \in \widehat{S}$, by similar arguments, we have $\frac{|\theta_j|}{v_j^{1/2}} > \delta_{N,T}/3$ on $A_1 \cap A_2$. Hence $\widehat{S} \setminus S(\boldsymbol{\theta}) \subset \{j : \delta_{N,T}/3 < \frac{|\theta_j|}{v_j^{1/2}} < 2\delta_{N,T}\} \subset \mathcal{G}(\boldsymbol{\theta})$. In fact, we have

proved that $S(\boldsymbol{\theta}) \subset \widehat{S}$ and $\widehat{S} \setminus S(\boldsymbol{\theta}) \subset \mathcal{G}(\boldsymbol{\theta})$ on the event $A_1 \cap A_2$ uniformly for $\boldsymbol{\theta} \in \Theta$. This yields

$$\inf_{\boldsymbol{\theta}\in\Theta} P(S(\boldsymbol{\theta}) \subset \widehat{S}|\boldsymbol{\theta}) \to 1, \quad \text{and} \quad \inf_{\boldsymbol{\theta}\in\Theta} P(\widehat{S} \setminus S(\boldsymbol{\theta}) \subset \mathcal{G}(\boldsymbol{\theta})) \to 1.$$

Moreover, it is readily seen that, under $H_0 : \boldsymbol{\theta} = 0$, by Assumption 1.3.1,

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

In addition, $\inf_{\boldsymbol{\theta}\in\Theta} P(J_0 > \sqrt{N}|S(\boldsymbol{\theta}) \neq \emptyset)$ is bounded from below by

$$\inf_{\boldsymbol{\theta}\in\Theta} P(\sqrt{N}\sum_{j\in\widehat{S}} \delta_{N,T}^2 > \sqrt{N}|S(\boldsymbol{\theta}) \neq \emptyset) \geq \inf_{\boldsymbol{\theta}\in\Theta} P(\sqrt{N}\delta_{N,T}^2 > \sqrt{N}|S(\boldsymbol{\theta}) \neq \emptyset) - o(1) \to 1.$$

Note that the last convergence holds uniformly in $\boldsymbol{\theta} \in \Theta$ because $\delta_{N,T} \to \infty$. This completes the proof.

Proof of Theorem 1.3.2

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

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

This together with the fact

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

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

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

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

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

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

This completes the proof.

Proof of Theorem 1.3.3

Proof. It suffices to verify conditions (i)-(iii) in Theorem 1.3.2 for $J_1 = J_Q$. Condition (i) follows from Assumption 1.3.3. Condition (iii) is fulfilled for $c > 2/\xi$, since

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

by using $F_q = O(1)$, $\xi_{N,T} \to \xi$, and $\mu_{N,T} \to 0$. We now verify condition (ii) for the $\Theta(J_Q)$ defined in the theorem. Let $\mathbf{D} = \text{diag}(v_1, ..., v_N)$. Then $\|\mathbf{D}\|_2 < C_3/T$ by Assumption 1.3.3(iv). On the event $A = \{\|(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{D}^{-1/2}\|^2 < \delta_{N,T}^2 N/4\}$, we have

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

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

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

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

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

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

Proof of Theorem 1.3.4

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

$$\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2') = (\mathbf{0}', \boldsymbol{\theta}_2'), \qquad \widehat{\boldsymbol{\theta}}' = (\widehat{\boldsymbol{\theta}}_1', \widehat{\boldsymbol{\theta}}_2'),$$

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

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

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

By the matrix inversion formula,

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

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

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

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

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

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

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

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

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

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

Note that the uniform convergence in Assumption 1.3.1 and boundness of $\|\boldsymbol{\theta}\|_{\max}$ imply that $P(\|\widehat{\boldsymbol{\theta}}\|_{\max} \le C) \to 1$ for a sufficient large constant C. For $\mathbf{G} = (g_{ij})$, note that $\max_{i\le r}\sum_{j=1}^{N-r}|g_{ij}| \le \|\mathbf{V}\|_1$. Hence, by using $\boldsymbol{\theta}_1 = 0$ again, with probability approaching one,

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

$$\leq \quad Cr_N\delta_{N,T}\sqrt{T}.$$

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

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

We denote $\text{var}(\widehat{\boldsymbol{\theta}}), \text{var}(\widehat{\boldsymbol{\theta}}_1), \text{var}(\widehat{\boldsymbol{\theta}}_2)$ to be the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$. Then $\frac{1}{T}\mathbf{V}^{-1} = \text{var}(\widehat{\boldsymbol{\theta}})$ and $\frac{1}{T}\mathbf{M}_1 = \text{var}(\widehat{\boldsymbol{\theta}}_1)$. It then follows from (1.10) that

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

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

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

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

□

5.1.2 Proofs for Section 1.4

Lemma 5.1.1. *When* $\text{cov}(\mathbf{f}_t)$ *is positive definite,* $E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t < 1$.

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

Proof of Proposition 1.4.1

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

Simple calculations yield

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

We first prove the second statement. Note that there is $\sigma_{\min} > 0$ (independent of $\boldsymbol{\theta}$) so that $\min_j \sigma_j > \sigma_{\min}$. By Lemma 5.1.11, there is $C > 0$, $\inf_\Theta P(\max_{j\le N}|\widehat{\sigma}_j - \sigma_j| < C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta}) \to 1$. On the event $\{\max_{j\le N}|\widehat{\sigma}_j - \sigma_j| < C\sqrt{\frac{\log N}{T}}\}$,

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

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

Note that v_j is independent of $\boldsymbol{\theta}$, so there is C_1 (independent of $\boldsymbol{\theta}$) so that $\max_{j\le N} v_j^{-1/2} < C_1\sqrt{T}$. On the event $\{\max_{j\le N} v_j^{1/2}/\widehat{v}_j^{1/2} < 2\} \cap \{\max_{j\le N}|\widehat{\theta}_j - \theta_j| < C\sqrt{\frac{\log N}{T}}\}$,

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

The constants C, C_1 appeared are independent of $\boldsymbol{\theta}$, and Lemma 5.1.11 holds uniformly in $\boldsymbol{\theta}$. Hence the desired result also holds uniformly in $\boldsymbol{\theta}$.

Proof of Proposition 1.4.2

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

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

Therefore, we only need to show

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

The left hand side is equal to

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

It was shown by Fan et al. (2011) that $\|\widehat{\boldsymbol{\Sigma}}_u - \boldsymbol{\Sigma}_u\|_2 = O_P(m_N\sqrt{\frac{\log N}{T}}) = \|\widehat{\boldsymbol{\Sigma}}_u^{-1} - \boldsymbol{\Sigma}_u^{-1}\|_2$. In addition, under H_0, $\|\widehat{\boldsymbol{\theta}}\|^2 = O_P(N\log N/T)$. Hence $b = O_P(\frac{m_N^2\sqrt{N}(\log N)^2}{T}) = o_P(1)$.

The challenging part is to prove $a = o_P(1)$ when $N > T$. As is described in the main text, simple inequalities like Cauchy-Schwarz accumulate estimation errors, and hence do not work. Define $\mathbf{e}_t = \boldsymbol{\Sigma}_u^{-1}\mathbf{u}_t = (e_{1t}, ..., e_{Nt})'$, 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 this proposition is to establish the following two convergences:

$$\frac{1}{T}E|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1 - \mathbf{f}_s'\bar{\mathbf{w}}))^2|^2 = o(1), \tag{5.1}$$

$$\frac{1}{T}E|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1-\mathbf{f}_s'\bar{\mathbf{w}})][\frac{1}{\sqrt{T}}\sum_{k=1}^{T}e_{jk}(1-\mathbf{f}_k'\bar{\mathbf{w}})]|^2 = o(1), \tag{5.2}$$

where

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

The sparsity condition assumes that most of the off-diagonal entries of $\mathbf{\Sigma}_u$ 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 (5.1) and (5.2) are given in the supplementary material in Appendix D.

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

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

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

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

We first examine a_3. Note that

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

Obviously,

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

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, $P(a_3 > T^{-1}) < \epsilon$, this implies $a_3 = o_P(1)$.

The proof is finished once we establish $a_i = o_P(1)$ for $i = 1, 2$, which are given in Lemmas 5.1.6 and 5.1.7 respectively in the supplementary material.

Proof of Theorem 1.4.1 Part (i) follows from Proposition 1.4.2 and that $P(J_0 = 0|H_0) \to 1$. Part (ii) follows immediately from Theorem 1.3.3.

5.1.3 Proofs for Section 1.5

Proof of Proposition 1.5.1

Lemma 5.1.2. *Under Assumption 1.5.1,* $\inf_{\boldsymbol{\theta}\in\Theta} P(\sqrt{nT}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| < \sqrt{\log n}|\boldsymbol{\theta}) \to 1$.

Proof. Note that

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

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

$$\begin{aligned} E\|\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{it}\widetilde{u}_{it}\|^2 &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{n}\sum_{s=1}^{T} E\widetilde{\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{js}\widetilde{u}_{it}\widetilde{u}_{js} \\ &= \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} E\widetilde{\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{it}E\widetilde{u}_{it}\widetilde{u}_{it} + \frac{1}{nT}\sum_{i\neq j}\sum_{t=1}^{T} E\widetilde{\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{jt}E\widetilde{u}_{it}\widetilde{u}_{jt} \\ &\le C_1 + \frac{1}{n}\sum_{i\neq j}|E\widetilde{\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{jt}||E\widetilde{u}_{it}\widetilde{u}_{jt}| \le C. \end{aligned}$$

Hence the result follows from the Chebyshev inequality and that $\lambda_{\min}(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{\mathbf{x}}_{it}\widetilde{\mathbf{x}}_{it}')$ is bounded away from zero with probability approaching one, uniformly in $\boldsymbol{\theta}$.

Lemma 5.1.3. *Suppose* $\max_{j\le n}\|\frac{1}{T}\sum_t \widetilde{\mathbf{x}}_{jt}\widetilde{\mathbf{x}}_{jt}'\|_2 < C'$ *with probability approaching one and* $\sup_\theta E(u_{jt}^4|\boldsymbol{\theta}) < C'$. *There is* $C > 0$, *so that*

(i) $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| < C\sqrt{\log n/T}|\boldsymbol{\theta}) \to 1$

(ii) $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{i,j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| < C\sqrt{\log n/T}|\boldsymbol{\theta}) \to 1$

(iii) $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq n} \frac{1}{T}\sum_{t=1}^{T} (u_{jt} - \widehat{u}_{jt})^2 < C\log n/T|\boldsymbol{\theta}) \to 1$

(iv) $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{i,j\leq n} |\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt}| < C\sqrt{\log n/T}|\boldsymbol{\theta}) \to 1$

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

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| \geq C\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\Theta} n \max_{j\leq n} P(|\frac{1}{T}\sum_{t=1}^{T} u_{jt}| \geq C\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta}) \\
&\leq \exp(\log n - \frac{C^2\log n}{4\max_{j\leq n}\sup_{\boldsymbol{\theta}\in\Theta} E(u_{jt}^2|\boldsymbol{\theta})}) = \frac{1}{n}.
\end{aligned}$$

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

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

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{i,j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| \geq C\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta}) \\
&\leq \sup_{\boldsymbol{\theta}\in\Theta} n^2 \max_{i,j\leq n} P(|\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| \geq C\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta}) \\
&\leq \exp(2\log n - \frac{C^2\log n}{4\max_{j\leq n}\sup_{\boldsymbol{\theta}\in\Theta} E(u_{jt}^4|\boldsymbol{\theta})}) = \frac{1}{n}.
\end{aligned}$$

(iii) Note that $\widehat{u}_{jt} - u_{jt} = -\frac{1}{T}\sum_{t=1}^{T} u_{jt} - \widetilde{\mathbf{x}}_{jt}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, and $\max_{j\leq n} \|\frac{1}{T}\sum_t \widetilde{\mathbf{x}}_{jt}\widetilde{\mathbf{x}}_{jt}'\|_2 < C$ with probability approaching one. The result then follows from part (i) and Lemma 5.1.2.

(iv) Observe that

$$\begin{aligned}
&|\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt}| \leq |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| + |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \widehat{u}_{it}\widehat{u}_{jt}| \\
&\leq |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| + \frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{jt} - u_{jt})^2 + (\frac{2}{T}\sum_t u_{jt}^2)^{1/2}(\frac{2}{T}\sum_t (\widehat{u}_{jt} - u_{jt})^2)^{1/2}
\end{aligned}$$

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

$$\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{j\le n}\frac{1}{T}\sum_t u_{jt}^2 < M|\boldsymbol{\theta})\to 1.$$

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

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

Hence we can pick up M so that $M-\sup_{\boldsymbol{\theta}\in\Theta}\max_{j\le n}E(u_{jt}^2|\boldsymbol{\theta})>C\sqrt{\log n/T}$, and

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\le n}\frac{1}{T}\sum_t u_{jt}^2 \ge M|\boldsymbol{\theta}) \le \sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\le n}|\frac{1}{T}\sum_t u_{jt}^2 - Eu_{jt}^2| \ge M - \max_{j\le n} Eu_{jt}^2|\boldsymbol{\theta})\\ &\le \sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\le n}|\frac{1}{T}\sum_t u_{jt}^2 - Eu_{jt}^2| \ge C\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta})\to 0.\end{aligned}$$

This proves the desired result.

Lemma 5.1.4. *Under Assumption 1.5.1, there is $C>0$, $\inf_{\boldsymbol{\theta}\in\Theta}P(\max_{ij}|\widehat{\rho}_{ij}-\rho_{ij}| < C\sqrt{\log n/T}|\boldsymbol{\theta})\to 1$.*

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

$$\begin{aligned}|\widehat{\rho}_{ij}-\rho_{ij}| \le &\underbrace{\frac{|\frac{1}{T}\sum_t\widehat{u}_{it}\widehat{u}_{jt}-u_{it}u_{jt}|}{(\frac{1}{T}\sum_{t=1}^T\widehat{u}_{it}^2)^{1/2}(\frac{1}{T}\sum_{t=1}^T\widehat{u}_{jt}^2)^{1/2}}}_{X_1}\\ &+\underbrace{|\frac{1}{T}\sum_t u_{it}u_{jt}||(\frac{1}{T}\sum_{t=1}^T\widehat{u}_{it}^2\frac{1}{T}\sum_{t=1}^T\widehat{u}_{jt}^2)^{-1/2}-(\frac{1}{T}\sum_{t=1}^T u_{it}^2\frac{1}{T}\sum_{t=1}^T u_{it}^2)^{-1/2}|}_{X_2}\end{aligned}$$

By part (iv) of Lemma 5.1.3, $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{i,j\leq n} |\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_{it}\widehat{u}_{jt} - Eu_{it}u_{jt}| < C\sqrt{\log n/T}|\boldsymbol{\theta}) \to 1$. Hence for sufficiently large $M > 0$ such that $\inf_{\boldsymbol{\theta}} \min_j E(u_{jt}^2|\boldsymbol{\theta}) - C/M > C\sqrt{\log n/T}$,

$$\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{ij} |X_1| > M\sqrt{\frac{\log n}{T}}|\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta}\in\Theta} P(\min_j \frac{1}{T}\sum_t \widehat{u}_{jt}^2 < C/M|\boldsymbol{\theta}) + o(1)$$
$$\leq \sup_{\boldsymbol{\theta}\in\Theta} P(\max_j |\frac{1}{T}\sum_t \widehat{u}_{jt}^2 - Eu_{jt}^2| > \min_j Eu_{jt}^2 - C/M|\boldsymbol{\theta}) + o(1) = o(1).$$

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

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

Proof of Proposition 1.5.1

Proof. As $1 - \rho_{ij}^2 > 1 - c$ uniformly for (i, j) and $\boldsymbol{\theta}$, the second convergence follows from Lemma 5.1.4. Also, with probability approaching one,

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

Proof of Theorem 1.5.1

Lemma 5.1.5. *There is $C > 0$ so that J_1 has power uniformly on $\Theta(J_1) = \{\sum_{i<j} \rho_{ij}^2 \geq Cn^2 \log n/T\}$.*

Proof. By Lemma 5.1.4, there is $C > 0$, $\inf_{\boldsymbol{\theta}\in\Theta} P(\max_{ij} |\widehat{\rho}_{ij} - \rho_{ij}| < C\sqrt{\log n/T}|\boldsymbol{\theta}) \to 1$. If we define

$$A = \{\sum_{i<j} (\widehat{\rho}_{ij} - \rho_{ij})^2 < C^2 n^2 (\log n/T)\},$$

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

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

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

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

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

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

Proof of Theorem 1.5.1

It suffices to verify conditions (i)-(iii) of Theorem 1.3.2. Condition (i) follows from Theorem 1 of Baltagi et al. (2012). As for condition (ii), note that $J_1 \geq -\frac{\sqrt{n(n-1)}}{2} - \frac{n}{2(T-1)}$ almost surely. Hence as $n, T \to \infty$,

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

Finally, condition (iii) follows from Lemma 5.1.5.

5.1.4 Supplementary Material

Auxiliary lemmas for the proof of Proposition 1.4.2

Define $\mathbf{e}_t = \mathbf{\Sigma}_u^{-1}\mathbf{u}_t = (e_{1t}, ..., e_{Nt})'$, which is an N-dimensional vector with mean zero and covariance $\mathbf{\Sigma}_u^{-1}$, whose entries are stochastically bounded. Let $\bar{\mathbf{w}} = (E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t$. Also recall that

$$a_1 = \frac{T}{\sqrt{N}}\sum_{i=1}^{N}(\widehat{\boldsymbol{\theta}}'\mathbf{\Sigma}_u^{-1})_i^2(\widehat{\sigma}_{ii} - \sigma_{ii}),$$

$$a_2 = \frac{T}{\sqrt{N}}\sum_{i\neq j,(i,j)\in S_U}(\widehat{\boldsymbol{\theta}}'\mathbf{\Sigma}_u^{-1})_i(\widehat{\boldsymbol{\theta}}'\mathbf{\Sigma}_u^{-1})_j(\widehat{\sigma}_{ij} - \sigma_{ij}).$$

One of the key steps of proving $a_1 = o_P(1), a_2 = o_P(1)$ is to establish the following two convergences:

$$\frac{1}{T}E|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1 - \mathbf{f}_s'\bar{\mathbf{w}}))^2|^2 = o(1), \tag{5.3}$$

$$\frac{1}{T}E|\frac{1}{\sqrt{NT}}\sum_{i\neq j,(i,j)\in S_U}\sum_{t=1}^{T}(u_{it}u_{jt} - Eu_{it}u_{jt})[\frac{1}{\sqrt{T}}\sum_{s=1}^{T}e_{is}(1-\mathbf{f}_s'\bar{\mathbf{w}})][\frac{1}{\sqrt{T}}\sum_{k=1}^{T}e_{jk}(1-\mathbf{f}_k'\bar{\mathbf{w}})]|^2 = o(1), \tag{5.4}$$

where $S_U = \{(i,j) : (\mathbf{\Sigma}_u)_{ij} \neq 0\}$. The proofs of (5.3) and (5.4) are given later below.

Lemma 5.1.6. *Under H_0, $a_1 = o_P(1)$.*

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

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

For a_{12}, note that $(\widehat{\boldsymbol{\theta}}'\mathbf{\Sigma}_u^{-1})_i = (1 - \bar{\mathbf{f}}'\mathbf{w})^{-1}\frac{1}{T}\sum_{s=1}^{T}(1 - \mathbf{f}_s'\mathbf{w})(\mathbf{u}_s'\mathbf{\Sigma}_u^{-1})_i = c\frac{1}{T}\sum_{s=1}^{T}(1 - \mathbf{f}_s'\mathbf{w})e_{is}$, 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 (5.3), $Ea_{12}^2 = o(1)$. On the other hand,

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

Note that $\max_{i \le N} \frac{1}{T} \sum_{t=1}^{T} (\widehat{u}_{it} - u_{it})^2 = O_P(\frac{\log N}{T})$ by Lemma 3.1 of Fan et al. (2011). Since $\|\widehat{\boldsymbol{\theta}}\|^2 = O_P(\frac{N \log N}{T})$, $\|\boldsymbol{\Sigma}_u^{-1}\|_2 = O(1)$ and $N(\log N)^3 = o(T^2)$,

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

To bound a_{112}, note that

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

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

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

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

Lemma 5.1.7. *Under H_0, $a_2 = o_P(1)$.*

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

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

where

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

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

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

By (5.4), $Ea_{22}^2 = o(1)$.

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

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

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

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

Similar to the proof of term a_{112} in Lemma 5.1.6, $\max_{ij}|\frac{1}{T}\sum_{t=1}^T u_{it}(\widehat{u}_{jt}-u_{jt})| = O_P(\frac{\log N}{T})$.

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

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

Proof of (5.3) and (5.4)

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

Lemma 5.1.8. *Recall that* $\mathbf{e}_t = \mathbf{\Sigma}_u^{-1}\mathbf{u}_t$. 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

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

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

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

This completes the proof. □

Proof of (5.3)

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 (5.3) is the same regardless of the type of sparsity in Assumption 1.4.2. 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,

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

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

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

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 $\frac{1}{T}\mathrm{var}(X) = o(1)$. The proof of (5.3) is then completed because $\frac{1}{T}EX^2 = \frac{1}{T}(EX)^2 + \frac{1}{T}\mathrm{var}(X) = o(1)$.

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 $\mathrm{cov}(\xi_{it}\zeta_{is}\zeta_{ik},\xi_{jl}\zeta_{jv}\zeta_{jp}) = 0$. 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 5.1.8, 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 $B_{22} = 0$.

$$B_{23} = \frac{1}{NT^4} \sum_{i\neq j} \sum_{t=1}^{T} \sum_{s\neq t} \sum_{l\neq s,t} E\xi_{it}\zeta_{it}E\zeta_{is}\zeta_{js}E\xi_{jl}\zeta_{jl} = O(\frac{1}{NT}) \sum_{i\neq j} |E\zeta_{is}\zeta_{js}|$$

$$= O(\frac{1}{NT}) \sum_{i \neq j} |E e_{is} e_{js} E(1 - \mathbf{f}_s' \mathbf{w})^2| = O(\frac{1}{NT}) \sum_{i \neq j} |E e_{is} e_{js}|.$$

By the definition $\mathbf{e}_s = \mathbf{\Sigma}_u^{-1} \mathbf{u}_s$, $\text{cov}(\mathbf{e}_s) = \mathbf{\Sigma}_u^{-1}$. Hence $E e_{is} e_{js} = (\mathbf{\Sigma}_u^{-1})_{ij}$, 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 5.1.8. Therefore, $\frac{1}{T} B_2 = o(1) + O(\frac{N}{T^2}) = o(1)$.

Proof of (5.4)

For notational simplicity, let $\xi_{ijt} = u_{it} u_{jt} - E u_{it} u_{jt}$. 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

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

The goal is to show $\frac{1}{T} E Z^2 = o(1)$. 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 1.4.2, and is very similar to that of (5.3). 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

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

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 1.4.2 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 5.1.9. *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

$\mathrm{cov}(u_{it}u_{jt}, u_{kt}u_{pt}) \neq 0$, 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 $\mathrm{cov}(u_{kt}, u_{pt}) \neq 0$, which implies $k \in A(p)$. Hence,

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

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

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

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

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 5.1.9. Hence $A_{21} = o(1)$.

Similarly, applying Lemma 5.1.9,

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

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

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

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

Term $a = O(\frac{1}{T})$. For b, note that Lemma 5.1.8 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

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

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

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

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

Further technical lemmas for Section 4

We cite a lemma that will be needed throughout the proofs.

Lemma 5.1.10. *Under Assumption 1.4.1, there is $C > 0$,*

(i) $P(\max_{i,j\leq N} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$

(ii) $P(\max_{i\leq K,j\leq N} |\frac{1}{T}\sum_{t=1}^{T} f_{it}u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$

(iii) $P(\max_{j\leq N} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| > C\sqrt{\frac{\log N}{T}}) \to 0.$

Proof. The proof follows from Lemmas A.3 and B.1 in Fan et al. (2011). □

Lemma 5.1.11. *When the distribution of $(\mathbf{u}_t, \mathbf{f}_t)$ is independent of $\boldsymbol{\theta}$, there is $C > 0$,*

(i) $\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq N} |\widehat{\theta}_j - \theta_j| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta}) \to 0$

(ii) $\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{i,j\leq N} |\widehat{\sigma}_{ij} - \sigma_{ij}| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta}) \to 0,$

(iii) $\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{i\leq N} |\widehat{\sigma}_i - \sigma_i| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta}) \to 0.$

Proof. Note that $\widehat{\theta}_j - \theta_j = \frac{1}{a_{f,T}T}\sum_{t=1}^{T} u_{jt}(1 - \mathbf{f}_t'\mathbf{w})$. Here $a_{f,T} = 1 - \bar{\mathbf{f}}'\mathbf{w} \to^p 1 - E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t > 0$, hence $a_{f,T}$ is bounded away from zero with probability approaching one. Thus by Lemma 5.1.10, there is $C > 0$ independent of $\boldsymbol{\theta}$, such that

$$\sup_{\boldsymbol{\theta}\in\Theta} P(\max_{j\leq N} |\widehat{\theta}_j - \theta_j| > C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta}) = P(\max_{j} |\frac{1}{a_{f,T}T}\sum_{t=1}^{T} u_{jt}(1 - \mathbf{f}_t'\mathbf{w})| > C\sqrt{\frac{\log N}{T}}) \to 0$$

(ii) There is C independent of $\boldsymbol{\theta}$, 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 $\boldsymbol{\theta}$ 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

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

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

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

which yields the desired result.

(iii): Recall $\widehat{\sigma}_j^2=\widehat{\sigma}_{jj}/a_{f,T}$, and $\sigma_j^2=\sigma_{jj}/(1-E\mathbf{f}_t'(E\mathbf{f}_t\mathbf{f}_t')^{-1}E\mathbf{f}_t)$. Moreover, $a_{f,T}$ is independent of $\boldsymbol{\theta}$. The result follows immediately from part (ii). □

Lemma 5.1.12. *For any $\epsilon>0$, $\sup_{\boldsymbol{\theta}}P(\|\widehat{\boldsymbol{\Sigma}}_u^{-1}-\boldsymbol{\Sigma}_u^{-1}\|>\epsilon|\boldsymbol{\theta})=o(1)$.*

Proof. By Lemma 5.1.11 (ii), $\sup_{\boldsymbol{\theta}\in\Theta}P(\max_{i,j\leq N}|\widehat{\sigma}_{ij}-\sigma_{ij}|>C\sqrt{\frac{\log N}{T}}|\boldsymbol{\theta})\to 1$. By Fan et al. (2011), 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 $\boldsymbol{\theta}$, $\|\widehat{\boldsymbol{\Sigma}}_u^{-1}-\boldsymbol{\Sigma}_u^{-1}\|\leq C'm_N(\frac{\log N}{T})^{1/2}$. Hence the result follows due to the sparse condition $m_N(\frac{\log N}{T})^{1/2}=o(1)$. □

5.2 Proofs for Chapter 2

In this section, we provide theoretical proofs in developing the theory of sufficient forecasting. We first cite a few lemmas from Fan et al. (2013), which are needed subsequently in the proofs.

Lemma 5.2.1. *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are two symmetric, semi-positive definite matrices, and that* $\lambda_{\min}(\mathbf{A}) > c_{p,T}$ *for some sequence* $c_{p,T} > 0$. *If* $||\mathbf{A} - \mathbf{B}|| = o_p(c_{p,T})$, *then*

$$||\mathbf{A}^{-1} - \mathbf{B}^{-1}|| = O_p(c_{p,T}^{-2})||\mathbf{A} - \mathbf{B}||.$$

Lemma 5.2.2. *Let* $\{\lambda_i\}_{i=1}^p$ *be the eigenvalues of* $\mathbf{\Sigma}$ *in descending order and* $\{\boldsymbol{\xi}\}_{i=1}^p$ *be their associated eigenvectors. Correspondingly, let* $\{\widehat{\lambda}_i\}_{i=1}^p$ *be the eigenvalues of* $\widehat{\mathbf{\Sigma}}$ *in descending order and* $\{\widehat{\boldsymbol{\xi}}\}_{i=1}^p$ *be their associated eigenvectors. Then,*

(a)(Weyl's theorem) $|\widehat{\lambda}_i - \lambda_i| \leq ||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}||$.

(b) ($\sin(\theta)$ *theorem)*

$$||\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i|| \leq \frac{||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}||/\sqrt{2}}{\min(|\widehat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \widehat{\lambda}_{i+1}|)}$$

5.2.1 Proof of Proposition 2.2.1

It suffices to show that $\widehat{\mathbf{f}}_t = \widehat{\mathbf{\Lambda}}_b \mathbf{x}_t$, or $\widehat{\mathbf{F}}' = \widehat{\mathbf{\Lambda}}_b \mathbf{X}$ in matrix form. First let $\mathbf{M} = \text{diag}(\lambda_1, ..., \lambda_K)$, where λ_i are the largest K eigenvalues of $\mathbf{X}'\mathbf{X}$. By construction, we have $(\mathbf{X}'\mathbf{X})\widehat{\mathbf{F}} = \widehat{\mathbf{F}}\mathbf{M}$, or $\mathbf{M}^{-1}\widehat{\mathbf{F}}'(\mathbf{X}'\mathbf{X}) = \widehat{\mathbf{F}}'$. Since $\widehat{\mathbf{B}}'\widehat{\mathbf{B}} = T^{-2}\widehat{\mathbf{F}}'(\mathbf{X}'\mathbf{X})\widehat{\mathbf{F}} = T^{-2}\widehat{\mathbf{F}}'\widehat{\mathbf{F}}\mathbf{M} = T^{-1}\mathbf{M}$, it follows that $(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}(T^{-1}\widehat{\mathbf{F}}'\mathbf{X}')\mathbf{X} = (\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}'\mathbf{X} = \widehat{\mathbf{F}}'$. This concludes the proof.

5.2.2 Proof of Theorem 2.3.1

Recall that $\mathbf{H} = (1/T)\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$. A preliminary result about $\mathbf{H}$ is as follows, which can be proved analogously to Lemma 11 in Fan et al. (2013).

Lemma 5.2.3. *Under assumptions 2.3.1-2.3.3, we have*

(a) $\mathbf{H}\mathbf{H}' = \mathbf{I}_K + O_p(\omega_{p,T})$,

(b) $\mathbf{H}'\mathbf{H} = \mathbf{I}_K + O_p(\omega_{p,T})$.

The next lemma shows that the normalization matrix $\mathbf{\Lambda}_b$ can be consistently estimated under operator norm.

Lemma 5.2.4. *Under assumptions 2.3.1-2.3.3,*

(a) $||\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'|| = O_p(p^{1/2}\omega_{p,T})$,

(b) $||\widehat{\mathbf{\Lambda}}_b - \mathbf{H}\mathbf{\Lambda}_b|| = O_p(p^{-1/2}\omega_{p,T})$.

Proof. (a) We outline the procedure as follows.

First, under assumptions 2.3.1-2.3.3, we have the following convergence of factors,

$$\frac{1}{T}\sum_{t=1}^{T}||\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t|| = O_p(\omega_{p,T}).$$

This result can be similarly obtained from Theorem 1 of Bai and Ng (2002).

Next, lemma 5.2.3 leads to $||\mathbf{H}|| = O_p(1)$. Note that $\widehat{\mathbf{b}}_i = (1/T)\sum_{t=1}^{T} x_{it}\widehat{\mathbf{f}}_t$ and that $(1/T)\sum_{t=1}^{T}\widehat{\mathbf{f}}_t'\widehat{\mathbf{f}}_t = \mathbf{I}_K$. As a result,

$$\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t u_{it} + \frac{1}{T}\sum_{t=1}^{T}x_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) + \mathbf{H}(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t'\mathbf{f}_t - \mathbf{I}_K)\mathbf{b}_i$$

The three terms on the right-hand side can be bounded as follows. For the first term, we have

$$||\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}\mathbf{f}_t u_{it}|| \leq ||\mathbf{H}|| \cdot ||\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t u_{it}||.$$

For the second term, since $E(x_{it}^2) = O(1)$, $T^{-1}\sum_{t=1}^{T}x_{it}^2 = O_p(1)$. Hence

$$||\frac{1}{T}\sum_{t=1}^{T}x_{it}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t)|| \leq (\frac{1}{T}\sum_{t=1}^{T}x_{it}^2)\frac{1}{T}\sum_{t=1}^{T}||\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t||^{1/2} = O_p(\omega_{p,T}).$$

And lastly, $||T^{-1}\sum_{t=1}^{T}(\mathbf{f}_t'\mathbf{f}_t - \mathbf{I}_K)|| = O_p(T^{-1/2})$ and $||\mathbf{b}_i|| = O(1)$ imply that the third term is $O_p(T^{-1/2})$.

Therefore we have

$$\begin{aligned} ||\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'||^2 &\leq ||\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'||_F^2 = \sum_{i=1}^{p} ||\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i||^2 \\ &\leq 3||\mathbf{H}||^2 \cdot (T^{-1} \sum_{i=1}^{p} ||\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{f}_t u_{it}||^2)) + pO_p(\omega_{p,T}^2), \end{aligned}$$

where we used the fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$. Since $p^{-1}\sum_{i=1}^{p} ||\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{f}_t u_{it}||^2 = O_p(1)$ by assumption 2.3.3, it follows that

$$||\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}'||^2 = O_p(p/T) + O_p(p\omega_{p,T}^2) = O_p(p\omega_{p,T}^2).$$

(b) Since $||\mathbf{B}|| = O_p(\sqrt{p})$, from part (a) we have $||\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{H}\mathbf{B}'\mathbf{B}\mathbf{H}'|| \leq ||(\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}')'|| \cdot ||\widehat{\mathbf{B}} + \mathbf{B}\mathbf{H}'|| = O_p(p\omega_{p,T})$. In addition, $\lambda_{\min}(\mathbf{B}'\mathbf{B}) > p/2$, by lemma 5.2.1,

$$||(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1} - (\mathbf{H}\mathbf{B}'\mathbf{B}\mathbf{H}')^{-1}|| = O_p(p^{-1}\omega_{p,T}).$$

According to lemma 5.2.3 , the effect of replacing $\mathbf{H}^{-1}$ by $\mathbf{H}'$ is negligible, as $||\mathbf{H}^{-1} - \mathbf{H}'|| = O_p(\omega_{p,T})$. From part (a), it follows that $||\mathbf{H}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{H}^{-1} - (\mathbf{H}\mathbf{B}'\mathbf{B}\mathbf{H}')^{-1}|| = O_p(p^{-1}\omega_{p,T})$. Hence

$$||(\mathbf{B}'\mathbf{B})^{-1} - \mathbf{H}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{H}^{-1}|| = O_p(p^{-1}\omega_{p,T}).$$

Consequently,

$$\begin{aligned} ||\widehat{\mathbf{\Lambda}}_b - \mathbf{H}\mathbf{\Lambda}_b|| &= ||(\widehat{\mathbf{B}}'\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}' - \mathbf{H}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{H}^{-1}\mathbf{H}\mathbf{B}'|| \\ &\leq ||(\mathbf{B}'\mathbf{B})^{-1} - \mathbf{H}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{H}^{-1}|| \cdot ||\widehat{\mathbf{B}}'|| + ||\mathbf{H}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{H}^{-1}|| \cdot ||\widehat{\mathbf{B}}' - \mathbf{H}\mathbf{B}'|| \\ &= O_p(p^{-1/2}\omega_{p,T}). \end{aligned}$$

□

The following lemma lays the foundation of inverse regression, which can be found in Li (1991).

Lemma 5.2.5. *Under model (2.1) and Assumption 2.3.1 (3), the centered inverse regression curve* $E(\mathbf{f}_t|y_{t+1}) - E(\mathbf{f}_t)$ *is contained in the linear subspace spanned by* $\boldsymbol{\phi}_k'\text{cov}(\mathbf{f}_t), k = 1, ..., L$.

We are now ready to complete the proof of Theorem 2.3.1.

Proof of Theorem 2.3.1

Let $\widehat{\mathbf{m}}_h = \frac{1}{c}\sum_{l=1}^{c}\mathbf{x}_{(h,l)}$ denote the average of the predictors within a particular slice I_h, and $\mathbf{m}_h = E(\mathbf{x}_t|y_{t+1} \in I_h)$ be its population version. We immediately have

$$\begin{aligned} ||\widehat{\mathbf{m}}_h - \mathbf{m}_h|| &= ||\frac{1}{c}\sum_{l=1}^{c}(\mathbf{B}\mathbf{f}_{(h,l)} + \mathbf{u}_{(h,l)}) - \mathbf{B}E(\mathbf{f}_t|y_{t+1} \in I_h)|| \\ &\leq ||\mathbf{B}|| \cdot ||\frac{1}{c}\sum_{l=1}^{c}\mathbf{f}_{(h,l)} - E(\mathbf{f}_t|y_{t+1} \in I_h)|| + ||\frac{1}{c}\sum_{l=1}^{c}\mathbf{u}_{(h,l)}|| \\ &= O_p(p^{1/2})O_p(T^{-1/2}) + O_p(\sqrt{p/T}) = O_p(\sqrt{p/T}). \end{aligned}$$

Here, we use the fact that the sample mean of $E(\mathbf{f}_t|y_{t+1} \in I_h)$ converges at the rate of $O_p(T^{-1/2})$. This holds true as the random variable $\mathbf{f}_t|y_{t+1} \in I_h$ is still stationary with finite second moments, and the sum of the α-mixing coefficients converges. This applies to $\mathbf{u}_t|y_{t+1} \in I_h$ as well.

In addition to the inequality above, we have $||\mathbf{m}_h|| = O_p(||E(\mathbf{x}_t|y_{t+1} \in I_h)||) \leq O_p(||\mathbf{B}|| \cdot ||E(\mathbf{f}_t|y_{t+1} \in I_h)||) = O_p(p^{1/2})$, so $\widehat{\mathbf{m}}_h = O_p(p^{1/2})$. It follows that

$$\begin{aligned} ||\widehat{\boldsymbol{\Lambda}}_b\widehat{\mathbf{m}}_h - \mathbf{H}\boldsymbol{\Lambda}_b\mathbf{m}_h|| &\leq ||\widehat{\boldsymbol{\Lambda}}_b - \mathbf{H}\boldsymbol{\Lambda}_b|| \cdot ||\widehat{\mathbf{m}}_h|| + ||\mathbf{H}\boldsymbol{\Lambda}_b|| \cdot ||\widehat{\mathbf{m}}_h - \mathbf{m}_h|| \\ &= O_p(\omega_{p,T}) + O_p(T^{-1/2}) = O_p(\omega_{p,T}). \end{aligned}$$

By definition, $\mathbf{\Sigma}_{f|y} = H^{-1}\sum_{h=1}^{H}(\mathbf{\Lambda}_b\mathbf{m}_h)(\mathbf{\Lambda}_b\mathbf{m}_h)'$. For fixed H, note that

$$\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{H}\mathbf{\Sigma}_{f|y}\mathbf{H}' = H^{-1}\sum_{h=1}^{H}[(\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h)(\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h)' - (\mathbf{H}\mathbf{\Lambda}_b\mathbf{m}_h)(\mathbf{H}\mathbf{\Lambda}_b\mathbf{m}_h)'],$$

and that both $||\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h||$ and $\widehat{\mathbf{\Lambda}}_b\widehat{\mathbf{m}}_h$ are $O_p(1)$, we reach the desired result that $||\widehat{\mathbf{\Sigma}}_{f|y} - \mathbf{H}\mathbf{\Sigma}_{f|y}\mathbf{H}'|| = O_p(\omega_{p,T})$.

A direct application of $\sin(\theta)$ theorem shows that $||\widehat{\boldsymbol{\phi}}_j - \mathbf{H}\boldsymbol{\phi}_j|| = O_p(\omega_{p,T})$. Since we have the normalization $\text{cov}(\mathbf{f}_t) = \mathbf{I}_K$ and $E(\mathbf{f}_t) = \mathbf{0}$, the eigenvalue $\boldsymbol{\phi}_j$'s of $\mathbf{\Sigma}_{f|y}$ constitute the SDR directions for model (2.1).

5.2.3 Proof of Proposition 2.3.1

First we write $\widehat{\boldsymbol{\phi}}$ in terms of the true factors $\mathbf{f}_t$,

$$\begin{aligned}\widehat{\boldsymbol{\phi}} &= \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}\widehat{\mathbf{f}}_t = \frac{1}{T-1}\widehat{\mathbf{\Lambda}}_b\sum_{t=1}^{T-1} y_{t+1}\mathbf{x}_t \\ &= \frac{1}{T-1}\widehat{\mathbf{\Lambda}}_b\sum_{t=1}^{T-1}(\mathbf{B}\mathbf{f}_t + \mathbf{u}_t)y_{t+1},\end{aligned}$$

where we used the fact that $\mathbf{f}_t = \widehat{\mathbf{\Lambda}}_b\mathbf{x}_t$. Using triangular inequality,

$$\begin{aligned}||\widehat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}|| &= ||\widehat{\boldsymbol{\phi}} - (\widehat{\mathbf{\Lambda}}_b\mathbf{B})\bar{\boldsymbol{\phi}} + (\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I})\bar{\boldsymbol{\phi}}|| \\ &\leq ||(\widehat{\mathbf{\Lambda}}_b\mathbf{B})\frac{1}{T-1}(\sum_{t=1}^{T-1} y_{t+1}\mathbf{f}_t - \bar{\boldsymbol{\phi}})|| + ||(\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I})\bar{\boldsymbol{\phi}}|| + ||\frac{1}{T-1}\widehat{\mathbf{\Lambda}}_b\sum_{t=1}^{T-1}\mathbf{u}_t y_{t+1}||.\end{aligned}$$

By lemma 5.2.3 and 5.2.4, we have $||\widehat{\mathbf{\Lambda}}_b\mathbf{B}|| = O_p(1)$ and $||\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{\Lambda}_b\mathbf{B}|| = ||\widehat{\mathbf{\Lambda}}_b\mathbf{B} - \mathbf{I}|| = O_p(\omega_{p,T})$. Since $||\bar{\boldsymbol{\phi}}|| = O_p(1)$, the second term on the right hand side of the inequality is $O_p(\omega_{p,T})$. For the third term, note that $\mathbf{u}_t$ is independent of y_{t+1}, hence $E(\mathbf{u}_t y_{t+1}) = 0$. By law of large numbers and $||\widehat{\mathbf{\Lambda}}_b|| = O_p(p^{-1/2})$, the third term is $O_p(T^{-1/2})$. It remains to bound $||\frac{1}{T-1}(\sum_{t=1}^{T-1} y_{t+1}\mathbf{f}_t - \bar{\boldsymbol{\phi}})||$ in the first term.

We express $\mathbf{f}_t$ along the basis $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L$ and their orthogonal hyperplane,

$$\mathbf{f}_t = \sum_{j=1}^{L} \langle \mathbf{f}_t, \boldsymbol{\phi}_j \rangle \boldsymbol{\phi}_j + \mathbf{f}_t^{\perp}.$$

By the orthogonal decomposition of normal distribution, $\langle \mathbf{f}_t, \boldsymbol{\phi}_j \rangle$ and $\mathbf{f}_t^{\perp}$ are independent. In addition, y_{t+1} depends on $\mathbf{f}_t$ only through $\boldsymbol{\phi}_1' \mathbf{f}_t, ..., \boldsymbol{\phi}_L' \mathbf{f}_t$, and is therefore conditionally independent of $\mathbf{f}_t^{\perp}$. It follows from contraction property that y_{t+1} and $\mathbf{f}_t^{\perp}$ are independent, unconditionally. $E(y_{t+1}\mathbf{f}_t^{\perp}) = E(y_{t+1})E(\mathbf{f}_t^{\perp}) = \mathbf{0}$. Now it is easy to see that

$$\begin{aligned} ||\frac{1}{T-1}(\sum_{t=1}^{T-1} y_{t+1}\mathbf{f}_t - \bar{\boldsymbol{\phi}})|| &= ||\sum_{j=1}^{L}[\frac{1}{T-1}\sum_{t=1}^{T-1}(\boldsymbol{\phi}_j'\mathbf{f}_t)y_{t+1} - E((\boldsymbol{\phi}_j'\mathbf{f}_t)y_{t+1})] + \frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}\mathbf{f}_t^{\perp}|| \\ &\leq \sum_{j=1}^{L} ||\frac{1}{T-1}\sum_{t=1}^{T-1}(\boldsymbol{\phi}_j'\mathbf{f}_t)y_{t+1} - E((\boldsymbol{\phi}_j'\mathbf{f}_t)y_{t+1})|| + ||\frac{1}{T-1}\sum_{t=1}^{T-1} y_{t+1}\mathbf{f}_t^{\perp}|| \end{aligned}$$

and each term is $O_p(T^{-1/2})$ by law of large numbers. This concludes the proof.

Bibliography

Ahn, S. C. and Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227.

Alessi, L., Barigozzi, M. and Capasso, M. (2010), 'Improved penalization for determining the number of factors in approximate factor models', *Statistics & Probability Letters* **80**(23), 1806–1813.

Andrews, D. (1998), 'Hypothesis testing with a restricted parameter space', *Journal of Econometrics* **84**, 155–199.

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

Antoniadis, A. and Fan, J. (2001), 'Regularized wavelet approximations', *Journal of the American Statistical Association* **96**, 939–967.

Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171.

Bai, J. and Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bai, J. and Ng, S. (2008), 'Forecasting economic time series using targeted predictors', *Journal of Econometrics* **146**(2), 304–317.

Bai, J. and Ng, S. (2009), 'Boosting diffusion indices', *Journal of Applied Econometrics* **24**(4), 607–629.

Bai, Z. D. and Saranadasa, H. (1996), 'Effect of high dimension: by an example of a two sample problem', *Statistica Sinica* **6**(2), 311–329.

Baltagi, B. (2008), *Econometric Analysis of Panel Data*, the fourth edition edn, 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.

Beaulieu, M., Dufour, J. 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.

Bernanke, B. S., Boivin, J. and Eliasz, P. (2005), 'Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach', *The Quarterly journal of economics* **120**(1), 387–422.

Bickel, P. and Levina, E. (2008), 'Covariance regularization by thresholding', *Annals of Statistics* **36**(6), 2577–2604.

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., Liu, W. and Xia, Y. (2013), 'Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings', *Journal of the American Statistical Association* **108**, 265–277.

Cai, T., Zhang, C. and Zhou, H. (2010), 'Optimal rates of convergence for covariance matrix estimation', *Annals of Statistics* **38**(4), 2118–2144.

Campbell, J. Y. and Shiller, R. J. (1988), 'Stock prices, earnings, and expected dividends', *Journal of Finance* **43**(3), 661–676.

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of finance* **52**(1), 57–82.

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

Chen, C.-H. and Li, K.-C. (1998), 'Can sir be as popular as multiple linear regression?', *Statistica Sinica* **8**(2), 289–316.

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**(2), 808–835.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013), Testing many moment inequalities, Technical report, MIT.

Connor, G., Hagmann, M. and Linton, O. (2012), 'Efficient semiparametric estimation of the fama–french model and extensions', *Econometrica* **80**(2), 713–754.

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

Cook, R. D. (2009), *Regression Graphics: ideas for Studying Regressions through Graphics*, Vol. 482, John Wiley & Sons.

Cook, R. D. and Li, B. (2002), 'Dimension reduction for conditional mean in regression', *The Annals of Statistics* **30**(2), 455–474.

Cook, R. D. and Weisberg, S. (1991), 'Sliced inverse regression for dimension reduction (with discussion)', *Journal of the American Statistical Association* **86**(414), 328–332.

Donald, S. G., Imbens, G. W. and Newey, W. K. (2003), 'Empirical likelihood estimation and consistent tests with conditional moment restrictions', *Journal of Econometrics* **117**(1), 55–93.

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 the American Statistical Association* **91**, 674–688.

Fan, J., Fan, Y. and Lv, J. (2008), 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* **147**(1), 186–197.

Fan, J., Han, X. and Gu, W. (2012), 'Estimating false discovery proportion under arbitrary covariance dependence (with discussion)', *Journal of the American Statistical Association* **107**(499), 1019–1035.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011), 'High dimensional covariance matrix estimation in approximate factor models', *Annals of Statistics* **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, Series B* **75**, 603–680.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000), 'The generalized dynamic-factor model: identification and estimation', *Review of Economics and statistics* **82**(4), 540–554.

Gagliardini, P., Ossola, E. and Scaillet, O. (2011), Time-varying risk premium in large cross-sectional equidity datasets, Technical report, Swiss Finance Institute.

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

Hall, P. and Li, K.-C. (1993), 'On almost linearity of low dimensional projections from high dimensional data', *The Annals of Statistics* **21**(2), 867–889.

Hallin, M. and Liška, R. (2007), 'Determining the number of factors in the general dynamic factor model', *Journal of the American Statistical Association* **102**(478), 603–617.

Hansen, P. (2003), Asymptotic tests of composite hypotheses, Technical report, CREATES.

Hansen, P. (2005), 'A test for superior predictive ability', *Journal of Business and Economic Statistics* **23**, 365–380.

Hotelling, H. (1957), 'The relations of the newer multivariate statistical methods to factor analysis', *British Journal of Statistical Psychology* **10**(2), 69–79.

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

Jiang, B. and Liu, J. S. (2014), 'Variable selection for general index models via sliced inverse regression', *The Annals of Statistics* **42**(5), 1751–1786.

Kelly, B. and Pruitt, S. (2013), 'Market expectations in the cross-section of present values', *Journal of Finance* **68**(5), 1721–1756.

Kelly, B. and Pruitt, S. (2014), 'The three-pass regression filter: a new approach to forecasting using many predictors', *Journal of Econometrics, conditionally accepted* .

Kendall, M. (1957), *A Course in Multioariate Analysis.*, London: Griffin.

Lam, C., Yao, Q. et al. (2012), 'Factor modeling for high-dimensional time series: inference for the number of factors', *The Annals of Statistics* **40**(2), 694–726.

Leek, J. T. and Storey, J. D. (2008), 'A general framework for multiple testing dependence', *Proceedings of the National Academy of Sciences* **105**(48), 18718–18723.

Li, B. and Dong, Y. (2009), 'Dimension reduction for nonelliptically distributed predictors', *The Annals of Statistics* **37**(3), 1272–1298.

Li, B. and Wang, S. (2007), 'On directional regression for dimension reduction', *Journal of the American Statistical Association* **102**(479), 997–1008.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction (with discussion)', *Journal of the American Statistical Association* **86**(414), 316–327.

Li, K.-C. (1992), 'On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma', *Journal of the American Statistical Association* **87**(420), 1025–1039.

Li, K.-C. and Duan, N. (1989), 'Regression analysis under link violation', *The Annals of Statistics* **17**(3), 1009–1052.

Lintner, J. (1965), 'The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets', *The Review of Economics and Statistics* **47**(1), 13–37.

Ludvigson, S. and Ng, S. (2007), 'The empirical risk return relation: A factor analysis approach', *Journal of Financial Economics* **83**(1), 171–222.

Ludvigson, S. and Ng, S. (2009), 'Macro factors in bond risk premia', *Review of Financial Studies* **22**(12), 5027–5067.

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

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

Paul, D., Tibshirani, R., Bair, E. and Hastie, T. (2006), 'Prediction by supervised principal components', *Journal of the American Statistical Association* **101**(473), 119–137.

Pesaran, H., Ullah, A. and Yamagata, T. (2008), 'A bias-adjusted lm test of error cross section independence', *Econometrics Journal* **11**, 105–127.

Pesaran, H. and Yamagata, T. (2012), Testing capm with a large number of assets, Technical report, University of South California.

Polk, C., Thompson, S. and Vuolteenaho, T. (2006), 'Cross-sectional forecasts of the equity premium', *Journal of Financial Economics* **81**(1), 101–141.

Rajan, R. G. and Zingales, L. (1998), 'Financial dependence and growth', *American Economic Review* **88**(3), 559–586.

Ross, S. (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', *Journal of the American Statistical Association* **104**(485), 177–186.

Schott, J. R. (1994), 'Determining the dimensionality in sliced inverse regression', *Journal of the American Statistical Association* **89**(3), 141–148.

Sharpe, W. F. (1964), 'Capital asset prices: a theory of market equilibrium under conditions of risk', *Journal of Finance* **19**(3), 425–442.

Stock, J. H. and Watson, M. W. (1989), New indexes of coincident and leading economic indicators, *in* 'NBER Macroeconomics Annual 1989, Volume 4', Vol. 4, Mit Press, pp. 351–409.

Stock, J. H. and Watson, M. W. (2002*a*), 'Forecasting using principal components from a large number of predictors', *Journal of the American Statistical Association* **97**(460), 1167–1179.

Stock, J. H. and Watson, M. W. (2002*b*), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**(2), 147–162.

Stock, J. H. and Watson, M. W. (2012), 'Generalized shrinkage methods for forecasting using many predictors', *Journal of Business & Economic Statistics* **30**(4), 481–493.

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

Zhu, L., Miao, B. and Peng, H. (2006), 'On sliced inverse regression with high-dimensional covariates', *Journal of the American Statistical Association* **101**(474), 630–643.