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
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 $H : \theta = 0$ 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 nonlinearity 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 Çınlar, 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
List of Tables

3.1 Means and covariances used to generate $b_t$ and $f_t$ ......................... 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 $J_{\text{wald}}$ 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:

$$H_0 : \theta = 0,$$

where $N = \text{dim}(\theta)$ is allowed to grow faster than the sample size $T$. We are particularly interested in boosting the power in sparse alternatives under which $\theta$ 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 $f_t = (f_{1t}, ..., f_{Kt})'$ be the excess returns of $K$
tradable market risk factors. Then, the excess return has the following decomposition:

\[ y_{it} = \theta_i + b_i'f_t + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \]

where \( b_i = (b_{i1}, \ldots, b_{iK})' \) is a vector of factor loadings and \( u_{it} \) represents the idiosyncratic error. The key implication from the multi-factor pricing theory is that the intercept \( \theta_i \) 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 \( H_0 : \theta = 0 \), where \( \theta = (\theta_1, \ldots, \theta_N)' \) is the vector of intercepts for all \( N \) financial assets. As the factor pricing model is derived from theories of financial economics (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 \( \theta \) is sparse.

Most of the conventional tests for \( H_0 : \theta = 0 \) are based on a quadratic form:

\[ W = \hat{\theta}' \hat{V} \hat{\theta}. \]

Here \( \hat{\theta} \) is an element-wise consistent estimator of \( \theta \), and \( \hat{V} \) is a high-dimensional positive definite weight matrix, often taken to be the inverse of the asymptotic covariance matrix of \( \hat{\theta} \) (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 \( \hat{V} \) 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:

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

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

$$y_{it} = \alpha + x'_{it}\beta + \mu_i + u_{it}, \quad i \leq n, t \leq T.$$
Let \( \rho_{ij} \) 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 \( \Sigma_u \) of \( \{u_{it}\}_{i \leq n} \) is diagonal. In empirical applications, weak cross-sectional correlations are often present, which results in a sparse covariance \( \Sigma_u \) with just a few nonzero off-diagonal elements. This leads to a sparse vector \( \theta = (\rho_{12}, \rho_{13}, \ldots, \rho_{n-1,n}) \). 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 \( (u_{1t}, \ldots, u_{nt}) \). 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 \( A \), let \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) represent its minimum and maximum eigenvalues. Let \( \|A\|_2 \) and \( \|A\|_1 \) denote its operator norm and \( l_1 \)-norm respectively, defined by \( \|A\|_2 = \lambda_{\max}^{1/2}(A'A) \) and \( \max_i \sum_j |A_{ij}| \). For a vector \( \theta \), define \( \|\theta\| = (\sum_j \theta_j^2)^{1/2} \) and \( \|\theta\|_{\max} = \max_j |\theta_j| \). 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:

\[ H_0 : \theta = 0, \quad H_a : \theta \in \Theta_a, \]

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

\[ J_1|H_0 \to^d F. \]  

(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 \( \{D : J_1 > F_q\} \) and satisfies

\[ \lim \sup_{T,N \to \infty} P(J_1 > F_q|H_0) = q. \]  

(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,

\[ \lim \inf_{T,N \to \infty} \inf_{\theta \in \Theta(J_1)} P(J_1 > F_q|\theta) \to 1. \]  

(1.3)

Typically, \( \Theta(J_1) \) consists of those \( \theta \)'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 \( \text{dim}(\theta) = N \) can grow fast with the sample size \( T \). We are particularly interested in \textit{sparse alternatives} \( \Theta_s \subset \Theta_a \) under which \( H_0 \) is violated only on a couple of exceptional components of \( \theta \). Specifically, when \( \theta \in \Theta_s \), 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 $\Theta_s$, 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

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

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

2. When $\theta \in \Theta_s$ is a sparse high-dimensional vector under the alternative, the “classical” test $J_1$ may have low power as $\|\theta\|$ is typically relatively small. On the other hand, for $\theta \in \Theta_s$, $J_0$ stochastically dominates $J_1$. As a result, $P(J > F_q|\theta) > P(J_1 > F_q|\theta)$ strictly holds, so the power of $J_1$ over the set $\Theta_s$ is enhanced after adding $J_0$. Often $J_0$ diverges fast under sparse alternatives $\Theta_s$, which ensures $P(J > F_q|\theta) \to 1$ for $\theta \in \Theta_s$. In contrast, the classical test only has $P(J_1 > F_q|\theta) < c < 1$ for some $c \in (0,1)$ and $\theta \in \Theta_s$, and when $\|\theta\|$ is sufficiently small, $P(J_1 > F_q|\theta)$ 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 \( \Theta_s \). Such a component can be constructed via screening as follows. Suppose we have a consistent estimator \( \hat{\theta} \) such that \( \max_{j \leq N} |\hat{\theta}_j - \theta_j| = o_p(1) \). For some slowly growing sequence \( \delta_{N,T} \to \infty \) (as \( T, N \to \infty \)), define a screening set:

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

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

\[
\inf_{\theta \in \Theta_a \cup \{0\}} P(\max_{j \leq N} |\hat{\theta}_j - \theta_j|/\hat{v}_j^{1/2} < \delta_{N,T}/2|\theta) \to 1
\]
for \( \theta \) under both null and alternate hypotheses. The screening statistic \( J_0 \) is then defined as

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

By (1.4) and (1.5), under \( H_0 : \theta = 0 \),

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

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

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

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

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

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

Thus, the power of \( J_1 \) is enhanced on the subset

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

As a byproduct, the screening set consistently identifies the elements of \( \theta \) 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 $\hat{\theta}$, denoted by $\hat{\text{var}}(\hat{\theta})^{-1}$, then $J_1$ can be chosen as the standardized Wald-statistic:

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

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

In sparse alternatives where $\parallel \theta \parallel$ may not grow fast with $N$ but $\theta \in \Theta_s$, 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 $\parallel \theta \parallel$ 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 $\Theta(J_1) \subset \mathbb{R}^N/\{0\}$ as the set of alternative $\theta$’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 $\theta$ on

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

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

### 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 $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_N)$, and forms the statistic based on $\max_{j \leq N} |\hat{\theta}_j|^{\delta}$ 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 $t_T$,

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

(1.7)

The accumulation of estimation errors is prevented due to the threshold $1\{|\hat{\theta}_j| > t_T w_j\}$ (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 $\hat{\theta}_j$ and $\hat{\theta}_{j+k}$ decay fast enough as $k \to \infty$ (Zhong et al. 2013). In cross-sectional testing problems, this essentially requires an explicit ordering among the cross-sectional units which is, however, often unavailable in panel data applications. In addition, as (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 \( \max_{j \leq N} |\hat{\theta}_j| / \hat{v}_j \), \( R \) 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 \( P(\cdot; \theta) \) to denote the probability measure defined from the sampling distribution with parameter \( \theta \). Let \( \Theta \subset \mathbb{R}^N \) be the parameter space of \( \theta \). When we write \( \inf_{\theta \in \Theta} P(\cdot; \theta) \), 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 \( \{\hat{\theta}_j, \hat{v}_j\}_{j \leq N} \) are such that

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

The normalizing constant \( v_j \) is often taken as the asymptotic variance of \( \hat{\theta}_j \), with \( \hat{v}_j \) being its consistent estimator. The constants \( 4/9 \) and \( 16/9 \) in condition (ii) are not optimally chosen, as this condition only requires \( \{\hat{v}_j\}_{j \leq N} \) 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, \( \max_{j \leq N} |\hat{\theta}_j - \theta_j| / \hat{v}_j^{1/2} = \)
O_P(\sqrt{\log N}). Therefore, we shall fix

\[ \delta_{N,T} = \log(\log T)\sqrt{\log N}, \]  

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 \( \hat{S} \) and \( S(\theta) \) are defined by (1.4) and (1.6) respectively for a given \( \theta \in \Theta \) and its consistent estimator \( \hat{\theta} \). In particular, \( S(\theta) = \{ j : |\theta_j| > 2v_j^{1/2}\delta_{N,T}, j = 1, ..., N \} \), so under

\[ H_0 : \theta = 0, \quad S(\theta) = \emptyset. \]

Recall that \( \Theta \) denotes the parameter space containing both the null and alternative hypotheses. The following theorem characterizes the asymptotic behavior of \( J_0 = \sqrt{N} \sum_{j \in \hat{S}} \hat{\theta}_j^2 v_j^{-1} \) under both the null and alternative hypotheses.

Define the “grey area set” as

\[ \mathcal{G}(\theta) = \{ j : |\theta_j|/v_j^{1/2} \approx \delta_{N,T}, j = 1, ..., N \}. \]

**Theorem 1.3.1.** Let Assumption 1.3.1 hold. As \( T, N \to \infty \), we have, under \( H_0 : \theta = 0 \),

\[ P(\hat{S} = \emptyset | H_0) \to 1. \]  

Hence

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

In addition,

\[ \inf_{\theta \in \Theta} P(S(\theta) \subset \hat{S}| \theta) \to 1 \quad \text{and} \quad \inf_{\theta \in \Theta} P(\hat{S} \setminus S(\theta) \subset \mathcal{G}(\theta)| \theta) \to 1. \]

Besides the asymptotic behavior of \( J_0 \), Theorem 1.3.1 also provides a “sure screening” property of \( \hat{S} \). Sometimes we wish to find out the identities of the elements in \( S(\theta) \), which
represent the components of $\theta$ that deviate from zero. Therefore, we are particularly interested in a type of alternative hypothesis that satisfies the following \textit{empty grey area} condition.

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

Theorem 1.3.1 shows that the “large” $\theta_j$’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.

\textbf{Corollary 1.3.1.} Under Assumptions 1.3.1, 1.3.2, as $T, N \to \infty$,

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

\textit{Proof.} Corollary 1.3.1 follows immediately from Theorem 1.3.1 and Assumption 1.3.2:

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

\hfill $\square$

\textbf{Remark 1.3.1.} Corollary 1.3.1 and its required assumptions (Assumptions 1.3.1 and 1.3.2) are stated uniformly over $\theta \in \Theta$. The empty grey area condition (Assumption 1.3.2) rules out $\theta$’s that have components on the boundary of the screening set. Intuitively, when a component $\theta_j$ 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:

$$\Theta_s = \{ \theta \in \Theta : \max_{j \leq N} \left| \frac{\theta_j}{v_j^{1/2}} \right| > 2\delta_{N,T} \}. \quad (1.9)$$
In particular, if \( \hat{\theta}_j \) is \( \sqrt{T} \)-consistent, and \( v_j^{1/2} \) is the asymptotic standard deviation of \( \hat{\theta}_j \), then \( \sigma_j = \sqrt{Tv_j} \) is bounded away from both zero and infinity. Using (1.8), we have

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

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 \( \Theta^* \subset \mathbb{R}^N \setminus \{0\} \) if

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

For a given distribution function \( F \), let \( F_q \) denote its \( q \)th 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 \( \{ D : J_1 > F_q \} \) 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 \( \inf_{\theta \in \Theta_s} P(c\sqrt{N} + J_1 > F_q|\theta) \to 1 \), 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 \( T, N \to \infty \)

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

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 \( \Theta_s \). 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 $\{D : J_1 > F_q\}$ and $\{D : J > F_q\}$ 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:

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

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 $V$ is positive definite, whose eigenvalues are bounded away from both zero and infinity. Here $TV$ is often taken to be the inverse of the asymptotic covariance matrix of $\hat{\theta}$. Other popular choices are $V = \text{diag}(\sigma_1^{-2}, \ldots, \sigma_N^{-2})$ with $\sigma_j = \sqrt{T v_j}$, $\text{Bai and Saranadasa [1996]}$, $\text{Chen and Qin [2010]}$, $\text{Pesaran and Yamagata [2012]}$ and $V = I_N$, 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 $V$ to be known, and just focus on the power enhancement properties. We will deal with unknown $V$ 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 $\{D : J_Q > F_q\}$,

(iii) $V$ is positive definite, and there exist two positive constants $C_1$ and $C_2$ such that $C_1 \leq \lambda_{\min}(V) \leq \lambda_{\max}(V) \leq C_2$.

(iv) $C_3 \leq Tv_j \leq C_4$, $j = 1, \ldots, N$ 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 $\Theta_s$ 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 \( H_0 : \theta = 0 \), \( J \to^d F \),

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

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

that is, \( \inf_{\theta \in \Theta_s \cup \Theta(J_Q)} P(J > F_q|\theta) \to 1 \) 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, \( T \hat{\theta}' V \hat{\theta} \) 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 \( TV \) being the inverse of the asymptotic covariance matrix of \( \hat{\theta} \), assumed to exist. Specifically, we assume the standardized \( T \hat{\theta}' V \hat{\theta} \) to be asymptotically normal under \( H_0 \):

\[
\frac{T \hat{\theta}' V \hat{\theta} - N}{\sqrt{2N}} \bigg|_{H_0} \to^d \mathcal{N}(0,1). \quad (1.10)
\]

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

Theorem 1.3.4. Suppose that (1.10) holds with \( \|V\|_1 < C \) and \( \|V^{-1}\|_1 < C \) 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 $\Theta_b$ given by

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

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

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

where $z_q$ is the $q$th 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: $T \sum_{i: \theta_i = 0} \hat{\theta}_i^2$. In contrast, the nonzero components of $\theta$ (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 $f_t = (f_{1t}, ..., f_{Kt})'$ be the observable excess returns of $K$ market risk factors. Then, the excess return has the following decomposition:

$$y_{it} = \theta_i + b_i' f_t + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T,$$
where \( b_i = (b_{i1}, \ldots, b_{iK})' \) is a vector of factor loadings and \( u_{it} \) represents the idiosyncratic error. To make the notation consistent, we pertain to use \( \theta \) 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 \( \theta_i \) should be zero for any asset \( i \) \cite{Ross1976, Merton1973, Chamberlain1983}. An important question is then testing the null hypothesis

\[
H_0 : \theta = 0,
\]

namely, whether the factor pricing model is consistent with empirical data, where \( \theta = (\theta_1, \ldots, \theta_N)' \) is the vector of intercepts for all \( N \) financial assets. One typically picks five-year monthly data, because the factor pricing model is technically a one-period model whose factor loadings can be time-varying; see \cite{Gagliardini2011} 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 \( \theta \), 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 \( \theta \) compared to \( N \).

Most existing tests to the problem (1.11) are based on the quadratic statistic

\[
W = T \hat{\theta}'V\hat{\theta},
\]

where \( \hat{\theta} \) is the OLS estimator for \( \theta \), and \( V \) is some positive definite matrix. Prominent examples are given by \cite{Gibbons1989, MacKinlay1991, Beaulieu2007}. When \( N \) is possibly much larger than \( T \), \cite{Pesaran2012} showed that, under regularity conditions (Assumption 1.4.1 below),

\[
J_1 = \frac{a_{f,T} T \hat{\theta}' \Sigma_{\hat{\theta}}^{-1} \hat{\theta} - N}{\sqrt{2N}} \rightarrow_d N(0,1).
\]
where $a_{f,T} > 0$ is a constant that depends only on factors’ empirical moments, and $\Sigma_u$ is the $N \times N$ covariance matrix of $u_t = (u_{1t}, ..., u_{Nt})'$, 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 $\Sigma_u^{-1}$, which controls the size under the null even when $N/T \to \infty$.

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

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

The OLS estimator of $\theta$ can be expressed as

$$\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_N)', \quad \hat{\theta}_j = \frac{1}{T a_{f,T}} \sum_{t=1}^{T} y_{jt} (1 - f_t' w). \quad (1.12)$$

When $\text{cov}(f_t)$ 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 $\hat{\theta}_j$ (given $\{f_t\}$) converges in probability to

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

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

$$\hat{v}_j = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{jt}^2/(Ta_{f,T})$$

$$\text{where } \hat{u}_{jt} = y_{jt} - \hat{\theta}_j - \hat{b}_j^r f_t.$$ 

We show in Proposition 1.4.1 below that $\max_{j \leq N} |\hat{\theta}_j - \theta_j|/\hat{v}_j^{1/2} = O_P(\sqrt{\log N})$. 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

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

and

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

### 1.4.3 Feasible Wald test in high dimensions

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

$$\frac{Ta_{f,T} \hat{\theta}' \Sigma_u^{-1} \hat{\theta} - N}{\sqrt{2N}}.$$  \hspace{1cm} (1.13)

Under $H_0 : \theta = 0$, it converges in distribution to $N(0, 1)$. Note that the idiosyncratic errors $(u_{1t}, ..., u_{Nt})$ are often cross-sectionally correlated, which leads to a non-diagonal inverse covariance matrix $\Sigma_u^{-1}$. When $N/T \to \infty$, it is practically difficult to estimate $\Sigma_u^{-1}$, as there are $O(N^2)$ free off-diagonal parameters.
To consistently estimate $\Sigma_u^{-1}$ when $N/T \to \infty$, without parametrizing the off-diagonal elements, we assume $\Sigma_u = \text{cov}(u_t)$ 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 $\Sigma_u^{-1}$ via thresholding: let $s_{ij} = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt}$. Define the covariance estimator as

\[
(\hat{\Sigma}_u)_{ij} = \begin{cases} 
  s_{ij}, & \text{if } i = j, \\
  h_{ij}(s_{ij}), & \text{if } i \neq j,
\end{cases}
\]

where $h_{ij}(\cdot)$ is a generalized thresholding function (Antoniadis and Fan, 2001; Rothman et al., 2009), with threshold value $\tau_{ij} = C(s_{ii}s_{jj} \log N)^{1/2}$ for some constant $C > 0$, designed to keep only the sample correlation whose magnitude exceeds $C(\log N)^{1/2}$. 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 $(\frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt})$ 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 $\hat{\Sigma}_u$ is strictly positive definite in finite sample even when $N > T$. 

23
With $\hat{\Sigma}_u^{-1}$, we are ready to define the feasible standardized Wald statistic:

$$J_{\text{wald}} = \frac{Ta_{T,T} \hat{\theta} \hat{\Sigma}_u^{-1} \hat{\theta} - N}{\sqrt{2N}}, \quad (1.14)$$

whose power can be enhanced under sparse alternatives by:

$$J = J_0 + J_{\text{wald}}. \quad (1.15)$$

### 1.4.4 Does the thresholded covariance estimator affect the size?

A natural but technical question to address is that when $\Sigma_u$ indeed admits a sparse structure, is the thresholded estimator $\hat{\Sigma}_u^{-1}$ accurate enough so that the feasible $J_{\text{wald}}$ 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 $\Sigma_u$ is a sparse matrix, under regularity conditions (Assumption 1.4.2 below), Fan et al. (2011) showed that

$$\|\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}\|_2 = O_P(\sqrt{\frac{\log N}{T}}). \quad (1.16)$$

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 $\Sigma_u^{-1}$ in (1.13) by $\hat{\Sigma}_u^{-1}$, one needs to show that the effect of such a replacement is asymptotically negligible, namely, under $H_0$,

$$T \hat{\theta}' (\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}) \hat{\theta} / \sqrt{N} = o_P(1). \quad (1.17)$$

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

$$|T \hat{\theta}' (\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}) \hat{\theta} / \sqrt{N}| = O_P(\sqrt{\frac{N \log N}{T}}).$$
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, $\hat{\theta}' (\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}) \hat{\theta}$ is a weighted estimation error of $\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}$, where the weights $\hat{\theta}$ “average down” the accumulated estimation errors in estimating elements of $\Sigma_u^{-1}$, 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 $\hat{\theta} - \theta$ as required in Assumption 1.3.1, (ii) defining the sparsity of $\Sigma_u$ so that $\hat{\Sigma}_u^{-1}$ is consistent, and (iii) showing (1.17), so that the errors from estimating $\Sigma_u^{-1}$ do not affect the size of the test.

Let $\mathcal{F}_{-\infty}$ and $\mathcal{F}_T^\infty$ denote the $\sigma$-algebras generated by $\{f_t : -\infty \leq t \leq 0\}$ and $\{f_t : T \leq t \leq \infty\}$ respectively. In addition, define the $\alpha$-mixing coefficient

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

**Assumption 1.4.1.** (i) $\{u_t\}_{t \leq T}$ is i.i.d. $\mathcal{N}(0, \Sigma_u)$, where both $\|\Sigma_u\|_1$ and $\|\Sigma_u^{-1}\|_1$ are bounded;

(ii) $\{f_t\}_{t \leq T}$ is strictly stationary, independent of $\{u_t\}_{t \leq T}$, and there are $r_1, b_1 > 0$ so that

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

(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) \( \text{cov}(f_t) \) is positive definite, and \( \max_{i \leq N} \|b_i\| < c_1 \) 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 \( u_t \) is required to be serially uncorrelated across \( t \). Under this condition, the conditional covariance of \( \hat{\theta} \), given the factors, has a simple expression \( \Sigma_u/(Ta_{f,T}) \). On the other hand, if serial correlations are present in \( u_t \), there would be additional autocovariance terms in the covariance matrix, which need to be further estimated via regularizations. Moreover, given that \( \Sigma_u \) is a sparse matrix, the Gaussianity ensures that most of the idiosyncratic errors are cross-sectionally independent so that \( \text{cov}(u^2_{it}, u^l_{jt}) = 0, l = 1, 2 \), for most of the pairs in \( \{(i,j) : i \neq j\} \).

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

**Remark 1.4.2.** The conditional homoskedasticity \( E(u_tu'_t|f_t) = E(u_tu'_t) \) 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 \( E(u_tu'_t|f_t) \) 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 \( \{E(u_tu'_t|f_t)^{-1} : t = 1, 2, \ldots\} \) 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 \( \Sigma_u^{-1} \) to be first-order negligible when \( N/T \to \infty \) requires delicate technical analysis.

To characterize the sparsity of \( \Sigma_u \) in our context, define

\[
m_N = \max_{i \leq N} \sum_{j=1}^{N} \mathbb{1}\{(\Sigma_u)_{ij} \neq 0\}, \quad D_N = \sum_{i \neq j} \mathbb{1}\{(\Sigma_u)_{ij} \neq 0\}.
\]
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) \( \min(\Sigma_{u})_{ij} \neq 0 \mid (\Sigma_{u})_{ij} \gg \sqrt{\log N}/T; \)

(ii) at least one of the following cases holds:

(a) \( D_N = O(N^{1/2}) \), and \( m_N^2 = O\left(\frac{T}{N^{1/2}(\log N)^{\gamma}}\right) \)

(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)), \( \Sigma_{u} \) 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 \( \Sigma_{u} \) 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 \( \Sigma_{u} \).

### 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 \((f_t, u_t)\) is independent of \(\theta\). Under Assumption 1.4.1, for \(\delta_{N,T} = \log(\log T) \sqrt{\log N}\), as \(T, N \to \infty\),

\[
\inf_{\theta \in \Theta} P(\max_{j \leq N} |\hat{\theta}_j - \theta_j| / \hat{v}_j^{1/2} < \delta_{N,T}/2|\theta) \to 1.
\]

\[
\inf_{\theta \in \Theta} P(4/9 < \hat{v}_j/\var{u_{jt}} < 16/9, \forall j = 1, \ldots, N|\theta) \to 1.
\]

Proposition 1.4.2. Under Assumptions 1.3.2, 1.4.1, and \(H_0\),

\[
J_{\text{wald}} = \frac{T a_{f,T} \hat{\Sigma}^{-1} \hat{\theta} - N}{\sqrt{2N}} \to^d N(0,1).
\]

As shown, the effect of replacing \(\Sigma_u^{-1}\) 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 \(J = J_0 + J_{\text{wald}}\) as follows. For \(\delta_{N,T} = \log(\log T) \sqrt{\log N}\), let

\[
\Theta_s = \{\theta \in \Theta : \max_{j \leq N} T^{1/2} |\theta_j| / \var{u_{jt}}^{1/2} > 2a_{f}^{-1/2} \delta_{N,T}\},
\]

\[
\Theta(J_{\text{wald}}) = \{\theta \in \Theta : ||\theta||^2 > C\delta_{N,T}^2 N/T\}.
\]

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

(i) Under the null hypothesis \(H_0 : \theta = 0\), as \(T, N \to \infty\),

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

and hence

\[
J = J_0 + J_{\text{wald}} \to^d N(0,1).
\]

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

\[
\inf_{\theta \in \Theta_s} P(J_0 > \sqrt{N}|\theta) \to 1, \quad \inf_{\theta \in \Theta(J_{\text{wald}})} P(J_{\text{wald}} > q|\theta) \to 1,
\]

28
and hence

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

where $z_q$ denotes the $q$th 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 $\Theta(J_{\text{wald}})$ to $\Theta_s \cup \Theta(J_{\text{wald}})$.

1.5 Application: Testing Cross-Sectional Independence

1.5.1 The model

Consider a mixed effect panel data model

$$y_{it} = \alpha + x_{it}' \theta + \mu_i + u_{it}, \quad i \leq n, t \leq T,$$

where the idiosyncratic error $u_{it}$ is assumed to be Gaussian. The regressor $x_{it}$ could be correlated with the individual random effect $\mu_i$, but is uncorrelated with $u_{it}$. Let $\rho_{ij}$ 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 $\theta = (\rho_{12}, \ldots, \rho_{1n}, \rho_{23}, \ldots, \rho_{2n}, \ldots, \rho_{n-1,n})'$ be an $N \times 1$ vector stacking all the mutual correlations, then the problem is equivalent to testing about a high-dimensional vector $H_0 : \theta = 0$. Note that often the cross-sectional dependences are weakly present. Hence the alternative
hypothesis of interest is often a sparse vector $\theta$, corresponding to a sparse covariance matrix $\Sigma_u$ of $u_{it}$.

Most of the existing tests are based on the quadratic statistic $W = \sum_{i<j} T \hat{\rho}^2_{ij} = T \hat{\theta}' \hat{\theta}$, where $\hat{\rho}_{ij}$ is the sample correlation between $u_{it}$ and $u_{jt}$, estimated by the within-OLS [Baltagi 2008], and $\hat{\theta} = (\hat{\rho}_{12}, ..., \hat{\rho}_{n-1,n})$. Pesaran et al. (2008) and Baltagi et al. (2012) studied the rescaled $W$, and showed that after a proper standardization, the rescaled $W$ is asymptotically normal when both $n, T \to \infty$. However, the quadratic test suffers from a low power if $\Sigma_u$ 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 $\Sigma_u$ 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 $\tilde{y}_{it} = y_{it} - \frac{1}{T} \sum_{t=1}^{T} y_{it}$, $\tilde{x}_{it} = x_{it} - \frac{1}{T} \sum_{t=1}^{T} x_{it}$, and $\tilde{u}_{it} = u_{it} - \frac{1}{T} \sum_{t=1}^{T} u_{it}$. Then $\tilde{y}_{it} = \tilde{x}_{it}' \hat{\beta} + \tilde{u}_{it}$. The within-OLS estimator $\hat{\beta}$ is obtained by regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$, which leads to the estimated residual $\hat{u}_{it} = \tilde{y}_{it} - \tilde{x}_{it}' \hat{\beta}$. Then $\rho_{ij}$ is estimated by

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

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

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

where $\delta_{N,T} = \log(\log T) \sqrt{\log N}$ as before. The set $\hat{S}$ screens off most of the estimation errors.
To control the size, we employ Baltagi et al. (2012)'s bias-corrected quadratic statistic:

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

(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

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

(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 \( \hat{S} \). 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 \( \theta = (\rho_{12}, ..., \rho_{1n}, \rho_{23}, ..., \rho_{2n}, ..., \rho_{n-1,n})' \) is required to be: \( \theta \) is element-wise bounded away from \( \pm 1 \): there is \( \rho_{\text{max}} \in (0,1) \),

\[ \Theta = \{ \theta \in \mathbb{R}^N : \|\theta\|_{\text{max}} \leq \rho_{\text{max}} \}. \]

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

Assumption 1.5.1. There are \( C_1, C_2 > 0 \), so that

(i) \( \sup_{\theta \in \Theta} \sum_{i \neq j \leq n} |E\tilde{x}_{it} \tilde{x}_{jt} E(u_{it} u_{jt} | \theta)| < C_1 n, \)

(ii) \( \sup_{\theta \in \Theta} \max_{j \leq n} E(u_{jt}^4 | \theta) < C_1, \inf_{\theta \in \Theta} \min_{j \leq n} E(u_{jt}^2 | \theta) > C_2, \)

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 \( \theta \in \Theta \), 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) \( \{u_t\}_{t \leq T} \) are i.i.d. \( N(0, \Sigma_u) \), \( E(u_t|\{f_t\}_{t \leq T}, \theta) = 0 \) almost surely.

(ii) With probability approaching one, all the eigenvalues of \( \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{jt} \tilde{x}_{jt}' \) 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 \( T, N \rightarrow \infty \),

\[
\inf_{\theta \in \Theta} P(\max_{ij} |\hat{\rho}_{ij} - \rho_{ij}|/\hat{\sigma}_{ij}^{1/2} < \delta_{N,T}/2|\theta) \rightarrow 1
\]

\[
\inf_{\theta \in \Theta} P(4/9 < \hat{\sigma}_{ij}^{1/2}/\sigma_{ij} < 16/9, \forall i \neq j|\theta) \rightarrow 1.
\]

Define

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

For \( J_1 \) defined in (1.19), let

\[
J = J_0 + J_1.
\]

The main result is presented as follows.

**Theorem 1.5.1.** Suppose Assumptions 1.3.2, 1.5.1, 1.5.2 hold. As \( T, N \rightarrow \infty \),

(i) under the null hypothesis \( H_0 : \theta = 0 \),

\[
P(J_0 = 0|H_0) \rightarrow 0, \quad J_1 \rightarrow^d \mathcal{N}(0,1),
\]

and hence

\[
J = J_0 + J_1 \rightarrow^d \mathcal{N}(0,1);
\]

32
(ii) there is $C > 0$ in the definition of $\Theta(J_1)$ so that for any $q \in (0, 1)$,

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

and hence

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

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 $\Theta_s$ in (1.20) is mild: the maximum cross-sectional correlation is only required to exceed a magnitude of $\log(\log T)\sqrt{(\log N)/T}$. 

33
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, 2002a, 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 [2002a, 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, a 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:

$$y_{t+1} = h(\phi'_1 f_t, \ldots, \phi'_L f_t, \epsilon_{t+1}), \quad (2.1)$$

$$x_{it} = b'_i f_t + u_{it}, \quad 1 \leq i \leq p, \quad 1 \leq t \leq T, \quad (2.2)$$

where $x_{it}$ is the $i$-th predictor observed at time $t$, $b_i$ is a vector of factor loadings, $f_t$ 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 $f_t = (f_1, \ldots, f_K)'$ through $L (\leq K)$ linear combinations $\phi_1, \ldots, \phi_L$, which are orthogonal unit $K \times 1$ vectors.

In (2.1), we assume that $h(\cdot)$ is an unknown link function and $\epsilon_{t+1}$ is some stochastic error independent of $f_t$ 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

$$y_{t+1} = \phi'_1 f_t + \epsilon_{t+1}.$$ 

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 $f_t$ that is target-relevant, i.e., $\phi'_1 f_t, \ldots, \phi'_L f_t$. 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

\[
    y_{t+1} = f_{(K-1)t} + f_{(K-1)t}f_{Kt} + \epsilon_{t+1},
\]

\[
    x_{it} = b_i'f_t + u_{it}, \quad 1 \leq i \leq p, \quad 1 \leq t \leq T,
\]

where \(\{\lambda_k = (b_{1k}, ..., b_{pk})'\}_{k=1}^{K}\) are orthogonal unit vectors and \(\{f_{kt}\}_{k=1}^{K}\) are uncorrelated (i.e. \(\text{cov}(f_t)\) 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 \(k \geq 2\). The covariance matrix of \(x_t = (x_{1t}, ..., x_{pt})'\) can be decomposed as

\[
    \text{cov}(x_t) = \sum_{k=1}^{K} \text{var}(f_{kt})\lambda_k\lambda_k' + \text{cov}(u_t).
\]

When \(\text{cov}(u_t)\) 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 \(\Sigma_x\). Writing \(x_t = (x_{1t}, ..., x_{pt})', B = (b_1, ..., b_p)'\) and \(u_t = (u_{1t}, ..., u_{pt})'\), we have

\[
    \Sigma_x = B\text{cov}(f_t)B' + \Sigma_u,
\]

(2.3)
where $\Sigma_u$ is the covariance matrix of $u_t$ 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

$$\text{cov}(f_t) = I_K \text{ and } B'B \text{ is diagonal}, \quad (2.4)$$

which serves as an identifiability condition because $Bf_t = B\Omega\Omega^{-1}f_t$ holds for any nonsingular matrix $\Omega$. Also assume for simplicity that $x_{it}$’s and $f_t$’s in (2.2) have already been de-meaned. If the common factors $f_t$ 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 $\phi_i$’s in the link function form the central subspace $S_{y/f}$ (Cook 2009), given which $y_{t+1}$ is independent of $f_t$. Li (1991) developed a sliced inverse regression (SIR) method to effectively estimate these SDR directions. Under model (2.1), Li (1991) showed that if $E(b'f_t|\phi_1'f_t,...,\phi_L'f_t)$ is a linear function of $\phi_1'f_t,...,\phi_L'f_t$ for any $b \in \mathbb{R}^p$, $E(f_t|y_{t+1})$ is contained in $S_{y/f}$. Thus, SDR directions can be obtained by the eigenvectors corresponding to the $L$ largest eigenvalues of $\text{cov}(E(f_t|y_{t+1}))$, which we denote by $\Sigma_{f|y}$.

Since the factors $f_t$ 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 \( \text{cov}(E(f_t | y_{t+1})) \), which leads to

\[
\Sigma^1_{f|y} := \text{cov}(E(\hat{f}_t | y_{t+1})),
\]  

(2.5)

where \( \hat{f}_t \) is some consistent estimator for \( f_t \).

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

\[
\text{cov}(E(x_t | y_{t+1})) = B \text{cov}(E(f_t | y_{t+1})) B'.
\]

Here, the error covariance matrix \( \Sigma_u \) in \([2.3]\) disappears since \( u_t \) and \( y_{t+1} \) are independent. This allows a flexible structure on \( \Sigma_u \), and includes the approximate factor model \([\text{Chamberlain and Rothschild, 1983}]\) as a special case. As is well-known that principal component analysis is not scale-invariant, we cannot directly deal with \( \text{cov}(E(x_t | y_{t+1})) \). Letting \( \Lambda_b = (B'B)^{-1}B' \) be a \( K \times p \) matrix, the following linear transformation connects \( x_t \) to the factors \( f_t \),

\[
\text{cov}(E(\Lambda_b x_t | y_{t+1})) = \text{cov}(E(f_t | y_{t+1})).
\]

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

\[
\Sigma^2_{f|y} := \hat{\Lambda}_b \text{cov}(E(x_t | y_{t+1})) \hat{\Lambda}_b'.
\]  

(2.6)

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

\[
\max_{b,T} \text{Corr}^2(T(y_{t+1}), b'f_t),
\]
where the maximum is taken over any transformation $T(\cdot)$ and $b \in \mathbb{R}^p$. This further justifies the need in considering $\text{cov}(E(f_t|y_{t+1}))$ and the corresponding SDR directions $\phi_j$'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 $E(y_{t+1}|f_t)$ 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

\[
\begin{align*}
\arg \min_{B,F} & \quad ||X - BF'||_F^2, \\
\text{subject to} & \quad T^{-1}F'F = I_K, \quad B'B \text{ is diagonal},
\end{align*}
\]

(2.7)

(2.8)

where $X = (x_1, \ldots, x_T)$ and $F' = (f_1, \ldots, f_T)$. This is a classical principal components problem, and has been used by many researchers to extract underlying factors (Stock and Watson, 2002a; Fan et al., 2013). The constraints (2.8) correspond to the normalization (2.4). The
minimizers $\hat{F}_K, \hat{B}_K$ are such that the columns of $\hat{F}_K / \sqrt{T}$ are the eigenvectors corresponding to the $K$ largest eigenvalues of the $T \times T$ matrix $X'X$ and $\hat{B}_K = T^{-1}X\hat{F}_K$.

To fully estimate $\text{cov}(E(f_t|y_{t+1}))$, we follow the sliced inverse regression scheme in Li (1991), replacing the expectation and covariance by their sample counterparts. Denote the order statistics of $\{(y_{t+1}, \hat{f}_t)\}_{t=1}^{T-1}$ by $\{(y_{(t+1)}, \hat{f}_{(t)})\}_{t=1}^{T-1}$ 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

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

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

$$\hat{\Sigma}^1_{f|y} = \frac{1}{H} \sum_{h=1}^{H} \left[ \frac{1}{c} \sum_{l=1}^{c} \hat{f}_{(h,l)} \right] \left[ \frac{1}{c} \sum_{l=1}^{c} \hat{f}_{(h,l)} \right]' . \quad (2.9)$$

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 $\Sigma^2_{f|y}$, we have

$$\hat{\Sigma}^2_{f|y} = \hat{\Lambda}_b \left( \frac{1}{H} \sum_{h=1}^{H} \left[ \frac{1}{c} \sum_{l=1}^{c} x_{(h,l)} \right] \left[ \frac{1}{c} \sum_{l=1}^{c} x_{(h,l)} \right]' \right) \hat{\Lambda}_b' . \quad (2.10)$$

The following proposition shows that the estimates of $\text{cov}(E(f_t|y_{t+1}))$ 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 $\hat{f}_t$ and $\hat{B}$ be estimated from the method of principal components. $\hat{\Lambda}_b$ is obtained by substitution. Then, the two estimators (2.9) and (2.10) for $\text{cov}(E(f_t|y_{t+1}))$ are equivalent, i.e.

$$\hat{\Sigma}^1_{f|y} = \hat{\Sigma}^2_{f|y}.$$
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 $\hat{\Sigma}_{f|y}$. We shall show that under mild conditions, $\hat{\Sigma}_{f|y}$ consistently estimate $\Sigma_{f|y}$ as $p, T \to \infty$. As a result, the eigenvectors of $\hat{\Sigma}_{f|y}$, denoted as $\hat{\phi}_j (j = 1, ..., K)$, converge to the corresponding eigenvectors of $\Sigma_{f|y}$, which span the central space discussed before. This will yield consistent estimates of sufficient predictive indices $\phi'_i f_t$, 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 $X'X$ arranged in descending order, i.e.

$$\hat{K} = \arg \max_{1 \leq i \leq k_{\text{max}}} \hat{\lambda}_i / \hat{\lambda}_{i+1},$$

where $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_T$ are the eigenvalues. The estimator enjoys good finite-sample performances and was motivated by the following observation: the $K$ largest eigenvalues of $X'X$ 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 \( \{ (x_t, y_t) \}_{t=1}^T \) are observable.

Assumption 2.3.1 (Factors and Loadings). (1) \( \| b_i \| \leq M \) for some \( M > 0 \) \( (i = 1, \ldots, p) \).

And as \( p \to \infty \), there exists positive constants \( c_1 \) and \( c_2 \) such that

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

(2) \( E(\| f_t \|^4) < \infty \), and \( T^{-1}F'F \to \text{cov}(f_t) = I \) as \( T \to \infty \).

(3) Linearity: \( E(b'f_t|\phi'_1 f_t, \ldots, \phi'_L f_t) \) is a linear function of \( \phi'_1 f_t, \ldots, \phi'_L f_t \) for any \( b \in \mathbb{R}^p \), where \( \phi_i \)'s come from model (2.1).

Condition (1) is often known as the pervasive condition \( \text{Bai and Ng, 2002; Fan et al., 2013} \) 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 \( E(f_t|y_{t+1}) \) is contained in the central space, and is satisfied when the distribution of \( f_t \) is elliptically symmetric \( \text{Hall and Li, 1993} \). If the distribution of \( f_t \) is non-elliptically distributed, we can follow \( \text{Li and Dong, 2009} \) to greatly relax the linearity condition in Assumption 2.3.1 and assume that \( E(f_t|\phi'_1 f_t, \ldots, \phi'_L f_t) \) is a polynomial function of \( \phi'_1 f_t, \ldots, \phi'_L f_t \), where \( \phi_i \)'s come from model (2.1).

We impose the strong mixing condition on the data generating process. Let \( \mathcal{F}_\infty^t \) and \( \mathcal{F}_T^\infty \) denote the \( \sigma \)-algebras generated by \( \{(f_t, y_{t+1}) : t \leq 0\} \) and \( \{(f_t, y_{t+1}) : t \geq T\} \) 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). \( \{f_t, u_t, \epsilon_{t+1}\}_{t \geq 1} \) is strictly stationary, \( E(||f_t||^2|y_{t+1}) < \infty \) 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 \( f_t|y_{t+1} \) 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. \( E(u_t) = 0 \), and \( E|u_{it}|^8 \leq M \).
2. \( ||\Sigma_u||_1 \leq M \), 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 \( (t, s) \), \( E|p^{-1/2} \sum_{i=1}^p (u'_s u_t - E(u_s u_t))|^4 \leq M \).
4. Weak dependence between factors and idiosyncratic errors

\[
E \left( \frac{1}{p} \sum_{i=1}^p \left| \frac{1}{\sqrt{T}} \sum_{t=1}^T f_t u_{it} \right|^2 \right) \leq M
\]

2.3.2 Convergence of \( \hat{\Sigma}_{f|y} \)

As with many other applications, the factor structure (2.2) often involves estimation of the unknown factors. Note that the factor loadings \( B \) and the common factors \( f_t \) are not separably identifiable. Let \( V \) denote the \( K \times K \) diagonal matrix of the first \( K \) largest eigenvalues of the sample covariance matrix \( T^{-1}X'X \) in descending order. Define a \( K \times K \) matrix \( H = (1/T)V^{-1}\hat{F}'FB'B \), where \( F' = (f_1, ..., f_T) \). Since \( Hf_t = (1/T)V^{-1}\hat{F}'(BF')Bf_t \) depends only on an identifiable part \( (BF')Bf_t \) and the data \( V^{-1}F' \), \( H \) eliminates identifiability issues when estimating \( (B, f_t) \) simultaneously.

The following theorem gives the rate of convergence of the estimated covariance of inverse regression curve \( \hat{\Sigma}_{f|y} \).
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

\[
\| \hat{\Sigma}_{f|y} - \Sigma_{f|y} \Sigma_{f|y}' \| = O_p(\omega_{p,T}).
\] (2.11)

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

\[
\| \hat{\phi}_j - \mathbf{H}\phi_j \| = O_p(\omega_{p,T})
\] (2.12)

for any \( j \leq L \).

The proof of this theorem relies on the fact that \( \hat{\Lambda}_b \) 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 \( \hat{\phi}_j f_t \rightarrow^p \phi_j f_t \) for any \( j \). The sufficient predictive indices can therefore be consistently estimated. Define \( \tilde{\xi}_j := \hat{\Lambda}_b' \hat{\phi}_j \), then similarly \( \hat{\xi}_j' x_t \rightarrow^p \phi_j f_t \). This supplies linear combinations of observed data \( x_t \) 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 $\mathbf{f}_t$.

$$\hat{\phi} = \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1} \hat{f}_t, \quad (2.13)$$

where $\hat{f}_t$’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 $h(\cdot)$, $\hat{\phi}$ falls into the central space spanned by $\phi_1, \ldots, \phi_L$ 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,

$$||\hat{\phi} - \bar{\phi}|| = O_p(\omega_{p,T}), \quad (2.14)$$

where $\bar{\phi} = \sum_{i=1}^{L} E((\phi_i^t \mathbf{f}_t) y_{t+1}) \phi_i$.

It is interesting to see that when $L = 1$, the coefficient $\hat{\phi}$ 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 $\phi_i$’s, and its coordinates depend on the correlation between the target and the sufficient predictive indices. This subspace, however, is entirely contained in $\Sigma_{f|y}$. By estimating $\Sigma_{f|y}$ 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

\[
y_{t+1} = \phi_1' f_t + \epsilon_{t+1}, \quad t \leq T.
\]

By Theorem 2.3.1 and Proposition 2.3.1, the eigenvector corresponding to the largest eigenvalue of \( \hat{\Sigma}_{f|y} \) 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

\[ y_{t+1} = (\phi'_1 f_t)(\phi'_2 f_t) + \epsilon_{t+1}, \]

where the interaction terms are formed by the two directions \( \phi_{1,2} \), 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 \( \phi_{1,2} \). 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:

\[ y_{it} = \theta_i + b_i f_t + u_{it}. \]
We simulate \( \{b_i\}_{i=1}^N, \{f_t\}_{t=1}^T \) and \( \{u_t\}_{t=1}^T \) independently from \( N_3(\mu_B, \Sigma_B), N_3(\mu_f, \Sigma_f) \), and \( N_N(0, \Sigma_u) \) 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 1st, 2008 to June 29th, 2012. These parameters are listed in the following table.

**Table 3.1: Means and covariances used to generate \( b_i \) and \( f_t \)**

<table>
<thead>
<tr>
<th>( \mu_B )</th>
<th>( \Sigma_B )</th>
<th>( \mu_f )</th>
<th>( \Sigma_f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9833</td>
<td>-0.0178</td>
<td>0.0436</td>
<td>0.0260</td>
</tr>
<tr>
<td>-0.1233</td>
<td>0.0862</td>
<td>-0.0211</td>
<td>0.0211</td>
</tr>
<tr>
<td>0.0839</td>
<td>-0.0211</td>
<td>0.7624</td>
<td>-0.0043</td>
</tr>
</tbody>
</table>

Set \( \Sigma_u = \text{diag}\{A_1, ..., A_{N/4}\} \) to be a block-diagonal covariance matrix. Each diagonal block \( A_j \) is a \( 4 \times 4 \) positive definite matrix, whose correlation matrix has equi-off-diagonal entry \( \rho_j \), generated from Uniform\([0, 0.5]\). The diagonal entries of \( A_j \) are obtained via \( (\Sigma_u)_{ii} = 1 + \|v_i\|^2 \), where \( v_i \) is generated independently from \( N_3(0, 0.01I_3) \).

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

- **sparse alternative** \( H^1_a \): \( \theta_i = \begin{cases} 0.3, & i \leq \frac{N}{T} \\ 0, & i > \frac{N}{T} \end{cases} \)
- **weak theta** \( H^2_a \): \( \theta_i = \begin{cases} \sqrt{\frac{\log N}{T}}, & i \leq N^{0.4} \\ 0, & i > N^{0.4} \end{cases} \)

Under \( H^1_a \), there are only a few nonzero \( \theta \)’s with a relative large magnitude. Under \( H^2_a \), there are many non-vanishing \( \theta \)’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 \( H^1_a \), \( P(\hat{S} = \emptyset) \) is close to zero because most of the first \( N/T \) estimated \( \theta \)’s should survive from the screening step. These survived \( \hat{\theta} \)’s contribute importantly to the rejection of the null hypothesis. In contrast, \( P(\hat{S} = \emptyset) \) should be much larger under \( H^2_a \) because the non-vanishing \( \theta \)’s are too weak to be detected.
For each test, we calculate the relative frequency of rejection under $H_0$, $H^1_a$ and $H^2_a$ based on 2000 replications, with significance level $q = 0.05$. We also calculate the relative frequency of $\hat{S}$ being empty, which approximates $P(\hat{S} = \emptyset)$. 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

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

Notes: This table reports the frequencies of rejection and $\hat{S} = \emptyset$ based on 2000 replications. Here $J_{\text{wald}}$ 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 $J_{\text{wald}}$ as well as those of the power enhanced test $J = J_0 + J_{\text{wald}}$. First of all, the size of $J_{\text{wald}}$ is close to the significance level. Under $H_0$, $P(\hat{S} = \emptyset)$ 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 $J_{\text{wald}}$. Under $H^1_a$, the PE test significantly improves the power of the standardized Wald-test. In this case, $P(\hat{S} = \emptyset)$ is nearly zero because the screening set manages to capture the big thetas. Under $H^2_a$, as the non-vanishing thetas are very week, it follows that $\hat{S}$ 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,

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

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

Note that we model $\{x_i\}$’s as AR(1) processes, so that $x_{it}$ is possibly correlated with $\mu_i$, 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: $\mu_i$ is drawn from $N(0,0.25)$ for $i = 1,\ldots,n$. The parameters $\alpha$ and $\beta$ are set $-1$ and $2$ respectively. In regression (3.2), $\xi = 0.7$ and $\varepsilon_{it} \sim N(0,1)$.

We generate $\{u_t\}_{t=1}^T$ from $N_n(0,\Sigma_u)$. Under the null hypothesis, $\Sigma_u$ is set to be a diagonal matrix $\Sigma_{u,0} = \text{diag}\{\sigma_1^2,\ldots,\sigma_n^2\}$. Following Baltagi et al. (2012), consider the heteroscedastic errors

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

with $\kappa = 0.5$, where $\bar{x}_i$ is the average of $x_{it}$ across $t$. Here $\sigma^2$ is scaled to fix the average of $\sigma_i^2$’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 $\Sigma_{u,1} = \text{diag}\{\Sigma_1,\ldots,\Sigma_{n/4}\}$ as a block-diagonal matrix with $4 \times 4$ blocks located along the main diagonal. Each $\Sigma_i$ is assumed to be $I_4$ initially. We then randomly choose $\lfloor n^{0.3} \rfloor$ blocks among them and make them non-diagonal by setting $\Sigma_i(m,n) = \rho^{|m-n|}(m,n \leq 4)$, with $\rho = 0.2$. To allow for error
cross-sectional heteroscedasticity, we set $\Sigma_u = \Sigma_u^{1/2} \Sigma_u^{1/2}$, where $\Sigma_u = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ 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 $\hat{S} = \emptyset$ as in (1.18) are recorded.

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

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>$T$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
<th>$n = 600$</th>
<th>$n = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$J_1/\text{PE} / P(\hat{S} = \emptyset)$</td>
<td>$J_1/\text{PE} / P(\hat{S} = \emptyset)$</td>
<td>$J_1/\text{PE} / P(\hat{S} = \emptyset)$</td>
<td>$J_1/\text{PE} / P(\hat{S} = \emptyset)$</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>4.7/5.5 /99.1</td>
<td>4.9/5.3 /99.6</td>
<td>5.5/5.7 /99.7</td>
<td>4.9/5.2 /99.7</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>5.3/5.3 /100.0</td>
<td>5.5/5.9 /99.6</td>
<td>4.7/5.1 /99.4</td>
<td>4.9/5.1 /99.8</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>5.2/5.2 /100.0</td>
<td>5.2/5.2 /100.0</td>
<td>4.6/4.6 /100.0</td>
<td>4.9/4.9 /100.0</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>4.7/4.7 /100.0</td>
<td>5.5/5.5 /100.0</td>
<td>5.0/5.0 /100.0</td>
<td>5.1/5.1 /100.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$H_a$</th>
<th>$T$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
<th>$n = 600$</th>
<th>$n = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
<td>26.4/95.5 /5.0</td>
<td>19.8/98.0 /2.3</td>
<td>13.5/98.2 /2.0</td>
<td>12.2/99.2 /0.9</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>54.6/98.8 /1.6</td>
<td>40.3/99.6 /0.5</td>
<td>24.8/99.6 /0.4</td>
<td>21/99.7 /0.3</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>78.9/99.25 /1.1</td>
<td>65.3/100.0 /0.1</td>
<td>41.7/99.9 /0.2</td>
<td>37.2/100.0 /0.1</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>93.5/99.85 /0.2</td>
<td>89.0/100.0 /0.0</td>
<td>69.1/100.0 /0.0</td>
<td>61.8/100.0 /0.0</td>
</tr>
</tbody>
</table>

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 $\hat{S}$ 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 $\Sigma_u$ 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 $\theta = 0$ (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 $\tau = 1984.12, \ldots, 2012.12$, we run the regressions

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

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


Table 3.4: Summary of descriptive statistics and testing results

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Std dev.</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_T$</td>
<td>617.70</td>
<td>26.31</td>
<td>621</td>
<td>574</td>
<td>665</td>
</tr>
<tr>
<td>$</td>
<td>\bar{S}</td>
<td>_0$</td>
<td>5.20</td>
<td>3.50</td>
<td>5</td>
</tr>
<tr>
<td>$\hat{\theta}_i^\tau$ (%)</td>
<td>0.9767</td>
<td>0.1519</td>
<td>0.9308</td>
<td>0.7835</td>
<td>1.3816</td>
</tr>
<tr>
<td>$\hat{\theta}_{1\in \bar{S}}^\tau$ (%)</td>
<td>4.5569</td>
<td>1.4305</td>
<td>4.1549</td>
<td>1.7839</td>
<td>10.8393</td>
</tr>
<tr>
<td>p-value of $J_{\text{wald}}$</td>
<td>0.2351</td>
<td>0.2907</td>
<td>0.0853</td>
<td>0</td>
<td>0.9992</td>
</tr>
<tr>
<td>p-value of $J_{\text{PE}}$</td>
<td>0.1148</td>
<td>0.2164</td>
<td>0.0050</td>
<td>0</td>
<td>0.9982</td>
</tr>
</tbody>
</table>

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 $\hat{S}$, which directly implies the presence of sparse alternatives. The threshold $\delta_{N,T} = \sqrt{\log(N)} \log(\log(T))$ varies as the panel size $N$ changes at the end of each month, and is about 3.5 on average, a high-criticism thresholding. The selected stocks have much larger alphas ($\theta$) than other stocks do. In addition, 64.05% of all the estimated alphas are positive, whereas 87.33% of the selected alphas in $\hat{S}$ 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 $J_{\text{wald}}$ are generally smaller than those of the power enhanced test $J_{\text{PE}}$.

Similar to Pesaran and Yamagata (2012), we plot the running $p$-values of $J_{\text{wald}}$ and the PE test from December 1984 to December 2012. We also add the dynamics of the percentage of selected stocks ($|\hat{S}|_0/N$) 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 $J_{\text{wald}}$ 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 $\theta = 0$ 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 cross-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

$$y_{t+1} = \phi' f_t + \sigma_y \epsilon_{t+1},$$

$$x_{it} = b'_i f_t + u_{it},$$

where we let $K = 5$ and $\phi = (0.8, 0.5, 0.3, 0, 0)'$. 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 $\alpha_j, \rho_i$ from $\sim U[0.2, 0.8]$ and fix them during simulations, while the shocks $e_{jt}, \nu_{it}$ and $\epsilon_{t+1}$ are standard normal respectively. $\sigma_y$ is adjusted to equal the variance of the factors, so that the infeasible best forecast when knowing $\phi'f_t$ 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>
<thead>
<tr>
<th>$p$</th>
<th>$T$</th>
<th>In-sample</th>
<th>Out-of-sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SIR</td>
<td>PCR</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>46.9</td>
<td>47.7</td>
</tr>
<tr>
<td>50</td>
<td>200</td>
<td>46.3</td>
<td>46.5</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>49.3</td>
<td>50.1</td>
</tr>
<tr>
<td>100</td>
<td>500</td>
<td>47.8</td>
<td>47.8</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>48.5</td>
<td>48.8</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>48.2</td>
<td>48.3</td>
</tr>
</tbody>
</table>

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 \( \epsilon_{t+1} \) 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 \( \hat{\phi}_{1,2} \) and the central subspace \( S_{f|y} \) spanned by \( \phi_1 = (1, 0, \ldots, 0) \) and \( \phi_2 = (0, 1, 1, \ldots, 0) \), we first rotate \( \hat{\phi}_i \) by left multiplying \( H' \) as in theorem 2.3.1 to obtain consistent estimators of \( \phi_{1,2} \). Following Li (1991), we use the squared multiple correlation coefficient \( R^2(\hat{\phi}_j) \) as an affine invariant criterion for each \( j \), where

\[
R^2(\hat{\phi}_j) = \max_{\phi \in S_{f|y}} \frac{[\hat{\phi}_j' \Sigma_f \phi]^2}{(H' \hat{\phi}_j)' \Sigma_f (H' \hat{\phi}_j) \cdot (\phi' \Sigma_f \phi)}.
\]

Note that we have the convenience \( \Sigma_f = I \) 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 \( R^2(\hat{\phi}_j) \). When \( T \) gets larger from 100 to 500, so are the \( R^2(\hat{\phi}_j) \)'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 \( R^2(\hat{\phi}_j) \). 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 \( \hat{\phi}_1' \hat{f}_1, \hat{\phi}_2' \hat{f}_2 \) and \( (\hat{\phi}_1' \hat{f}_1) \cdot (\hat{\phi}_2' \hat{f}_2) \) 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)

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

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)](http://example.com), 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](http://example.com) [Ludvigson and Ng, 2009](http://example.com)). 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 $t + 1$. 

62
### Table 3.7: Out-of-sample Macroeconomic Forecasting

<table>
<thead>
<tr>
<th>Category</th>
<th>Label</th>
<th>SIR</th>
<th>SIR(2)</th>
<th>PCR-ER</th>
<th>PC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP components</td>
<td>GDP261</td>
<td>7.4</td>
<td>8.6</td>
<td>2.6</td>
<td>-2.5</td>
</tr>
<tr>
<td>IP</td>
<td>IPS13</td>
<td>21.2</td>
<td>31.8</td>
<td>24.4</td>
<td>13.3</td>
</tr>
<tr>
<td>Employment</td>
<td>CES033</td>
<td>38.8</td>
<td>27.5</td>
<td>38.6</td>
<td>40.4</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>LHU15</td>
<td>28.7</td>
<td>33.0</td>
<td>30.1</td>
<td>23.7</td>
</tr>
<tr>
<td>Housing</td>
<td>HSNE</td>
<td>33.6</td>
<td>28.8</td>
<td>31.8</td>
<td>30.5</td>
</tr>
<tr>
<td>Inventories</td>
<td>PMDEL</td>
<td>27.4</td>
<td>20.1</td>
<td>16.8</td>
<td>17.1</td>
</tr>
<tr>
<td>Prices</td>
<td>GDP276,3</td>
<td>8.3</td>
<td>10.9</td>
<td>8.2</td>
<td>4.9</td>
</tr>
<tr>
<td>Wages</td>
<td>CES278.R</td>
<td>22.0</td>
<td>15.4</td>
<td>20.3</td>
<td>19.6</td>
</tr>
<tr>
<td>Interest rates</td>
<td>FYFF</td>
<td>8.0</td>
<td>13.9</td>
<td>10.4</td>
<td>10.6</td>
</tr>
<tr>
<td>Money</td>
<td>CCINRV</td>
<td>3.1</td>
<td>11.8</td>
<td>0.5</td>
<td>-0.1</td>
</tr>
<tr>
<td>Exchange rates</td>
<td>EXRCAN</td>
<td>1.4</td>
<td>3.5</td>
<td>1.3</td>
<td>-1.8</td>
</tr>
<tr>
<td>Stock prices</td>
<td>FSPCOM</td>
<td>19.0</td>
<td>15.5</td>
<td>16.0</td>
<td>15.0</td>
</tr>
<tr>
<td>Consumer expectations</td>
<td>HHSNTN</td>
<td>6.1</td>
<td>8.2</td>
<td>8.2</td>
<td>7.6</td>
</tr>
</tbody>
</table>

**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 $\hat{\Sigma}_{f|y}$ 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 $\hat{\Sigma}_{f|y}$, the estimated regression surface and the running out-of-sample $R^2$'s. As can be 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 $\hat{\Sigma}_{f|y}$. 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 \( H : \theta = 0 \) 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 different places.

5.1.1 Proofs for Section 1.3

Proof of Theorem 1.3.1

Proof. Define events

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

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

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

This implies that $j \in \hat{S}$, hence $S(\theta) \subset \hat{S}$. If $j \in \hat{S}$, by similar arguments, we have $\frac{|\theta_j|}{\hat{v}_j^{1/2}} > \delta_{N,T}/3$ on $A_1 \cap A_2$. Hence $\hat{S} \setminus S(\theta) \subset \{ j : \delta_{N,T}/3 < \frac{|\theta_j|}{v_j^{1/2}} < 2\delta_{N,T} \} \subset G(\theta).$ In fact, we have
proved that $S(\theta) \subset \hat{S}$ and $\hat{S} \setminus S(\theta) \subset G(\theta)$ on the event $A_1 \cap A_2$ uniformly for $\theta \in \Theta$. This yields
\[
\inf_{\theta \in \Theta} P(S(\theta) \subset \hat{S}|\theta) \to 1, \quad \text{and} \quad \inf_{\theta \in \Theta} P(\hat{S} \setminus S(\theta) \subset G(\theta)) \to 1.
\]
Moreover, it is readily seen that, under $H_0 : \theta = 0$, by Assumption 1.3.1
\[
P(J_0 = 0|H_0) \geq P(\hat{S} = \emptyset|H_0) = P(\max_{j \leq N} \{|\hat{\theta}_j|/\hat{\nu}_j^{1/2}\} < \delta_{N,T}|H_0) \to 1.
\]
In addition, $\inf_{\theta \in \Theta} P(J_0 > \sqrt{N}|S(\theta) \neq \emptyset)$ is bounded from below by
\[
\inf_{\theta \in \Theta} P(\sqrt{N} \sum_{j \in \hat{S}} \delta_{N,T}^2 > \sqrt{N}|S(\theta) \neq \emptyset) \geq \inf_{\theta \in \Theta} P(\sqrt{N} \delta_{N,T}^2 > \sqrt{N}|S(\theta) \neq \emptyset) - o(1) \to 1.
\]
Note that the last convergence holds uniformly in $\theta \in \Theta$ 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 $\{D : J > F_q\}$ has size $q$. Moreover, by the power condition of $J_1$ and $J_0 \geq 0$,
\[
\inf_{\theta \in \Theta(J_1)} P(J > F_q|\theta) \geq \inf_{\theta \in \Theta(J_1)} P(J_1 > F_q|\theta) \to 1.
\]
This together with the fact
\[
\inf_{\theta \in \Theta_s \cup \Theta(J_1)} P(J > F_q|\theta) \geq \min\{ \inf_{\theta \in \Theta_s} P(J > F_q|\theta), \inf_{\theta \in \Theta(J_1)} P(J > F_q|\theta) \},
\]
establish the theorem, if we show $\inf_{\theta \in \Theta_s} P(J > F_q|\theta) \to 1$. 

68
By the definition of $\hat{S}$ and $J_0$, we have $\{J_0 < \sqrt{N}\delta^2_{N,T}\} = \{\hat{S} = \emptyset\}$. Since $\inf_{\theta \in \Theta} P(S(\theta) \subset \hat{S}|\theta) \to 1$ and $\Theta_s = \{\theta \in \Theta : S(\theta) \neq \emptyset\}$, we have

$$\sup_{\theta \in \Theta_s} P(J_0 < \sqrt{N}\delta^2_{N,T}|\theta) = \sup_{\theta \in \Theta_s} P(\hat{S} = \emptyset|\theta) \leq \sup_{\theta \in \Theta_s} P(\hat{S} = \emptyset, S(\theta) \subset \hat{S}|\theta) + o(1),$$

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

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

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

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

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 $D = \text{diag}(v_1, \ldots, v_N)$. Then $\|D\|_2 < C_3/T$ by Assumption 1.3.3 iv). On the event $A = \{\|[(\hat{\theta} - \theta)'D^{-1/2}]^2 < \delta^2_{N,T,N}/4\}$, we have

$$|[(\hat{\theta} - \theta)'V\theta| \leq \|[(\hat{\theta} - \theta)'D^{-1/2}]\|\|D^{1/2}V\theta\|$$

$$\leq \delta_{N,T}\sqrt{N}\|\theta\|_2^{1/2}\|V\|_2^{1/2}(\theta'V\theta)^{1/2}/2$$

$$\leq \delta_{N,T}\sqrt{N}(C_3/T)^{1/2}\|V\|_2^{1/2}(\theta'V\theta)^{1/2}/2.$$
For \( \|\theta\|^2 > C\delta_{N,T}^2 N/T \) with \( C = 4C_3\|V\|_2/\lambda_{\min}(V) \), we can bound further that

\[
\|\hat{\theta} - \theta\|V\theta| \leq \theta'V\theta/4.
\]

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

\[
\sup_{\theta \in \Theta(J_Q)} P(J_Q \leq F_q|\theta) \leq \sup_{\theta(J_Q)} P(T\theta'V\theta/2 - 2N \leq F_q|\theta) + \sup_{\theta(J_Q)} P(A^c|\theta)
\]

\[
\leq \sup_{\theta(J_Q)} P(T\lambda_{\min}(V)\|\theta\|^2 < 2F_q\xi\sqrt{N} + 4N|\theta) + o(1)
\]

\[
\leq \sup_{\theta(J_Q)} P(\lambda_{\min}(V)C\delta_{N,T}^2 N < 5N|\theta) + o(1),
\]

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

\[
\theta' = (\theta'_1, \theta'_2) = (0', \theta'_2), \quad \hat{\theta}' = (\hat{\theta}'_1, \hat{\theta}'_2),
\]

where \( \dim(\theta_1) = N - r_N \) and \( \dim(\theta_2) = r_N \). Corresponding to \( (\theta'_1, \theta'_2) \), we partition \( V^{-1} \) and \( V \) into:

\[
V^{-1} = \begin{pmatrix} M_1 & \beta' \\ \beta & M_2 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} M_1^{-1} + A & G' \\ G & C \end{pmatrix},
\]

where \( M_1 \) and \( A \) are \((N - r_N) \times (N - r_N)\); \( \beta \) and \( G \) are \( r_N \times (N - r_N) \); \( M_2 \) and \( C \) are \( r_N \times r_N \).

By the matrix inversion formula,

\[
A = M_1^{-1}\beta'(M_2 - \beta M_1^{-1}\beta')^{-1}\beta M_1^{-1}.
\]
Let $\Delta = T\hat{\theta}'V\hat{\theta} - T\hat{\theta}_1' M_{1}^{-1}\hat{\theta}_1$. Note that

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

We first look at $T\hat{\theta}_1' A\hat{\theta}_1$. Let $\lambda_{N,T} = T\lambda_{\text{max}}((M_2 - \beta M_1^{-1}\beta')^{-1})$ and $D_1 = \text{diag}(\frac{1}{T}M_1)$. Note that the diagonal entries of $\frac{1}{T}V^{-1}$ are given by $\text{diag}(\frac{1}{T}V^{-1}) = \{v_j\}_{j \leq N}$. Therefore $D_1$ is a diagonal matrix with entries $\{v_j\}_{j \leq N - r_N}$, and $\max_j v_j = O(T^{-1})$.

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

$$
T\hat{\theta}_1' A\hat{\theta}_1 \leq \lambda_{N,T}\|\beta M_1^{-1}\hat{\theta}_1\|^2
\leq \lambda_{N,T} r_N \|M_1^{-1}(\hat{\theta}_1 - \theta_1)\|^2 \max_{i \leq r_N} \sum_{j \leq N - r} |\beta_{ij}|^2
\leq \lambda_{N,T} \|D_1^{1/2}\|_1^2 \|\hat{\theta}_1 - \theta_1\|_2^2 \|V^{-1}\|_1^2,
$$

where we used $\theta_1 = 0$ in the second inequality and the fact that $\max_{i \leq r_N} \sum_{j \leq N - r} |\beta_{ij}| \leq \|V^{-1}\|_1$. Note that $\|V\|_1 = O(1) = \|V^{-1}\|_1$. Hence,

$$
\|M_1^{-1} D_1^{1/2}\|_1^2 = O(T^{-1}), \quad \text{and} \quad \lambda_{N,T} = O(T).
$$

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

$$
T\hat{\theta}_1' A\hat{\theta}_1 \leq C r_N \|D_1^{1/2}(\hat{\theta}_1 - \theta_1)\|_2^2 \leq C r_N \delta_{N,T}^2.
$$

Note that the uniform convergence in Assumption 1.3.1 and boundness of $\|\theta\|_{\text{max}}$ imply that $P(\|\hat{\theta}\|_{\text{max}} \leq C) \to 1$ for a sufficient large constant $C$. For $G = (g_{ij})$, note that $\max_{i \leq r} \sum_{j = 1}^{N - r} |g_{ij}| \leq \|V\|_1$. Hence, by using $\theta_1 = 0$ again, with probability approaching one,

$$
|T\hat{\theta}_2' G\hat{\theta}_1| = T|\hat{\theta}_2' GD_1^{1/2}D_1^{-1/2}(\hat{\theta}_1 - \theta_1)|
\leq T\|\hat{\theta}_2\|_{\text{max}} \|D_1^{-1/2}(\hat{\theta}_1 - \theta_1)\|_{\text{max}} \sum_{i = 1}^{r_N} \sum_{j = 1}^{N - r} |g_{ij}| \sqrt{v_j}
$$
\[ \leq C_{TN} \delta_{N,T} \sqrt{T}. \]

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

\[ \Delta = O_P(r_N \delta^2_{N,T} + r_N T). \]

We denote \( \text{var}(\hat{\theta}), \text{var}(\hat{\theta}_1), \text{var}(\hat{\theta}_2) \) to be the asymptotic covariance matrix of \( \hat{\theta}, \hat{\theta}_1 \) and \( \hat{\theta}_2 \). Then \( \frac{1}{p} V^{-1} = \text{var}(\hat{\theta}) \) and \( \frac{1}{p} M_1 = \text{var}(\hat{\theta}_1) \). It then follows from (1.10) that

\[ Z \equiv \frac{T\hat{\theta}'_1 M_1^{-1} \hat{\theta}_1 - (N - r_N)}{\sqrt{2(N - r_N)}} \xrightarrow{d} \mathcal{N}(0, 1). \]

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

\[
P(J_Q > F_q) = P\left( T\hat{\theta}'_1 M_1^{-1} \hat{\theta}_1 + \frac{\Delta - N}{\sqrt{2N}} > F_q \right)
= P\left( Z \sqrt{\frac{N - r_N}{N}} + \frac{\Delta - r_N}{\sqrt{2N}} > F_q \right)
\leq P\left( Z \sqrt{\frac{N - r_N}{N}} + \epsilon > F_q \right) + P(A^c),
\]

which is further bounded by \( 1 - \Phi(F_q - \epsilon) + P(A^c) + o(1) \). Since \( 1 - \Phi(F_q) = q \), for small enough \( \epsilon \), \( 1 - \Phi(F_q - \epsilon) = q + O(\epsilon) \). By letting \( \epsilon \to 0 \) 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.

\[ \square \]

5.1.2 Proofs for Section 1.4

Lemma 5.1.1. When \( \text{cov}(f_i) \) is positive definite, \( Ef'_i(Ef_i f'_i)^{-1}Ef_i < 1 \).
Proof. If \( Ef_i = 0 \), then \( Ef_i'(Ef_if_i')^{-1}Ef_i < 1 \). If \( Ef_i \neq 0 \), because \( \text{cov}(f_i) \) is positive definite, let \( c = (Ef_if_i')^{-1}Ef_i \), then \( c'(Ef_if_i' - Ef_if_i)c > 0 \). Hence \( c'Ef_if_i'c < c'Ef_if_i'c \) implies \( Ef_i'(Ef_if_i')^{-1}Ef_i > (Ef_i'(Ef_if_i')^{-1}Ef_i)^2 \). This implies \( Ef_i'(Ef_if_i')^{-1}Ef_i < 1 \).

**Proof of Proposition 1.4.1**

Recall that \( v_j = \text{var}(u_{jt})/(T - T Ef'_i(Ef_if_i')^{-1}Ef_i) \), and \( \hat{v}_j = \frac{1}{T} \sum T u^2_{jt}/(T a_{j,T}) \). Write \( \sigma_{ij} = (\Sigma_u)_{ij}, \hat{\sigma}_{ij} = \frac{1}{T} \sum T \hat{u}_{jt} \hat{u}_{jt}, \sigma_j^2 = T v_j, \) and \( \hat{\sigma}_j^2 = T \hat{v}_j \).

Simple calculations yield

\[
\hat{\theta}_i = \theta_i + a_{j,T}^{-1} \frac{1}{T} \sum T u_{jt}(1 - f_i'w).
\]

We first prove the second statement. Note that there is \( \sigma_{\min} > 0 \) (independent of \( \theta \)) so that \( \min_j \sigma_j > \sigma_{\min} \). By Lemma 5.1.1.1 there is \( C > 0 \), \( \inf_{\theta} P(\max_{j \leq N} |\hat{\sigma}_j - \sigma_j| < \sqrt{\frac{\log N}{T}}) \to 1 \). On the event \( \{\max_{j \leq N} |\hat{\sigma}_j - \sigma_j| < \sqrt{\frac{\log N}{T}}\} \),

\[
\max_{j \leq N} \left| \frac{\hat{v}_{j}^{1/2}}{v_j^{1/2}} - 1 \right| \leq \max_{j \leq N} \left| \frac{\hat{\sigma}_j - \sigma_j}{\sigma_j} \right| \leq \frac{C\sqrt{\log N}}{\sigma_{\min} \sqrt{T}}.
\]

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

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

\[
\max_{j \leq N} \frac{|\hat{\theta}_j - \theta_j|}{\hat{v}_j^{1/2}} \leq C \sqrt{\frac{\log N}{T}} \max_{j \leq N} v_j^{-1/2} \leq 2CC_1 \sqrt{\log N} < \delta_{N,T}.
\]

The constants \( C, C_1 \) appeared are independent of \( \theta \), and Lemma 5.1.1.1 holds uniformly in \( \theta \). Hence the desired result also holds uniformly in \( \theta \).
Proof of Proposition 1.4.2

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

\[ \frac{(Ta_f,T^\prime \hat{\theta} - \Sigma_u^{-1} \hat{\theta} - N)}{\sqrt{2N}} \rightarrow^d N(0,1). \]

Therefore, we only need to show

\[ \frac{T^\prime \hat{\theta}' (\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}) \hat{\theta}}{\sqrt{2N}} = o_P(1). \]

The left hand side is equal to

\[ \frac{T^\prime \hat{\theta}' (\Sigma_u^{-1} - \hat{\Sigma}_u^{-1}) \hat{\theta}}{\sqrt{2N}} + \frac{T^\prime (\hat{\Sigma}_u^{-1} - \Sigma_u^{-1}) (\hat{\Sigma}_u - \Sigma_u) \Sigma_u^{-1} \hat{\theta}'}{\sqrt{N}} \equiv a + b. \]

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

The challenging part is to prove \( a = o_P(1) \) when \( N > T. \) As is described in the main text, simple inequalities like Cauchy-Schwarz accumulate estimation errors, and hence do not work. Define \( e_t = \Sigma_u^{-1} u_t = (e_{1t}, ..., e_{Nt})', \) which is an \( N \)-dimensional vector with mean zero and covariance \( \Sigma_u^{-1}, \) whose entries are stochastically bounded. Let \( \bar{w} = (Ef_t f_t')^{-1} Ef_t. \)

A key step of proving this proposition is to establish the following two convergences:

\[ \frac{1}{T} E \left| \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^{T} e_{is}(1 - f_{is}' \bar{w}))^2 \right| = o(1), \quad (5.1) \]

\[ \frac{1}{T} E \left| \sum_{i \neq j, (i,j) \in S_U} \sum_{t=1}^{T} (u_{it} u_{jt} - Eu_{it} u_{jt}) \left[ \frac{1}{\sqrt{T}} \sum_{s=1}^{T} e_{is}(1 - f_{is}' \bar{w}) \right] \left[ \frac{1}{\sqrt{T}} \sum_{k=1}^{T} e_{jk}(1 - f_{jk}' \bar{w}) \right] \right| = o(1), \quad (5.2) \]

where

\[ S_U = \{ (i, j) : (\Sigma_u)_{ij} \neq 0 \}. \]
The sparsity condition assumes that most of the off-diagonal entries of $\Sigma_u$ are outside of $S_U$.

The above two convergences are weighted cross-sectional and serial double sums, where the weights satisfy $\frac{1}{T} \sum_{t=1}^{T} e_i(1 - f_t \tilde{w}) = O_p(1)$ 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 $s_{ij} = \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt}$ and $\sigma_{ij} = (\Sigma_u)_{ij}$. Under hard-thresholding,

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

Write $(\hat{\theta}' \Sigma_u^{-1})_i$ to denote the $i$th element of $\hat{\theta}' \Sigma_u^{-1}$, and $S_{U_i} = \{(i, j) : (\Sigma_u)_{ij} = 0\}$. For $\sigma_{ij} = (\Sigma_u)_{ij}$ and $\hat{\sigma}_{ij} = (\hat{\Sigma}_u)_{ij}$, we have

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

$$+ \frac{T}{\sqrt{N}} \sum_{(i, j) \in S_{U_i}} (\hat{\theta}' \Sigma_u^{-1})_i (\hat{\theta}' \Sigma_u^{-1})_j (\hat{\sigma}_{ij} - \sigma_{ij})$$

$$= a_1 + a_2 + a_3$$

We first examine $a_3$. Note that

$$a_3 = \frac{T}{\sqrt{N}} \sum_{(i, j) \in S_{U_i}} (\hat{\theta}' \Sigma_u^{-1})_i (\hat{\theta}' \Sigma_u^{-1})_j \hat{\sigma}_{ij}.$$
Because $s_{ii}$ is uniformly (across $i$) bounded away from zero with probability approaching one, and $\max_{(i,j) \in S_i} |s_{ij}| = O_P(\sqrt{\frac{\log N}{T}})$. 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 $P(J_0 = 0 \mid H_0) \to 1$. 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, $\inf_{\theta \in \Theta} P(\sqrt{nT} \left\| \hat{\beta} - \beta \right\| < \sqrt{\log n} | \theta) \to 1$.

**Proof.** Note that

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

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

$$E \left\| \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{u}_{it} \right\|^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{n} \sum_{s=1}^{T} E \tilde{x}_{it}' \tilde{x}_{it} \tilde{x}_{js}' \tilde{x}_{js} E \tilde{u}_{it} \tilde{u}_{it} \tilde{u}_{js} \tilde{u}_{js}$$

$$= \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} E \tilde{x}_{it}' \tilde{x}_{it} E \tilde{u}_{it} \tilde{u}_{it} + \frac{1}{nT} \sum_{i \neq j} \sum_{t=1}^{T} E \tilde{x}_{it}' \tilde{x}_{jt} E \tilde{u}_{it} \tilde{u}_{jt}$$

$$\leq C_1 + \frac{1}{n} \sum_{i \neq j} |E x_{it}' x_{jt} E \tilde{u}_{it} \tilde{u}_{jt}| \leq C.$$

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

**Lemma 5.1.3.** Suppose $\max_{j \leq n} M_j = \frac{1}{nT} \sum_{t=1}^{T} \tilde{x}_{jt} \tilde{x}_{jt}' < C'$ with probability approaching one and $\sup_{\theta} E(u_{jt}' | \theta) < C'$. There is $C > 0$, so that
(i)  inf_{\theta}\{P(\max_{j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| < C \sqrt{\log n/T}\theta) \to 1}

(ii)  inf_{\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}\theta) \to 1}

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

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

Proof. (i) By the Bernstein inequality, for 

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

(ii) For \( C = (8 \max_{j\leq n} \sup_{\theta\in\Theta} E(u_{jt}^2|\theta))^{1/2} \), we have

\[ \sup_{\theta\in\Theta} P(\max_{j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{jt}| \geq C \sqrt{\log n/T}\theta) \leq \exp(\log n - \frac{C^2 \log n}{4 \max_{j\leq n} \sup_{\theta\in\Theta} E(u_{jt}^2|\theta)}) = \frac{1}{n}. \]

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

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

\[ \sup_{\theta\in\Theta} P(\max_{i,j\leq n} |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - EU_{it}u_{jt}| \geq C \sqrt{\log n/T}\theta) \leq \exp(2 \log n - \frac{C^2 \log n}{4 \max_{j\leq n} \sup_{\theta\in\Theta} E(u_{jt}^4|\theta)}) = \frac{1}{n}. \]

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

(iv) Observe that

\[ |\frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it}\hat{u}_{jt} - EU_{it}u_{jt}| \leq |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - EU_{it}u_{jt}| + |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - \hat{u}_{it}\hat{u}_{jt}| \]

\[ \leq |\frac{1}{T}\sum_{t=1}^{T} u_{it}u_{jt} - EU_{it}u_{jt}| + \frac{1}{T}\sum_{t=1}^{T} (\hat{u}_{jt} - u_{jt})^2 + \frac{2}{T}\sum_{t=1}^{T} u_{jt}^2(\frac{2}{T}\sum_{t=1}^{T} (\hat{u}_{jt} - u_{jt})^2)^{1/2} \]

77
The first two terms and \( \left( \frac{1}{T} \sum_t (\hat{u}_{jt} - u_{jt})^2 \right)^{1/2} \) 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

\[
\inf_{\theta \in \Theta} P(\max_{j \leq n} \frac{1}{T} \sum_t u_{jt}^2 < M|\theta) \rightarrow 1.
\]

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

\[
\inf_{\theta \in \Theta} P(\max_{j \leq n} \left| \frac{1}{T} \sum_t u_{jt}^2 - E u_{jt}^2 \right| < C \sqrt{\log n/T}|\theta) \rightarrow 1.
\]

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

\[
\sup_{\theta \in \Theta} P(\max_{j \leq n} \frac{1}{T} \sum_t u_{jt}^2 \geq M|\theta) \leq \sup_{\theta \in \Theta} P(\max_{j \leq n} \left| \frac{1}{T} \sum_t u_{jt}^2 - E u_{jt}^2 \right| \geq M - \max_{j \leq n} E u_{jt}^2|\theta)
\]

\[
\leq \sup_{\theta \in \Theta} P(\max_{j \leq n} \left| \frac{1}{T} \sum_t u_{jt}^2 - E u_{jt}^2 \right| \geq C \sqrt{\log n/T}|\theta) \rightarrow 0.
\]

This proves the desired result.

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

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

\[
|\tilde{\rho}_{ij} - \rho_{ij}| \leq \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt} - u_{it} u_{jt} \right) \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it}^2 \right)^{1/2} \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{jt}^2 \right)^{1/2} \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt} \right)
\]

\[
+ \left| \frac{1}{T} \sum_{t=1}^{T} u_{it} u_{jt} \right| \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it}^2 \right)^{1/2} \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{jt}^2 \right)^{1/2} - \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it}^2 \right)^{1/2} \left( \frac{1}{T} \sum_{t=1}^{T} \hat{u}_{jt}^2 \right)^{1/2}.
\]

\[
X_1 + X_2
\]

78
By part (iv) of Lemma 5.1.3, \( \inf_{\theta \in \Theta} P(\max_{i,j \leq n} |\frac{1}{T} \sum_{t=1}^{T} \hat{u}_{it} \hat{u}_{jt} - Eu_{it}u_{jt}| < C \sqrt{\log n / T} | \theta) \to 1 \). Hence for sufficiently large \( M > 0 \) such that \( \inf_{\theta \in \Theta} \min_{j} E(u_{jt}^{2} | \theta) - C/M > C \sqrt{\log n / T} \),

\[
\sup_{\theta \in \Theta} P(\max_{ij} |X_{1}| > M \sqrt{\frac{\log n}{T} | \theta}) \leq \sup_{\theta \in \Theta} P(\min_{j} \frac{1}{T} \sum_{t} \hat{u}_{jt}^{2} < C/M | \theta) + o(1)
\]

\[
\leq \sup_{\theta \in \Theta} P(\max_{j} \frac{1}{T} \sum_{t} \hat{u}_{jt}^{2} - Eu_{jt}^{2} | \min_{j} Eu_{jt}^{2} - C/M | \theta) + o(1) = o(1).
\]

By a similar argument, there is \( M' > 0 \) so that \( \sup_{\theta \in \Theta} P(\max_{ij} |X_{2}| > M' \sqrt{\frac{\log n}{T} | \theta}) = o(1) \).

The result then follows as,

\[
\sup_{\theta} P(\max_{ij} |\hat{\rho}_{ij} - \rho_{ij}| \geq 2(M + M') \sqrt{\log n / T})
\]

\[
\leq \sup_{\theta} P(\max_{ij} |X_{1}| \geq (M + M') \sqrt{\log n / T}) + \sup_{\theta} P(\max_{ij} |X_{2}| \geq (M + M') \sqrt{\log n / T}) = o(1).
\]

**Proof of Proposition 1.5.1**

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

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

**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 \( C > 0 \), \( \inf_{\theta \in \Theta} P(\max_{ij} |\hat{\rho}_{ij} - \rho_{ij}| < C \sqrt{\log n / T} | \theta) \to 1 \). If we define

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

79
then \( \inf_\theta P(A|\theta) \to 1 \). On the event \( A \), we have, uniformly in \( \theta = \{\rho_{ij}\} \),

\[
\sum_{i<j} (\hat{\rho}_{ij} - \rho_{ij}) \rho_{ij} \leq \left( \sum_{i<j} (\hat{\rho}_{ij} - \rho_{ij})^2 \right)^{1/2} \left( \sum_{i<j} \rho_{ij}^2 \right)^{1/2} \leq \frac{Cn\sqrt{\log n}}{\sqrt{T}} \left( \sum_{i<j} \rho_{ij}^2 \right)^{1/2}.
\]

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

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

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

\[
\sup_{\Theta(J_1)} P(J_1 < F_q|\theta) \leq \sup_{\Theta(J_1)} P\left( \sum_{i<j} \hat{\rho}_{ij}^2 < \frac{n(n-1)}{2T} + \left( F_q + \frac{n}{2(T-1)} \right) \frac{\sqrt{n(n-1)}}{T} |\theta \right)
\]

\[
\leq \sup_{\Theta(J_1)} P\left( \frac{1}{2} \sum_{i<j} \rho_{ij}^2 < \frac{n(n-1)}{2T} + \left( F_q + \frac{n}{2(T-1)} \right) \frac{\sqrt{n(n-1)}}{T} |\theta \right) + \sup_{\Theta(J_1)} P(A^c|\theta) \to 0.
\]

**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 \( n, T \to \infty \),

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

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 \( e_t = \Sigma_u^{-1}u_t = (e_{1t}, ..., e_{Nt})' \), which is an \( N \)-dimensional vector with mean zero and covariance \( \Sigma_u^{-1} \), whose entries are stochastically bounded. Let \( \mathbf{w} = (Ef'_t)^{-1}Ef_t \). Also recall that

\[
a_1 = \frac{T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1})^2_i (\hat{\sigma}_{ii} - \sigma_{ii}),
\]

\[
a_2 = \frac{T}{\sqrt{N}} \sum_{i \neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1})_i (\hat{\theta}' \Sigma_u^{-1})_j (\hat{\sigma}_{ij} - \sigma_{ij}).
\]

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

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

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

where \( S_U = \{(i,j) : (\Sigma_u)_{ij} \neq 0\} \). 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 \( a_1 = \frac{T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1})^2_i \frac{1}{T} \sum_{t=1}^{T} (\hat{\sigma}_{it}^2 - Eu_{it}^2) \), which is

\[
\frac{T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1})^2_i \frac{1}{T} \sum_{t=1}^{T} (\hat{\sigma}_{it}^2 - u_{it}^2) + \frac{T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1})^2_i \frac{1}{T} \sum_{t=1}^{T} (u_{it}^2 - Eu_{it}^2) = a_{11} + a_{12}.
\]

For \( a_{12} \), note that \( (\hat{\theta}' \Sigma_u^{-1})_i = (1 - \bar{f}' \mathbf{w})^{-1} \frac{1}{T} \sum_{s=1}^{T} (1 - f'_s \mathbf{w})(u'_s \Sigma_u^{-1})_i = c \frac{1}{T} \sum_{s=1}^{T} (1 - f'_s \mathbf{w})e_{is}, \) where \( c = (1 - \bar{f}' \mathbf{w})^{-1} = O_P(1) \). Hence

\[
a_{12} = \frac{Tc}{\sqrt{N}} \sum_{i=1}^{N} (\frac{1}{T} \sum_{s=1}^{T} (1 - f'_s \mathbf{w})e_{is})^2 \frac{1}{T} \sum_{t=1}^{T} (u_{it}^2 - Eu_{it}^2)
\]
By (5.3), $Ea_{12}^2 = o(1)$. On the other hand,

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

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

$$a_{111} \leq O_P(\frac{\log N}{T}) \frac{T}{\sqrt{N}} \|\hat{\theta}' \Sigma_u^{-1}\|^2 = O_P(\frac{(\log N)^2 \sqrt{N}}{T}) = o_P(1),$$

To bound $a_{112}$, note that

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

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

$$a_{112} = \frac{2T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1} i)^2 \frac{1}{T} \sum_{t=1}^{T} u_{it} (\hat{\theta}_i - \theta_i) + \frac{2T}{\sqrt{N}} \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1} i)^2 (\hat{b}_i - b_i)' \frac{1}{T} \sum_{t=1}^{T} f_i u_{it}$$

$$\leq O_P(\frac{\log N}{\sqrt{N}}) \|\hat{\theta}' \Sigma_u^{-1}\|^2 = o_P(1).$$

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

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

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

$$\frac{T}{\sqrt{N}} \sum_{i\neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1} i)(\hat{\theta}' \Sigma_u^{-1} j) \left( \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})(\hat{u}_{jt} - u_{jt}) + \frac{1}{T} \sum_{t=1}^{T} (u_{it} - E u_{it})(u_{jt} - E u_{jt}) \right) = a_{21} + a_{22},$$

where

$$a_{21} = \frac{T}{\sqrt{N}} \sum_{i\neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1} i)(\hat{\theta}' \Sigma_u^{-1} j) \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})(\hat{u}_{jt} - u_{jt}).$$
Under $H_0$, $\Sigma_u^{-1}\hat{\theta} = \frac{1}{T}(1 - f'w)^{-1}\sum_{t=1}^{T} \Sigma_u^{-1}u_t(1 - f'_tu_t)$, and $e_t = \Sigma_u^{-1}u_t$, we have

$$a_{22} = \frac{T}{\sqrt{N}} \sum_{i \neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1})_i(\hat{\theta}' \Sigma_u^{-1})_j \frac{1}{T} \sum_{t=1}^{T} (u_{it} - E_{it}u_{jt})(u_{jt} - E_{jt}u_{jt})$$

$$= \frac{Tc}{\sqrt{N}} \sum_{i \neq j, (i,j) \in S_U} \frac{1}{T} \sum_{s=1}^{T} (1 - f'_sw)\hat{e}_{is} \frac{1}{T} \sum_{k=1}^{T} (1 - f'_kw)e_{jk} \frac{1}{T} \sum_{t=1}^{T} (u_{it}u_{jt} - E_{it}u_{jt}).$$

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

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

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

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

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

$$|a_{211}| \leq O_P\left(\frac{\log N}{\sqrt{N}}\right) \sum_{i \neq j, (i,j) \in S_U} |(\hat{\theta}' \Sigma_u^{-1})_i||(\hat{\theta}' \Sigma_u^{-1})_j|$$

$$\leq O_P\left(\frac{\log N}{\sqrt{N}}\right) \left( \sum_{i \neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1})_i^2 \right)^{1/2} \left( \sum_{i \neq j, (i,j) \in S_U} (\hat{\theta}' \Sigma_u^{-1})_j^2 \right)^{1/2}$$

$$= O_P\left(\frac{\log N}{\sqrt{N}}\right) \sum_{i=1}^{N} (\hat{\theta}' \Sigma_u^{-1})_i^2 \sum_{j: (\Sigma_u)_j \neq 0} 1 \leq O_P\left(\frac{\log N}{\sqrt{N}}\right) ||\hat{\theta}' \Sigma_u^{-1}||^2 m_N$$

$$= O_P\left(\frac{m_N\sqrt{N}(\log N)^2}{T}\right) = o_P(1).$$

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

$$|a_{212}| \leq O_P\left(\frac{\log N}{\sqrt{N}}\right) \sum_{i \neq j, (i,j) \in S_U} |(\hat{\theta}' \Sigma_u^{-1})_i||(\hat{\theta}' \Sigma_u^{-1})_j| = O_P\left(\frac{m_N\sqrt{N}(\log N)^2}{T}\right) = o_P(1).$$

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 $|A|$ denote its number of elements.

**Lemma 5.1.8.** Recall that $e_t = \sum_{-1}^0 u_{it}$, $e_{it}$ and $u_{jt}$ are independent if $i \neq j$.

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

$$A = \begin{pmatrix} \Sigma_u & 0 \\ 0 & \Sigma_u^{-1} \end{pmatrix}.$$  

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

$$\begin{pmatrix} I_N & 0 \\ 0 & \Sigma_u^{-1} \end{pmatrix} \begin{pmatrix} \Sigma_u & \Sigma_u \\ \Sigma_u & \Sigma_u \end{pmatrix} \begin{pmatrix} I_N & 0 \\ 0 & \Sigma_u^{-1} \end{pmatrix} = \begin{pmatrix} \Sigma_u & I_N \\ I_N & \Sigma_u^{-1} \end{pmatrix}.$$  

This completes the proof. □

**Proof of (5.3)**

Let $X = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T (u_{it}^2 - Eu_{it}^2)(\frac{1}{\sqrt{T}} \sum_{s=1}^T e_{is}(1 - f'_s w))^2$. The goal is to show $E X^2 = O(T)$. 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

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

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, $\xi_{it}$ is independent of $\zeta_{js}$ 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,
\[ EX = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \text{cov}(\xi_{it}, \left( \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is} \right)^2) = \frac{1}{T\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{k=1}^{T} \text{cov}(\xi_{it}, \zeta_{is}\zeta_{ik}) \]
\[ = \frac{1}{T\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \text{cov}(\xi_{it}, \zeta_{it}^2) + 2 \sum_{k \neq t} \text{cov}(\xi_{it}, \zeta_{it}\zeta_{ik}) \right) \]
\[ = \frac{1}{T\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \text{cov}(\xi_{it}, \zeta_{it}^2) = O\left( \frac{\sqrt{NT}}{T} \right), \]
where the second last equality follows since \( E\xi_{it} = E\zeta_{it} = 0 \) and when \( k \neq t \) \( \text{cov}(\xi_{it}, \zeta_{it}\zeta_{ik}) = E\xi_{it}\zeta_{it}\zeta_{ik} = E\xi_{it}\zeta_{it}E\zeta_{ik} = 0 \). It then follows that \( \frac{1}{T}(EX)^2 = O\left( \frac{N}{T^2} \right) = o(1) \), given \( N = o(T^2) \).

### Variance

Consider the variance. We have,
\[ \text{var}(X) = \frac{1}{N} \sum_{i=1}^{N} \text{var}\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_{it}\left( \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is} \right)^2 \right) \]
\[ + \frac{1}{NT^3} \sum_{i \neq j} \sum_{t,s,k,l,v,p \leq T} \text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jl}\zeta_{jv}\zeta_{jp}) = B_1 + B_2. \]

\( B_1 \) can be bounded by the Cauchy-Schwarz inequality. Note that \( E\xi_{it} = E\zeta_{js} = 0 \),
\[ B_1 \leq \frac{1}{N} \sum_{i=1}^{N} \text{var}\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_{it}\left( \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is} \right)^2 \right) \leq \frac{1}{N} \sum_{i=1}^{N} \left[ \text{var}\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \xi_{it} \right) \right]^{1/2} \left[ \text{var}\left( \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is} \right) \right]^{1/2}. \]

Hence \( B_1 = O(1) \).

We now show \( \frac{1}{T}B_2 = o(1) \). Once this is done, it implies \( \frac{1}{T}\text{var}(X) = o(1) \). The proof of (5.3) is then completed because \( \frac{1}{T}EX^2 = \frac{1}{T}(EX)^2 + \frac{1}{T}\text{var}(X) = o(1) \).

For two variables \( X, Y \), writing \( X \perp Y \) if they are independent. Note that \( E\xi_{it} = E\zeta_{is} = 0 \), and when \( t \neq s \), \( \xi_{it} \perp \zeta_{js}, \xi_{it} \perp \zeta_{js}, \zeta_{it} \perp \zeta_{js} \) for any \( i, j \leq N \). Therefore, it is straightforward to verify that if the set \( \{t, s, k, l, v, p\} \) contains more than three distinct elements, then \( \text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jl}\zeta_{jv}\zeta_{jp}) = 0 \). Hence if we denote \( \Xi \) 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

\[
\sum_{t,s,k,l,v,p \leq T} \text{cov}(\xi_{it}\zeta_{is}\zeta_{it}, \xi_{jt}\zeta_{js}\zeta_{jl}) = \sum_{(t,s,k,l,v,p) \in \Xi} \text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jt}\zeta_{js}\zeta_{jl}).
\]

Hence

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

Let us partition \( \Xi \) into \( \Xi_1 \cup \Xi_2 \) where each element \((t, s, k, l, v, p)\) in \( \Xi_1 \) contains exactly three distinct indices, while each element in \( \Xi_2 \) contains less than three distinct indices. We know that

\[
\frac{1}{NT^3} \sum_{i \neq j} \sum_{(t,s,k,l,v,p) \in \Xi_2} \text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jt}\zeta_{js}\zeta_{jl}) = O\left(\frac{1}{NT^3} N^2 T^2 \right) = O\left(\frac{N}{T} \right),
\]

which implies

\[
\frac{1}{T} B_2 = \frac{1}{NT^4} \sum_{i \neq j} \sum_{(t,s,k,l,v,p) \in \Xi_1} \text{cov}(\xi_{it}\zeta_{is}\zeta_{ik}, \xi_{jt}\zeta_{js}\zeta_{jl}) + O_p\left(\frac{N}{T^2} \right).
\]

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.

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

Note that if \((\Sigma_u)_{ij} = 0\), \(u_{ij}\) and \(u_{ji}\) are independent, and hence \(E\xi_{ij}\zeta_{ji} = 0\). This implies \(\sum_{i \neq j} |E\xi_{ij}\zeta_{ji}| \leq O(1) \sum_{i \neq j, (i,j) \in S_U} 1 = O(N)\). Hence \(B_{21} = o(1)\).

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

By Lemma 5.1.8, \(u_{js}\) and \(e_{is}\) are independent for \(i \neq j\). Also, \(u_{js}\) and \(f_s\) are independent, which implies \(\xi_{js}\) and \(\zeta_{is}\) are independent. So \(E\xi_{js}\zeta_{is} = 0\). It follows that \(B_{22} = 0\).

\[
B_{23} = \frac{1}{NT^4} \sum_{i \neq j} \sum_{t=1}^T \sum_{s \neq t} \sum_{l \neq s, t} E\xi_{it}\zeta_{jt} E\zeta_{is} E\xi_{js} E\zeta_{jl} = O\left(\frac{1}{NT} \right) \sum_{i \neq j} |E\zeta_{is}\zeta_{js}|.
\]
By the definition $e_s = \Sigma_u^{-1} u_s$, $\text{cov}(e_s) = \Sigma_u^{-1}$. Hence $Ee_is e_js = (\Sigma_u^{-1})_{ij}$, which implies $B_{23} \leq O(\frac{N}{NT}) \|\Sigma_u^{-1}\|_1 = o(1)$.

Finally, $B_{25} = \frac{1}{NT^4} \sum_{i \neq j} \sum_{t=1}^{T} \sum_{s \neq t} \sum_{l \neq s} \sum_{t \neq l} E\xi_{it} \xi_{jt} E\zeta_{is} \zeta_{js} E\zeta_{il} \zeta_{jl} = 0$, because $E\zeta_{is} \zeta_{js} = 0$ when $i \neq j$, following from Lemma 5.1.8. Therefore, $\frac{1}{T} B_2 = o(1) + O(\frac{N}{NT^2}) = o(1)$.

**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

$$Z = \frac{1}{\sqrt{NT}} \sum_{(i,j) \in H} \sum_{t=1}^{T} (u_{it} u_{jt} - E u_{it} u_{jt}) \left[ \frac{1}{\sqrt{T}} \sum_{s=1}^{T} e_{is} (1 - f'_s w) \right] \left[ \frac{1}{\sqrt{T}} \sum_{k=1}^{T} e_{jk} (1 - f'_k w) \right]$$

$$= \frac{1}{\sqrt{NT}} \sum_{(i,j) \in H} \sum_{t=1}^{T} \xi_{ijt} \left[ \frac{1}{\sqrt{T}} \sum_{s=1}^{T} \zeta_{is} \right] \left[ \frac{1}{\sqrt{T}} \sum_{k=1}^{T} \zeta_{jk} \right] = \frac{1}{T \sqrt{NT}} \sum_{(i,j) \in H} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{k=1}^{T} \xi_{ijt} \zeta_{is} \zeta_{jk}.$$ 

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

**Expectation**

The proof for the expectation is the same regardless of the type of sparsity in Assumption 1.4.2 and is very similar to that of (5.3). In fact,

$$EZ = \frac{1}{T \sqrt{NT}} \sum_{(i,j) \in H} \sum_{t=1}^{T} \sum_{s=1}^{T} \sum_{k=1}^{T} \text{cov}(\xi_{ijt}, \zeta_{is} \zeta_{jk}) = \frac{1}{T \sqrt{NT}} \sum_{(i,j) \in H} \sum_{t=1}^{T} \sum_{k=1}^{T} \text{cov}(\xi_{ijt}, \zeta_{il} \zeta_{jl}).$$
Because $\sum_{(i,j)\in H} 1 = O(N)$, $EZ = O(\sqrt{N/T})$. Thus $\frac{1}{T}(EZ)^2 = o(1)$.

**Variance**

For the variance, we have

$$\text{var}(Z) = \frac{1}{T^3N} \sum_{(i,j)\in H} \text{var}(\sum_{t=1}^T \sum_{s=1}^T \sum_{k=1}^T \xi_{ijt}\zeta_{is}\zeta_{jk})$$

$$+ \frac{1}{T^3N} \sum_{(i,j)\in H} \sum_{(m,n)\in H, (m,n)\neq (i,j), t,s,k,l,v,p\leq T} \text{cov}(\xi_{ijt}\zeta_{is}\zeta_{jk}, \xi_{mnl}\zeta_{mv}\zeta_{np})$$

$$= A_1 + A_2.$$ 

By the Cauchy-Schwarz inequality and the serial independence of $\xi_{ijt}$,

$$A_1 \leq \frac{1}{N} \sum_{(i,j)\in H} E \left[ \frac{1}{\sqrt{T}} \sum_{t=1}^T \xi_{ijt}\frac{1}{\sqrt{T}} \sum_{s=1}^T \zeta_{is}\frac{1}{\sqrt{T}} \sum_{k=1}^T \zeta_{jk} \right]^2$$

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

So $A_1 = O(1)$.

Note that $E\xi_{ijt} = E\zeta_{is} = 0$, and when $t \neq s$, $\xi_{ijt} \perp \zeta_{ms}$, $\xi_{ijt} \perp \zeta_{ms}$, $\zeta_{it} \perp \zeta_{js}$ (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 $\Xi$ 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

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

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

Similar to the proof of the first statement, for the same set $\Xi_1$ that contains exactly three distinct indices in each of its element, (recall $|H|_0 = O(N)$)

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

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

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

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

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

$$A_{21} = \frac{1}{NT^4} \sum_{(i,j) \in H} \sum_{(m,n) \in H, (m,n) \neq (i,j)} \sum_{t=1}^T \sum_{s \neq t} \sum_{l \neq s} E\xi_{ijt}\xi_{mnt}E\xi_{is}\zeta_{js}E\zeta_{ml}\zeta_{nt}$$

$$\leq O\left(\frac{1}{NT^4}\right) \sum_{(i,j) \in H} \sum_{(m,n) \in H, (m,n) \neq (i,j)} |E\xi_{ijt}\xi_{mnt}|$$

$$\leq O\left(\frac{1}{NT^4}\right) \sum_{(i,j) \in H} \sum_{m \in A(i) \cup A(j)} \sum_{n \in A(m)} \sum_{m \notin A(i) \cup A(j)} \sum_{n \in A(m)} |\text{cov}(u_{it}u_{jt}, u_{mt}u_{nt})|.$$  

The first term is $O\left(\frac{1}{T^4}\right)$ 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 $\text{cov}(u_{it}u_{jt}, u_{mt}u_{nt}) \neq 0$, then $m \in B(i, j)$. Hence the second term is bounded by $O\left(\frac{1}{NT^4}\right) \sum_{(i,j) \in H} \sum_{m \in B(i, j)} \sum_{n \in A(m)} |\text{cov}(u_{it}u_{jt}, u_{mt}u_{nt})|$, which is also $O\left(\frac{1}{T^4}\right)$ by Lemma 5.1.9. Hence $A_{21} = o(1)$.

Similarly, by Lemma 5.1.9

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

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

90
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

\[
A_{23} = \frac{1}{NT^4} \sum_{(i,j) \in H, (m,n) \in H, (m,n) \neq (i,j)} \sum_{t=1}^T \sum_{s \neq t} \sum_{l \neq t,s} E\xi_{ijt} \zeta_{ilt} E\xi_{js} \xi_{mns} E\xi_{ml} \zeta_{nl}
\]

\[
\leq O\left( \frac{1}{NT} \right) \sum_{j=1}^N \sum_{i \in A(j)} 1 \sum_{(m,n) \in H, (m,n) \neq (i,j)} \left| E\xi_{js} \xi_{mns} \right|
\]

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

Term \( a = O\left( \frac{1}{T} \right) \). 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 \( u_s \) and \( f_s \) are independent, hence \( \zeta_{js} \) and \( \xi_{mms} \) are independent, which implies that \( b = 0 \). Hence \( A_{23} = o(1) \).

The same argument as of \( A_{23} \) also implies

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

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

\[
A_{25} = \frac{1}{NT^4} \sum_{(i,j) \in H, (m,n) \in H, (m,n) \neq (i,j)} \sum_{t=1}^T \sum_{s \neq t} \sum_{l \neq t,s} E\xi_{ijt} \zeta_{ilt} E\xi_{is} \xi_{mns} E\xi_{ml} \zeta_{nl}
\]

\[
\leq O\left( \frac{1}{NT} \right) \sum_{(i,j) \in H, (m,n) \in H, (m,n) \neq (i,j)} \left| E\xi_{ijt} \zeta_{ilt} E\xi_{is} \xi_{mns} E\xi_{ml} \zeta_{nl} \right|
\]

\[
\leq O\left( \frac{1}{NT} \right) \sum_{i=1}^N \sum_{m=1}^N \left| E\xi_{is} \xi_{mns} \right| \leq O\left( \frac{1}{NT} \right) \sum_{i=1}^N \sum_{m=1}^N \left| (\Sigma^{-1})_{im} \right| E(1 - f'_s w)^2
\]

\[
\leq O\left( \frac{N}{NT} \right) \| \Sigma^{-1} \|_1 = o(1).
\]

In summary, \( \frac{1}{T} A_2 = o(1) + O\left( \frac{N}{T^2} \right) = o(1) \). 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) $P(\max_{i,j \leq N} \left| \frac{1}{T} \sum_{t=1}^{T} u_{it}u_{jt} - Eu_{it}u_{jt} \right| > C \sqrt{\frac{\log N}{T}}) \to 0$.

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

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

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

Lemma 5.1.11. When the distribution of $(u_t, f_t)$ is independent of $\theta$, there is $C > 0$,

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

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

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

Proof. Note that

$$\hat{\theta}_j - \theta_j = \frac{1}{af, T} \sum_{t=1}^{T} u_{jt}(1 - f_t^w).$$

Here $a_{f,T} = 1 - \bar{f}^w \rightarrow 1 - Ef_t(Ef_t f_t)^{-1}Ef_t > 0$, 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 $\theta$, such that

$$\sup_{\theta \in \Theta} P(\max_{j \leq N} |\hat{\theta}_j - \theta_j| > C \sqrt{\frac{\log N}{T}}(\theta)) = P(\max_{j} \left| \frac{1}{a_{f,T}T} \sum_{t=1}^{T} u_{jt}(1 - f_t^w) \right| > C \sqrt{\frac{\log N}{T}}) \to 0$$

(ii) There is $C$ independent of $\theta$, such that the event

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

has probability approaching one. Also, there is $C_2$ also independent of $\theta$ 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,

$$|\hat{\sigma}_{ij} - \sigma_{ij}| \leq C \sqrt{\frac{\log N}{T}} + 2 \max_i \sqrt{\frac{1}{T} \sum_{t}(\hat{u}_{it} - u_{it})^2}C_2 + \max_i \frac{1}{T} \sum_{t}(u_{it} - \hat{u}_{it})^2.$$
It can be shown that
\[
\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 \leq \max_{i} (\|\hat{b}_i - b_i\|^2 + (\hat{\theta}_i - \theta_i)^2)(\frac{1}{T} \sum_{t=1}^{T} \|f_t\|^2 + 1).
\]

Note that \(\hat{b}_i - b_i\) and \(\hat{\theta}_i - \theta_i\) only depend on \((f_t, u_t)\) (independent of \(\theta\)). By Lemma 3.1 of Fan et al. (2011), there is \(C_3 > 0\) such that \(\sup_{b, \theta} P(\max_{i \leq N}\|\hat{b}_i - b_i\|^2 + (\hat{\theta}_i - \theta_i)^2 > C_3 \frac{\log N}{T}) = o(1)\). Combining the last two displayed inequalities yields, for \(C_4 = (C + 1)C_3\),
\[
\sup_{\theta} P(\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} (\hat{u}_{it} - u_{it})^2 > C_4 \frac{\log N}{T} |\theta) = o(1),
\]
which yields the desired result.

(iii): Recall \(\hat{\sigma}^2_j = \hat{\sigma}_{jj}/a_{f,T}\), and \(\sigma^2_j = \sigma_{jj}/(1 - Ef_t(Ef_t'f_t)^{-1}Ef_t)\). Moreover, \(a_{f,T}\) is independent of \(\theta\). The result follows immediately from part (ii). \(\square\)

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

**Proof.** By Lemma 5.1.11 (ii), \(\sup_{\theta \in \Theta} P(\max_{i,j \leq N} |\hat{\sigma}_{ij} - \sigma_{ij}| > C \frac{\log N}{T} |\theta) \to 1\). By Fan et al. (2011), on the event \(\max_{i,j \leq N} |\hat{\sigma}_{ij} - \sigma_{ij}| \leq C' \frac{\log N}{T}\), there is constant \(C'\) that is independent of \(\theta\), \(\|\hat{\Sigma}^{-1}_u - \Sigma^{-1}_u\| \leq C' m_N(\frac{\log N}{T})^{1/2}\). Hence the result follows due to the sparse condition \(m_N(\frac{\log N}{T})^{1/2} = o(1)\). \(\square\)

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

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

Lemma 5.2.2. Let \( \{\lambda_i\}_{i=1}^p \) be the eigenvalues of \( \Sigma \) in descending order and \( \{\xi_i\}_{i=1}^p \) be their associated eigenvectors. Correspondingly, let \( \{\hat{\lambda}_i\}_{i=1}^p \) be the eigenvalues of \( \hat{\Sigma} \) in descending order and \( \{\hat{\xi}_i\}_{i=1}^p \) be their associated eigenvectors. Then,

(a) (Weyl’s theorem) \(|\hat{\lambda}_i - \lambda_i| \leq \|\hat{\Sigma} - \Sigma\|\).

(b) (sin(\theta) theorem)

\[
\|\hat{\xi}_i - \xi_i\| \leq \frac{\|\hat{\Sigma} - \Sigma\|/\sqrt{2}}{\min(|\hat{\lambda}_{i-1} - \lambda_i|, |\lambda_i - \hat{\lambda}_{i+1}|)}
\]

5.2.1 Proof of Proposition 2.2.1

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

5.2.2 Proof of Theorem 2.3.1

Recall that \( H = (1/T)V^{-1}\hat{F}'FB'B \). A preliminary result about \( H \) 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) \( HH' = I_K + O_p(\omega_{p,T}) \),

(b) \( H'H = I_K + O_p(\omega_{p,T}) \).

The next lemma shows that the normalization matrix \( \Lambda_b \) can be consistently estimated under operator norm.
Lemma 5.2.4. Under assumptions 2.3.1-2.3.3,

(a) \( ||\hat{B} - BH'|| = O_p(p^{1/2}\omega_{p,T}) \),

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

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

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

\[
\frac{1}{T} \sum_{t=1}^{T} ||\hat{f}_t - Hf_t|| = O_p(\omega_{p,T}).
\]

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

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

\[
\hat{b}_i - Hb_i = \frac{1}{T} \sum_{t=1}^{T} Hf_t u_{it} + \frac{1}{T} \sum_{t=1}^{T} x_{it}(\hat{f}_t - Hf_t) + H(\frac{1}{T} \sum_{t=1}^{T} \hat{f}_t'\hat{f}_t - I_K)b_i
\]

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

\[
||\frac{1}{T} \sum_{t=1}^{T} Hf_t u_{it}|| \leq ||H|| \cdot ||\frac{1}{T} \sum_{t=1}^{T} f_t u_{it}||.
\]

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

\[
||\frac{1}{T} \sum_{t=1}^{T} x_{it}(\hat{f}_t - Hf_t)|| \leq (\frac{1}{T} \sum_{t=1}^{T} x^2_{it}) \frac{1}{T} \sum_{t=1}^{T} ||\hat{f}_t - Hf_t||^{1/2} = O_p(\omega_{p,T}).
\]

And lastly, \( ||T^{-1} \sum_{t=1}^{T} (f_t'f_t - I_K)|| = O_p(T^{-1/2}) \) and \( ||b_i|| = O(1) \) imply that the third term is \( O_p(T^{-1/2}) \).
Therefore we have

\[
\|\hat{B} - BH'\|^2 \leq \|\hat{B} - BH\|^2 = \sum_{i=1}^{p} \|\hat{b}_i - Hb_i\|^2 \\
\leq 3\|H\|^2 \cdot (T^{-1} \sum_{i=1}^{p} \|\frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_{iut}\|^2) + pO_p(\omega_{p,T}^2),
\]

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

\[
\|\hat{B} - BH'\|^2 = O_p(p/T) + O_p(p\omega_{p,T}^2) = O_p(p\omega_{p,T}^2).
\]

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

\[
\|(\hat{B}'\hat{B})^{-1} - (HB'BH')^{-1}\| = O_p(p^{-1}\omega_{p,T}).
\]

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

Hence

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

Consequently,

\[
\|\hat{A}_b - HA_b\| = \|(\hat{B}'\hat{B})^{-1}\hat{B}' - H(B'B)^{-1}H^{-1}HB'\|
\leq \|(B'B)^{-1} - H(B'B)^{-1}H^{-1}\| \cdot \|\hat{B}'\| + \|H(B'B)^{-1}H^{-1}\| \cdot \|\hat{B}' - HB'\|
\leq O_p(p^{-1/2}\omega_{p,T}).
\]
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 \( E(f_t|y_{t+1}) - E(f_t) \) is contained in the linear subspace spanned by \( \phi_k^k \text{cov}(f_t), k = 1, ..., L \).

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

**Proof of Theorem 2.3.1**

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

\[
||\hat{m}_h - m_h|| = \left| \frac{1}{c} \sum_{l=1}^{c} (Bf_{(h,l)} + u_{(h,l)}) - BE(f_t|y_{t+1} \in I_h) \right| \\
\leq ||B|| \cdot \left| \frac{1}{c} \sum_{l=1}^{c} f_{(h,l)} - E(f_t|y_{t+1} \in I_h) \right| + ||\frac{1}{c} \sum_{l=1}^{c} u_{(h,l)}|| \\
= O_p(p^{1/2})O_p(T^{-1/2}) + O_p(\sqrt{p/T}) = O_p(\sqrt{p/T}).
\]

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

In addition to the inequality above, we have \( ||m_h|| = O_p(||E(x_t|y_{t+1} \in I_h)||) \leq O_p(||B|| \cdot ||E(f_t|y_{t+1} \in I_h)||) = O_p(p^{1/2}), \) so \( \hat{m}_h = O_p(p^{1/2}) \). It follows that

\[
||\hat{A}_b \hat{m}_h - \hat{A}_b m_h|| \leq ||\hat{A}_b - \hat{A}_b|| \cdot ||\hat{m}_h|| + ||\hat{A}_b|| \cdot ||\hat{m}_h - m_h|| \\
= O_p(\omega_{p,T}) + O_p(T^{-1/2}) = O_p(\omega_{p,T}).
\]
By definition, \( \Sigma_{f|y} = H^{-1} \sum_{h=1}^{H} (\Lambda_h m_h)(\Lambda_h m_h)' \). For fixed \( H \), note that

\[
\hat{\Sigma}_{f|y} - H \Sigma_{f|y} H' = H^{-1} \sum_{h=1}^{H} [(\hat{\Lambda}_h \hat{m}_h)(\hat{\Lambda}_h \hat{m}_h)' - (H \Lambda_h m_h)(H \Lambda_h m_h)'],
\]

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

A direct application of \( \sin(\theta) \) theorem shows that \( ||\hat{\phi}_j - H \phi_j|| = O_p(\omega_{p,T}) \). Since we have the normalization \( \text{cov}(f_t) = I_K \) and \( E(f_t) = 0 \), the eigenvalue \( \phi_j \)'s of \( \Sigma_{f|y} \) constitute the SDR directions for model (2.1).

### 5.2.3 Proof of Proposition 2.3.1

First we write \( \hat{\phi} \) in terms of the true factors \( f_t \),

\[
\hat{\phi} = \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1} \hat{f}_t = \frac{1}{T-1} \hat{\Lambda}_b \sum_{t=1}^{T-1} y_{t+1} x_t
\]

\[
= \frac{1}{T-1} \hat{\Lambda}_b \sum_{t=1}^{T-1} (B f_t + u_t) y_{t+1},
\]

where we used the fact that \( f_t = \hat{\Lambda}_b x_t \). Using triangular inequality,

\[
||\hat{\phi} - \bar{\phi}|| = ||\hat{\phi} - (\hat{\Lambda}_b B) \phi + (\hat{\Lambda}_b B - I) \bar{\phi}||
\]

\[
\leq ||(\hat{\Lambda}_b B) \frac{1}{T-1} (\sum_{t=1}^{T-1} y_{t+1} f_t - \bar{\phi})|| + ||(\hat{\Lambda}_b B - I) \bar{\phi}|| + \frac{1}{T-1} ||\hat{\Lambda}_b \sum_{t=1}^{T-1} u_t y_{t+1}||.
\]

By lemma 5.2.3 and 5.2.4, we have \( ||\hat{\Lambda}_b B|| = O_p(1) \) and \( ||\hat{\Lambda}_b B - \Lambda_b B|| = ||\hat{\Lambda}_b B - I|| = O_p(\omega_{p,T}) \). Since \( ||\bar{\phi}|| = O_p(1) \), the second term on the right hand side of the inequality is \( O_p(\omega_{p,T}) \). For the third term, note that \( u_t \) is independent of \( y_{t+1} \), hence \( E(u_t y_{t+1}) = 0 \). By law of large numbers and \( ||\hat{\Lambda}_b|| = O_p(p^{-1/2}) \), the third term is \( O_p(T^{-1/2}) \). It remains to bound \( ||\frac{1}{T-1} (\sum_{t=1}^{T-1} y_{t+1} f_t - \bar{\phi})|| \) in the first term.
We express \( f_t \) along the basis \( \phi_1, ..., \phi_L \) and their orthogonal hyperplane,

\[
f_t = \sum_{j=1}^{L} \langle f_t, \phi_j \rangle \phi_j + f_t^\perp.
\]

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

\[
|| \frac{1}{T-1} \sum_{t=1}^{T-1} (y_{t+1} f_t - \bar{\phi}) || = || \sum_{j=1}^{L} \left[ \frac{1}{T-1} \sum_{t=1}^{T-1} (\phi'_j f_t) y_{t+1} - E((\phi'_j f_t) y_{t+1}) \right] + \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1} f_t^\perp ||
\]

\[
\leq \sum_{j=1}^{L} || \frac{1}{T-1} \sum_{t=1}^{T-1} (\phi'_j f_t) y_{t+1} - E((\phi'_j f_t) y_{t+1}) || + || \frac{1}{T-1} \sum_{t=1}^{T-1} y_{t+1} f_t^\perp ||
\]

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


Bernanke, B. S., Boivin, J. and Eliasz, P. (2005), ‘Measuring the effects of monetary pol-
icy: 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


recovery in high-dimensional and sparse settings’, *Journal of the American Statistical


**52**(1), 57–82.


Chen, C.-H. and Li, K.-C. (1998), ‘Can sir be as popular as multiple linear regression?’,

Chen, S. X. and Qin, Y.-L. (2010), ‘A two-sample test for high-dimensional data with ap-

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


Vol. 482, John Wiley & Sons.


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


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


