UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays on estimation and inference in high-dimensional models with applications to finance and economics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Management

by

Yinchu Zhu

Committee in charge:

Professor Allan Timmermann, Chair
Professor Brendan Beare
Professor Ivana Komunjer
Professor Jun Liu
Professor Rossen Valkanov

2017

The dissertation of Yinchu Zhu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

To my family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am forever indebted to Allan Timmermann, my advisor. I have been truly fortunate to have your guidance and support as I pursue my interest in econometrics and statistics as a finance student. You are an amazing mentor and constantly inspire me to grow as a scholar.

I would like to thank Ross Valkanov and Jun Liu for their support and help. I have learned a lot from our conversations. Thanks to Ivana Komunjer for being a wonderful mentor, whose advice has been invaluable to my academic development. I benefit a lot intellectually from our discussions and the experience of working together. Thanks to Brendan Beare for introducing me to the world of econometrics and for guiding me with my first research project. Thanks to Andres Santos for all the advice and discussions; thank you for not kicking me out when my 5-minute questions routinely took half of the afternoon. Thanks to Larry Schmidt for being a great friend and co-author. Thanks to Jelena Bradic, a fantastic friend, co-author and mentor. Although we have only known each other for less than two years, my research interests in high-dimensional statistics are inspired by interactions with you, mostly via our usual pattern of having multiple discussions per day. I would also like to thank faculty, graduate students and staff at UCSD for all of their help and suggestions, especially Yixiao Sun, Jim Hamilton, Kaspar Wuthrich, Patrik Guggenberger, Lei Ni, Wei Chen, Riccardo Sabbatucci, Alberto Rossi, Jungbin Hwang, Claudio Labanca and Marisol Nierva-Magnano.

Chapters 1 and 2, in full, are currently being prepared for submission for publication of the material. Zhu, Yinchu. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, is joint work with Jelena Bradic and has been submitted for publication of the material as it may appear in Zhu, Yinchu; Bradic, Jelena, Journal of the American Statistical Association, 2017. The dissertation author was the primary investigator and author of this paper.

VITA

2009	B. M in Accounting, Zhongnan University of Economics and Law, China
2011	M. S. in Finance *cum laude*, Bocconi University, Italy
2014	M. A. in Mathematics (Applied), University of California, San Diego
2017	Ph. D. in Management, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Essays on estimation and inference in high-dimensional models with applications to finance and economics

by

Yinchu Zhu

Doctor of Philosophy in Management

University of California, San Diego, 2017

Professor Allan Timmermann, Chair

Economic modeling in a data-rich environment is often challenging. To allow for enough flexibility and to model heterogeneity, models might have parameters with dimensionality growing with (or even much larger than) the sample size of the data. Learning these high-dimensional parameters requires new methodologies and theories. We consider three important high-dimensional models and propose novel methods for estimation and inference. Empirical applications in economics and finance are also studied.

In Chapter 1, we consider high-dimensional panel data models (large cross sections and long time horizons) with interactive fixed effects and allow the covariate/slope coefficients to vary over time without any restrictions. The parameter

of interest is the vector that contains all the covariate effects across time. This vector has dimensionality tending to infinity, potentially much faster than the cross-sectional sample size. We develop methods for the estimation and inference of this high-dimensional vector, i.e., the entire trajectory of time variation in covariate effects. We show that both the consistency of our estimator and the asymptotic accuracy of the proposed inference procedure hold uniformly in time. Our methodology can be applied to several important issues in econometrics, such as constructing confidence bands for the entire path of covariate coefficients across time, testing the time-invariance of slope coefficients and estimation and inference of patterns of time variations, including structural breaks and regime switching. An important feature of our method is that it provides inference procedures for the time variation in pre-specified components of slope coefficients while allowing for arbitrary time variation in other components. Computationally, our procedures do not require any numerical optimization and are very simple to implement. Monte Carlo simulations demonstrate favorable properties of our methods in finite samples. We illustrate our methods through empirical applications in finance and economics.

In Chapter 2, we consider large factor models with unobserved factors. We formalize the notion of common factors between different groups of variables and propose to use it as a general approach to study the structure of factors, i.e., which factors drive which variables. The spanning hypothesis, which states that factors driving one group are spanned by those driving another group, can be studied as a special case under our framework. We develop a statistical procedure for testing the number of common factors. Our inference procedure is built upon recent results on high-dimensional bootstrap and is shown to be valid under the asymptotic framework of large n and large T. In Monte Carlo simulations, our procedure performs well in finite samples. As an empirical application, we construct confidence sets for the number of common factors between the macroeconomy and the financial markets.

Chapter 3 is joint work with Jelena Bradic. We propose a methodology for testing linear hypothesis in high-dimensional linear models. The proposed test does not impose any restriction on the size of the model, i.e. model sparsity or

the loading vector representing the hypothesis. Providing asymptotically valid methods for testing general linear functions of the regression parameters in high-dimensions is extremely challenging – especially without making restrictive or unverifiable assumptions on the number of non-zero elements. We propose to test the moment conditions related to the newly designed restructured regression, where the inputs are transformed and augmented features. These new features incorporate the structure of the null hypothesis directly. The test statistics are constructed in such a way that lack of sparsity in the original model parameter does not present a problem for the theoretical justification of our procedures. We establish asymptotically exact control on Type I error without imposing any sparsity assumptions on model parameter or the vector representing the linear hypothesis. Our method is also shown to achieve certain optimality in detecting deviations from the null hypothesis. We demonstrate the favorable finite-sample performance of the proposed methods, via a number of numerical and a real data example.

Chapter 1

High-dimensional panel data with time heterogeneity: estimation and inference

1.1 Introduction

How heterogeneity is modeled plays a key role in many empirical studies in economics and finance. Although linear panel data models have been extensively employed to account for cross-sectional and temporal heterogeneity, such heterogeneity is usually restricted to the error terms by various specifications of fixed effects and random effects. In contrast, slope coefficients are typically assumed to be homogeneous in cross sections and over time.

Allowing for heterogeneity in both error terms and slope coefficients can be very important in applied research. For example, consider the literature on predictability of stock returns. Most work applies linear regressions of stock returns against predictors such as the lagged dividend yield.[1] Suppose that we have a panel dataset containing observations of returns and dividend yields for a large number of stocks over a long time horizon. Heterogeneity in the error terms may arise as different stocks have different sensitivities to common shocks (e.g.,

[1]See Campbell and Shiller (1988b), Keim and Stambaugh (1986), Campbell and Thompson (2008), Goyal and Welch (2003, 2008) and Rapach, Strauss, and Zhou (2010).

macroeconomic activity and market-wide shocks) and firm-specific components (e.g., firm fixed effects). Meanwhile, it is also reasonable to expect time heterogeneity in the relationship between expected stock returns and the dividend yield since the instability of this relationship is well documented in the finance literature.[2] Researchers are often interested in whether return predictability from the dividend yield is stable across time, how such predictability evolves and whether the state of the macro economy affects such predictability. For example, an important question is how macroeconomic and/or financial turmoil, such as the Great Recession, affects the predictability of stock returns. Does the Great Recession only amount to shocks in the error terms or does it fundamentally change the relationship between stock returns and dividend yields?

The main contribution of this paper is to address these types of questions using a linear panel data model with general time-heterogeneous covariate effects. Suppose that for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, we observe dependent variables $y_{i,t} \in \mathbb{R}$ and covariates/regressors $x_{i,t} \in \mathbb{R}^k$ from the following model

$$y_{i,t} = x_{i,t}'\beta_t + \alpha_{i,t} + u_{i,t}, \tag{1.1.1}$$

where $\beta_t \in \mathbb{R}^k$ is the vector containing unobserved covariate effects at time t, $\alpha_{i,t}$ is unobserved fixed effects of individual i at time t and $u_{i,t}$ is an idiosyncratic error with $\mathbb{E}u_{i,t} = 0$ and $\mathbb{E}x_{i,t}u_{i,t} = 0$. We consider the interactive fixed effects for $\alpha_{i,t}$ and impose a factor structure on the regressors (similar to Pesaran (2006)); see Section 1.2 for details. We allow for dynamic structures since components of $x_{i,s}$ can be correlated with $u_{i,t}$ for $s \neq t$.

The most important feature of our model (1.1.1) is that the covariate effects $\{\beta_t\}_{t=1}^T$ are allowed to vary across time without any restrictions. The sequence $\{\beta_t\}_{t=1}^T$ can be viewed as either a deterministic sequence or a stochastic process with arbitrary correlation with observed variables.[3] Throughout the paper, we assume that k is fixed and both n and T tend to infinity. Let β denote the high-dimensional

[2] See Paye and Timmermann (2006), Ang and Bekaert (2007), Pettenuzzo and Timmermann (2011) and Lettau and Van Nieuwerburgh (2008).

[3] See Remark 1.3.3 for more discussion.

vector representing the trajectory of β_t over time

$$\beta := (\beta_1', \ldots, \beta_T')' \in \mathbb{R}^{kT}.$$

We treat the high-dimensional vector β as the model parameter of interest and develop procedures for its estimation and inference. In particular, we propose a new methodology that can be used to construct confidence sets for β and simultaneously test multiple (or many) linear hypotheses of β.

Our general model eliminates the risk of misspecification in time-varying pattern of $\{\beta_t\}_{t=1}^T$. Time variation in model parameters has been recognized in many areas of applied research, such as macroeconomic forecasting (Stock and Watson 1996, 2007; Giacomini and Rossi 2006, 2009, 2010; Rossi 2013). Most empirical work that addresses the issue of time-varying parameters uses a random coefficient approach and assumes that parameters evolve according to a particular stochastic process.[4] However, imposing parametric or non-parametric structures introduces the risk of misspecification, which might deliver misleading or even spurious results.[5] Our proposed methodology does not require any restriction on the time variation in the slope coefficients.

Moreover, the flexibility in our setup provides a natural framework of estimating parametric specifications for the time variation in β_t and testing the validity of these specifications. For example, one of the most popular models accounting for time-varying parameters is the structural break model in which β_t is assumed to have a piecewise constant pattern across t. If the true underlying trajectory of $\{\beta_t\}_{t=1}^T$ indeed follows such a pattern, then our method can be used to estimate the number and locations of structural breaks. Testing the validity of the structural break model is also straight-forward. Under the null hypothesis of correct specification, estimates of the break points are consistent and thus the stability of β_t between two breaks can be rewritten as multiple linear hypotheses of β, which

[4]Popular specifications either impose parametric models, such as piecewise constant parameters (structural breaks), Markov chains, Bernoulli distributions, random walks, autoregressive models, or assume non-parametric time variation with smooth paths.

[5]Even the flexible non-parametric specification that assumes only smoothness in time variation can fail to capture brief temporary changes, which may be due to momentary shocks in the economy and weather.

can be tested using the proposed methodology. Similarly, our methodology can be used to estimate regime-switching models and test their specifications. If β_t is regime-dependent, then the proposed procedures can consistently recover the time series of regime membership and thus the hypothesis of homogeneity within each regime can be again formulated as multiple linear hypotheses on β. In addition, since our regime estimator does not assume any structure on the time variation of regimes, the time series of estimated regime membership can be used to test the validity of candidate specifications, e.g., whether the regimes evolve as a Markov chain (Hamilton 1989) or a Bernoulli process.

A distinct feature of the proposed method is that our results can be used for sub-vector (partial) inference allowing for flexible structures in the nuisance parameter. In practice, applied researchers are often interested in only a subset of the slope coefficients. For example, suppose $\beta_t = (\beta_{1,t}, \beta_{2,t})'$, where $\beta_{1,t}$ is of empirical interest and $\beta_{2,t}$ corresponds to a control variable. Our method can be used to test specifications of $\{\beta_{1,t}\}_{t=1}^T$ without imposing any restrictions on the time variation of the nuisance parameter $\{\beta_{2,t}\}_{t=1}^T$. Many existing specification tests, such as the popular test by Bai and Perron (1998) for structural breaks, can only handle the null hypothesis that specify the time variation in the entire $k \times 1$ vector β_t. In our example, a typical existing test for lack of structural breaks has the null hypothesis that both $\{\beta_{1,t}\}_{t=1}^T$ and $\{\beta_{2,t}\}_{t=1}^T$ are constant across time. Hence, our methodology provides specification tests that are robust to misspecifications of nuisance parameters.

Our results offer an intuitive setup to study and explain the time variation in the slope coefficients. For example, suppose that the researcher is interested in testing whether the slope coefficients vary with the business cycle. This question can be formulated in terms of the average value of β_t in economic recessions and expansions and thus can be phrased as inference of linear hypothesis of β. Alternatively, a regression-based approach can be applied. Since our methodology deliver consistent estimators for the entire path $\{\beta_t\}_{t=1}^T$, we can fit the estimated β_t in a time series regression against other explanatory variables.

Our paper contributes to econometric theories in several ways. First, we

propose a new strategy for identification and estimation, overcoming difficulties due to flexibility in fixed effects and the general specification of β. Since the fixed effects in (1.1.1) is potentially correlated with the regressors, running the ordinary least square (OLS) estimation for each t does not guarantee consistent estimation of β_t. Even under strict exogeneity of the regressors, potential cross-sectional dependence in error terms could render traditional methods invalid; see e.g., Phillips and Sul (2003), Andrews (2005) and Pesaran and Tosetti (2011). This problem is illustrated in Appendix A.1.

Second, our methodology can be used for inference on the high-dimensional vector β. To the best of our knowledge, our work is the first in the literature on panel data models to address the inference problem of the entire path of unrestricted time variation in coefficients. Although model (1.2.1) with time or individual-specific covariate effects has been studied by authors such as Pesaran (2006)[6], inference results are only available for low-dimensional components of β (e.g., β_t for a fixed t). In contrast, our results deal with inference on the entire vector β by capitalizing on recent advances in high-dimensional statistics and probability. In existing work, inference on individual β_t's is based on the classical central limit theorem (CLT). Since T tends to infinity, $\beta \in \mathbb{R}^{kT}$ is a high-dimensional object and thus the classical CLT is not suitable for our purposes. One might attempt to construct a confidence set for the whole trajectory of β_t over time from confidence sets for each β_t. However, constructing confidence bands for the whole trajectory of β_t amounts to approximating the distribution of the maximal estimation error of β_t over all $t = 1, \ldots, T$. This is not straight-forward when T tends to infinity.[7] Building upon the recent results by Chernozhukov, Chetverikov, and Kato (2013, 2014), we develop a multiplier bootstrap procedure, which is shown to be asymptotically

[6]In fact, Pesaran (2006) considers panel data models with individual-specific covariate coefficients, but his method can be applied to models with time-heterogeneity by swapping the time and individual indices.

[7]To illustrate these issues, suppose that $k = 1$ and that for each t, there is an estimator $\hat{\beta}_t$ such that $\sqrt{n}(\hat{\beta}_t - \beta_t) \to^d N(0, 1)$. Constructing a confidence band for all β_t's amounts to finding $c > 0$ such that $\mathbb{P}(\max_{1 \le t \le T} \sqrt{n}|\hat{\beta}_t - \beta_t| > c) \approx \eta$ for some pre-specified $\eta \in (0, 1)$. For large T, the difficulties of conducting inference based on existing methods arise as the validity of approximating $\sqrt{n}(\hat{\beta}_t - \beta_t)$ with Gaussian distributions might not be uniform in t and it is not straight-forward to account for the interdependence across t.

exact in terms of size control.[8]

Finally, the estimation procedures proposed in this paper are computationally simple. For high-dimensional models, computational burden is often a key concern as naively extending algorithms designed for low-dimensional problems might not be computationally feasible. As will be introduced in Section 1.2, the fixed effect assumes a factor structure $\alpha_{i,t} = L'_{\alpha,i}F_{\alpha,t}$. Then the least squared estimator minimizes $\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{i,t} - x'_{i,t}\beta_t - L'_{\alpha,i}F_{\alpha,t})^2$ over $\{\beta_t\}_{t=1}^{T}$, $\{L_{\alpha,i}\}_{i=1}^{n}$ and $\{F_{\alpha,t}\}_{t=1}^{T}$. This estimator has been applied intensively for low-dimensional problems (i.e., time-homogeneous β_t); see Bai (2009) and Moon and Weidner (2015). Since there is no closed-end solution to this optimization problem and the objective function is not jointly convex in $\{\beta_t\}_{t=1}^{T}$, $\{L_{\alpha,i}\}_{i=1}^{n}$ and $\{F_{\alpha,t}\}_{t=1}^{T}$, most numerical algorithms are not guaranteed to return the global maximizer. The usual remedy of trying many starting points is virtually infeasible since β is high-dimensional. We develop alternative identification and estimation strategies and derive procedures that only involve matrix multiplications and singular value decompositions, thereby considerably reducing the computational burden. Moreover, unlike most nonparametric methods, the methodology proposed in this paper does not require choosing any tuning parameters, except for the number of factors, which can also be consistently estimated in a manner free of tuning parameters. Our theoretical results still hold when the true number of factors are replaced by consistent estimators.

We demonstrate the advantage of the proposed methodology via three empirical studies in finance and economics. The first study is concerned with the predictability of stock returns using the lagged dividend yield and volatility as predictors. We find that the predictive power of both the dividend yield and volatility exhibits very different patterns of time variation; in particular, return predictability is linked to the macroeconomy but in different manners. We also find seasonality patterns in predictability, which is different from seasonality in the error term, often referred to as calendar effects. The second empirical study uses

[8]One may address the issue of inter-temporal dependence from the perspective of multiple testing problems and use the Bonferroni method to control the family-wise error rate. Unfortunately, this approach usually results in a great loss of power and leads to conservative tests, especially in our case where the number of tests (in the multiple testing problem) can be much larger than the sample size.

panel data on firms and focuses on the effects of several variables on firms' capital structure. We find that the patterns of time variation in the slope coefficients can be quite different from what is generated by simply applying time-homogeneous models to subsamples of the data. The third empirical application studies the effect of investment on economic growth. Using a multi-country panel dataset, we find strong evidence of time variation in this effect. Our methodology also finds group patterns in the fixed effects, suggesting that developing and developed countries have different trends that are likely to be driven by the same factor but with different factor loadings. In all these studies, we find that time heterogeneity in the slope coefficients exists and displays complicated patterns that are difficult to capture by parametric models. Since no restrictions are imposed on the time heterogeneity in β_t, our findings are not subject to the misspecification risk in this regard.

Related literature

Our work builds upon the literature on large dynamic panel data models with fixed effects. The asymptotic framework in this literature allows both n and T to tend to infinity. The most common specification for the fixed effects is time-invariant individual-specific fixed effects (sometimes plus a time-specific component), e.g., see Phillips and Moon (1999), Hahn and Kuersteiner (2002), Alvarez and Arellano (2003) and Hahn and Moon (2006). Bonhomme and Manresa (2015) propose a structure under which individuals are classified into several groups and the fixed effects are allowed to have unconstrained time variations but are homogeneous among individuals in the same group. Factor structures in fixed effects have also been considered, e.g., Andrews (2005), Bai (2009), Ahn, Lee, and Schmidt (2013), Su, Jin, and Zhang (2015) and Moon and Weidner (2015).

Although most empirical work that uses panel data models assumes homogeneous covariate effects, numerous authors, such as Phillips and Sul (2003), Pesaran and Yamagata (2008) and Su and Chen (2013), have developed tests for assessing the reasonableness of this popular specification. In addition, the literature has seen work that directly considers models with heterogeneous slope coefficients. Just

as the heterogeneity in the error terms can be treated as fixed or random effects, counterparts of these two approaches are also found in the study of heterogeneity in covariate effects. Under one approach, slope coefficients for different i and/or t are viewed as fixed parameters to be estimated, see e.g., Pesaran (2006), Zaffaroni (2009) and Lin and Ng (2012); under the other approach, the slope coefficients are assumed to be random variables generated from parametric models and the focus is the estimation and inference of these parametric models, see e.g., Swamy (1970), Rosenberg (1972) and Hsiao, Appelbe, and Dineen (1993). Beyond the usual parametric/linear specification, several authors study nonparametric estimation and inference for heterogeneous covariate effects; see Qian and Wang (2012), Chen, Gao, and Li (2013) and Boneva, Linton, and Vogt (2015). An excellent survey for heterogeneous parameters in panel data models can be found in Chapter 6 of Hsiao (2014). Existing results on estimation and inference mainly focus on the average (across i or t) covariate effects and pointwise covariate effects (for given i or t).

An interesting paper by Freyberger (2012) considers heterogeneous nonparametric panel data models with interactive fixed effects. He treats the factor loadings as random variables and exploits their distributional properties to achieve nonparametric identification; the estimation strategy relies on the assumption that distribution of observables are identical in the cross section. In contrast, our result focuses on the linear models but can deal with non-random factor loadings and heterogeneous distributions across units.

Our specification of β falls into the category of high-dimensional models . Our theoretical results are based on the recent advances by Chernozhukov, Chetverikov, and Kato (2013, 2014) on high-dimensional central limit theorems and bootstrap. To handle the high-dimensional nuisance parameter (fixed effects), we borrow tools from random matrix theory, see Vershynin (2010), and the literature on large factor models, see e.g., Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (2002b), Bai and Ng (2002) and Bai (2003).

Organization of the paper and notations

The rest of the paper is organized as follows. The formal setup of our model is introduced in Section 1.2. We provide the details of the main results in Section 1.3. In Section 1.4, we discuss several related econometric problems. Finite-sample properties of our procedures are demonstrated via Monte Carlo simulations in Section 1.5. We apply our methods to several empirical studies in Section 1.6. The appendix contains the proofs of theoretical results.

For any vector $x = (x_1, \ldots, x_{n_1})' \in \mathbb{R}^{n_1}$, $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2} = \sqrt{x'x}$, $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_\infty = \max_{1\le i\le n_1} |x_i|$ and $\|x\|_0$ denotes the number of nonzero entries in x. For any matrix $A \in \mathbb{R}^{n_1\times n_2}$, $\|A\|$ denotes the spectral norm of A and we say that $A = U_A S_A V_A'$ is a singular value decomposition (SVD) if $U_A \in \mathbb{R}^{n_1\times n_1}$ and $V_A \in \mathbb{R}^{n_2\times n_2}$ are both orthogonal matrices and $S_A \in \mathbb{R}^{n_1\times n_2}$ is a (rectangular) diagonal matrix with singular values of A on the diagonal in the non-increasing order. We also introduce the low rank approximation operator: for a non-negative integer r, define $\mathcal{T}_r(A) := U_A \bar{S}_r V_A'$, where $A = U_A S_A V_A'$ is an SVD and $\bar{S}_r$ is equal to S_A with all the diagonal entries of S_A set to zero except the first r diagonal entries. $s_j(A)$ denotes the jth largest singular value of A, counting multiplicity. For two positive sequences a_n and b_n, we use $a_n \asymp b_n$ to denote the condition that there exist constant $c_1, c_2 > 0$ such that $c_1 a_n \le b_n \le c_2 a_n$. We use $\sigma(\cdot)$ to denote the σ-algebra generated by random variables.

1.2 Model Setup and Assumptions

Suppose that for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, we observe dependent variables $y_{i,t} \in \mathbb{R}$ and covariates/regressors $x_{i,t} \in \mathbb{R}^k$ from the following model

$$y_{i,t} = x_{i,t}'\beta_t + \alpha_{i,t} + u_{i,t} \qquad \text{with} \qquad \alpha_{i,t} = F_{\alpha,t}' L_{\alpha,i}, \tag{1.2.1}$$

where $\beta_t \in \mathbb{R}^k$ is the vector containing unobserved covariate effects at time t, $\alpha_{i,t}$ is unobserved fixed effects of individual i at time t with $F_{\alpha,t} \in \mathbb{R}^{r_\alpha}$ and $L_{\alpha,i} \in \mathbb{R}^{r_\alpha}$ being the unobserved factor and its loading, and $u_{i,t}$ is an idiosyncratic error with

$\mathbb{E}u_{i,t} = 0$ and $\mathbb{E}x_{i,t}u_{i,t} = 0$. We assume that k, r_α and r_Q are fixed and $T = T_n \to \infty$ as $n \to \infty$.

To achieve identification in this general model, we introduce assumptions on the regressors. Similar to Pesaran (2006), we assume a factor structure

$$x_{i,t} = Q_{i,t} + v_{i,t} \qquad \text{with} \qquad Q_{i,t} = F'_{Q,t}L_{Q,i}, \tag{1.2.2}$$

where $F_{Q,t} \in \mathbb{R}^{r_Q \times k}$ and $L_{Q,i} \in \mathbb{R}^{r_Q}$ are unobserved factors and their loadings, r_Q is fixed and $v_{i,t} \in \mathbb{R}^k$ is the idiosyncratic errors. Arbitrary correlations between $\{F_{Q,t}\}_{t=1}^T$ and $\{F_{\alpha,t}\}_{t=1}^T$ are permitted. The model (1.2.2) can be justified in many applications. Factor structures have been motivated on both theoretical and empirical grounds and have been widely used to model financial and macroeconomic data[9], to account for unobserved abilities (e.g., Lord, Novick, and Birnbaum (1968), Hansen, Heckman, and Mullen (2004)) and to study consumer theory (e.g., Gorman (1981) and Lewbel (1991)).

The factor structure in $\alpha_{i,t}$, often referred to as interactive fixed effects, allows for a rich class of unobserved common effects and nests popular fixed effects models as special cases, see Bai (2009). The interactive fixed effects also allow for flexible cross-sectional and inter-temporal dependence among the regression residuals $\alpha_{i,t} + u_{i,t}$, see e.g., Andrews (2005) and Pesaran (2006).

The goal of this paper is to build a confidence set for $\beta \in \mathbb{R}^{kT}$ (a confidence band for β_t that is uniformly valid over t) and test hypotheses of the form

$$H_0 : \ J\beta = a, \tag{1.2.3}$$

where $J \in \mathbb{R}^{m_J \times kT}$ and $a \in \mathbb{R}^{m_J}$ are nonrandom and m_J can be as large as $O(n^l)$ for some constant $0 \leq l < \infty$.

We introduce the following definition, which is satisfied by a large class of random variables including polynomials of sub-Gaussian random variables as well as finite mixtures of random variables with thin-tailed distributions.

[9]See e.g., Ross (1976), Campbell, Lo, and MacKinlay (1997), Fama and French (1992, 2016), Ludvigson and Ng (2007), Forni and Lippi (1997), Stock and Watson (1998, 2002b, 2006)

Definition 1.2.1. A random variable Z is said to have an exponential-type tail with parameter (b, γ) if $\forall z > 0$, $\mathbb{P}(|Z| > z) \leq \exp\left[1 - (z/b)^{\gamma}\right]$.

We impose the following conditions for model (1.2.1) and (1.2.2).

Assumption 1. *Assume that the following hold:*

(i) There exist constants $b_*, \gamma_* > 0$ such that $\forall (i,t) \in \{1,\ldots,n\} \times \{1,\ldots,T\}$, each entry of $F_{\alpha,t}$, $L_{\alpha,i}$, $F_{Q,t}$, $L_{Q,i}$, $u_{i,t}$ and $v_{i,t}$ has an exponential-type tail with parameter (b_*, γ_*).

(ii) There exist constants $c_*, \gamma_{**} > 0$ such that $\alpha_{mixing}(t) \leq c_* \exp(-t^{\gamma_{**}})$ $\forall t \geq 1$, where

$$\alpha_{mixing}(t) := \sup\Big\{ |\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \bigcap B)| : A \in \sigma\left(\{(F_{Q,s}, F_{\alpha,s}, v_s, u_s) : s \leq \tau\}\right), B \in \sigma\left(\{(F_{Q,s}, F_{\alpha,s}, v_s, u_s) : s \geq \tau + t\}\right) \text{ and } \tau \in \mathbb{Z} \Big\}.$$

(iii) There exist constants $\kappa_1, \kappa_2 > 0$ and $\xi \in (6/7, 2)$ such that $\kappa_1 n^{\xi} \leq T \leq \kappa_2 n^{\xi}$.

(iv) There exist constants $C_1, C_2 > 0$ such that, with probability approaching one, all the eigenvalues of $n^{-1}L_Q'L_Q$, $T^{-1}F_Q'F_Q$, $n^{-1}L_\alpha'L_\alpha$ and $T^{-1}F_\alpha'F_\alpha$ lie in $[C_1, C_2]$.

(v) $\{(\underline{v}_i, \underline{u}_i)\}_{i=1}^n$ is independent across i, where $\underline{v}_i = (v_{i,1}, \ldots, v_{i,T})' \in \mathbb{R}^{T \times k}$ and $\underline{u}_i = (u_{i,1}, \ldots, u_{i,T})' \in \mathbb{R}^T$.

(vi) $\{u, v\}$ is independent of $\{L_Q, F_Q, L_\alpha, F_\alpha\}$ and $\forall i, t$, $\mathbb{E}v_{i,t}u_{i,t} = 0$.

(vii) There exists a constant $C_5 > 0$ such that $\min_{1 \leq t \leq T} s_k\left(n^{-1}\sum_{i=1}^n \mathbb{E}v_{i,t}v_{i,t}'\right) > C_5$.

Assumption 1(i) and (ii) enable us to apply large deviation theory, which is convenient in deriving bounds for the maximum of a large number of sums of random variables. Assumption 1(i) allows for thicker tails than the Gaussian and exponential distribution, although it rules out fat-tailed distributions such as

student t distribution or the stationary distribution of GARCH processes. However, in Monte Carlo simulations, our procedure performs well with these fat-tailed distributions. With more careful arguments, it is possible that we can invoke the moderate deviation theory for self-normalized sums, such as Chen, Shao, and Wu (2016), and replace the exponential-type tails in Assumption 1(i) with bounded moment conditions. Assumption 1(ii) allows weak dependence across t and is satisfied in many situations.[10] Assumption 1(i) and (ii) are also imposed by Bonhomme and Manresa (2015) in their Assumption 2.

Assumption 1(iii) specifies the relative magnitude between n and T. Recent literature on dynamic panel data models considers three cases of sample size: $n/T \to 0$, $T/n \to 0$ and $n \asymp T$; see Hahn and Kuersteiner (2002), Moon and Phillips (2004), Arellano and Hahn (2007) and Bai (2009) among many others. We allow for all these three cases, which correspond to $\xi < 1$, $\xi > 1$ and $\xi = 0$, respectively. Assumption 1(iv) assumes strong factors in α and Q and is a standard condition in the large factor model literature; see Bai and Ng (2002), Bai (2003, 2009) and Moon and Weidner (2015).

Assumption 1(v) and (vi) say that the idiosyncratic terms are independent across i and are independent of the factors and their loadings. Similar conditions are routinely imposed in the literature on large factor models, e.g., see Bai (2003) and Bai and Ng (2006a). Notice that Assumption 1(v) and (vi) still allow for arbitrary dependence across i for $L_{\alpha,i}$ and $L_{Q,i}$, as well as serial dependence within u, v, F_α and F_Q. Contemporaneous exogeneity of $v_{i,t}$ in Assumption 1(vi) is required for the identification of β_t. Heteroskedasticity is also allowed in $v_{i,t}$ and $u_{i,t}$ under Assumption 1. Finally, Assumption 1(vii) rules out asymptotically vanishing variances in the idiosyncratic terms of the regressors.

We now demonstrate Assumption 1 with a concrete example.

Example 1.2.1 (Time-heterogeneous dynamic panel data model)**.** Let $y_{i,t} = L'_{\alpha,i}(\sum_{j=0}^{\infty} \gamma_{t,j} F_{\alpha,t-j}) + \sum_{j=0}^{\infty} \gamma_{t,j} u_{i,t-j}$, where $\gamma_{t,j}$ is defined as the following: $\gamma_{t,0} = 1$, $\gamma_{t,j} = \Pi_{l=1}^{j} \beta_{t-l+1}$ for $j > 0$ and $\gamma_{t,j} = 0$ for $j < 0$. For simplicity, let $u_{i,t}, F_{\alpha,t}, L_{\alpha,i} \sim$

[10] For linear processes and GARCH processes, see Gorodetskii (1978) and Carrasco and Chen (2002). For Markov processes, one can actually show geometric decay of β-mixing coefficients using the so-called V-ergodicity property; see Meyn and Tweedie (2012).

$i.i.d$ $N(0,1)$ and assume that $\sup_{t\geq 0}|\beta_t| \leq c$ for some constant $c \in (0,1)$. Then one can easily verify that $y_{i,t}$ defined above satisfies

$$y_{i,t} = L'_{\alpha,i}F_{\alpha,t} + \beta_t y_{i,t-1} + u_{i,t}.$$

Thus, in the notations of (1.2.1) and (1.2.2), $x_{i,t} = y_{i,t-1}$, $L_{Q,i} = L_{\alpha,i}$, $F_{Q,t} = \sum_{j=0}^{\infty}\gamma_{t-1,j}F_{\alpha,t-1-j}$ and $v_{i,t} = \sum_{j=0}^{\infty}\gamma_{t-1,j}u_{i,t-1-j}$. Assumption 1(i) holds by the Gaussianity and Assumptions 1(iii)-(vii) obviously hold. In Lemma A.2.16 of Appendix A.2.4, we show that Assumption 1(ii) also holds.

1.3 Main Results

In this section, we present the main results for estimation and inference of $\{\beta_t\}_{t=1}^T$. In Section 1.3.1, we discuss the key idea behind our identification strategy. Sections 1.3.2 and 1.3.3 develop the main methodology for estimation and inference and establish theoretical properties of the proposed procedures. Section 1.3.4 deals with the issue of determining the number of factors.

1.3.1 Identification strategy

Given the model (1.2.1) and (1.2.2), our estimation strategy is based on the following observation:

$$y_{i,t} = x'_{i,t}\beta_t + \alpha_{i,t} + u_{i,t} = v'_{i,t}\beta_t + \left(Q'_{i,t}\beta_t + \alpha_{i,t} + u_{i,t}\right).$$

We shall assume that $v_{i,t}$ is uncorrelated with $Q_{i,t}$, $\alpha_{i,t}$ and $u_{i,t}$. Therefore, at time t, we can view $Q'_{i,t}\beta_t + \alpha_{i,t} + u_{i,t}$ as the error term and simply use the cross-sectional variation to identify β_t:

$$\beta_t = \left(\sum_{i=1}^{n}\mathbb{E}v_{i,t}v'_{i,t}\right)^{-1}\left(\sum_{i=1}^{n}\mathbb{E}v_{i,t}y_{i,t}\right).$$

In other words, for each t, we run a cross-sectional regression of $y_{i,t}$ against

$v_{i,t}$. Notice that $v_{i,t}$ is unobserved. To make this approach feasible, we exploit the factor structure (1.2.2) again and employ the technique of principal component analysis (PCA) to identify $v_{i,t}$.

Remark 1.3.1. Pesaran (2006) proposes the common correlated effect estimator (CCE), which can be adapted to our model. The strategy is the following. If $L_Q := (L_{Q,1}, \ldots, L_{Q,n})' \in \mathbb{R}^{n \times r_Q}$ were observed, then $v_{i,t}$ could be estimated as the residuals from projecting columns $X_t = (x_{1,t}, \ldots, x_{n,t})' \in \mathbb{R}^{n \times k}$ onto L_Q; since we do not observe L_Q, we need to replace it with an observed matrix $\tilde{L}$. Therefore, the plan is (1) to construct $\tilde{L}$ whose columns span a space that approximately contains columns of L_Q and (2) to take as estimates of $\{v_{i,t}\}_{i=1}^n$ the residuals of projecting columns of X_t onto $\tilde{L}$. To illustrate the idea of CCE, consider $\tilde{L} = (\bar{x}_{(1)}, \ldots, \bar{x}_{(n)})' \in \mathbb{R}^{n \times k}$ with $\bar{x}_{(i)} = T^{-1} \sum_{t=1}^T x_{i,t}$. Notice that under the specification (1.2.2), if we assume that the law of large numbers (LLN) applies across t, then $\bar{x}_{(i)} = A_T L_{Q,i} + T^{-1} \sum_{t=1}^T v_{i,t} \approx A_T L_{Q,i}$, where $A_T = T^{-1} \sum_{t=1}^T F'_{Q,t} \in \mathbb{R}^{k \times r_Q}$. Ignoring the approximation error due to LLN, we have $\tilde{L} = L_Q A'_T$. Then columns of $\tilde{L}$ span a space that contains columns of L_Q if and only if rank$A_T = r_Q$, which, in Pesaran (2006), is referred to as the rank condition. A necessary condition for the rank condition is $k \geq r_Q$, which may or may not hold in practice. In contrast, our method uses PCA and does not require this rank condition.

We now introduce some notations that will be used in the rest of the paper: $Y = [Y_1, \ldots, Y_T] \in \mathbb{R}^{n \times T}$, $X = [X_1, \ldots, X_T] \in \mathbb{R}^{n \times kT}$, $\alpha = [\alpha_1, \ldots, \alpha_T] \in \mathbb{R}^{n \times T}$, $u = [u_1, \ldots, u_T] \in \mathbb{R}^{n \times T}$, $v = [v_1, \ldots, v_T] \in \mathbb{R}^{n \times kT}$, $Q = [Q_1, \ldots, Q_T] \in \mathbb{R}^{n \times kT}$, $F_\alpha = [F_{\alpha,1}, \ldots, F_{\alpha,T}]' \in \mathbb{R}^{T \times r_\alpha}$, $F_Q = [F_{Q,1}, \ldots, F_{Q,T}]' \in \mathbb{R}^{kT \times r_Q}$ $L_Q = (L_{Q,1}, \ldots, L_{Q,n})' \in \mathbb{R}^{n \times r_Q}$ and $L_\alpha = (L_{\alpha,1}, \ldots, L_{\alpha,n})' \in \mathbb{R}^{n \times r_\alpha}$, where $y_t = (y_{1,t}, \ldots, y_{n,t})' \in \mathbb{R}^n$, $X_t = (x_{1,t}, \ldots, x_{n,t})' \in \mathbb{R}^{n \times k}$, $\alpha_t = (\alpha_{1,t}, \ldots, \alpha_{n,t})' \in \mathbb{R}^n$, $u_t = (u_{1,t}, \ldots, u_{n,t})' \in \mathbb{R}^n$, $v_t = (v_{1,t}, \ldots, v_{n,t})' \in \mathbb{R}^{n \times k}$ and $Q_t = (Q_{1,t}, \ldots, Q_{n,t})' \in \mathbb{R}^{n \times k}$. Notice that $Q = L_Q F'_Q$ and $\alpha = L_\alpha F'_\alpha$.

1.3.2 Estimation of β

For now, we assume that the values of r_Q and r_α are known and we will provide consistent estimators for r_Q and r_α later in Section 1.3.4. Since v_t is

unknown, we first estimate it and use the estimated v_t to obtain an initial estimator for β_t. We define

$$\hat{\beta}_t = (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'Y_t, \tag{1.3.1}$$

where $\hat{Q} = [\hat{Q}_1, \ldots, \hat{Q}_T] = \mathcal{T}_{r_Q}(X)$ and $\hat{v} = [\hat{v}_1, \ldots, \hat{v}_T] = X - \hat{Q}$. The following result establishes the theoretical properties of the above estimator.

Theorem 1.3.1 (Uniform estimation of β)**.** *Under Assumption 1, we have*

$$\|\hat{\beta} - \beta\|_\infty = O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right]\log^{c_0} n\right),$$

where $c_0 > 0$ is a constant and $\hat{\beta} := (\hat{\beta}_1', \ldots, \hat{\beta}_T')' \in \mathbb{R}^{kT}$ with $\hat{\beta}_t$ defined in (1.3.1).

This result says that $\hat{\beta}_t$ is a consistent estimator for β_t uniformly over t and the rate of convergence depends on the relative size of n and T. If $\xi \geq 1$ ($n/T = O(1)$), then the convergence rate is the parametric rate up to a logarithm factor, $n^{-1/2}\log^{c_0} n$. The logarithm factor is the price we pay for the high dimensionality of β and is common in the literature on high-dimensional statistics.[11] The exact value of c_0 is not important for our purposes. If $\xi < 1$ (n much larger than T), then the rate of convergence is strictly slower than $n^{-1/2}\log^{c_0} n$.

It turns out that the non-standard rate of convergence of $\hat{\beta}$ is due to the bias in the estimator; we now show that once the bias is removed, the rate of convergence in ℓ_∞-norm is $\sqrt{n^{-1}\log n}$. Notice that by the properties of SVD, $\hat{Q}_t'\hat{v}_t = 0$. Thus, it is not hard to see that

$$\sqrt{n}(\hat{\beta}_t - \beta_t) = (n^{-1}\hat{v}_t'\hat{v}_t)^{-1}n^{-1/2}\hat{v}_t'(\alpha_t + u_t). \tag{1.3.2}$$

Our strategy is to remove the effect of $n^{-1/2}\hat{v}_t'\alpha_t$ by subtracting $(\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'\hat{\alpha}_t$ from $\hat{\beta}_t$, where $\hat{\alpha}_t$ is an estimator for α_t such that $n^{-1/2}\max_{1\leq t\leq T}\|\hat{v}_t'\alpha_t - \hat{v}_t'\hat{\alpha}_t\| = o_P(1)$. As we shall show, this is can be done in an intuitive manner. Since $\hat{\beta}_t$ is a consistent estimator for β_t, $y_t - X_t\hat{\beta}_t = \alpha_t + u_t + X_t(\beta_t - \hat{\beta}_t)$ is a consistent

[11] For example, see Bickel, Ritov, and Tsybakov (2009), Bühlmann and Van De Geer (2011) and Belloni and Chernozhukov (2011).

estimator for $\alpha_t + u_t$. Heuristically speaking, we have a consistent estimator for $\alpha + u$ and can simply apply PCA again to obtain an estimator for α.

Algorithm 1. *Implement the following steps:*

1. *Compute* $[\hat{\alpha}_1, \ldots, \hat{\alpha}_T] = \mathcal{T}_{r_\alpha}([y_1 - X_1\hat{\beta}_1, \ldots, y_T - X_T\hat{\beta}_T])$*, where* $\hat{\beta}_t$ *is defined in (1.3.1).*

2. *Compute* $\tilde{\beta}_t = \hat{\beta}_t - (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'\hat{\alpha}_t$.

The following result establishes the rate of convergence for the estimator in Algorithm 1.

Theorem 1.3.2. *Under Assumption 1, we have*

$$\|\tilde{\beta} - \beta\|_\infty = O_P\left(\sqrt{n^{-1}\log n}\right),$$

where $c_0 > 0$ *is a constant and* $\tilde{\beta} := (\tilde{\beta}_1', \ldots, \tilde{\beta}_T')' \in \mathbb{R}^{kT}$ *with* $\tilde{\beta}_t$ *defined in Algorithm 1.*

A comparison between Theorems 1.3.1 and 1.3.2 demonstrates the advantage of bias correction. When $\xi < 1$ (i.e., $n/T \to \infty$), $\tilde{\beta}$ is a strictly better estimator than $\hat{\beta}$ in terms of the rate of convergence in the ℓ_∞-norm; when $\xi \geq 1$ (i.e., $n = O(T)$), $\tilde{\beta}$ and $\hat{\beta}$ have the same rates of convergence up to logarithm factors.

1.3.3 Inference on β

Now we turn to the problem of testing high-dimensional linear combinations of β in the form (3.1.2). The idea is to approximate $\tilde{\beta}_t$ with an average of independent high-dimensional vectors. Let $G_i = (G_{i,1}', ..., G_{i,T}')' \in \mathbb{R}^{kT}$ with $G_{i,t} = \Sigma_t^{-1} v_{i,t} u_{i,t}$ and $\Sigma_t = n^{-1}\sum_{i=1}^n Ev_{i,t}v_{i,t}'$. We show, in the appendix, that

$$\left\| J\tilde{\beta} - J\beta - n^{-1}\sum_{i=1}^n JG_i \right\|_\infty = O_P(n^{-1/2-c})$$

for some constant $c > 0$, where $\tilde{\beta} = (\tilde{\beta}_1', \ldots, \tilde{\beta}_T')'$. The above display suggests the "obvious" strategy of approximating the distribution of $\sqrt{n}\|J\tilde{\beta} - J\beta\|_\infty$ by that of

$\|N(0,\Omega)\|_\infty$, where $\Omega = n^{-1}\sum_{i=1}^n E(JG_iG_i'J')$. Since Ω is unknown, we replace it with a plug-in estimator. We will show that this intuitive approach can be justified even if the dimension of Ω is much larger than n and T.

To simplify the presentation, we introduce the following notation. For a random vector $Z \sim N(0,\Sigma)$, we define $\Phi(z,\Sigma) = \mathbb{P}(\|Z\|_\infty \leq z)$ and denote by $\Phi^{-1}(\cdot,\Sigma)$ the inverse of $\Phi(z,\Sigma)$ as a function of z. For a given Σ, the function $\Phi^{-1}(\cdot,\Sigma)$ can be easily computed by simulation. Our inference procedure for testing H_0 in (3.1.2) can be formally summarized as follows.

Algorithm 2. *For a test for H_0 (3.1.2) with nominal size $\eta \in (0,1)$, implement the following steps:*

1. *Compute $\hat{u}_t = y_t - X_t\hat{\beta}_t - \hat{\alpha}_t$, where $\hat{\beta}_t$ and $\hat{\alpha}_t$ are defined in (1.3.1) and Algorithm 1, respectively.*

2. *Comupute $\hat{G}_i = (\hat{G}_{i,1}', \ldots, \hat{G}_{i,T}')' \in \mathbb{R}^{kT}$, where $\hat{G}_{i,t} = \hat{\tilde{v}}_{i,t}\hat{u}_{i,t}$, $\hat{\tilde{v}}_{i,t} = \hat{\Sigma}_t^{-1}\hat{v}_{i,t}$ and $\hat{\Sigma}_t = n^{-1}\hat{v}_t'\hat{v}_t$ with $\hat{v}_t$ defined in (1.3.1).*

3. *Generate $\{\zeta_i\}_{i=1}^n$ i.i.d $N(0,1)$ independent of the sample and compute $n^{-1/2}\sum_{i=1}^n J\hat{G}_i\zeta_i$.*

4. *Repeat the previous step as many times as computationally convenient to compute $\Phi^{-1}(1-\eta,\hat{\Omega})$, where $\hat{\Omega} = n^{-1}\sum_{i=1}^n J\hat{G}_i\hat{G}_i'J'$.*

5. *Reject H_0 in (3.1.2) if and only if $\|J\tilde{\beta} - a\|_\infty > \Phi^{-1}(1-\eta,\hat{\Omega})$, where $\tilde{\beta} = (\tilde{\beta}_1', \ldots, \tilde{\beta}_T')'$ and $\tilde{\beta}_t$ is defined in Algorithm 1.*

Although the parameter of interest β is high-dimensional, we establish the validity of such procedures for our problem using recent tools developed by Chernozhukov, Chetverikov, and Kato (2013). Even in light of their results, we still need to deal with the technical challenges arising due to the facts that $\tilde{\beta}$ is not exactly the mean of independent vectors and that the large-sample behavior of $\tilde{\beta}$ depends on the residuals, such as $u_{i,t}$, which are not observed and need to be replaced with estimates for the bootstrap procedure to be feasible. To justify Algorithm 2, we need some restrictions on J.

Assumption 2. *Assume that the following conditions hold for J in* (3.1.2)*:*

(i) $m_J = O(n^l)$ *for some constant* $0 \leq l < \infty$.

(ii) There exists a constant $A_1 > 0$ *such that* $\max_{1 \leq j \leq m_J} \|J_j\|_1 \leq A_1$, *where* J_j *is the transpose of the* j*th row of* J.

(iii) There exists a constant $b_1 > 0$ *such that* $J_j' \left(n^{-1} \sum_{i=1}^n \mathbb{E} G_i G_i'\right) J_j \geq b_1 \ \forall j \in \{1, \ldots, m_J\}$, *where* $G_i = (G_{i,1}', \ldots, G_{i,T}')' \in \mathbb{R}^{kT}$, $G_{i,t} = \bar{v}_{i,t} u_{i,t}$, $\bar{v}_{i,t} = \Sigma_t^{-1} v_{t,i}$ *and* $\Sigma_t = n^{-1} \sum_{i=1}^n \mathbb{E} v_{i,t} v_{i,t}'$.

Assumption 2(i) allows us to test m_J linear transformations of β, where m_J can be fixed or grow polynomially fast in n. Notice that this allows for $m_J \gg \max\{n, T\}$. Building a confidence set for all the entries of β implies that $m_J = kT = O(n)$; inference on the individual β_t or on the average of β_t over t corresponds to a fixed m_J. Assumption 2(ii) can be viewed as a "near-sparsity" assumption on the rows of J, while it still allows $\|J_j\|_0 = kT \ \forall 1 \leq j \leq m_J$. This is needed to control the bias of $J\tilde{\beta}$: although the maximal bias of all $\tilde{\beta}_t$'s can be shown to be small, the bias of each row of $J\tilde{\beta}$ is a linear combination of all the biases of $\tilde{\beta}_t$'s. Assumption 2(ii) allows us to control the bias of $J\tilde{\beta}$ via Holder's inequality. Assumption 2(iii) rules out "degenerate" linear combinations of G_i. This is needed for the theory of high-dimensional bootstrap.

The following theorem is our main theoretical result and establishes the validity of Algorithm 2.

Theorem 1.3.3 (High-dimensional inference)**.** *Under Assumptions 1 and 2, we have*

$$\limsup_{n \to \infty} \sup_{\eta \in (0,1)} \left| \mathbb{P}\left(\sqrt{n} \|J\tilde{\beta} - J\beta\|_\infty > \Phi^{-1}(1 - \eta, \hat{\Omega}) \right) - \eta \right| = 0,$$

where $\tilde{\beta}$ *and* $\Phi^{-1}(1 - \eta, \hat{\Omega})$ *are defined in Algorithm 2.*

As our main result for inference, Theorem 1.3.3 says that Algorithm 2 can be used to test hypotheses that involve m_J linear combinations of a kT-dimensional vector, where both m_J and T can grow polynomial fast with n. We can easily invert the test to obtain confidence sets for $J\beta$.

Corollary 1.3.1. *Let Assumptions 1 and 2 hold. For any fixed* $\eta \in (0,1)$*, let* $\Phi^{-1}(1-\eta, \hat{\Omega})$ *and* $\tilde{\beta}$ *be defined as in Algorithm 2. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(J\beta \in \mathcal{C}_{1-\eta}(J)\right) = 1-\eta,$$

where $\mathcal{C}_{1-\eta}(J) = \left\{ J\tilde{\beta} + v \,\middle|\, v \in \mathbb{R}^{m_J} \text{ and } \|v\|_\infty \leq \Phi^{-1}(1-\eta, \hat{\Omega})\}/\sqrt{n} \right\}$.

In the following result, we show that the width of the above confidence set is $O_P\left(\sqrt{n^{-1}\log n}\right)$.

Theorem 1.3.4. *Suppose that Assumptions 1 and 2 hold. Then there exists a constant* $M > 0$ *such that,* $\forall \eta \in (0,1)$, $\mathbb{P}\left(\Phi^{-1}(1-\eta, \hat{\Omega}) \leq M\sqrt{\log n}\right) \to 1$.

When the entries of $n^{-1/2}\sum_{i=1}^n JG_i$ have no correlation among each other, it can be shown that there exists a constant $M_0 > 0$ such that $\mathbb{P}(\Phi^{-1}(1-\eta, \hat{\Omega}) \geq M_0\sqrt{\log n}) \to 1$. This highlights the nonstandard nature of the problem as $\sqrt{n}\|J\tilde{\beta} - J\beta\|_\infty$ might not have a well-defined limiting distribution. Under certain conditions that guarantee weak dependence among entries of JG_i, one can employ tools from the extreme value theory and obtain a well-defined asymptotic distribution for a properly scaled version of $\|J\tilde{\beta} - J\beta||_\infty$, such as $n\|J\tilde{\beta} - J\beta\|_\infty^2 + A_1 \log n + A_2 \log\log n$ with constants $A_1, A_2 \in \mathbb{R}$; see Cai and Jiang (2011). Although this alternative approach can provide a procedure with analytical critical values, it might require additional assumptions as well as very different theoretical techniques; we shall leave this possibility to future research.

Remark 1.3.2. Notice that the confidence set $\mathcal{C}_{1-\eta}(J)$ defined in Corollary 1.3.1 is a rectangle in $\mathbb{R}^{m_J}$, similar in spirit to a Kolmogorov-Smirnov-type test. One might wonder whether it is possible to build Cramer-von-Mises-type tests by considering $\|J\tilde{\beta} - J\beta\|_2$. Unfortunately, this is technically challenging since the tools from probability theories still appear inadequate in handling the ℓ_2-norm of the sum of independent high-dimensional vectors. To the best of our knowledge, existing tools can only handle the problems in which the dimensionality is much smaller than the sample size, e.g., $T = o(n^{1/4})$; see Peng and Schick (2012) and Pouzo (2015).

Remark 1.3.3. Our results are easy to understand when we treat $\{\beta_t\}_{t=1}^T$ as a deterministic sequence. However, it is worth pointing out that all the theoretical results so far still hold even if $\{\beta_t\}_{t=1}^T$ is stochastic process that is allowed to have arbitrary correlation with the observed variables Y and X. To see how this flexibility is possible, notice that deriving (1.3.2) the estimation error of the original estimator $\hat{\beta}$ is merely algebraic computations and does not require any knowledge of randomness in the data. Moreover, subsequent analysis used to derive Theorems 1.3.1, 1.3.2, 1.3.3 and 1.3.4 only involves properties of the factors, factor loadings and the idiosyncratic terms. Therefore, the estimation error of $\tilde{\beta}$ still decays at the rate $O_P(\sqrt{n^{-1}\log n})$ in ℓ_∞-norm and $\mathcal{C}_{1-\eta}(J)$ still contains the (random) vector $J\beta$ with probability approaching $1-\eta$. When β is random, the object of interest is typically parameters governing the randomness of β. In Sections 1.4.5 and 1.4.7, we illustrate how our results can be used for this purpose.

Sometimes, an applied researcher might be interested in the fixed effects. For example, when the fixed effects are assumed to be group-specific, see Bonhomme and Manresa (2015), consistent estimators for the fixed effects can be used to determine the group membership via k-means clustering (Forgy 1965; Lloyd 1982). The following result says that the fixed effects can be consistently estimated uniformly over i and t.

Theorem 1.3.5 (Uniform estimation of fixed effects)**.** *Under Assumption 1, for $\hat{\alpha}_{i,t}$ defined in Algorithm 1, we have that for some constant $c_0 > 0$,*

$$\max_{1\le i\le n,\ 1\le t\le T} |\hat{\alpha}_{i,t} - \alpha_{i,t}| = O_P\left(\left[n^{\xi/2-1} + n^{2-5\xi/2}\right]\log^{c_0} n\right).$$

Similar to Theorem 1.3.1, Theorem 1.3.5 says that the rate of convergence for the fixed effect depends on the relative size of n and T. If n and T have the same order of magnitude, then the convergence rate is $n^{-1/2}\log^{c_0} n$; if $\xi \neq 1$, then the rate would be strictly slower.

1.3.4 Determining the number of factors r_α and r_Q

So far, all the results are derived with the knowledge of the true values of r_Q and r_α for PCA. In practice, these values are often unknown. Now we derive two consistent estimators for these values. Although existing methods, such as Bai and Ng (2002), Onatski (2009) and Ahn and Horenstein (2013), can be used for estimating r_Q, these methods cannot be directly applied for the estimation of r_α due to the estimation errors in $\hat{\beta}_t$. We invoke results from random matrix theory and construct two simple estimators that are consistent under Assumption 1.

Theorem 1.3.6 (Information criterion)**.** *Let Assumption 1 hold. Define* $\hat{r}_Q := \max\{r \mid s_r(X) \geq \mu_n\}$ *and* $\hat{r}_\alpha := \max\left\{r \mid s_r\left([y_1 - X_1\hat{\beta}_1, \ldots, y_T - X_T\hat{\beta}_T]\right) \geq \tilde{\mu}_n\right\}$, *where* $\mu_n, \tilde{\mu}_n \to \infty$. *Then*

(1) If $\mu_n/\left(\sqrt{n}\log^p n\right) \to \infty$ *for any constant* $p > 0$ *and* $\mu_n/\sqrt{nT} \to 0$, *then* $\mathbb{P}\left(\hat{r}_Q = r_Q\right) \to 1$.

(2) If $\tilde{\mu}_n/\left(\left[\sqrt{T} + n/\sqrt{T}\right]\log^p n\right) \to \infty$ *for any constant* $p > 0$ *and* $\tilde{\mu}_n/\sqrt{nT} \to 0$, *then* $\mathbb{P}\left(\hat{r}_\alpha = r_\alpha\right) \to 1$.

The above estimator for r_Q and r_α is based on information criteria. Similar estimators are proposed by Bai and Ng (2002). One needs to choose a sequence of tuning parameters that satisfy certain rate conditions; however, it might not always be clear how to choose these tuning parameters in finite samples. For this reason, we also provide the following alternative estimators based on the ratio of singular values. These estimators are similar to the ones studied in Ahn and Horenstein (2013) and the only input is an upper bound on r_α and r_Q. In many situations, economic theories can shed some light on these upper bounds. For $r_{\max} \geq 1$, we define

$$\begin{aligned}
\hat{r}_Q^{SV} &:= \underset{1\leq r\leq r_{\max}}{\arg\max} \frac{s_r(X)}{s_{r+1}(X)} \\
\hat{r}_\alpha^{SV} &:= \underset{1\leq r\leq r_{\max}}{\arg\max} \frac{s_r\left([y_1 - X_1\hat{\beta}_1, \ldots, y_T - X_T\hat{\beta}_T]\right)}{s_{r+1}\left([y_1 - X_1\hat{\beta}_1, \ldots, y_T - X_T\hat{\beta}_T]\right)}
\end{aligned}$$

Theorem 1.3.7 (Singular value ratio estimator)**.** *Let Assumption 1 hold. Suppose that* $1 \leq r_Q \leq r_{\max}$ *and* $1 \leq r_Q \leq r_{\max}$*. Then* $\mathbb{P}(\hat{r}_Q^{SV} = r_Q) \to 1$ *and* $\mathbb{P}(\hat{r}_\alpha^{SV} = r_\alpha) \to 1$*.*

Remark 1.3.4. In practice, researchers might need to take additional care in applying the above results. For datasets that contain variables with very different scales, standardization is recommended, similar to the empirical applications in Stock and Watson (2002b) and Boivin and Ng (2006).

1.4 Some Important Inference Problems

In this section, we discuss how several problems often encountered in applied research can be addressed using the methodology proposed in Section 1.3. It turns out that solving these problems reduces to finding the appropriate matrix J in Algorithm 2. Since empirical work typically focuses on single entries of β_t corresponding to variables of interest, we shall mainly discuss this case. Suppose that we are interested in inference on $\{\beta_{j_0,t}\}_{t=1}^T$, the trajectory of the j_0-th entry of $\beta_t \in \mathbb{R}^k$ across time. For $k \neq 1$, the inference problems only concern part of the parameter $\{\beta_t\}_{t=1}^T$ and shall be referred to as partial inference problems.

1.4.1 Uniform (over t) inference on $\beta_{j_0,t}$

In empirical research, the goal is often to find out whether some slope coefficient is zero (or some other pre-specified value of interest). When the slope coefficients are allowed to vary over time, the question often becomes whether the slope coefficient $\beta_{(j_0)} = (\beta_{j_0,1}, \ldots, \beta_{j_0,T})' \in \mathbb{R}^T$ is zero in all the time periods.

Notice that this is very different from the problem of testing simple hypotheses on β. Simple hypotheses completely specify the value for all the entries in β; as a result, one can plug-in the hypothesized value of β and test certain moment conditions, such as the orthogonality between $y_{i,t} - x_{i,t}'\beta_t$ and $x_{i,t}$. However, we are dealing with the more difficult problem of testing composite hypotheses on β. For example, consider the problem of testing $\beta_{(j_0)} = 0$. Since $\{\beta_{j,t}\}_{t=1}^T$ with $j \neq j_0$ are still allowed to take any values, the null hypothesis here does not determine the

vector β and thus the aforementioned approach for testing simple hypotheses does not apply.

We now demonstrate how our method can be used to solve this inference problem. Let $J = I_T \otimes \tau'_{j_0,k}$ and $\tau_{j_0,k}$ denote the j_0-th column of I_k. Then we have $\beta_{(j_0)} = J\beta$. Notice that Assumption 2(i)-(ii) hold as $m_J = T$ and $\max_{1\leq j\leq m_J} \|J_j\|_1 = 1$. Under the above notations, $\beta_{j_0,t} = 0\ \forall 1 \leq t \leq T$ if and only if $J\beta = 0$. Hence, we only need to implement Algorithm 2 with $a = 0$. The problem of building confidence bands for $\{\beta_{j_0,t}\}_{t=1}^T$ reduces to constructing a rectangular confidence set for $\beta_{(j_0)}$ and can be easily solved using Corollary 1.3.1.

1.4.2 Inference on temporal difference in $\beta_{j_0,t}$

One of the simplest ways of studying time variation in parameters is to compare $\beta_{j_0,t}$ in different time periods. To formalize the idea, let $A, B \subset \{1, \ldots, T\}$ be disjoint sets $A \bigcap B = \emptyset$. We construct confidence intervals for the difference in average parameter values between these two groups of time periods, i.e.,

$$d(A, B) = \frac{\sum_{t=1}^T \beta_{j_0,t}\mathbf{1}\{t \in A\}}{\sum_{t=1}^T \mathbf{1}\{t \in A\}} - \frac{\sum_{t=1}^T \beta_{j_0,t}\mathbf{1}\{t \in B\}}{\sum_{t=1}^T \mathbf{1}\{t \in B\}}. \tag{1.4.1}$$

As convention, we define $d(A, \emptyset) = \left[\sum_{t=1}^T \beta_t \mathbf{1}\{t \in A\}\right] / \left[\sum_{t=1}^T \mathbf{1}\{t \in A\}\right]$, which is the average parameter value for time periods in the set A. For example, A and B can denote the sets of time periods of economic recessions and expansions, respectively, and $d(A, B)$ is a measure of how the parameters differ across different stages of the business cycle.

We now phrase the problem as inference on a linear combination of β. Let M denote the $1 \times T$ row vector whose s-th entry is equal to $\mathbf{1}\{s \in A\}/|A| - \mathbf{1}\{s \in B\}/|B|$, where $|A|$ and $|B|$ denote the cardinality of the set A and B, respectively. Then it is not hard to see that $d(A, B) = J\beta$, where $J = M \otimes \tau'_{j_0,k}$. Therefore, a confidence interval can be used by implementing Algorithm 2 and computing $\mathcal{C}_{1-\eta}(J)$ defined in Corollary 1.3.1.

1.4.3 Estimation and inference of partial parameter instability

The estimate and the confidence set for $\beta_{(j_0)} \in \mathbb{R}^T$ give some indication on whether the slope coefficient is time-varying. We shall refer to changes in parameter values from one period to the next as parameter instability. Define the set of time periods of parameter instability as $\mathcal{B} = \{t \mid 2 \leq t \leq T \text{ and } \beta_{j_0,t} \neq \beta_{j_0,t-1}\}$. Our method can be used for the estimation and inference on $\mathcal{B}$.

This problem can be easily formulated into our framework. Notice that

$$\begin{pmatrix} \beta_{j_0,2} - \beta_{j_0,1} \\ \beta_{j_0,3} - \beta_{j_0,2} \\ \vdots \\ \beta_{j_0,T} - \beta_{j_0,T-1} \end{pmatrix} = J\beta \qquad \text{with} \qquad \underset{(T-1)\times kT}{J} = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -1 \end{pmatrix}}_{(T-1)\times T} \otimes \tau'_{j_0,k}. \tag{1.4.2}$$

Clearly, the hypothesis of lack of parameter instability can be stated as $J\beta = 0$. Since Assumption 2(i)-(ii) are satisfied (due to $m_J = T - 1$ and $\max_{1\leq j\leq m_J} \|J_j\|_1 = 2$), Theorem 1.3.3 says that the hypothesis of absence of parameter instability can be tested by applying Algorithm 2 with $a = 0$.

Remark 1.4.1. As explained in Section 1.1, a major advantage of our approach is that no assumptions are placed on the time variation in parameters not under testing, i.e., $\{\beta_{j,t}\}_{t=1}^T$ for $j \neq j_0$. Hence, the common approach of imposing the null hypothesis of $\beta_{j_0,1} = \ldots = \beta_{j_0,T}$, such as Su and Chen (2013), does not apply to the partial inference problem here.

Algorithm 2 also provides a natural estimate for the set $\mathcal{B}$. For $\eta \in (0, 1)$, consider

$$\widehat{\mathcal{B}}(1-\eta) = \left\{ t \,\middle|\, 2 \leq t \leq T \text{ and } |\tilde{\beta}_{j_0,t} - \tilde{\beta}_{j_0,t-1}| > \Phi^{-1}(1-\eta, \hat{\Omega})/\sqrt{n} \right\},$$

where $\tilde{\beta}_{j_0,t}$ denotes the j_0-th entry of $\tilde{\beta}_t$ (defined in Algorithm 2). By Theorem

1.3.3, it follows that

$$\liminf_{n\to\infty} \mathbb{P}\left(\widehat{\mathcal{B}}(1-\eta) \subseteq \mathcal{B}\right) \geq 1-\eta.$$

This means that, $\widehat{\mathcal{B}}(1-\eta)$, as an estimate for the set of instability periods, has asymptotic control on the false discovery rates: the probability of $\widehat{\mathcal{B}}(1-\eta)$ containing points that are not in $\mathcal{B}$ is asymptotically at most η.

Under additional assumptions, $\widehat{\mathcal{B}}$ can also serve as an estimator for $\mathcal{B}$. By Theorem 1.3.4, $\Phi^{-1}(1-\eta, \hat{\Omega}) = O_P(\sqrt{\log n})$. Therefore, if we assume that the structural breaks are not too small[12] (i.e., $\min_{t\in\mathcal{B}} |\beta_{j_0,t} - \beta_{j_0,t-1}|\sqrt{n/\log n} \to \infty$), then $\mathbb{P}(\mathcal{B} \subseteq \widehat{\mathcal{B}}(1-\eta)) \to 1$ and thus $\mathbb{P}(\mathcal{B} = \widehat{\mathcal{B}}(1-\eta))$ is asymptotically at least $1-\eta$.

Remark 1.4.2. To achieve consistent estimation of $\mathcal{B}$, we propose to replace $\widehat{\mathcal{B}}(1-\eta)$ with $\widehat{\mathcal{B}} = \{2 \leq t \leq T : \ |\tilde{\beta}_{j_0,t} - \tilde{\beta}_{j_0,t-1}| > z_n/\sqrt{n}\}$, where z_n is a sequence such that $\forall \eta \in (0,1)$, $\Phi^{-1}(1-\eta, \hat{\Omega}) \leq z_n$ for large n and $z_n \asymp \sqrt{\log n}$. The above analysis implies that $\liminf \mathbb{P}(\mathcal{B} = \widehat{\mathcal{B}}) \geq 1-\eta$ for any $\eta \in (0,1)$ and therefore $\lim \mathbb{P}(\mathcal{B} = \widehat{\mathcal{B}}) = 1$. Now we propose a simple choice of z_n. By Lemma 2 in Chapter 7 of Feller (1968) and the union bound, we can easily show that $\forall x > 0$, $1 - \Phi(x, \hat{\Omega}) \leq 2k(T-1)x^{-1}\|\hat{\Omega}\|_\infty^{1/2}\phi(x^{-1}\|\hat{\Omega}\|_\infty^{1/2})$, where $\phi(\cdot)$ is the p.d.f of $N(0,1)$. Hence, it is straight-forward to show that $\forall \eta \in (0,1)$, we have that for large enough n, $\Phi^{-1}(1-\eta, \hat{\Omega}) \leq \sqrt{2\|\hat{\Omega}\|_\infty \log[k(T-1)]}$ almost surely. Therefore, a natural choice is $z_n = \sqrt{2\|\hat{\Omega}\|_\infty \log[k(T-1)]}$.

Remark 1.4.3. We note that power properties of tests for time-invariance in our panel setup might not follow existing results that deal with models for single time series. For example, consider the problem of testing $\beta_1 = \ldots = \beta_T$ versus $\beta_1 \neq 0$ and $\beta_2 = \ldots = \beta_T = 0$. When the sample consists of one time series, it is quite hard to detect the deviation that only occurs in one time period, regardless of how large T is. However, in our panel data setting, β_t is identified by the cross-sectional variations across n units and thus can be expected to estimated accurately for large

[12] In a sense, estimation of the set $\mathcal{B}$ is similar to the problem of model selection and thus requires similar regularity conditions, such as the so-called beta-min condition in high-dimensional models; see Bühlmann and Van De Geer (2011).

n.

1.4.4 Partial inference on structural breaks

Sporadic changes in parameter values, often referred to as structural breaks, can be viewed as piecewise constant patterns of $\{\beta_{j_0,t}\}_{t=1}^{T}$. Unlike the setup considered in Section 1.4.3, structural breaks are viewed as infrequent jumps in parameter values, which remain stable for extended periods of time, say at least $2q$ periods. Inference on structural breaks for the full vector β_t has been widely studied, e.g., Andrews (1993), Bai and Perron (1998, 2003) and Qu and Perron (2007) among others. However, the partial inference problem of testing for structural breaks in $\{\beta_{j_0,t}\}_{t=1}^{T}$ without imposing any restrictions on $\{\beta_{j,t}\}_{t=1}^{T}$ for $j \neq j_0$ is rarely discussed.

Suppose that there are m structural breaks, which occur in periods $B_1 < \ldots < B_m$:

$$\begin{aligned} \beta_{j_0,1} &= \ldots = \beta_{j_0,B_1} \\ \beta_{j_0,B_1+1} &= \ldots = \beta_{j_0,B_2} \quad \text{and} \quad \beta_{j_0,B_1} \neq \beta_{j_0,B_1+1} \\ &\vdots \\ \beta_{j_0,B_m+1} &= \ldots = \beta_{j_0,T} \quad \text{and} \quad \beta_{j_0,B_m} \neq \beta_{j_0,B_m+1}. \end{aligned} \tag{1.4.3}$$

We follow the convention in the literature by setting $B_0 = 1$ and $B_{m+1} = T$. We consider the problem of testing the hypothesis that there are no structural breaks. Although we can simply use the test discussed in Section 1.4.3, we might obtain a more powerful test by taking into account the block structure of structural breaks. We consider a block average scheme under which, for any pair of two adjacent blocks of time periods, the average parameter values in these two blocks are compared.

Consider two adjacent blocks of time periods of length q and compute the

difference in the block average of parameter values (DBA), i.e.,

$$\mathrm{DBA}(t;\beta,q) := \frac{1}{q}\sum_{s=t-q+1}^{t}\beta_{j_0,s} - \frac{1}{q}\sum_{s=t+1}^{t+q}\beta_{j_0,s} = \left(B'_{(t,q)}\otimes\tau'_{j_0,k}\right)\beta,$$

where the s-th entry of the vector $B_{(t,q)}\in\mathbb{R}^T$ is defined as

$$B_{(t,q),s} = \frac{1}{q}\Big[\mathbf{1}\{t-q+1\le s\le t\} - \mathbf{1}\{t+1\le s\le t+q\}\Big].$$

The hypothesis of lack of structural breaks can be restated as $\mathrm{DBA}(t;\beta,q) = 0$, $\forall 2q\le t\le T-2q$, which corresponds to $J\beta = 0$, where the rows of J are $B'_{(t,q)}\otimes\tau'_{j_0,k}$ for all $t\in\{2q,\dots,T-2q\}$. Notice that Assumption 2(i)-(ii) hold since $m_J = T-4q < T$ and $\max_{1\le j\le m_J}\|J_j\|_1 = 1$. By Theorem 1.3.3, we can test the hypothesis of lack of structural breaks by implementing Algorithm 2 with $a=0$.

Under the alternative that $\{\beta_{j_0,t}\}_{t=1}^T$ has a piecewise constant structure as in (1.4.3), $\mathrm{DBA}(B_j;\beta,q) = \beta_{j_0,B_j+1} - \beta_{j_0,B_j}$; on the other hand, it is natural to expect $\mathrm{DBA}(B_j;\beta,q)$ to be better estimated than $\beta_{j_0,B_j+1}-\beta_{j_0,B_j}$ simply because the former is the difference between two averages, especially for large q. Hence, we expect that for large q, the test proposed in this subsection be more powerful in detecting structural breaks than the test discussed in Section 1.4.3.

Our block average setup also yields a natural estimator. The basic idea is the following. Suppose that we observe the true sequence $\{\beta_{j_0,t}\}_{t=1}^T$. Then $\mathrm{DBA}(t;\beta,q)\neq 0$ if and only if $B_j-q\le t\le B_j+q-1$ for some $j\in\{1,\dots,m\}$. Also notice that $t\mapsto|\mathrm{DBA}(t;\beta,q)|$ reaches the maximum (over $\{B_j-q,\dots,B_j+q-1\}$) at $t=B_j$. This suggests a recursive strategy. Suppose that we already found B_{j-1}. Let s_j denote the smallest number s such that $s\ge B_{j-1}+2q-1$ and $|\mathrm{DBA}(s;\beta,q)|>0$. Then $B_j = \arg\max\{|\mathrm{DBA}(t;\beta,q)| \mid s_j\le t\le s_j+2q-1\}$.

Since $\|\tilde\beta-\beta\|_\infty = O_P(\sqrt{n^{-1}\log n})$, we consider a similar strategy with β replaced by $\tilde\beta$. Let $\delta_n = \Phi^{-1}(\hat\Omega,1-\eta)/\sqrt{n}$, where $\Phi^{-1}(\hat\Omega,1-\eta)$ is defined in Algorithm 2 using J described above. Starting with $\hat B_0 = 1$, we compute $\hat B_j$

recursively:

$$\hat{s}_j = \min\left\{ s \mid s \geq \hat{B}_{j-1} + 2q - 1 \text{ and } |\mathrm{DBA}(s; \tilde{\beta}, q)| > \delta_n \right\}$$

and

$$\hat{B}_j = \underset{\hat{s}_j \leq t \leq \hat{s}_j + 2q - 1}{\arg\max} \left| \mathrm{DBA}(t; \tilde{\beta}, q) \right|.$$

The iteration continues until $j = \hat{m}$, where $|\mathrm{DBA}(s; \tilde{\beta}, q)| \leq \delta_n$, $\forall s \geq \hat{m} + 2q - 1$. When the true $\{\beta_{j_0,t}\}_{t=1}^T$ follows the structural break pattern in (1.4.3) and the breaks are pronounced enough ($\sqrt{n/\log n} \min_{1 \leq l \leq m} |\beta_{j_0,B_l} - \beta_{j_0,B_l+1}| \to \infty$), then Theorems 1.3.3 and 1.3.4 imply that both $\mathbb{P}(\hat{m} = m)$ and $\mathbb{P}(\{\hat{B}_1, \ldots, \hat{B}_{\hat{m}}\} = \{B_1, \ldots, B_m\})$ are asymptotically at least $1 - \eta$. For these probabilities to tend to one, we can simply choose $\delta_n = \sqrt{2n^{-1} \|\hat{\Omega}\|_\infty \log m_J}$, see Remark 1.4.2.

1.4.5 Estimating partial regime-dependence

Popular models for the time variation in parameter values often specify a pattern in which parameters take values in a small set, whose elements are often referred to as regimes. For example, models with structural breaks have parameters staying in one regime between breaks; Markov switching models, such as Hamilton (1989), often assume that the parameters follow Markov chain with a few states. Due to the flexibility of our setup, if the underlying parameters indeed follow such regime patterns and these regimes are different enough (from each other), then our results can be used to estimate the membership of these regimes, i.e., which regime contains which time periods. Notice that $\{\beta_{j_0,t}\}_{t=1}^T$ is allowed to be random here; see Remark 1.3.3.

Suppose that there m regimes for $\beta_{j_0,t}$, which can take value in $\{a_1, \ldots, a_m\}$. Since $\beta_{j_0,t}$ is a scalar, we assume, without loss of generality, that $a_1 < a_2 < \ldots < a_m$. For $1 \leq r \leq m$, define the set of time periods corresponding to the r-th regime: $\mathcal{Q}(r) = \{t \mid 1 \leq t \leq T \text{ and } \beta_{j_0,t} = a_r\}$. The goal is to estimate m as well as $\mathcal{Q}(r)$ for each $1 \leq r \leq m$.

Due to the monotonicity of $\{a_r\}_{r=1}^m$, we consider a simple sorting strategy. The basic idea is quite simple. If we could sort the true values $\{\beta_{j_0,t}\}_{t=1}^T$, then we

would obtain a piecewise constant and non-decreasing path and different regimes are separated by jumps in the sorted sequence, leading to a structural break pattern in the sorted sequence. Hence, we could simply apply the techniques outlined in Section 1.4.4 to the sorted sequence. Since the true values $\{\beta_{j_0,t}\}_{t=1}^T$ are unknown, we simply implement this idea with $\{\tilde{\beta}_{j_0,t}\}_{t=1}^T$.

Formally, let $\pi : \{1,\ldots,T\} \mapsto \{1,\ldots,T\}$ be a permutation (bijective mapping) such that $\tilde{\beta}_{j_0,\pi(1)} \leq \ldots \leq \tilde{\beta}_{j_0,\pi(T)}$. Suppose that each regime contains at least $2q$ time periods. Discussions in Section 1.4.4 allow us to identify $\hat{\mu}$ breaks in the sequence $\{\tilde{\beta}_{j_0,\pi(s)}\}_{s=1}^T$, say $\varsigma_1,\ldots,\varsigma_{\hat{\mu}}$. Then our estimate for m is $\hat{\mu}+1$. We define

$$\widehat{\mathcal{Q}}(r) = \begin{cases} \{t \mid \pi(t) \geq \varsigma_{\hat{\mu}}\} & r = \hat{\mu}+1 \\ \{t \mid \varsigma_{r-1} \leq \pi(t) < \varsigma_r\} & 2 \leq r \leq \hat{\mu} \\ \{t \mid \pi(t) < \varsigma_1\} & r = 1 \end{cases} \tag{1.4.4}$$

If $\min_{2\leq j\leq m}(a_j - a_{j-1})\sqrt{n}/\log n \to \infty$, then it follows, by Theorems 1.3.3 and 1.3.4, that $\mathbb{P}(\hat{\mu}+1=m) \to 1$ and $\mathbb{P}(\bigcap_{r=1}^m\{\widehat{\mathcal{Q}}(r) = \mathcal{Q}(r)\}) \to 1$. In other words, when the regimes are different enough ($\min_{2\leq j\leq m}(a_j - a_{j-1})$ not too small), $\widehat{\mathcal{Q}}(r)$ can recover the regime pattern and thus be used to assess the specification of the time variation of parameters. For example, structural breaks should correspond to large blocks of time periods in which $\beta_{j_0,t}$ takes the same value, implying that, for each $1 \leq r \leq m$, $\mathcal{Q}(r)$ should contain consecutive time periods; regime switching patterns would imply the opposite for $\mathcal{Q}(r)$. We can conduct specifications tests. Consider, as an example, the problem of testing whether the switching pattern follows an i.i.d Bernoulli process or a first-order Markov chain. This problem reduces to testing the restrictions on transition probabilities using a sample of observed Markov chains.

1.4.6 Detecting general patterns of partial time-variation

In practice, big sudden shifts are not the only pattern of time variation. For example, changes in the parameters might be small at each point in time but accumulate to a large value over a long time horizon. In this case, the test discussed

in previous subsections might not reveal evidence of structural breaks, but this does not mean that we should conclude that slope coefficients are time-invariant. Again, this is a partial inference problem in that the null hypothesis of time invariance only concerns $\beta_{(j_0)}$ and allows arbitrary time variation in $\{\beta_{j,t}\}_{t=1}^T$ for $j \neq j_0$.

To test for general patterns of time variation, we consider the maximal time variation. Notice that time invariance means that $\beta_{j_0,t_1} = \beta_{j_0,t_2}$, $\forall 1 \leq t_1 < t_2 \leq T$. Hence, we can consider $|\beta_{j_0,t_1} - \beta_{j_0,t_2}|$ all combinations of $t_1, t_2 \in \{1, \ldots, T\}$ with $t_1 < t_2$. We define $\iota_{t_1,t_2} \in \mathbb{R}^T$ as a vector of zeros, except that the t_1-th entry is 1 and the t_2-th entry is -1. Clearly, there are $T(T-1)/2$ vectors of this form and we form the matrix J as follows: the row $Tt_1 + t_2$ of J is $\iota'_{t_1,t_2} \otimes \tau'_{j_0,k}$. Under this notation, the null hypothesis of time invariance becomes $J\beta = 0$. Since Assumption 2(i)-(ii) hold with $m_J = T(T-1)/2$ and $\max_{1 \leq j \leq m_J} \|J_j\|_1 = 2$, Theorem 1.3.3 guarantees the validity of testing for time invariance using Algorithm 2 with $a = 0$.

Notice that the test statistics $\|J\tilde{\beta}\|_\infty$ is equal to $\max_{1 \leq t_1 < t_2 \leq T} |\tilde{\beta}_{j_0,t_1} - \tilde{\beta}_{j_0,t_2}|$, which is an estimate for $\max_{1 \leq t_1 < t_2 \leq T} |\beta_{j_0,t_1} - \beta_{j_0,t_2}|$, the distance between the peak and trough in $\tilde{\beta}_{j,t}$. Therefore, it follows, by Theorem 1.3.4, that this test can detect any time variation resulting in a parameter trajectory that cannot be contained in a band of constant width of $O(\sqrt{n^{-1} \log n})$. For this reason, this procedure can be used as a test for time invariance whenever the alternative does not specify a particular pattern of parameter changes.

1.4.7 Explaining time variations in the slope coefficients

One common method of explaining variations is to use regression analysis. Here, we treat $\{\beta_{j_0,t}\}_{t=1}^T$ as a stochastic process; see Remark 1.3.3. For example, applied researchers can analyze the randomness of $\{\beta_{j_0,t}\}_{t=1}^T$ using linear regressions:

$$\beta_{j_0,t} = z'_t \underset{m \times 1}{\theta} + \varepsilon_t, \tag{1.4.5}$$

where $z_t \in \mathbb{R}^m$ is the vector of observed explanatory variables with fixed m and ε_t is the error term. The goal is to conduct inference on θ. Notice that we still allow $\{\beta_{j,t}\}_{t=1}^T$ with $j \neq j_0$ to have completely different time variation patterns. Therefore,

we cannot state the model (1.2.1) in terms of the low-dimensional parameter θ by imposing (1.4.5).

Were the process $\{\beta_{j_0,t}\}_{t=1}^T$ observed, one could simply estimate θ by the ordinary least squared estimator

$$\hat{\theta} = \left(\sum_{t=1}^T z_t z_t'\right)^{-1} \left(\sum_{t=1}^T z_t \beta_{j_0,t}\right).$$

However, in practice, the process $\{\beta_{j_0,t}\}_{t=1}^T$ is not observed and thus the estimator $\hat{\theta}$ is not feasible. Since Algorithm 2 delivers the estimated process $\{\tilde{\beta}_{j_0,t}\}_{t=1}^T$, we can consider the following estimator

$$\tilde{\theta} = \left(\sum_{t=1}^T z_t z_t'\right)^{-1} \left(\sum_{t=1}^T z_t \tilde{\beta}_{j_0,t}\right).$$

The following result says that, under certain conditions, $\tilde{\theta}$ and $\hat{\theta}$ are asymptotically equivalent.

Theorem 1.4.1. *Let Assumptions 1 and 2 hold with $J = I_{kT}$ and $\xi < 3/2$. Suppose that $\{z_t\}_{t=1}^T$ is independent of v and u. If $(T^{-1}\sum_{t=1}^T z_t z_t')^{-1} = O_P(1)$ and $\max_{1\le t\le T} \mathbb{E}\|z_t\|^2 = O(1)$. Then*

$$\sqrt{T}(\tilde{\theta} - \hat{\theta}) = o_P(1).$$

For inference of θ under the above linear regression framework, Theorem 1.4.1 says that it suffices to derive the limiting distribution of $\sqrt{T}(\hat{\theta} - \theta)$ since $\sqrt{T}(\tilde{\theta} - \theta)$ and $\sqrt{T}(\hat{\theta} - \theta)$ differ by $o_P(1)$. Therefore, the estimation error in $\tilde{\beta}$ does not contribute to the asymptotic distribution of the estimator $\tilde{\theta}$.

Remark 1.4.4. The key condition of Theorem 1.4.1 is the independence between $\{z_t\}_{t=1}^\infty$ and (v, u). Notice that we do not impose any assumption on the error term ε_t in (1.4.5). Our assumption here is similar in spirit to Assumption E in Bai and Ng (2006a), who consider a related problem: factors are first estimated from a large panel dataset and then used as covariates in a separate regression. However, their conclusion is quite different from ours. Bai and Ng (2006a) show that the

estimation errors of the factors would in general influence the asymptotics in the latter regression, while Theorem 1.4.1 says that the estimation error in $\tilde{\beta}_t$ does not contribute to the limiting distribution of $\sqrt{T}(\tilde{\theta} - \theta)$. This is because the estimation errors of $\tilde{\beta}$ are noise in the dependent variable in regression (1.4.5), whereas in Bai and Ng (2006a), the estimation errors of the factors affect the covariates in the regression of interest. In the regression setup, measurement errors in regressors cause a bigger problem than those in the response variable.

1.5 Monte Carlo Simulations

We consider both static (STA) and dynamic (DYN) models, which are specified as follows. The static model reads

$$y_{i,t} = F'_{\alpha,t} L_{\alpha,i} + \beta_{1,t} x_{1,i,t} + \beta_{2,t} x_{2,i,t} + u_{i,t}, \qquad \text{(STA)}$$

where $x_{i,t} = (x_{1,i,t}, x_{2,i,t})' = F'_{Q,t} L_{Q,i} + v_{i,t}$. Columns of F_α, F_Q and rows of v and u are generated as independent stochastic processes, which can take three specifications, denoted by GAUSS, STU-T and ARMA (specified later). We generate entries of $L_{\alpha,i}$ and $L_{Q,i}$ from i.i.d. $N(0, 1/2)$ and set the first column of $F_{Q,t}$ equal the first column of $F_{\alpha,t}$. $\beta_{1,t}$ and $\beta_{2,t}$ are drawn from i.i.d uniform distribution on $[-1, 1]$.

The dynamic model reads

$$y_{i,t} = F'_{\alpha,t} L_{\alpha,i} + \beta_{1,t} y_{i,t-1} + \beta_{2,t} \tilde{x}_{i,t} + u_{i,t}, \qquad \text{(DYN)}$$

where $\tilde{x}_{i,t} = F'_{\tilde{Q},t} L_{\tilde{Q},i} + \tilde{v}_{i,t}$ with $F_{\tilde{Q},t} \in \mathbb{R}^{r_Q - r_\alpha}$. Thus, the number of factors in the regressors are r_Q (from both $F_{\alpha,t}$ and $F_{\tilde{Q},t}$)[13]. As before, columns of F_α, $F_{\tilde{Q}} = (F_{\tilde{Q},1}, \ldots, F_{\tilde{Q},T})'$, $\{\tilde{v}_{i,t}\}_{t=1}^T$ and rows of u are independent stochastic processes, which can take three specifications, denoted by GAUSS, STU-T and GARCH. We generate entries of $L_{\alpha,i}$ and $L_{\tilde{Q},i}$ from i.i.d. $N(0, 1)$ and draw $\beta_{1,t}$ and $\beta_{2,t}$ from i.i.d

[13]Notice that when lagged $y_{i,t}$ is included as the regressors, we always have $r_Q \geq r_\alpha$. Since r_α factors drive $y_{i,t}$ and thus drive lagged $y_{i,t}$, which is only part of the regressors, the total number of factors driving the regressors is at least r_α.

uniform distribution on $[-0.7, 0.7]$.

Here, GAUSS denotes the process of i.i.d $N(0,1)$; STU-T denotes the process of i.i.d Student's t-distribution with 6 degrees of freedom normalized to have variance one. ARMA denotes the ARMA(1,1) zero-mean process with autoregressive and moving average coefficient being 0.924 and 0.592 (calibrated to quarterly data of real U.S. GDP growth), where the innovations are i.i.d zero-mean Gaussian with variance chosen such that the long-run variance of the ARMA process is one. GARCH denotes the GARCH(1,1) zero-mean process with ARCH and GARCH parameters being 0.12 and 0.85 (calibrated to monthly returns of the S&P500 index), where the standardized innovations are i.i.d zero-mean Gaussian with variance chosen such that the long-run mean variance of the GARCH process is one.

In all of our simulations, r_Q and r_α are estimated using $\hat{r}_Q^{SV}$ and $\hat{r}_\alpha^{SV}$ as discussed in Theorem 1.3.7. In all the tables and figures, the coverage probabilities of confidence bands and the rejection probabilities of tests are based on 2000 random samples.

For each simulated sample, we construct a 95% confidence band for the trajectory of the first entry of β_t; see Corollary 1.3.1 and Section 1.4.1. The results are reported in Table 1.1. As we can see, these results demonstrate decent finite-sample performance of the proposed confidence bands. The 95% confidence bands has empirical coverage probabilities around the nominal level, even for a sample size as small as $n = T = 60$. Strictly speaking, STU-T and GARCH processes do not satisfy the condition of exponential-type tails, but our procedures still perform quite well. For dynamic models, certain under coverage could occur for large n and relatively small T; this is only a finite sample problem since in our unreported results with larger sample sizes (e.g., $n = 900$ and $T = 200$), we find the coverage probability of the confidence bands close to their nominal levels.

We also consider the test for structural breaks. We keep the same specifications STA and DYN, except that $\beta_{1,t}$ is generated as

$$\beta_{1,1} = \ldots = \beta_{1,\lfloor \lambda T \rfloor} = w \qquad \text{and} \qquad \beta_{1,\lfloor \lambda T \rfloor + 1} = \ldots = \beta_{1,T} = w + \delta,$$

where $\lambda \in (0,1)$ is a fixed parameter and $\lfloor \lambda T \rfloor$ denotes the largest integer not

exceeding λT. For STA and DYN specifications, w is from the uniform distribution on $[-1, 1]$ and on $[-0.7, 0.7]$, respectively. The null hypothesis of $\beta_1 = \ldots = \beta_T$ corresponds to $\delta = 0$. The deviation from the null hypothesis is measured in δ. Notice that this we only have structural breaks in $\{\beta_{1,t}\}_{t=1}^T$ since $\{\beta_{2,t}\}_{t=1}^T$ is still drawn from i.i.d uniform distributions as in STA and DYN.

We consider the test discussed in Section 1.4.3. The size properties of a 5% test are reported in Table 1.2. For static models, our test has decent size control in finite samples; for dynamic models, slight over-rejection could occur for large n and small T. In Figures 1.1 and 1.2, we plot the power curves of 5% tests under the STA-GAUSS and DYN-GAUSS specifications, respectively. As expected, the power increases with the sample size and the magnitude of δ. Interestingly, the power function is not sensitive to λ, the location of the structural break. Since we identify $\beta_{1,t}$ through cross-sectional units, rather than the time dimension, we do not need many time periods for each regime (i.e., before and after the break), a similar situation as discussed in Remark 1.4.3.

1.6 Empirical Applications

In this section, we illustrate the proposed methodology via three empirical problems: (1) stock return predictability, (2) firms' capital structure and (3) the effect of investment on economic growth.

1.6.1 Stock return predictability

A question of fundamental interest in finance is whether the equity risk premium is time-varying and, if so, can be predicted ahead of time as suggested by studies such as Campbell and Cochrane (1999) and Bansal and Yaron (2004). Two of the most popular predictors are the dividend yield[14] and volatility[15]. Here, we

[14]See e.g., Campbell and Shiller (1988b, 1988a), Fama and French (1988), Hodrick (1992) and Koijen and Van Nieuwerburgh (2011)

[15]See e.g., Goyal and Santa-Clara (2003), Bakshi and Kapadia (2003), Ang, Hodrick, Xing, and Zhang (2006) and Bollerslev, Gibson, and Zhou (2011).

study the following regression using panel data

$$r_{i,t} = L'_{\alpha,i} F_{\alpha,t} + \theta_t d_{i,t-1} + \gamma_t \text{VOL}_{i,t-1} + u_{i,t}, \tag{1.6.1}$$

where $r_{i,t}$ is the log excess return in period t on asset i, $d_{i,t-1}$ is the dividend yield in period $t-1$ for asset i and $\text{VOL}_{i,t-1}$ denotes the variance of asset i in period t conditional on the information in period $t-1$.

We interpret θ_t (and γ_t) as capturing predictability in stock returns by means of time-variation in the dividend yield or conditional variance. In the specification in (1.6.1), we use a factor structure to model potential cross-sectional dependence among the error terms. These common factors include financial and macroeconomic shocks that drive the returns of all stocks, as well as time-specific and asset-specific fixed effects.[16] Due to the presence of these factors, methods based on OLS, such as the Fama-MacBeth regression, might provide inconsistent estimators even under strict exogeneity; see Appendix A.1 for a simple example.

We use annual data on 100 equity portfolios sorted by size and book-to-market ratio and compute the dividend yield from the cum-dividend and ex-dividend return series.[17] The conditional volatility is computed by fitting on AR(1) model with annual realized volatility. We specify the sampling period to run from 1960 to 2015 ($T = 56$ years) and retain the observations of $n = 89$ portfolios after removing 11 portfolios with missing data. We apply the methodology proposed in Sections 1.3 and 1.4. The estimate of $\beta = (\beta_1, \ldots, \beta_T)' \in \mathbb{R}^{2T}$ with $\beta_t = (\theta_t, \gamma_t)'$ and the 95% uniform confidence band are displayed in Figure 1.3.

The two plots in Figure 1.3 suggest quite different patterns of time variation for $\{\theta_t\}_{t=1}^T$ and $\{\gamma_t\}_{t=1}^T$. Since the red horizontal line in Panel A of Figure 1.3 representing the vector of zeros does not lie in the confidence band, we reject the hypothesis that $\theta_1 = \ldots = \theta_T = 0$. Time variation in θ_t is quite evident in Figure 1.3. Sporadic spikes in θ_t occur in the late 1960's and around 2000. This pattern

[16] Well-known factors include the Fama-French factors (Fama and French (1992, 2016)) and macroeconomic factors in the large factor model literature, e.g., Stock and Watson (1998, 2002b, 2006).

[17] The data is obtained from the website of Kenneth French.

indicates parameter instability, which is also documented in existing work[18]; the p-value of testing $\theta_1 = \ldots = \theta_T$ using the framework discussed in Section 1.4.3 is 0.001. Panel B of Figure 1.3 displays the path of time variation in the predictive power of conditional volatility. Such predictive power is mainly concentrated in the 1960's and 1970's and seems to have disappeared after 1980. We also cluster slope coefficients using structural break models; in particular, we assume that there are at least four years between breaks and apply the methodology outlined in Section 1.4.4. The results are reported in Table 1.3 and Figure 1.3. We find that the only structural break in θ_t occurred in the late 1990's and that there are three structural breaks in γ_t, which occurred in the late 1960's, late 1970's and early 1990's, respectively. However, we also reject the hypothesis that the parameter values are stable between the estimated structural breaks; this suggests that models with structural breaks in parameters might not be flexible enough to reveal all the features in return predictability.

Our method separates shocks in the error terms from those in the return predictability. Figure 1.4 plots the time series of the average noise level $n^{-1}\sum_{i=1}^{n}\hat{u}_{i,t}^2$, where $\hat{u}_{i,t}$ is defined in Algorithm 2. We compare Figures 1.3 and 1.4. During the recent Great Recession, the return predictability from the dividend yield and the conditional volatility was quite stable whereas large spikes are found in the average noise level. This indicates that the Great Recession only contributed to the noise in the error terms and did not change the relationship between stock returns and predictors, such as the dividend yield and conditional volatility. However, the collapse of the dot-com bubble appears to be a different kind of shock; we find large spikes in θ_t and the average noise level but not in γ_t. It is perhaps not surprising to see changes in the relationship between stock returns and the dividend yield as companies in the information technology sector, known for low dividends and realized profits, saw their stock prices soar and then plummet.

To study any seasonality in return predictability as well as its link to the macroeconomy, we also estimate model (1.6.1) using quarterly data over the same time periods ($T = 224$ quarters)[19]. Switching to quarterly data makes it more

[18]See Paye and Timmermann (2006), Lettau and Van Nieuwerburgh (2008), and Viceira (1997).

[19]The conditional volatility is obtained by fitting the quarterly realized volatility to AR(4)

convenient to explore time variation related to macroeconomic variables, many of which are observed on a quarterly basis. We apply the framework outlined in Section 1.4.2. In Table 1.4, we construct confidence intervals for $d(A, B)$ (defined in (1.4.1)), where A and B are sets containing different time periods; average predictability corresponds to $A = \{1, \ldots, T\}$ and $B = \emptyset$. From Table 1.4, the average (across time) of return predictability from the dividend yield is estimated to be 0.49 and is not statistically significant from zero, while the average predictive power of volatility is negative and statistically significant, findings consistent with existing literature, see e.g., Glosten, Jaganathan, and Runkle (1993) and Goyal and Welch (2008).

Table 1.4 includes other intriguing findings. First, return predictability coefficients exhibit strong seasonality. A large literature has documented the presence of calendar effects in stock returns, i.e., different patterns of stock returns on certain days of the week, months of the year, etc.[20] Typically, these calendar effects are not conditional on other variables and thus should correspond to part of the fixed effects in (1.6.1). Our specification allows for both interactive fixed effects and time-heterogeneous slope coefficients and is thus flexible enough to distinguish seasonality in the error terms from seasonal changes in θ_t and γ_t. Table 1.4 and Figure 1.5 say that, on average, predictability using the dividend yield is particularly profound in the third quarter of the year and is not statistically different from zero in the other three quarters; on average, volatility has predictive power only in the second and third quarters. Our finding suggests that the calendar effects are present not only in the error terms but also in the slope coefficients.

Second, return predictability is related to the state of the macroeconomy. Numerous studies have found that stock returns are predictable only in certain stage of the business cycle, see e.g., Fama and French (1989), Rapach and Wohar (2006), Rapach, Strauss, and Zhou (2010) and Dangl and Halling (2012). Table 1.4 suggests that the dividend yield is informative only in economic recessions (defined by the NBER recession indicators); similar results hold if we treat as recessions

---

model.

[20] See e.g., Jones, Pearce, and Wilson (1987), Keim and Stambaugh (1986), Haugen and Lakonishok (1988) and Kramer (1994).

periods in which the real GDP growth is smaller than its median. The predictive power of volatility is strong in NBER expansions, but not in recessions; on the other hand, this predictive power is only significant in periods with slow GDP growth. Unlike most work in the literature, we do not fit a two-regime parameter model to the data and thus our findings are not driven by specific model assumptions on the time variation in return predictability.

1.6.2 Firms' choice of capital structure

The study of firms' capital structure decisions is of fundamental interest in corporate finance. A large body of theoretical and empirical work has emerged to explain how corporations make decisions on the use of debt, see Titman and Wessels (1988), Harris and Raviv (1991), Rajan and Zingales (1995), Graham and Harvey (2001) and Welch (2004) among many others. In a survey paper, Frank and Goyal (2009) investigate numerous variables that can affect firms' capital structure. Following this literature, we consider the following regression:

$$LV_{i,t+1} = L'_{\alpha,i} F_{\alpha,t+1} + x'_{i,t}\beta_t + u_{i,t+1},$$

where $LV_{i,t+1}$ is the leverage ratio of firm i at time $t+1$ and $x_{i,t}$ contains 11 covariates observed at time t for firm i.[21] We use the same data as Frank and Goyal (2009) and take the variables from Table II therein. We drop from $x_{i,t}$ variables that are either only time-specific (e.g. macroeconomic variables) or only firm-specific (e.g. whether the industry of the firm is regulated) since the effects of these variables are captured by the fixed effects. After removing missing data, we have a balanced panel with annual observations of $n = 167$ firms from 1963 to 2003 ($T = 41$).

We shall revisit the following conclusions of Frank and Goyal (2009):

(a) Firms with higher market-to-book ratios tend to have less leverage

[21] These 11 variables are profitability, book assets, market-to-book ratio, change in assets, capital expenditure, median industry leverage, median industry growth, tangible assets, R&D expense, uniqueness and SGA (selling, general and administration) expense. See Appendix B of Frank and Goyal (2009) for detailed definitions.

(b) Firms with more tangible assets tend to have more leverage

(c) Firms with more profits tend to have less leverage

(d) Firms with more book assets tend to have more leverage

These conclusions are statements on the components of β_t corresponding to the following four regressors: profitability, book assets, market-to-book ratio and tangible assets. In this exercise, we focus on (1) estimates for $\beta = (\beta_1', \ldots, \beta_T)'$ and its 95% confidence sets, (2) testing for time-invariance of β_t and (3) inference on the average effect, i.e., $T^{-1}\sum_{t=1}^{T}\beta_t$. We consider two measures of the leverage ratio: the ratio of total debt to market assets (DM) and the ratio of total debt to book assets (DB). In Figures 1.6 and 1.7, we report the confidence bands for DM and DB, respectively. In Table 1.5, we report inference results for the average effects and time-invariance.

We find clear evidence of time variation in β_t. This is visually discernible in Figures 1.6 and 1.7. We also notice that the time variations are mostly slow changes in β_t rather than sudden abrupt changes. Applying the test for time invariance described in Section 1.4.6, we conclude, at the 5% significance level, that time variations are present in β_t for assets, profit and tangible assets; time invariance for the effects of market-to-book is also rejected at the 5% significance level when we use DM as the leverage ratio.

From Figures 1.6 and 1.7, we can reject the hypothesis that $\beta_{j,1} = \ldots = \beta_{j,T} = 0$ at the 5% significance level, for tangible assets, profits and book assets. From Table 1.5, we also reject, at the 5% level, that the average effect is zero for all the four variables of interest. Interestingly, the average effects of market-to-book ratio have different signs, depending on whether we use DM or DB as the leverage ratio, a finding consistent with Table V of Frank and Goyal (2009).

Overall, we confirm the findings in Frank and Goyal (2009), but our results also suggest quite different patterns of time variation. For example, Figures 1.6 and 1.7 show that the effects of the tangible assets change considerably and might have declined to zero or even switched signs at some point, whereas Table V of Frank and Goyal (2009) shows that the corresponding component of β_t has stayed

away from zero in each decade and is relatively stable. Moreover, Frank and Goyal (2009) conclude that, for leverage measured by DM, the importance of profits has declined significantly since the 1950's and that of book assets has increased during that period, see Table V therein. From Figure 1.6, we see that the importance of profits has stayed stable if not increased. It is true that its importance might have temporarily dropped in the late 1980's, but quickly recovered in the early 1990's. Figure 1.6 also shows that the effect of book assets increased from zero to its peak in the late 1980's before it dropped to a level close to zero.

The above difference might suggest the benefit of our method, compared to the simple practice of dividing the sample into subsamples. In a sense, the approach adopted by Frank and Goyal (2009) amounts to specifying structural breaks that could occur only at the end of each decade for all the parameters. However, estimates from our model in Figures 1.6 and 1.7 indicate smooth and gradual changes for at least some of the parameters, such as book assets. We also see that certain trends in parameter values can reverse within one decade. These findings can serve as evidence supporting that a structural break model might not be a suitable specification for the parameters. Since different parameters can have completely different patterns of time variation, such as profits and book assets in Figure 1.6, it is advantageous to apply our flexible setup, which allows for any pattern of time variation in parameters.

1.6.3 Investment and economic growth

Our third application is related to the long-running debate on whether investment causes economic growth. Despite the obvious importance of this question, it appears that a consensus has yet to emerge. The literature contains studies that support such causality and perhaps equally many papers that conclude otherwise; see e.g., DeLong and Summers (1991), Mankiw, Romer, and Weil (1992), Islam (1995), Jones (1995), Blomström, Lipsey, and Zejan (1996) and Bond, Leblebicioglu, and Schiantarelli (2010).

To address this issue, we present a panel data analysis that allows for interactive fixed effects and unrestricted time-heterogeneous slope coefficients. Since

the fixed effects can account for the endogeneity of investment, our setup could help shed light on any (time-varying) effect of investment on economic growth. We consider the following regression equation similar to the one studied in Blomström, Lipsey, and Zejan (1996):

$$g_{i,t} = L'_{\alpha,i} F_{\alpha,t} + \theta_t INV_{i,t-1} + \gamma_t g_{i,t-1} + u_{i,t}, \tag{1.6.2}$$

where $g_{i,t}$ is the growth of real GDP per capita in country i in year t and $INV_{i,t-1}$ is the ratio of gross capital formation to GDP of country i in year $t-1$. The data is obtained from Penn World Table 9.0. After removing missing values, we have a balanced panel consisting of $n = 74$ countries over $T = 53$ years from 1962 to 2014.

In Table 1.6, we conduct inference regarding the average θ_t across time and test the time-invariance of θ_t. Figure 1.8 plots estimates for $\{\theta_t\}_{t=1}^T$ and its 95% confidence bands.

We find that the average value of θ_t across time is close to zero but that θ_t is not always zero. In Table 1.6, we see that the average θ_t across time is not statistically different from zero; however, time-invariance of θ_t is strongly rejected. From Figure 1.8, we see that the 95% confidence band for $\{\theta_t\}_{t=1}^T$ does not contain the red line representing the zero vector and does not contain any horizontal lines, implying time variation in $\{\theta_t\}_{t=1}^T$.

We also find that the average effect of investment on economic growth increased after the early 1990's. The methodology outlined in Section 1.4.4 is applied to the estimated $\{\theta_t\}_{t=1}^T$ in order to identify structural breaks. As shown in Figure 1.8, our method suggests that there is only one structural break, which occurred in the early 1990's. According to Table 1.6, the average effect of investment is not significantly different from zero in the pre-break periods and is significantly positive in the post-break periods. One explanation is related to advances in technology in the early 1990's. Several studies, such as Litan and Rivlin (2001) and Freund and Weinhold (2004), have found that the Internet has positive effects on productivity, management efficiency and international trade. Our findings are consistent with the possibility that the adoption of the Internet in the early 1990's increases the effect of investment on the economy. Moreover, in both pre-break and

post-break periods, we reject time homogeneity in θ_t. This suggests that the usual structural break model might not be sufficient to describe the time-varying pattern in θ_t, highlighting the advantage of the proposed methodology.

We also consider the grouped fixed effects (GFE) discussed by Bonhomme and Manresa (2015). GFE assumes that the cross-sectional units can be categorized into a small number of groups and the time variation of the fixed effects is the same among nations in the same group. Since this specification can be viewed as a special case of the interactive fixed effects, Theorem 1.3.5 implies that the estimator $\hat{\alpha}_{i,t}$ from Algorithm 2 is a consistent estimator for GFE. Similar to Bonhomme and Manresa (2015), we estimate the group membership by applying the k-means clustering algorithm (Forgy 1965; Lloyd 1982) to $\hat{\alpha}_{i,t}$.

We comment on two findings under the GFE specification. First, we find a separation that roughly divides the countries in the sample into developed and developing nations. The result is reported in Figure 1.9. The red group in Figure 1.9 contains mostly developed countries, such as nations in North America, Western Europe, Australia, Japan and South Korea; the blue group in Figure 1.9 contains primarily developing countries, such as China, India, nations in Africa and South America.

Second, the estimated number of factors in the fixed effects is one and we find evidence supporting that the two groups are driven by this factor but with different sensitivities. In Figure 1.10, we plot the estimated trajectories of fixed effects in the two groups. The two paths of fixed effects display substantial co-movement but possess different volatilities: the red group has a more volatile path in the fixed effects. This suggests that the fixed effects in the two groups are driven by the same factor but the factor loading of the red group (mainly developed countries) is larger in magnitude than that of the blue group (mainly developing countries). One explanation is that developed nations, compared to developing nations, are more involved in international economic/political activities and are thus more sensitive to world-wide economic/political shocks. This also explains why the red group contains some countries that are usually classified as developing countries. For example, since the economy of Iran and Venezuela heavily relies

on exporting petroleum-related products, which are closely connected to global economic trends, it is perhaps not surprising that these two countries are highly susceptible to international economic forces.

These results illustrate the benefit of our methodology. The patterns that admit economic interpretations, such as the group membership, are not results of a priori specifications that are explicitly imposed. In particular, we do not impose any restrictions on the group membership or on the co-movement of fixed effects between the two groups. Moreover, our results are robust to arbitrary time-heterogeneity in the slope coefficients. This is important since we find strong evidence of such time heterogeneity.

1.7 Conclusion

We consider panel data models with interactive fixed effects and time-heterogeneous slope coefficients. These models do not restrict the time-variation in the slope coefficients, while allowing for both cross-sectional and inter-temporal dependence in the error terms. As the data consists of a large number of cross-sectional observations over many time periods, the vector β containing all the slope coefficients across time has dimensionality tending to infinity.

We propose methods for estimating and conducting inference on β and establish their asymptotic properties. We treat the entire vector β as a high-dimensional parameter and provide tools for inference on the trajectory of the time-variation of slope coefficients. In particular, our results can be used to construct confidence bands for this trajectory of slope coefficients, to test for time-invariance and to conduct inference on specific patterns of time variations, including structural breaks and regime switching. Our methods are simple to implement and computationally convenient.

An interesting extension of our work is to allow covariate effects to be heterogeneous both across cross-sectional units and across time. Such a flexible framework could be quite natural in empirical applications. For example, certain treatments might have different effects on different individuals in different time

periods; applied researchers might be interested in questions such as how the average (across individuals) treatment effects vary over time, whether certain (groups of) individuals are always more responsive to the treatment and whether time variation in the treatment effects is synchronized across individuals. Estimation and inference of these models would probably require certain structures on the heterogeneity in slope coefficients. To this end, one might borrow from popular specifications of fixed effects, although formal analysis is likely to encounter additional technical challenges and is left for future research.

1.8 Acknowledgements

Chapters 1, in full, is currently being prepared for submission for publication of the material. Zhu, Yinchu. The dissertation author was the primary investigator and author of this material.

Tables and Figures

Table 1.1: Coverage probability of 95% confidence bands for $\{\beta_{1,t}\}_{t=1}^{T}$

Panel A: static model (STA)

$r_Q = r_\alpha = 1$

	GAUSS			STU-T			ARMA		
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.973	0.984	0.988	0.969	0.984	0.989	0.944	0.979	0.982
120	0.954	0.969	0.971	0.966	0.982	0.981	0.952	0.969	0.974
180	0.949	0.965	0.963	0.963	0.977	0.982	0.945	0.960	0.966
$r_Q = r_\alpha = 2$									
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.962	0.982	0.987	0.956	0.984	0.988	0.953	0.982	0.990
120	0.951	0.965	0.977	0.956	0.976	0.984	0.941	0.967	0.977
180	0.947	0.964	0.963	0.956	0.980	0.979	0.932	0.959	0.960
$r_Q = r_\alpha = 3$									
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.968	0.980	0.992	0.942	0.973	0.982	0.962	0.981	0.992
120	0.950	0.974	0.980	0.945	0.978	0.982	0.948	0.970	0.980
180	0.940	0.966	0.969	0.942	0.970	0.983	0.931	0.959	0.972

Panel B: dynamic model (DYN)

$r_\alpha = 1$ and $r_Q = 1$

	GAUSS			STU-T			GARCH		
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.929	0.962	0.964	0.929	0.960	0.980	0.923	0.957	0.969
120	0.903	0.945	0.954	0.900	0.957	0.966	0.942	0.968	0.977
180	0.881	0.942	0.947	0.865	0.946	0.959	0.942	0.970	0.982
$r_\alpha = 1$ and $r_Q = 2$									
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.941	0.963	0.970	0.941	0.972	0.975	0.919	0.955	0.968
120	0.904	0.942	0.958	0.893	0.960	0.965	0.944	0.972	0.974
180	0.860	0.948	0.955	0.845	0.943	0.962	0.941	0.970	0.984
$r_\alpha = 1$ and $r_Q = 3$									
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.944	0.967	0.969	0.942	0.977	0.975	0.909	0.939	0.962
120	0.916	0.957	0.962	0.911	0.953	0.975	0.927	0.961	0.967
180	0.868	0.933	0.947	0.855	0.947	0.965	0.914	0.966	0.978

Table 1.2: Rejection probability under the null hypothesis that $\beta_{1,1} = \ldots = \beta_{1,T}$

We report the rejection probabilities of tests for structural breaks with nominal size 5% under the null hypothesis that $\beta_{1,1} = \ldots = \beta_{1,T}$.

Panel A: static model (STA)

$r_Q = r_\alpha = 1$

	GAUSS			STU-T			ARMA		
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.054	0.042	0.024	0.038	0.028	0.019	0.046	0.022	0.019
120	0.052	0.043	0.040	0.044	0.033	0.024	0.044	0.036	0.034
180	0.055	0.041	0.038	0.055	0.034	0.032	0.053	0.037	0.038

$r_Q = r_\alpha = 2$

$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.044	0.023	0.019	0.059	0.026	0.019	0.044	0.014	0.012
120	0.057	0.039	0.031	0.056	0.030	0.020	0.048	0.030	0.033
180	0.055	0.048	0.037	0.056	0.046	0.035	0.061	0.033	0.032

$r_Q = r_\alpha = 3$

$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.042	0.020	0.015	0.063	0.027	0.019	0.033	0.017	0.007
120	0.070	0.040	0.029	0.066	0.029	0.022	0.047	0.023	0.015
180	0.072	0.047	0.040	0.057	0.033	0.031	0.050	0.027	0.024

Panel B: dynamic model (DYN)

$r_\alpha = 1$ and $r_Q = 1$

	GAUSS			STU-T			GARCH		
$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.078	0.049	0.041	0.055	0.032	0.031	0.073	0.056	0.047
120	0.072	0.054	0.048	0.070	0.042	0.038	0.046	0.039	0.033
180	0.084	0.062	0.056	0.086	0.051	0.040	0.047	0.028	0.027

$r_\alpha = 1$ and $r_Q = 2$

$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.057	0.048	0.041	0.046	0.038	0.023	0.067	0.052	0.046
120	0.070	0.057	0.056	0.077	0.043	0.025	0.053	0.034	0.029
180	0.091	0.051	0.048	0.105	0.042	0.041	0.050	0.029	0.020

$r_\alpha = 1$ and $r_Q = 3$

$n \setminus T$	60	120	180	60	120	180	60	120	180
60	0.058	0.035	0.039	0.039	0.028	0.022	0.107	0.060	0.045
120	0.059	0.055	0.041	0.060	0.042	0.030	0.074	0.051	0.036
180	0.082	0.049	0.050	0.091	0.044	0.032	0.069	0.041	0.028

Table 1.3: Forecasting stock returns (annual data).

Structural break points are estimated using the methodology outlined in Section 1.4.4 and assuming that there are at least four years between structural breaks. For $\{\theta_t\}_{t=1}^T$, we find one break point denoted by $T_{\theta,1}$; for $\{\gamma_t\}_{t=1}^T$, we find three break points, denoted by $T_{\gamma,1}$, $T_{\gamma,2}$ and $T_{\gamma,3}$, respectively. See Figure 1.3 for plots for these breaks.

	Estimate	**t-stat**	**Conf interval**		**P-value** (Time variation)
Panel A: return predictability from the dividend yield $\{\theta_t\}_{t=1}^T$					
$T^{-1}\sum_{t=1}^{T}\theta_t$	-0.79	-1.44	-1.87	0.28	0.00
$T_{\theta,1}^{-1}\sum_{t=1}^{T_{\theta,1}}\theta_t$	1.05	5.09	0.64	1.45	0.00
$(T-T_{\theta,1})^{-1}\sum_{t=T_{\theta,1}+1}^{T}\theta_t$	-0.79	-1.44	-1.87	0.28	0.00
Panel B: return predictability from the conditional variance $\{\gamma_t\}_{t=1}^T$					
$T^{-1}\sum_{t=1}^{T}\gamma_t$	-0.27	-0.71	-1.01	0.47	0.00
$T_{\gamma,1}^{-1}\sum_{t=1}^{T_{\gamma,1}}\gamma_t$	5.28	3.33	2.17	8.38	0.00
$(T_{\gamma,2}-T_{\gamma,1})^{-1}\sum_{t=T_{\gamma,1}+1}^{T_{\gamma,2}}\gamma_t$	-6.02	-5.09	-8.33	-3.70	0.00
$(T_{\gamma,3}-T_{\gamma,2})^{-1}\sum_{t=T_{\gamma,2}+1}^{T_{\gamma,3}}\gamma_t$	0.60	0.71	-1.05	2.24	0.00
$(T-T_{\gamma,3})^{-1}\sum_{t=T_{\gamma,3}+1}^{T}\gamma_t$	-0.46	-0.96	-1.41	0.48	0.00

Table 1.4: Forecasting stock returns (quarterly data): difference in predictability

For disjoint sets $A, B \subset \{1, \ldots, T\}$, we consider the difference in average predictability, i.e., $d(A, B)$ defined in (1.4.1). We report the estimates and 95% confidence intervals for $d(A, B)$, as well as the t-stat for testing $d(A, B) = 0$. The sets used in the table are defined as follows.

- For $j \in \{1, 2, 3, 4\}$, $\mathcal{Q}_j = \{t \mid 1 \leq t \leq T$ and time t is quarter j of some year$\}$. The results are also plotted in Figure 1.5.
- $\mathcal{R}_{NBER} = \{t \mid 1 \leq t \leq T \text{ and } NBER_t = 1\}$ and $\mathcal{E}_{NBER} = \{t \mid 1 \leq t \leq T \text{ and } NBER_t = 0\}$, where $NBER_t$ is the NBER indicator for economic recessions, which takes value one if the economy is in recession and takes value zero otherwise. Monthly data of the NBER indicators is obtained from the website of St. Louis Fed and the value of the indicator of the last month in a quarter is used as the value of that quarter.
- $\mathcal{L}_{GDP} = \{t \mid 1 \leq t \leq T \text{ and } GDP_t < \text{median}(GDP)\}$ and $\mathcal{H}_{GDP} = \{t \mid 1 \leq t \leq T \text{ and } GDP_t > \text{median}(GDP)\}$, where GDP_t denotes the real U.S. GDP growth in time period t and median(GDP) denotes the sample median of real GDP growth. We obtain the data from the website of St. Louis Fed.

		θ_t (dividend yield)				γ_t (conditional volatility)			
Set A	Set B	**Est**	**t-stat**	**Conf Interval**		**Est**	**t-stat**	**Conf Interval**	
$\{1, \ldots, T\}$	$\emptyset$	0.49	1.45	-0.17	1.15	-1.66	-3.29	-2.65	-0.67
$\mathcal{Q}_1$	$\emptyset$	0.20	0.33	-0.99	1.38	0.90	0.64	-1.84	3.63
$\mathcal{Q}_2$	$\emptyset$	-0.10	-0.14	-1.51	1.30	-2.76	-3.55	-4.29	-1.23
$\mathcal{Q}_3$	$\emptyset$	2.18	4.19	1.16	3.19	-4.79	-6.91	-6.14	-3.43
$\mathcal{Q}_4$	$\emptyset$	-0.31	-0.59	-1.35	0.73	0.00	0.00	-1.79	1.78
$\mathcal{Q}_1$	$\mathcal{Q}_2$	0.30	0.38	-1.26	1.86	3.66	2.99	1.26	6.05
$\mathcal{Q}_2$	$\mathcal{Q}_3$	-2.28	-2.49	-4.07	-0.48	2.03	1.86	-0.10	4.16
$\mathcal{Q}_3$	$\mathcal{Q}_4$	2.49	3.54	1.11	3.86	-4.78	-5.61	-6.45	-3.11
$\mathcal{R}_{NBER}$	$\emptyset$	2.20	2.56	0.52	3.89	1.22	1.37	-0.53	2.96
$\mathcal{E}_{NBER}$	$\emptyset$	0.22	0.62	-0.48	0.93	-2.11	-3.70	-3.23	-0.99
$\mathcal{R}_{NBER}$	$\mathcal{E}_{NBER}$	1.98	2.16	0.18	3.77	3.33	3.12	1.23	5.42
$\mathcal{L}_{GDP}$	$\emptyset$	1.48	2.67	0.39	2.56	-4.10	-5.75	-5.50	-2.70
$\mathcal{H}_{GDP}$	$\emptyset$	-0.50	-1.52	-1.14	0.14	0.77	1.43	-0.29	1.83
$\mathcal{L}_{GDP}$	$\mathcal{H}_{GDP}$	1.98	3.24	0.78	3.17	-4.87	-6.40	-6.36	-3.38

Table 1.5: Determinants of firms' capital structures

$\beta_{\text{Profit},t}$, $\beta_{\text{Assets},t}$, $\beta_{\text{Mktbk},t}$ and $\beta_{\text{Tang},t}$ represent the components of $\beta_t \in \mathbb{R}^{11}$ corresponding to profitability, assets, market-to-book ratio and tangibility, respectively. The above table reports the point estimate, t-statistic and confidence interval for the average β_t, as well as p-value of the test for lack of parameter instability of β_t described in Section 1.4.3.

	Estimate	**t-stat**	**Conf interval**		**P-value** (Time variation)
Panel A: LV measured as DM					
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Profit},t}$	-0.72	-10.80	-0.84	-0.59	0.01
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Assets},t}$	0.06	6.65	0.04	0.08	0.01
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Mktbk},t}$	-0.03	-5.22	-0.05	-0.02	0.02
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Tang},t}$	0.19	6.96	0.13	0.24	0.02
Panel B: LV measured as DB					
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Profit},t}$	-0.50	-6.87	-0.64	-0.35	0.00
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Assets},t}$	0.04	4.70	0.02	0.05	0.01
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Mktbk},t}$	0.02	2.79	0.00	0.03	0.25
$T^{-1}\sum_{t=1}^{T}\beta_{\text{Tang},t}$	0.18	7.67	0.13	0.23	0.00

Table 1.6: Fixed investment and economic growth

We consider the regression equation (1.6.2). In the above table, columns 1, 2 and 3 report the point estimate, t-statistic and confidence interval for $T^{-1}\sum_{t=1}^{T}\theta_t$. The last column reports the p-value of the test for lack of parameter instability of θ_t described in Section 1.4.3. T_0 is the structural break point estimated using the methodology outlined in Section 1.4.4; see Figure 1.8.

	Estimate	**t-stat**	**Conf interval**		**P-value** (Time variation)
$T^{-1}\sum_{t=1}^{T}\theta_t$	0.023	0.987	-0.022	0.067	0.000
$T_0^{-1}\sum_{t=1}^{T_0}\theta_t$	-0.055	-1.708	-0.117	0.008	0.000
$(T-T_0)^{-1}\sum_{t=T_0+1}^{T}\theta_t$	0.114	2.982	0.039	0.188	0.000

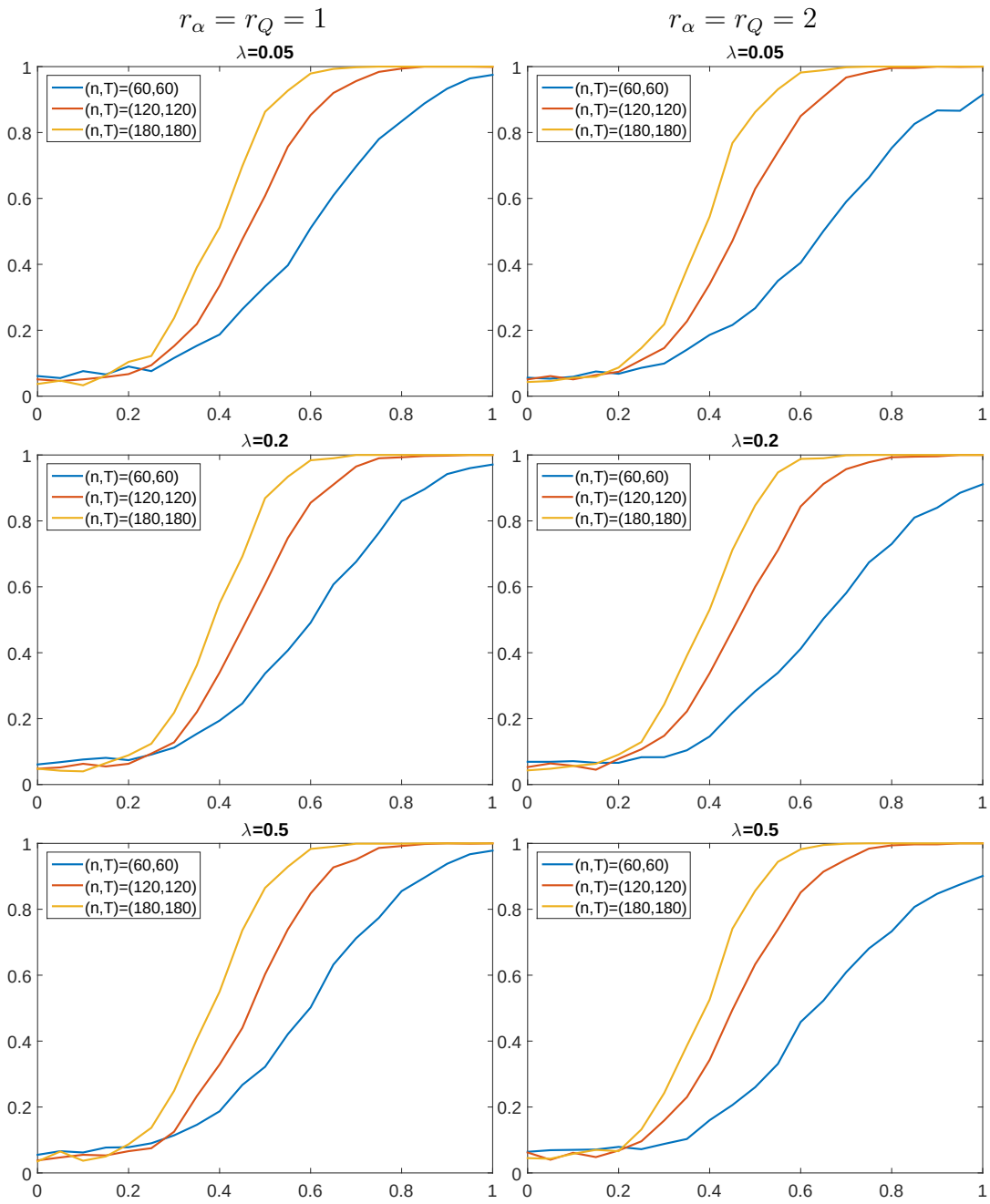


Figure 1.1: Power curves for testing structural breaks in $\{\beta_{1,t}\}_{t=1}^T$ (STA)

We generate $\beta_{1,1} = \ldots = \beta_{1,\lfloor \lambda T \rfloor}$ and $\beta_{1,\lfloor \lambda T \rfloor+1} = \ldots = \beta_{1,T}$ with $\delta = \beta_{1,\lfloor \lambda T \rfloor+1} - \beta_{1,\lfloor \lambda T \rfloor}$. In the above plots, we report the probability of rejecting $\beta_{1,1} = \ldots = \beta_{1,T}$ as a function of δ, for various values of (n, T) and λ.

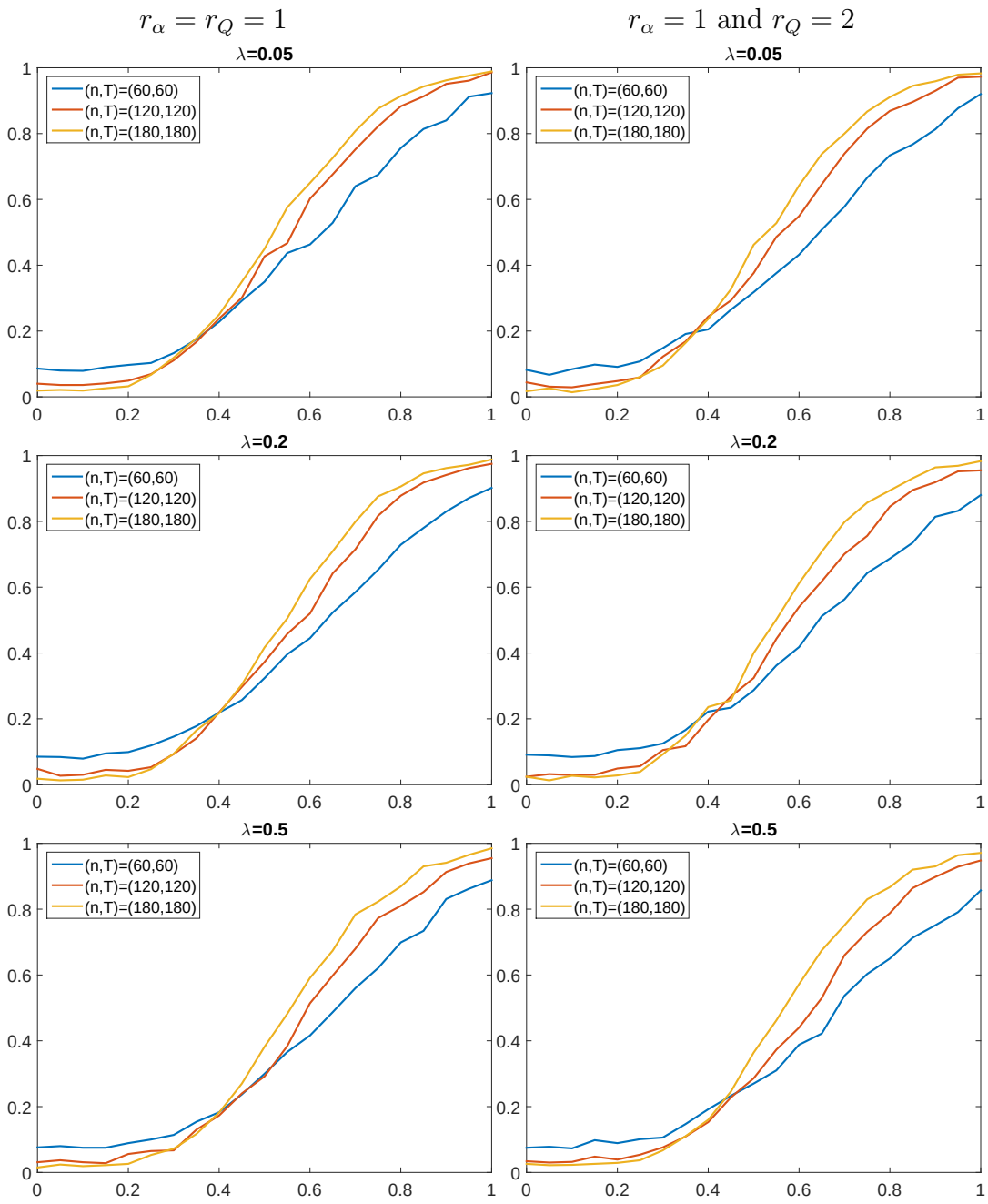
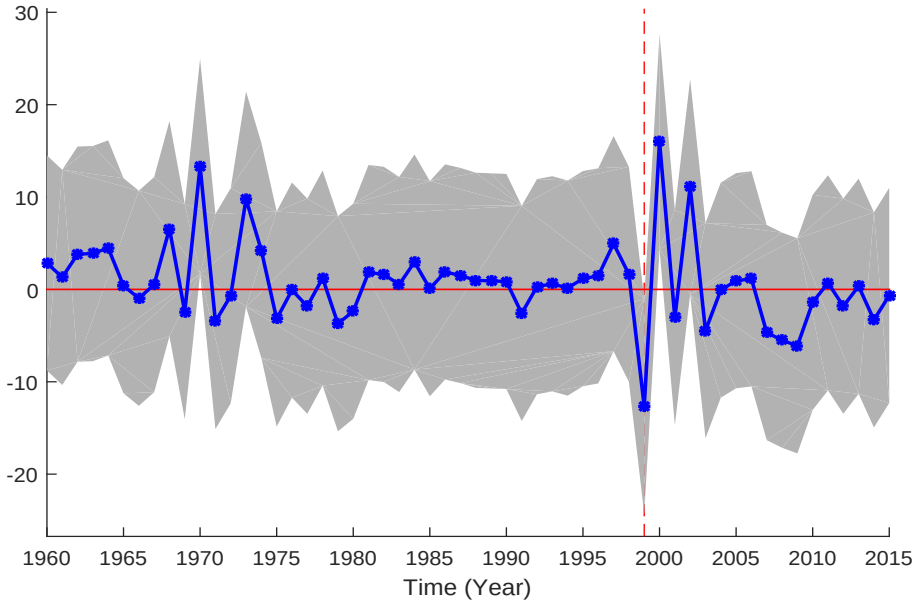


Figure 1.2: Power curves for testing structural breaks in $\{\beta_{1,t}\}_{t=1}^T$ (DYN)

We generate $\beta_{1,1} = \ldots = \beta_{1,\lfloor \lambda T \rfloor}$ and $\beta_{1,\lfloor \lambda T \rfloor+1} = \ldots = \beta_{1,T}$ with $\delta = \beta_{1,\lfloor \lambda T \rfloor+1} - \beta_{1,\lfloor \lambda T \rfloor}$. In the above plots, we report the probability of rejecting $\beta_{1,1} = \ldots = \beta_{1,T}$ as a function of δ, for various values of (n, T) and λ.

Panel A: Estimate and 95% confidence band for $\{\theta_t\}_{t=1}^{T}$ (predictive power of the dividend yield)



Panel B: Estimate and 95% confidence band for $\{\gamma_t\}_{t=1}^{T}$ (predictive power of volatility)

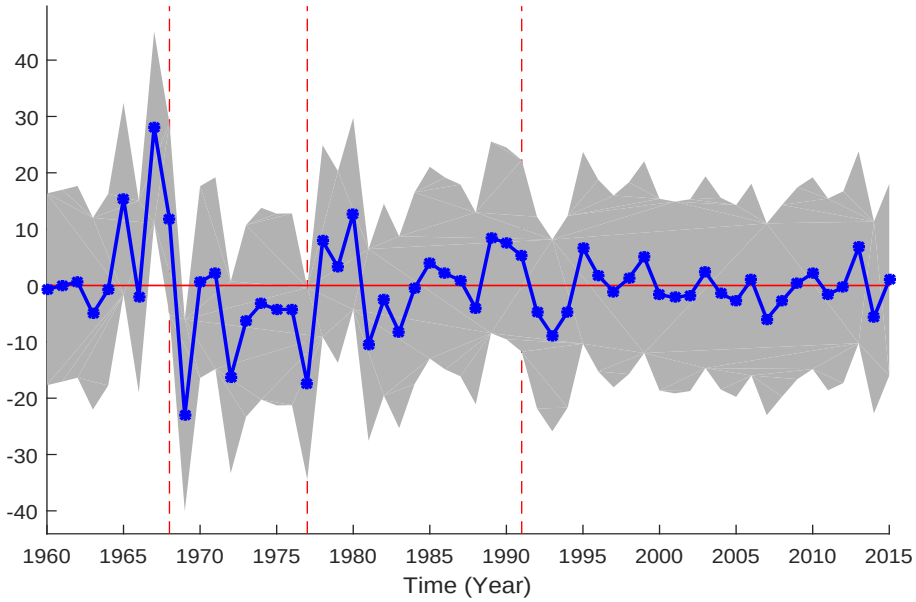


Figure 1.3: Predictability of stock returns (annual data)
The blue line represents the estimate for $\{\theta_t\}_{t=1}^{T}$ (or $\{\gamma_t\}_{t=1}^{T}$) and the shaded area is the 95% confidence band. The red dashed vertical lines are the structural break points estimated using the methodology outlined in Section 1.4.4 and assuming that there are at least four years between structural breaks.

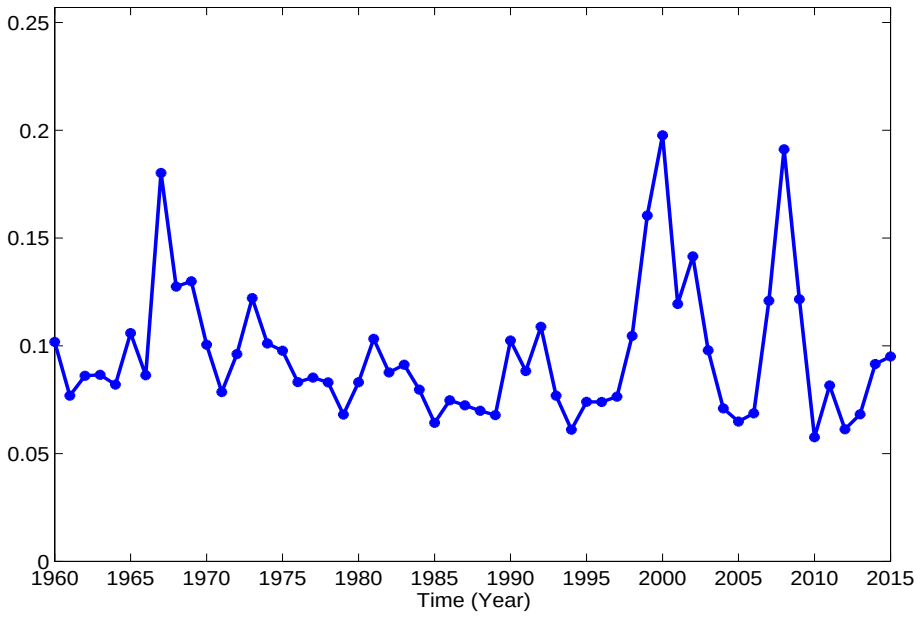
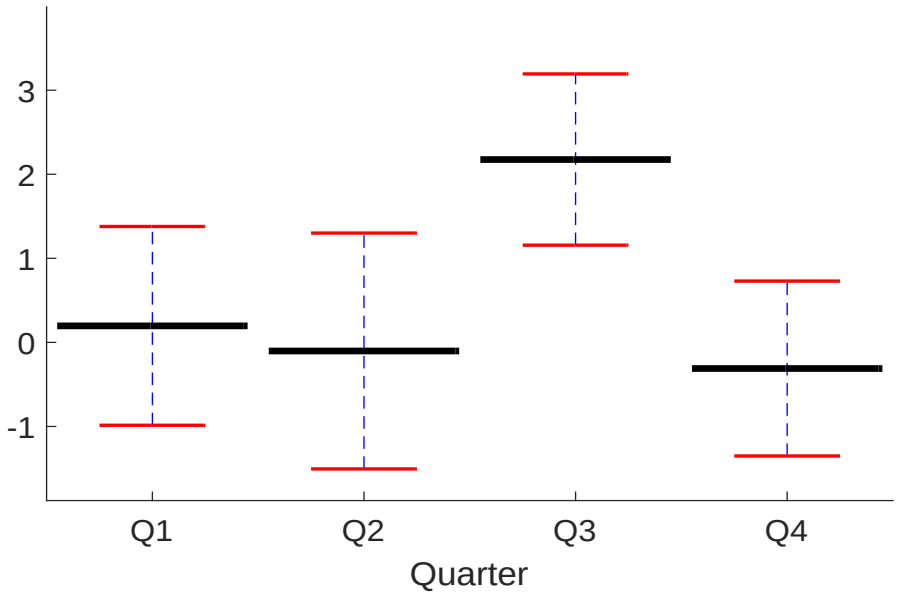


Figure 1.4: Predictability of stock returns (annual data): average noise level in error terms

We plot the average noise level in error terms $\{\hat{\bar{\sigma}}_{u,t}\}_{t=1}^{T}$ defined by $\hat{\bar{\sigma}}_{u,t}^2 = n^{-1}\sum_{i=1}^{n}\hat{u}_{i,t}^2$, where $\hat{u}_{i,t}$ is defined in Algorithm 2.

Panel A: Average θ_t in each quarter of the year (predictive power of the dividend yield)



Panel B: Average γ_t in each quarter of the year (predictive power of volatility)

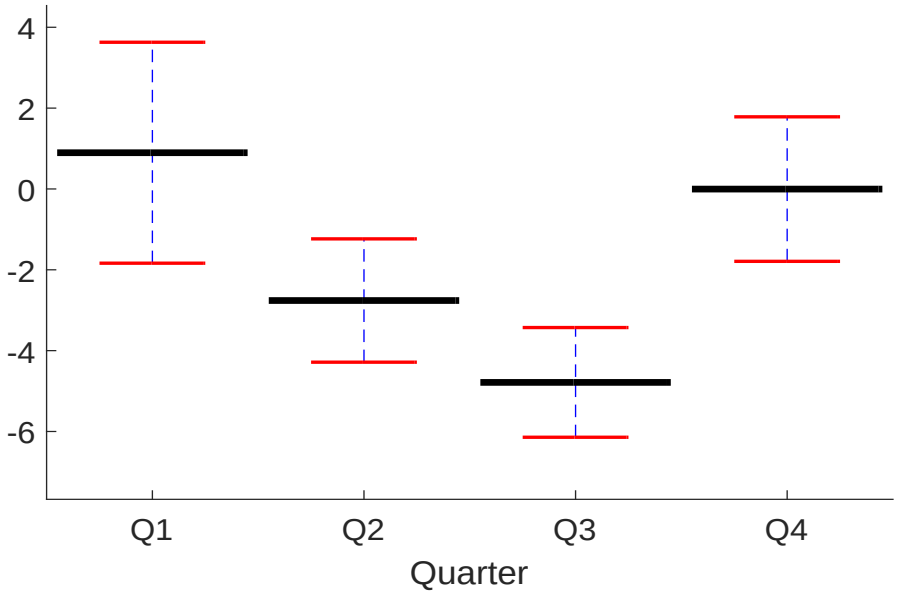


Figure 1.5: Seasonality of return predictability (quarterly data)

The black line represents the average θ_t (or γ_t) with one quarter of all the years and the red lines denote the 95% confidence interval of this average.

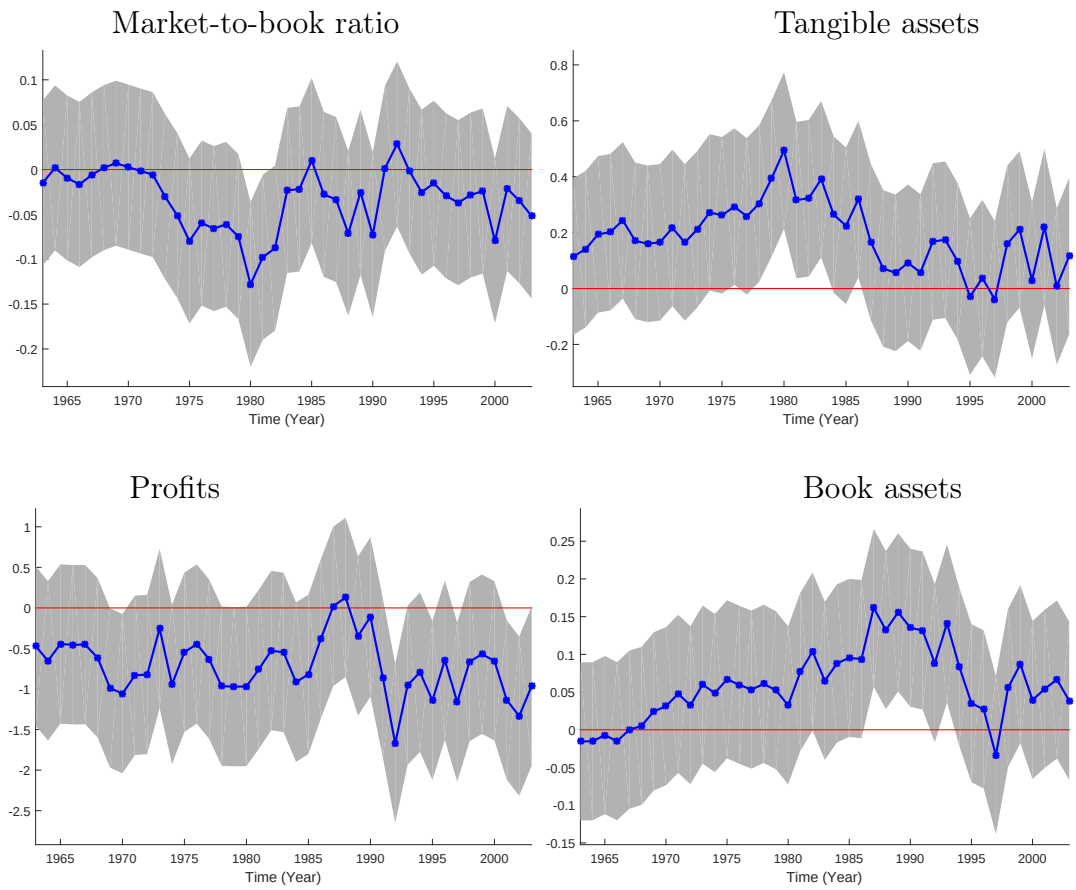


Figure 1.6: Firms' capital structure decisions (leverage ratio defined as DM)

We consider the components of β_t corresponding to market-to-book ratio, tangible assets, book assets and profits. The blue lines are estimates for β. The shaded area is the 95% confidence set for β.

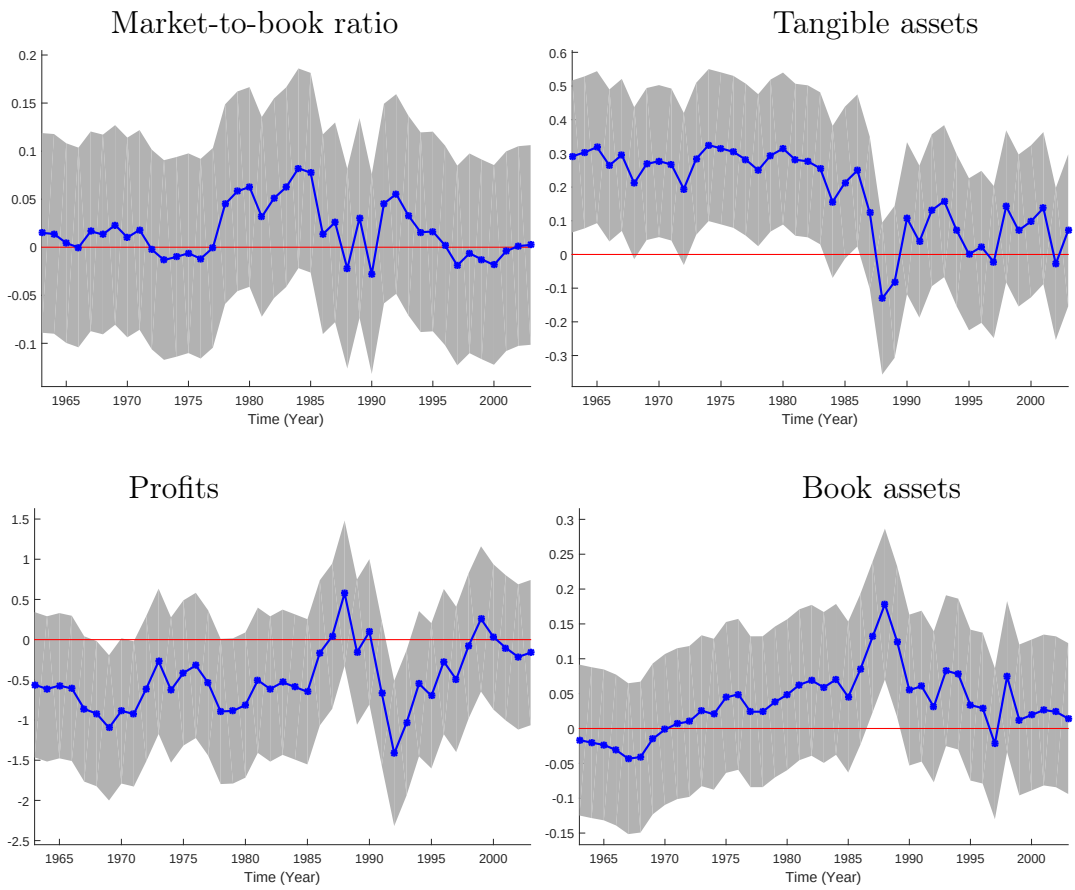


Figure 1.7: Firms' capital structure decisions (leverage ratio defined as DB)

We consider the components of β_t corresponding to market-to-book ratio, tangible assets, book assets and profits. The blue lines are estimates for β. The shaded area is the 95% confidence set for β.

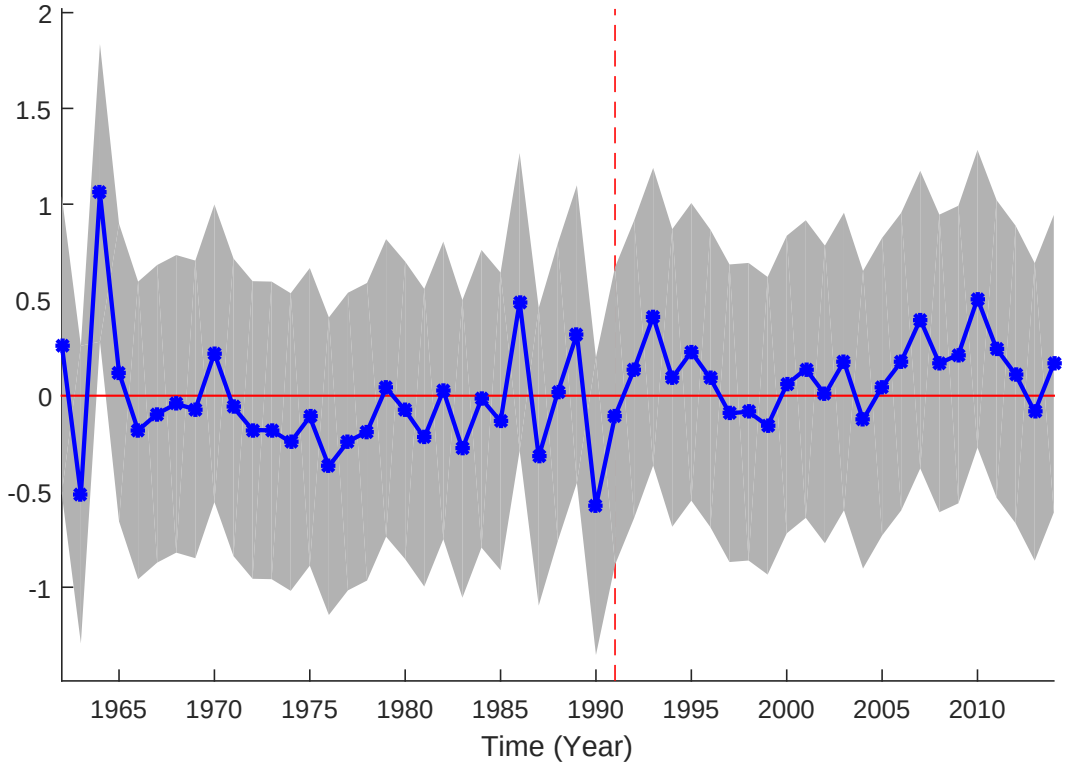


Figure 1.8: Fixed investment and economic growth

We consider the regression (1.6.2). The blue lines represent the estimate for $\{\theta_t\}_{t=1}^{T}$ and the shaded area is the 95% confidence band. The red dashed vertical line is the structural break point estimated using the methodology outlined in Section 1.4.4 and assuming that there are at least two years between structural breaks.

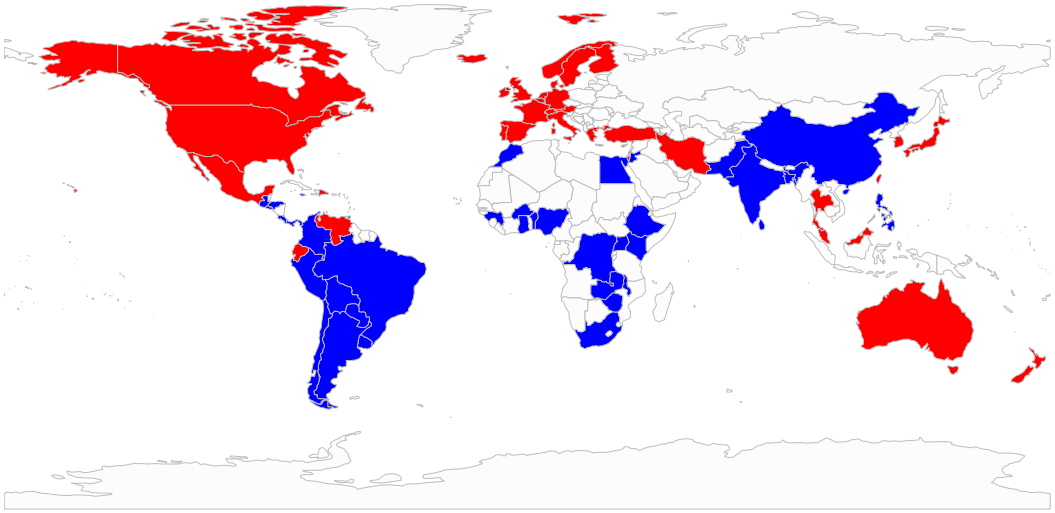


Figure 1.9: Fixed investment and economic growth: grouped pattern of fixed effects

We apply the the k-means clustering algorithm (Forgy 1965; Lloyd 1982) to the estimated fixed effects $\hat{\alpha}_{i,t}$ obtained in Algorithm 2. The estimated fixed effects are clustered into two groups, labeled by the red and blue colors.

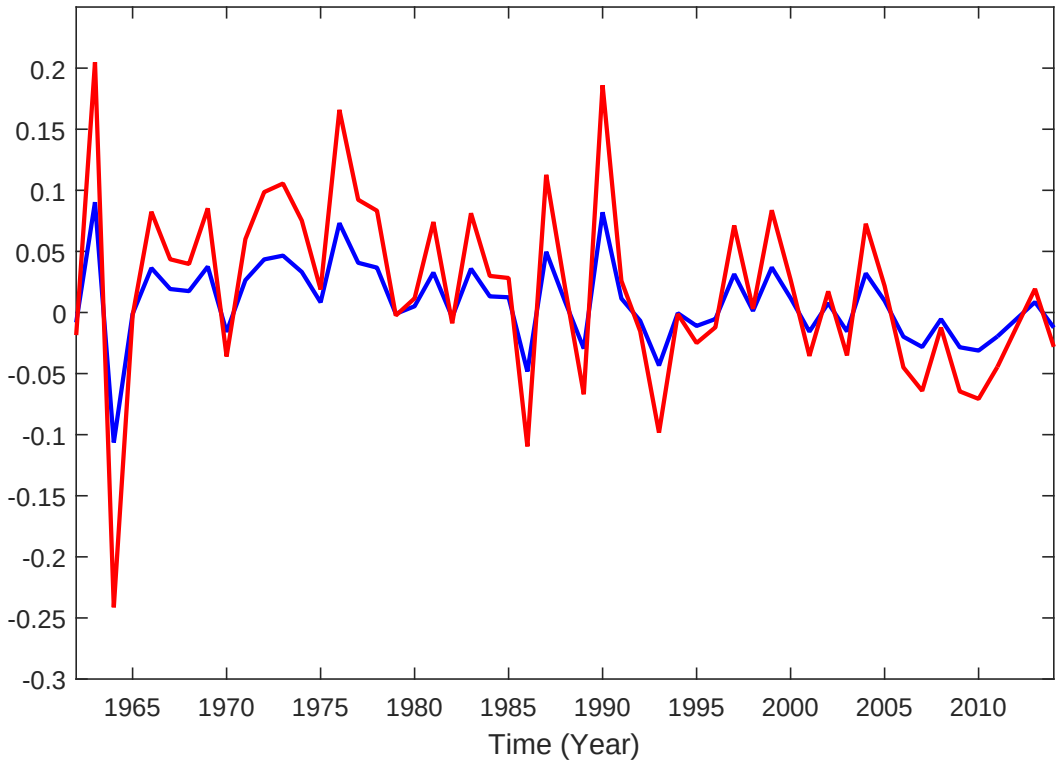


Figure 1.10: Fixed investment and economic growth: trajectories of grouped pattern of fixed effects

We apply the the k-means clustering algorithm (Forgy 1965; Lloyd 1982) to the estimated fixed effects $\hat{\alpha}_{i,t}$ obtained in Algorithm 2. The estimated fixed effects are clustered into two groups, labeled by the red and blue colors. Here, we plot trajectories of the average fixed effects in the two groups under the same color label as in Figure 1.9. For example, the red color represents the same group in both this figure and Figure 1.9.

Chapter 2

Testing for common factors in large factor models

2.1 Introduction

In economics and finance, large factor models are very popular due to their well-documented success in forecasting. A small number of factors can explain a large fraction of the variations in a large number of variables and also have high predicting power for economic and financial variables; see Stock and Watson (2002a, 2002b, 2015), Bernanke and Boivin (2003), Bernanke, Boivin, and Eliasz (2005), Ludvigson and Ng (2007, 2009), Foerster, Sarte, and Watson (2011) and McCracken and Ng (2015). To explain the predictive power of factor models and to interpret structural models, it is essential to determine the economic nature of the factors.

Studying the interpretation of factors often involves the question of which factors drive which variables. Consider, for example, a large dataset consisting of several groups of macroeconomic variables, such as in Stock and Watson (2002b) and McCracken and Ng (2015). To determine the nature of the factors, one would like to know (1) whether the same factors drive both labor market variables and output variables and (2) if not, how many factors drive both groups of variables and how many factors are unique to labor market variables. If one factor were all that drives both groups, then it might not make much sense to label one factor as

labor and another as output.

In the asset pricing literature, factor models are popular in explaining returns of a large number of financial assets. The factors are usually interpreted as risks that are priced in these assets in the sense of Ross (1976)'s arbitrage pricing theory. For factor-based asset allocation strategies (see Ang (2014)), a key ingredient is to maintain the desired level of exposure to certain risk factors. Hence, questions of both empirical interest and practical importance include which risks are priced in which assets. As a simple example, in order to understand what risks these factors represent, one would be interested in (1) whether equities and fixed income assets have any risk factors in common and (2) how many risk factors are peculiar to equities.

In this paper, we first formalize the notion of common factors given two groups of variables. We then propose a statistical test for the null hypothesis that the number of common factors in these two groups is equal to a given number.

The problem we study here cannot be addressed by the popular prediction-based method used by Stock and Watson (2002b) and McCracken and Ng (2015). Under this prediction-based method, one extracts as factors the principal components (PCs) from a large dataset and then assigns names to these PCs according to their predictive power for certain economic or financial variables. For example, if the first PC has very high predictive power for labor market variables and the second PC explains a large portion of variations in output variables, then these two PCs would be labeled as labor and output factors, respectively. The prediction-based method imposes a particular normalization for the PCs and studies the predictive power of these PCs for different groups of variables. In contrast, we aim to test hypotheses regarding the underlying factor, which may differ from PCs, and ask the question whether a certain number of the underlying factors driving one group of variables coincide (possibly after rotation) with some of those driving another group of variables. Since the imposed normalization might not take into account the potential correlation among the true underlying factors, using only the predictive ability of the PCs satisfying this normalization might not yield satisfactory answers to the questions we consider in this paper. We illustrate this point through the

following example.

Example 2.1.1. There are six groups of random variables denoted by price, output, labor, consumption, money and inventory. These groups are driven by three factors, denoted by F1, F2 and F3. The variables in "price" group only load on F1, output variables only on the F2, labor variables only on F2 and F3, consumption variables on F1 and F2, money variables on F1 and F3 and inventory variables load on all three factors. The non-zero loadings are generated as independent $N(0,1)$ random variables; we generate serially uncorrelated factors from $N(0,\Sigma_F)$ with $\Sigma_F \in \mathbb{R}^{3\times 3}$ having ones on the diagonal and $\rho = 0.6$ on the off diagonal entries. Idiosyncratic terms are zero.

Since there are no error terms, extracted PCs are exactly a rotated version of the true factors. We compute the R^2 of regressing each variable on each PC and average these R^2 values in each group. The means of these averaged R^2 values are reported in Table 2.1.

Table 2.1: Mean of the average R^2 for each PC in Example 2.1.1

These results are based on 5000 simulated samples. Each group has $n = 100$ variables and the time horizon is $T = 100$ periods. The jth entry in each row is computed as follows. In each simulated sample and for each group, we compute R^2 of regressing every variable in that group against the jth PC and take the average of these R^2's across variables in the group. Then we report the mean (across 5000 random samples) of this average in the jth entry of the row corresponding to the group.

Group (actual factors) \ PC	PC1	PC2	PC3
Price (F1)	0.756	0.173	0.071
Output (F2)	0.757	0.174	0.069
Labor (F2 and F3)	0.599	0.177	0.225
Consumption (F1 and F2)	0.634	0.274	0.092
Money (F1 and F3)	0.599	0.175	0.226
Inventory (F1, F2 and F3)	0.572	0.231	0.197

From Table 2.1, the first PC explains over 70% of the variations in variables in "price" and "output" groups.[1] The prediction-based approach might name the

[1]In practice, it is not uncommon for the researcher to attribute the unexplained variation to idiosyncratic errors, which are pervasive in real data.

first PC as price-output factor, which appears to be shared by these two groups. However, these two groups are actually driven by two different (although correlated) factors. Also notice that PCs have very similar predictive power for "money" and "labor" groups. The prediction-based approach might conclude that they have similar factor loadings, but they actually share only one factor. In contrast, the procedure proposed in this paper aim to reveal the underlying structure of the factors. For instance, in Example 2.1.1, the proposed procedure would conclude (1) that the factors in "price" group are different from those in the "output" group and (2) that "price" group and "money" group share exactly one factor. Based on only correlation between variables and PCs (as displayed in Table 2.1), the prediction-based method seems unlikely to yield such findings.

Our paper contributes to the literature on large factor models in several ways. First, we formalize the notion of common factors among subsets/subgroups in a large dataset and propose to use it as a general tool to study the structure of factors. The number of common factors sheds light on the nature of factors by providing information on whether or not a factor is specific to a certain subgroup of variables. We apply the proposed method to a large dataset consisting of many macroeconomic and financial variables used by Jurado, Ludvigson, and Ng (2015) and find (1) that there are at most three common factors driving both the macroeconomy and the financial markets, (2) that there are at least 2 factors that drive the macroeconomy but not the financial markets and (3) that there are at least 4 factors peculiar to the financial markets; see Section 2.5 for details.

Second, other hypotheses of economic interest can be phrased in terms of common factors. Consider the spanning hypothesis, which states that the factors driving one group of variables are linear combinations of those driving another group of variables. Testing the spanning hypothesis is equivalent to testing whether the number of common factors is equal to the number of factors in the former group.

Third, to the best of our knowledge, this paper provides the first statistical inference result regarding the number of common factors. Instead of using the usual tools, such as the classical central limit theorems (CLTs) or the random matrix

theory, we construct the procedure as a test for a high-dimensional parameter and build upon recent results by Chernozhukov, Chetverikov, and Kato (2014) on high-dimensional bootstrap. Our method can also be used to test the number of factors in models with n and T having the same order of magnitude; see Remark 2.3.5 for details.

Our work is related to several strands of the literature on large factor models. Early work focuses on estimation; see Bai and Ng (2002) and Stock and Watson (2002a). More recent work deals with the inference of large factor models. Bai (2003) provides pointwise[2] results on the inference of factors and factor loadings. Bai and Ng (2006a, 2008a) consider the inference problem of using estimated factors in regressions. Bai and Ng (2006b) and Chen (2012) propose tests on the relationship between the factors and observed variables. Tests for the number of factors are developed by Onatski (2009) and Kapetanios (2010). In this paper, we only consider the so-called "static" factor models; see, among many others, Forni, Hallin, Lippi, and Reichlin (2000, 2004, 2012), Amengual and Watson (2007), Bai and Ng (2012) and Doz, Giannone, and Reichlin (2012) for results on "dynamic" factor models. Excellent survey of the large factor model literature can be found in Bai and Ng (2008b), Bai and Wang (2016) and Stock and Watson (2015). Some recent work also applies high-dimensional statistical tools to large factor models. For example, Cheng, Liao, and Schorfheide (2016) proposed a method for detecting breaks in factors and/or their loadings using a shrinkage approach that combines the group Lasso (Yuan and Lin (2006)) and the adaptive Lasso (Zou (2006)). Bai and Liao (2012) developed a regularized estimation procedure that exploits sparsity in the covariance matrix of the idiosyncratic terms.

The rest of this paper is organized as follows. In Section 2.2, we introduce the concept of common factors and develop the testing methodology. In Section 2.3, we establish theoretical properties of the proposed method. Finite sample performance is assessed through Monte Carlo simulations in Section 2.4. In Section 2.5, we apply the proposed method to a real dataset and study common factors between macroeconomic and financial variables. Appendices B.1-B.3 contain the

[2]Here, the term "pointwise" means that the asymptotic distribution is established for estimators of factors at each time period or factor loadings of each individual/variable.

proofs of the theoretical results as well as technical tools used in the proofs.

Notation. *We introduce some notations that will be used in the rest of the paper. For a positive integer* q, $[q] = \{1, \cdots, q\}$*; for* $q \leq 0$*, the convention is* $[q] = \emptyset$*.* I_q *denotes the* $q \times q$ *identity matrix. For a vector* $v = (v_1, \cdots, v_q)' \in \mathbb{R}^q$, $\text{Diag}(v)$ *denotes the* $q \times q$ *diagonal matrix whose* i*th entry on the diagonal is* v_i*. For a matrix* $A \in \mathbb{R}^{q_1 \times q_2}$, $\|A\|_\infty = \|\text{vec}(A)\|_\infty$ *and* $\|A\|$ *denotes the spectral norm (largest singular value), where* $\text{vec}(\cdot)$ *denotes column-wise vectorization. We say that* $A = U_A S_A V_A'$ *is a singular value decomposition (SVD) if* $U_A \in \mathbb{R}^{q_1 \times q_1}$ *and* $V_A \in \mathbb{R}^{q_2 \times q_2}$ *are both orthogonal matrices and* $S_A \in \mathbb{R}^{q_1 \times q_2}$ *is a (rectangular) diagonal matrix with singular values of* A *on the diagonal in the non-increasing order. For a matrix* $A \in \mathbb{R}^{q_1 \times q_2}$ *of full column rank, let* $\Pi_A = A(A'A)^{-1}A'$ *and* $M_A = I_{q_1} - \Pi_A$*. For two matrices* A_1 *and* A_2, $\text{Blockdiag}(A_1, A_2)$ *denotes the block diagonal matrix with* A_1 *and* A_2 *on the diagonal. For* $a, b \in \mathbb{R}$, $a \vee b$ *and* $a \wedge b$ *denote* $\max\{a, b\}$ *and* $\min\{a, b\}$*, respectively. Let* $\sigma(\cdot)$ *denote the* σ*-algebra generated by random variables.*

2.2 Methodology

Consider the observed data Y and W generated from the following model:

$$\underset{n_Y \times T}{Y} = \underset{n_Y \times p_Y}{L} \ \underset{p_Y \times T}{F'} + \underset{n_Y \times T}{e} \tag{2.2.1}$$

and

$$\underset{n_W \times T}{W} = \underset{n_W \times p_W}{R} \ \underset{p_W \times T}{X'} + \underset{n_W \times T}{u}, \tag{2.2.2}$$

where L and R are factor loadings, F and X are factors and e and u are idiosyncratic terms. In this paper, n_Y, n_W and T tend to infinity. We assume that the factors are not redundant: $\text{rank}L = \text{rank}F = p_Y$ and $\text{rank}R = \text{rank}X = p_W$. We assume that p_Y and p_W are known or can be consistently estimated using methods such as those proposed in Bai and Ng (2002) and Ahn and Horenstein (2013). We introduce the following definition.

Definition 2.2.1 (Common factors)**.** Under the model (2.2.1)-(2.2.2), the number of common factors between Y and W, denoted by p_C, is the largest number in the set $\mathcal{C}_{F,X}$, where

$$\mathcal{C}_{F,X} = \{j \mid \exists R_1 \in \mathbb{R}^{p_Y \times j}, R_2 \in \mathbb{R}^{p_W \times j} \text{ satisfying } FR_1 = XR_2 \text{ and } \text{rank}R_1 = \text{rank}R_2 = j\}.$$

Remark 2.2.1. Notice that $\mathcal{C}_{F,X}$ always contains zero. If $\mathcal{C}_{F,X} = \{0\}$, then F and X do not have any common components in the sense that, under any rotation, no column in F coincides with one column in X. If p_C, the maximal element in $\mathcal{C}_{F,X} \neq \{0\}$, is not zero, then we can rotate F and X such that p_C columns of F match exactly p_C columns of X; we refer to these p_C columns as the common factors. Under this rotation, there are $p_Y - p_C$ factors specific to Y and there are $p_W - p_C$ factors that drive W but not Y.

The goal of this paper is to test

$$H_0 : \ p_C = k_0 \tag{2.2.3}$$

versus

$$H_1 : \ p_C < k_0 \tag{2.2.4}$$

Consider the combined dataset:

$$\underset{n\times T}{\chi} = \underset{n\times r}{\Lambda}\ \underset{r\times T}{Z'} + \underset{n\times T}{v}, \tag{2.2.5}$$

where $\chi = [Y', W']'$, $n = n_Y + n_W$, $r = \text{rank}([F, X])$, $v = [e', u']'$ and Λ and Z satisfy $\Lambda Z' = \text{Blockdiag}(L, R)[F, X]'$. We exploit the following equivalence relation in constructing our test.

Lemma 2.2.1. *Consider the model (2.2.1)-(2.2.2). Then, p_C, the number of common factors between Y and W is equal to j if and only if $r = p_Y + p_W - j$ and $M_Z[F, X] = 0$.*

The proof of Lemma 2.2.1 is provided in Appendix B.2. Roughly speaking,

our proposed methodology for testing H_0 is to estimate $M_Z[F, X]$ imposing $r = r_0$ and to check whether this estimate is close enough to zero, where $r_0 = p_Y + p_W - k_0$. Applying principal component analysis (PCA) to Y and W, we obtain $\hat{F}$ and $\hat{X}$, estimators of a rotated version of F and X, respectively. To estimate Z, we apply PCA to the combined dataset imposing $r = r_0$ and obtain $\hat{Z}$.

The idea of our test is the following. Suppose that $\hat{F}$, $\hat{X}$ and $\hat{Z}$ are reasonably good estimates of (a rotated version of) F, X and Z. Under H_0, Lemma 2.2.1 implies that all the entries in $M_{\hat{Z}}[\hat{F}, \hat{X}]$ should be close to zero; if H_1 holds, then Lemma 2.2.1 implies that $r = p_Y + p_W - p_C > r_0$. Therefore, under H_1, $\hat{Z}$ estimated with $r = r_0$ does not contain all the factors in the combined dataset and thus cannot span $[F, X]$. Therefore, under H_1, at least some entries in $M_{\hat{Z}}[\hat{F}, \hat{X}]$ do not converge to zero, leading to power against H_1.

We consider $\|\sqrt{n}M_{\hat{Z}}[\hat{F}, \hat{X}]\|_\infty$ as the test statistic. Under H_0, $\|\sqrt{n}M_{\hat{Z}}[\hat{F}, \hat{X}]\|_\infty$ is the maximal (scaled) estimation error of entries in $M_Z[F, X]$. Since $M_Z[F, X]$ is $(p_Y + p_W)T$-dimensional with T tending to infinity, the usual tools for low-dimensional problems, such as the classical CLT, cannot be used to obtain an asymptotic distribution for our test statistic. Instead, we exploit recent results on high-dimensional multiplier bootstrap proposed by Chernozhukov, Chetverikov, and Kato (2013, 2014).

To see the intuition behind our bootstrap scheme, we need some notations. For the matrices in (2.2.1), (2.2.2) and (2.2.5), the tth rows of F, X and Z are denoted by F_t', X_t' and Z_t', respectively. The ith rows in L, R, Λ and v are denoted by L_i', R_i', Λ_i' and v_i', respectively. We define $\Sigma_F = F'F/T$, $\Sigma_X = X'X/T$, $\Sigma_L = L'L/n_Y$ and $\Sigma_R = R'R/n_W$. We also define $\zeta = [F, X]$ and $\Sigma_\zeta = T^{-1}\zeta'\zeta$. PCA is used to estimate Z, F and X:

$$\begin{cases} \hat{\Lambda} = \sqrt{n}\hat{U}_{\chi,(r)}, \ \hat{Z} = \chi'\hat{\Lambda}/n \text{ and } \hat{v} = \chi - \hat{\Lambda}\hat{Z}' \\ \hat{L} = \sqrt{n_Y}\hat{U}_{Y,(k)}, \quad \hat{F} = Y'\hat{L}/n_Y \text{ and } \hat{Q}_F = (\hat{Z}'\hat{Z})^{-1}(\hat{Z}'\hat{F}) \\ \hat{R} = \sqrt{n_W}\hat{U}_{W,(p)}, \ \hat{X} = W'\hat{R}/n_W \text{ and } \hat{Q}_X = (\hat{Z}'\hat{Z})^{-1}(\hat{Z}'\hat{X}) \end{cases} \tag{2.2.6}$$

where $\chi = \hat{U}_\chi\hat{S}_\chi\hat{V}_\chi'$, $Y = \hat{U}_Y\hat{S}_Y\hat{V}_Y'$ and $W = \hat{U}_W\hat{S}_W\hat{V}_W'$ are SVDs, $\hat{U}_{\chi,(r)}$ the first

$r = p_Y + p_W - p_C$ columns of $\hat{U}_\chi$, $\hat{U}_{Y,(k)} \in \mathbb{R}^{n\times k}$ the first p_Y columns of $\hat{U}_Y$ and $\hat{U}_{W,(p)} \in \mathbb{R}^{n\times p}$ the first p_W columns of $\hat{U}_W$.

The critical value is obtained from a bootstrap procedure based on the following idea. As shown in Appendix B.1, under H_0, the $T\times p$ matrix $n^{1/2}M_{\hat{Z}}[\hat{F},\hat{X}]$ (up to sign changes in its columns) can be approximated in $\|\cdot\|_\infty$-norm by

$$n^{-1/2}\sum_{i=1}^{n} v_i\hat{\Gamma}_i', \tag{2.2.7}$$

where $\hat{\Gamma}_i'$ is the ith row of the matrix $\hat{\Gamma} = -\hat{\Lambda}[\hat{Q}_F,\hat{Q}_X] + \text{Blockdiag}(n_Y^{-1}n\hat{L}, n_W^{-1}n\hat{R})$. This means that the test statistic can be approximated by the $\|\cdot\|_\infty$-norm of the sum of n nearly independent terms, where each term has dimension $T(p_Y + p_W)$. This motivates a multiplier bootstrap scheme similar to the ones studied by Chernozhukov, Chetverikov, and Kato (2013, 2014). Our proposed procedure is summarized below.

Algorithm 3. *The test for H_0 in (2.2.3) of nominal size α is implemented as follows:*

1. *Compute $\hat{\Lambda}$, $\hat{Z}$, $\hat{L}$, $\hat{F}$, $\hat{R}$, $\hat{X}$, $\hat{v}$, $\hat{Q}_F$ and $\hat{Q}_X$ as in (2.2.6), as well as $\hat{\Gamma} = -\hat{\Lambda}[\hat{Q}_F,\hat{Q}_X] + \text{Blockdiag}(n_Y^{-1}n\hat{L}, n_W^{-1}n\hat{R})$.*

2. *Compute the test statistic $S_n = \|n^{1/2}M_{\hat{Z}}[\hat{F},\hat{X}]\|_\infty$.*

3. *Generate vectors $\xi^{(n)} \sim N(0, I_n)$ independent of the data and compute $S_n^{BS} = \|n^{-1/2}\hat{v}'\text{Diag}(\xi^{(n)})\hat{\Gamma} - (n^{-1/2}\mathbf{1}_n'\xi^{(n)})\hat{v}'\hat{\Gamma}\|_\infty$, where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of ones.*

4. *For a test of nominal size α, repeat the previous step as many times as computationally convenient and compute $\mathcal{Q}\left(1-\alpha, S_n^{BS}\right) = \inf\{x \in \mathbb{R} \mid \mathbb{P}\left(S_n^{BS} > x \mid W, Y\right) \le \alpha\}$.*

5. *Reject H_0 in (2.2.3) if and only if $S_n > \mathcal{Q}\left(1-\alpha, S_n^{BS}\right)$.*

This procedure is computationally simple and fast since it only involves performing SVDs and repeatedly generating standard normal random variables.

Remark 2.2.2. Equivalently, one can also implement the test whose output is the p-value. In Step 4, we can compute (by simulation) the empirical distribution function of S_n^{BS}: $F_n^{BS}(x) = \mathbb{P}(S_n^{BS} \leq x \mid W, Y)$. Then the p-value of the test is $U_n = 1 - F_n^{BS}(S_n)$ and the test is to reject H_0 if and only if $U_n \leq \alpha$.

Remark 2.2.3. In Step 3, we can also use $S_n^{BS} = \|n^{-1/2}\hat{v}'\text{Diag}(\xi^{(n)})\hat{\Gamma}\|_\infty$ without significantly changing the proofs. The rationale of the expression stated in Step 3 is to bootstrap recentered quantities: $n^{-1/2}\hat{v}'\text{Diag}(\xi^{(n)})\hat{\Gamma} - (n^{-1/2}\mathbf{1}_n'\xi^{(n)})\hat{v}'\hat{\Gamma} = n^{-1/2}\sum_{i=1}^n(\hat{v}_i\hat{\Gamma}_i' - \overline{v\Gamma})\xi_i^{(n)}$, where $\xi_i^{(n)}$ is the ith entry of $\xi^{(n)}$ and $\overline{v\Gamma} = n^{-1}\sum_{i=1}^n \hat{v}_i\hat{\Gamma}_i'$.

Remark 2.2.4. As mentioned in Section 2.1, if $p_Y \leq p_W$ and $k_0 = p_Y$, then Algorithm 3 becomes a test for the spanning hypothesis that columns of F are linear combinations of columns in X.

Remark 2.2.5. Since the triple (p_Y, p_W, p_C) tells us the number of common factors as well as the number of factors unique to each group, we can view (p_Y, p_W, p_C) as the structure of the factors. Notice that we can construct confidence sets for (p_Y, p_W, p_C) by inverting the test summarized in Algorithm 3. This approach is particular useful when reasonably good estimates for p_Y and p_W are not available. We apply this approach in Section 2.5.

2.3 Theoretical results

Before describing our assumptions, we introduce the following concept.

Definition 2.3.1 (Exponential-type tails)**.** A random variable M is said to have an exponential-type tail with parameter (b, h) if $\forall z > 0$, $\mathbb{P}(|M| > z) \leq \exp\left[1 - (z/b)^h\right]$.

Remark 2.3.1. Random variables with exponential-type tails include polynomials of Gaussian random vectors. Finite mixtures of random variables with exponential-type tails also have exponential-type tails.

We impose the following regularity conditions.

Assumption 3. *For some constants $\beta, \gamma, \kappa, \rho \in (0, \infty)$, the following conditions hold:*

(i) Each entry of X, R, L, F, u and e has an exponential-type tail with parameter (β, γ).
(ii) With probability approaching one, (κ^{-1}, κ) contains n_Y/T, n_W/T and all the eigenvalues of Σ_L, Σ_R, Σ_F and Σ_X, as well as the first r eigenvalues of Σ_ζ.
(iii) v is independent of (L, F, X, R) and $\{v_i\}_{i=1}^n$ is a sequence of n independent vectors in $\mathbb{R}^T$ with mean zero.
(iv) $\max_{(i,s)\in[n]\times[T]} \sum_{t=1}^T |\mathbb{E}v_{i,t}v_{i,s}| = O(\log^\kappa n)$ *and* $\max_{i\in[n]} \|\mathbb{E}v_i v_i'\| \leq \kappa$.
(v) $3\rho^{-1} + \gamma^{-1} > 1$ *and* $\alpha_{mixing}(t) \leq \exp(-\rho t^\rho)$ $\forall t \geq 1$, *where*

$$\alpha_{mixing}(t) := \sup\Big\{ \Big| \mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \bigcap B) \Big| : A \in \sigma\left(\{(X_s, F_s, e_s, u_s, R, L) \mid s \leq l\}\right), B \in \sigma\left(\{(X_s, F_s, e_s, u_s, R, L) \mid s \geq l + t\}\right) \text{ and } l \in \mathbb{Z} \Big\}.$$

(vi) $\min_{(i,t)\in[n]\times[T]} \mathbb{E}v_{i,t}^2 \geq \rho$
(vii) both $\Sigma_F\Sigma_L$ and $\Sigma_X\Sigma_R$ converge to (possibly different) matrices with distinct eigenvalues.

Remark 2.3.2. Assumption 3(i) imposes exponential-type tails. This allow us to apply large deviation theory, which provides finite-sample exponential bounds for sums of random variables. These inequalities are useful in bounding the maximum of a large number of sums of random variables and play an essential role in obtaining the approximation in (2.2.7). Although Assumption 3(i) rules out fat-tailed distributions such as student t distributions, our procedure works well under these fat-tailed distributions in Monte Carlo simulations in Section 2.4.

Remark 2.3.3. In Assumption 3(ii)-(iii), we require the factors to be strong and rule out cross-sectional dependence in the idiosyncratic terms. Assumption 3(iv) and (v) allow for weak dependence in the time dimension. These assumptions still allow for heteroskedasticity in the data. Similar conditions are commonly imposed in the literature of large factor models; see Stock and Watson (2002a) and Bai and Ng (2002, 2006a).

Remark 2.3.4. In Assumption 3(vi), asymptotically degenerate idiosyncratic errors are ruled out. This is needed by the anti-concentration inequalities of Gaussian

vectors that we use in the proof. Similar to some results in Bai (2003), Assumption 3(vii) imposes distinct singular values of $\Sigma_F\Sigma_L$ and $\Sigma_X\Sigma_R$. Notice that $\hat{L}$ and $\hat{R}$ are estimating the first p_Y left singular vectors of LF' and the first p_W left singular vectors of RX', respectively. Distinct eigenvalues guarantee that these singular vectors be identified up to a sign change. In our experience, even if this condition is violated, we do not find any evidence of failure of our method. Notice that we do not require the distinct singular value condition for $\Sigma_Z\Sigma_\Lambda$. This is because the test statistic involves $M_{\hat{Z}}$ and M_Z is identified regardless of whether or not individual singular vectors of $\Lambda Z'$ are identified.

Theorem 2.3.1. *Let Assumption 3 hold. Then, under H_0 in (2.2.3),*

$$\limsup_{n\to\infty} \sup_{\alpha\in(0,1)} \left|\mathbb{P}\left(S_n > \mathcal{Q}\left(1-\alpha, S_n^{BS}\right)\right) - \alpha\right| = 0.$$

Theorem 2.3.1 establishes the validity of Algorithm 3. Theorems 2.3.1 and 2.3.2 below are proved in Appendix B.2. The proof of Theorem 2.3.1 is still non-trivial despite the remarkable results by Chernozhukov, Chetverikov, and Kato (2014) on the validity of high-dimensional multiplier bootstrap. Besides deriving the uniform approximation of $\sqrt{n}M_{\hat{Z}}[\hat{F},\hat{X}]$ by the expression in (2.2.7), we need to deal with two complications in order to establish bootstrap validity. First, although v is assumed to be independent of (L,F,R,X), v is not independent of $\hat{\Gamma}$ defined in (2.2.7) due to the estimation errors. Second, existing high-dimensional bootstrap schemes require direct observations of the variables in the summation, but these variables involve the unobservable v. We also derive the power properties of our procedures.

Theorem 2.3.2. *Let Assumption 3 hold. Then under H_1 in (2.2.4),*

$$\mathbb{P}\left[S_n > \mathcal{Q}\left(1-\eta, S_n^{BS}\right)\right] \to 1.$$

Remark 2.3.5. Algorithm 3 can also be used to test the number of factors. Suppose that we are interested in testing whether the number of factors is p_0. We split the data into two subgroups and use them as W and Y in Algorithm 3 with

$p_Y = p_W = p_C = p_0$. If the number of factors is correctly specified, then the two subgroup of variables would have p_0 common factors and, by Theorem 2.3.1, the probability of rejecting H_0 in (2.2.3) converges to the nominal size of the test. Here, the data splitting needs to be done in a way that each subgroup is driven by all the factors. A natural way of doing this is to randomly split the data. The arbitrariness in splitting the sample can be dealt with using techniques similar to those in Meinshausen, Meier, and Bühlmann (2012). Inference on the number of factors is addressed by Onatski (2009) and Kapetanios (2010). Onatski (2009) requires $n/T = o(1)$, while Kapetanios (2010)'s subsampling method requires an abstract condition on the limiting distribution on the eigenvalues of large random matrices. In contrast, our test handles the more realistic case of n and T having the same order of magnitude and only imposes weak conditions listed in Assumption 3.

Remark 2.3.6. One might attempt to recast the problem of testing H_0 as inference on low-dimensional parameters. These methodologies might involve nonstandard situations where theoretical properties are much harder to obtain. We illustrate the difficulty through one example in Appendix B.4, where the test statistic is based on an estimator of Σ_ζ. In that example, due to the bias and the singularity of the asymptotic variance of this estimator, one needs to use higher order Edgeworth expansions with weakly dependent data. Another "side effect" of the singularity in the asymptotic variance is that it is unclear how many terms in the Edgeworth expansion and/or Taylor's expansion we need to consider. For these reasons, our bootstrap procedure seems theoretically more elegant. Moreover, the limiting distribution of the test statistic in example in Appendix B.4 is likely to be nonstandard and also involve unknown quantities that requires further estimation.

2.4 Monte Carlo simulations

In this section, we demonstrate the finite-sample performance of our procedures. All the columns of X, the last $p_Y - p_C$ columns of F and rows of e and u are generated as independent AR(1) processes whose AR coefficient is generated from

the uniform distribution on $[-0.9, 0.9]$ and whose residuals are generated from a student t distribution with 6 degrees of freedom normalized to have variance equal to one. The first p_C columns of F are equal to $XQ_{(p_C)}$, where $Q_{(p_C)}$ is the first p_C columns of Q and Q is simulated independently from from the uniform distribution (Haar measure) on the set of $p_W \times p_W$ orthogonal matrices. Rows of L and R are generated as i.i.d $N(0, 4I_{p_Y})$ and $N(0, 4I_{p_W})$, respectively.

To remove the impact of estimation errors in p_Y and p_W, we assume that their values are known. In Table 2.2, we report the rejection frequencies of tests for H_0 with nominal size 5% under different data-generating processes (DGPs). The rejection frequencies are computed using 500 random samples and Algorithm 3 is implemented using 200 bootstrap samples.

As we can see from Table 2.2, our test has decent size control; we also have good power against overstatements of p_C but not understatements. This is consistent with our theory in Section 2.3.

Table 2.2: Rejection frequency of H_0

Panel A: size properties									
	$H_0: p_C = 0$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 0)$								
	$T = 100$			$T = 150$			$T = 200$		
$n_W \backslash n_Y$	50	100	150	50	100	150	50	100	150
50	0.016	0.042	0.070	0.008	0.018	0.044	0.002	0.022	0.028
100	0.046	0.032	0.070	0.022	0.024	0.040	0.022	0.008	0.032
150	0.074	0.064	0.056	0.044	0.046	0.040	0.032	0.030	0.038
	$H_0: p_C = 1$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 1)$								
50	0.032	0.020	0.022	0.036	0.036	0.030	0.030	0.030	0.022
100	0.028	0.030	0.038	0.024	0.058	0.050	0.040	0.038	0.026
150	0.030	0.026	0.036	0.024	0.032	0.064	0.026	0.026	0.048
	$H_0: p_C = 2$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 2)$								
50	0.032	0.042	0.010	0.018	0.002	0.010	0.008	0.012	0.008
100	0.020	0.026	0.038	0.022	0.026	0.040	0.030	0.018	0.052
150	0.024	0.042	0.044	0.014	0.022	0.032	0.028	0.032	0.026
	$H_0: p_C = 3$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 3)$								
50	0.018	0.014	0.008	0.016	0.008	0.008	0.006	0.012	0.008
100	0.024	0.030	0.028	0.018	0.030	0.014	0.004	0.020	0.024
150	0.020	0.022	0.036	0.014	0.024	0.034	0.014	0.022	0.020
Panel B: power properties									
	$H_0: p_C = 0$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 1)$								
	$T = 100$			$T = 150$			$T = 200$		
$n_W \backslash n_Y$	50	100	150	50	100	150	50	100	150
50	0.042	0.024	0.042	0.044	0.048	0.030	0.032	0.028	0.018
100	0.036	0.050	0.050	0.034	0.034	0.034	0.038	0.030	0.032
150	0.026	0.054	0.050	0.024	0.030	0.040	0.020	0.036	0.040
	$H_0: p_C = 2$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 1)$								
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$H_0: p_C = 1$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 2)$								
50	0.026	0.012	0.024	0.038	0.018	0.024	0.026	0.012	0.022
100	0.026	0.026	0.038	0.022	0.036	0.026	0.018	0.040	0.040
150	0.020	0.048	0.048	0.014	0.026	0.040	0.014	0.042	0.044
	$H_0: p_C = 3$ under the DGP $(p_Y, p_W, p_C) = (3, 3, 2)$								
50	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

2.5 Empirical applications

2.5.1 Common factors between the macroeconomy and financial markets

Jurado, Ludvigson, and Ng (2015) propose an approach of measuring uncertainty using factors extracted from a large number of macroeconomic and financial variables. In this section, we investigate the structure of these factors. In particular, we are interested in whether the macroeconomy and the financial markets are driven by exactly the same factors and how many factors they share in common.

We obtain the data used in Jurado, Ludvigson, and Ng (2015) from the AEA website[3]. Of the 279 variables used in that paper, let Y denote the group of the 132 macroeconomic variables and W the group of the 147 financial variables. Here, p_Y and p_W (the numbers of factors driving the macroeconomy and the financial markets, respectively) are unknown and different estimators can give quite different results. Due to this difficulty in estimating p_Y and p_W, we implement Algorithm 3 with various choice of p_Y and p_W and interpret our results as tests for the triple (p_Y, p_W, p_C); see Remark 2.2.5.

We report the p-values for testing $H_0 : \ p_C = k_0$ for different values of k_0. The results for $k_0 \in \{0, 1\}$ and $k_0 \in \{2, 3\}$ are reported in Tables 2.3 and 2.4, respectively. P-values above 0.05 are highlighted in bold red font. Since the p-values for $k_0 > 3$ are always smaller than 0.05 for any values of (p_Y, p_W), we do not list them here. With these results, we invert the test in Algorithm 3 and construct a 95% confidence set for (p_Y, p_W, p_C). Since the number of factors in these 279 variables is estimated, by Jurado, Ludvigson, and Ng (2015), to be 12 and should equal $p_Y + p_W - p_C$, we set the parameter space for (p_Y, p_W, p_C) to be

$$\Pi = \{(p_Y, p_W, p_C) \mid p_Y + p_W - p_C \leq 12, \ p_C \leq p_Y \text{ and } p_C \leq p_W\}. \tag{2.5.1}$$

The constraint of $p_C \leq p_Y$ is natural because the number of common factors between Y and W cannot exceed the number of factors in Y. A similar constraint

[3]https://www.aeaweb.org/articles?id=10.1257/aer.20131193

of $p_C \leq p_Y$ is imposed in the parameter space.

We test each element in Π and report, in Table 2.5, the ones that are not rejected at 5% significance level. Table 2.5 allows us to conduct inference for any given function of (p_Y, p_W, p_C). For example, 95% confidence sets for p_Y, p_W and p_C are $\{2, \cdots, 7\}$, $\{6, 7\}$ and $\{0, 1, 2, 3\}$, respectively. Hence, the macroeconomy and the financial markets share at most 3 common factors. The number of factors driving the financial markets is found to be either 6 or 7, which suggests more factors than what popular asset pricing models find, e.g., 3-factor model (Fama and French (1992)) or 5-factor model (Fama and French (2015)).[4] Onatski (2009) found that the number of factors in the macroeconomic dataset used in Stock and Watson (2002b) is no more than 2. Here, we conclude that there are 2 to 7 factors driving the macroeconomy.

We also conduct inference on the number of factors specific to each group. Notice that there are $p_Y - p_C$ factors peculiar to the macroeconomy and $p_W - p_C$ factors unique to the financial markets. From Table 2.5, 95% confidence sets for $p_Y - p_C$ and $p_W - p_C$ are $\{2, \cdots, 6\}$ and $\{4, 5, 6, 7\}$, respectively. Therefore, there are at least 2 factors unique to the macroeconomy and 4 factors specific to the financial markets. This also means that the factors in the macroeconomy and the financial markets do not span each other.

2.5.2 Structure of macroeconomic factors

We now focus on the macroeconomic factors and study which factors drive which macroeconomic variables. In Jurado, Ludvigson, and Ng (2015), intuitive "labels" are given to the principal components (PCs). For example, the first three PCs are interpreted as "stock market", "manufacturing production, employment, total production and employment, and capcacity utilization" and "bond market". The last two labels are related to the macroeconomic dataset, of which the 132

[4]Some of these 147 financial variables are not tradable assets so the number of factors in these financial variables might exceed the number of factors in assets' returns. Moreover, factors in returns of financial assets could include aspects that might not count as risks. For example, Goyal, Pérignon, and Villa (2008) even find that the name of the stock exchange (whether it is NYSE or Nasdaq) could drive stocks' returns as a "factor".

variables are categorized into 8 groups detailed in the online appendix[5].

We apply the proposed methodology to the macro variables alone and study the structure of these macroeconomic factors. We classify the 8 groups of macroeconomic variables into two classes. Class Y contains 71 variables, consisting of, in terms of the classification in Jurado, Ludvigson, and Ng (2015), Groups 1 (output and income), 2 (labor market) and 6 (bond and exchange rates). Class W contains the other 5 groups with 61 macro variables. The number of factors in Y and W are still denoted by p_Y and p_W, respectively, while p_C denotes the number of common factors. Recalling from the previous results that there are 2 to 7 factors driving the macroeconomy, we define the "parameter" space for (p_Y, p_W, p_C) as

$$\Pi_{Macro} = \{(p_Y, p_W, p_C) \mid p_Y + p_W - p_C \leq 7,\ p_C \leq p_Y \text{ and } p_C \leq p_W\}. \quad (2.5.2)$$

As before, we invert the proposed test by applying it to all the elements in Π_{macro}. The resulting 95%-confidence set for (p_Y, p_W, p_C) is

$$\{(5,1,0),(5,1,1),(5,2,0)\}.$$

This indicates a "concentrated" structure in the macro factors: Y, which comprises 3 groups (as categorized in Jurado, Ludvigson, and Ng (2015)) of variables, contain 5 factors while W, which comprises the other 5 groups of variables, contain at most two factors. This finding has at least two implications regarding how to label the factors. First, since there are 5 factors in 3 groups, we have do not have enough labels if we name these 5 factors after the groups. Second, if we name more than 2 factors after the 5 groups of variables in W, then some labels are "spurious" in that some named factors might not really exist. In light of these findings, the methods based on correlation or preditive power might produce misleading interpretations.

[5]https://www.aeaweb.org/aer/app/10503/20131193_app.pdf

2.6 Acknowledgements

Chapters 2, in full, is currently being prepared for submission for publication of the material. Zhu, Yinchu. The dissertation author was the primary investigator and author of this material.

Table 2.3: Testing that the macroeconomy and financial markets have k_0 common factors

Y contains the 132 macroeconomic variables and W contains the 147 financial variables used in Jurado, Ludvigson, and Ng (2015). We report the p-values of testing H_0 : Y and W have k_0 common factors for various values of (p_Y, p_W) using Algorithm 3. The p-values exceeding 0.05 are reported in bold red font.

Panel A: $k_0 = 0$

$p_Y \backslash p_W$	1	2	3	4	5	6	7	8	9	10	11	12
1	0.007	0.000	0.000	0.000	0.000	0.000	0.011	0.012	0.012	0.018	0.016	0.011
2	0.000	0.000	0.000	0.000	0.000	**0.096**	0.012	0.011	0.016	0.026	0.016	0.013
3	0.000	0.000	0.000	0.000	0.013	**0.094**	0.017	0.022	0.019	0.011	0.006	0.002
4	0.000	0.000	0.000	0.000	0.007	**0.111**	0.046	0.024	0.007	0.013	0.001	0.001
5	0.000	0.000	0.000	0.000	0.027	**0.092**	**0.052**	0.006	0.008	0.005	0.000	0.007
6	0.000	0.000	0.000	0.000	0.009	**0.104**	0.033	0.005	0.001	0.000	0.000	0.002
7	0.000	0.000	0.000	0.006	0.005	**0.071**	0.023	0.002	0.001	0.000	0.003	0.000
8	0.005	0.000	0.000	0.010	**0.054**	**0.080**	0.006	0.000	0.000	0.003	0.001	0.001
9	0.003	0.000	0.002	0.023	**0.052**	**0.058**	0.001	0.000	0.000	0.000	0.002	0.000
10	0.000	0.000	0.005	0.007	**0.067**	0.025	0.001	0.000	0.000	0.000	0.001	0.000
11	0.000	0.007	0.001	0.007	0.039	0.029	0.013	0.000	0.000	0.000	0.003	0.004
12	0.001	0.000	0.002	0.011	0.036	0.029	0.006	0.000	0.000	0.004	0.003	0.000

Panel B: $k_0 = 1$

1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.007	0.008	0.015	0.021
2	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.010	0.010	0.020	0.016	0.012
3	0.001	0.000	0.000	0.000	0.000	**0.077**	0.021	0.009	0.015	0.015	0.011	0.011
4	0.000	0.000	0.000	0.000	0.010	**0.074**	0.014	0.015	0.016	0.009	0.009	0.005
5	0.000	0.000	0.000	0.000	0.006	**0.107**	0.045	0.019	0.009	0.007	0.000	0.003
6	0.000	0.000	0.000	0.000	0.000	**0.091**	**0.056**	0.007	0.007	0.002	0.001	0.000
7	0.000	0.000	0.000	0.000	0.009	**0.045**	0.019	0.009	0.007	0.000	0.003	0.001
8	0.001	0.003	0.001	0.000	0.005	**0.060**	0.028	0.004	0.000	0.001	0.001	0.002
9	0.006	0.006	0.000	0.005	0.046	**0.067**	0.006	0.000	0.000	0.001	0.001	0.001
10	0.008	0.000	0.000	0.008	0.047	**0.056**	0.002	0.000	0.000	0.000	0.001	0.003
11	0.000	0.000	0.006	0.002	**0.070**	0.023	0.003	0.001	0.000	0.001	0.002	0.003
12	0.000	0.001	0.000	0.004	0.007	0.041	0.004	0.001	0.000	0.002	0.004	0.005

Table 2.4: Testing that the macroeconomy and financial markets have k_0 common factors

Y contains the 132 macroeconomic variables and W contains the 147 financial variables used in Jurado, Ludvigson, and Ng (2015). We report the p-values of testing H_0 : Y and W have k_0 common factors for various values of (p_Y, p_W) using Algorithm 3. The p-values exceeding 0.05 are reported in bold red font. We only report the results where $\min\{p_Y, p_W\} \geq k_0$ because the number of common factors cannot be larger than the number of total factors in each group.

Panel A: $k_0 = 2$

$p_Y \backslash p_W$	2	3	4	5	6	7	8	9	10	11	12
2	0.000	0.000	0.000	0.000	0.000	0.001	0.015	0.009	0.007	0.011	0.017
3	0.000	0.000	0.000	0.000	0.000	0.013	0.015	0.008	0.02	0.025	0.008
4	0.000	0.000	0.000	0.000	**0.098**	0.013	0.004	0.012	0.016	0.011	0.01
5	0.000	0.000	0.000	0.000	**0.089**	0.018	0.009	0.019	0.008	0.011	0.006
6	0.000	0.000	0.000	0.000	**0.055**	**0.053**	0.012	0.007	0.006	0.005	0.001
7	0.000	0.000	0.000	0.000	0.043	**0.051**	0.008	0.004	0.003	0.000	0.001
8	0.004	0.000	0.000	0.000	**0.065**	0.029	0.006	0.002	0.003	0.001	0.002
9	0.005	0.003	0.000	0.005	**0.074**	0.027	0.001	0.001	0.003	0.001	0.001
10	0.005	0.000	0.000	0.006	**0.060**	0.01	0.000	0.001	0.000	0.001	0.000
11	0.000	0.000	0.005	0.003	**0.076**	0.002	0.001	0.001	0.000	0.000	0.002
12	0.000	0.001	0.000	0.001	0.013	0.003	0.000	0.001	0.000	0.002	0.001

Panel B: $k_0 = 3$

3	0.000	0.000	0.000	0.000	0.000	0.005	0.013	0.008	0.021	0.011
4	0.000	0.000	0.000	0.000	0.032	0.008	0.008	0.012	0.004	0.014
5	0.000	0.000	0.000	0.029	0.012	0.01	0.015	0.011	0.004	0.008
6	0.000	0.000	0.000	0.003	0.011	0.013	0.017	0.01	0.007	0.006
7	0.000	0.000	0.000	0.000	**0.054**	0.017	0.005	0.007	0.004	0.001
8	0.006	0.004	0.001	0.000	0.049	0.009	0.011	0.003	0.002	0.002
9	0.003	0.005	0.000	0.048	0.023	0.009	0.003	0.002	0.001	0.002
10	0.008	0.000	0.000	0.032	0.030	0.004	0.002	0.000	0.000	0.000
11	0.000	0.000	0.005	0.01	0.008	0.002	0.000	0.001	0.000	0.001
12	0.001	0.007	0.000	0.01	0.006	0.000	0.000	0.000	0.001	0.001

Table 2.5: 95% confidence set for (p_Y, p_W, p_C) in the combined dataset (both macro and financial variables)

Each element (triple) in the confidence set is represented by one row in the above matrix. Only elements in Π defined in (2.5.1) are considered.

p_Y	p_W	p_C
2	6	0
3	6	0
4	6	0
5	6	0
5	7	0
6	6	0
3	6	1
4	6	1
5	6	1
6	6	1
6	7	1
7	6	1
4	6	2
5	6	2
6	6	2
6	7	2
7	7	2
7	7	3

Chapter 3

Linear Hypothesis Testing in Dense High-Dimensional Linear Models

3.1 Introduction

A high-dimensional inference is a fundamental topic of interest in modern scientific problems that are widely recognized to be of high-dimensional nature, i.e., that require estimation of parameters with dimensionality exceeding the number of observations. Applications span a wide variety of scientific fields, such as biology, medicine, genetics, neuroscience, economics, and finance. Minimizing a suitably regularized (quasi-)likelihood function was developed (Tibshirani 1996; Fan and Li 2001) as a suitable approach for the estimation in such models. In particular, high-dimensional linear models have been studied extensively in recent years and take the following form

$$y_i = x_i^\top \beta_* + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{3.1.1}$$

for a response $y_i \in \mathbb{R}$, a feature vector $x_i \in \mathbb{R}^p$ and the noise $\varepsilon_i \in \mathbb{R}$, such that $E[\varepsilon_i] = 0$ and $E[\varepsilon_i^2] = \sigma_\varepsilon^2$ with $0 < \sigma_\varepsilon^2 < \infty$. The vector $\beta_* \in \mathbb{R}^p$ is the unknown model parameter and we allow for $p \gg n$. We consider a random design setting with the feature vectors satisfying $Ex_i = 0$ and $E[x_i x_i^\top] = \Sigma_X$. Under certain regularity conditions on the design matrix $X = (x_1, x_2, \ldots, x_n)^\top$, regularized methods with a

suitable choice of the tuning parameter have been shown to achieve the optimal rate of estimation as long as the vector β_* is sparse in that $\|\beta_*\|_0 = o(n/\log p)$.

The goal of the present article is to address the testing problem for linear hypotheses of the form

$$H_0 : \ a^\top \beta_* = g_0, \tag{3.1.2}$$

where the loading vector $a \in \mathbb{R}^p$ is pre-specified and $g_0 \in \mathbb{R}$ is given, and design an asymptotically valid test statistic that does not rely on sparsity assumptions. Some central challenges have hindered the systematic development of tools for statistical inference in such settings. The non-sparse nature of the model parameter β_* poses serious challenges to consistent estimation; moreover, the size and structure of the loading vector a introduce additional difficulty for the inference. However, in this article we consider potentially dense vectors β_* with $0 \leq \|\beta_*\|_0 \leq p$. We also allow for the non-sparse loadings with $1 \leq \|a\|_0 \leq p$. The inference problem for the mean of the response y_i conditional on $x_i = a$, is a prototypical case for the general functional $a^\top \beta_*$ and is a representative case for dense loading a.

We develop the principles of *restructured regression,* where a hypothesis-driven *feature synthetization* is introduced. The feature augmentation is done in such a way to separate useful inferential information from the useless one, by "projecting" the original feature space to the space spanned by the vector a and the space orthogonal to a. This orthogonal projection is introduced to achieve the above separation and avoid the curse of dimensionality. Then, an appropriate moment condition is invoked on the restricted regression and a suitable test statistic constructed. The structure of the moment condition and its test depend on whether or not the covariance of the features Σ_X is known. When prior knowledge of Σ_X is available, the synthesized features can be created in such a way that the resulting moment condition and testing procedure do not depend on β_*; thus, estimation of β_* is completely avoided. As a result, no assumption on the sparsity of β_* is required. We establish theoretical guarantees for Type I error control and show that the test can detect the deviation from the null hypothesis of the order $O(\|a\|_2/\sqrt{n})$. To the best of our knowledge, our approach provides the first result on testing general linear hypothesis (3.1.2) in high-dimensional linear models with potentially

non-sparse (dense) parameters.

When prior knowledge of Σ_X is unavailable, the orthogonalization and perfect separation is not achievable due to the unknown projection matrix. We design an estimator of the projection matrix and further condition the new and augmented features in such a way that their correlations are estimable and yet the format of the restructured regression remains unchanged. The developed hypothesis-driven feature separation diminishes the impact of the inaccuracy of an estimator of a transformation of β_*. Consequently, we can establish asymptotically exact control of Type I error. We believe there is currently no result on testing $a^\top \beta_*$ in the case where Σ_X is unknown, and both β_* and a are allowed to be dense. Moreover, when sparsity assumptions hold, our procedure is shown to achieve optimality guarantees; hence, it does not loose efficiency.

Since we do not assume sparsity in β_*, our work does not directly compare to the existing results, which are only valid for sparse β_*. However, in some cases, our work generalizes existing results to the non-sparse models. For example, Cai and Guo (2015) show that when Σ_X is known, the minimax length of the confidence interval for $a^\top \beta_*$ is of the order $O(\|a\|_2/\sqrt{n})$ if $\|\beta_*\|_0 = O(n/\log p)$. As confidence sets for $a^\top \beta_*$ can be easily constructed by inverting the proposed tests, our results indicate that their conclusion continues to hold for non-sparse models, where $\|\beta_*\|_0$ can be as large as p. For the case of dense a, we do not impose any constraint on a. However, existing work, such as Cai and Guo (2015), imposes a lower bound (in terms of $\|a\|_\infty$) on the minimal non-zero coordinate of a – a condition that is seldom satisfied for inference of conditional mean, when a is typically drawn from a continuous distribution (e.g. a is drawn from the same distribution as the distribution of the x_i's).

3.1.1 Relation to existing literature

Confidence intervals and hypothesis testing play a fundamental role in statistical theory and applications. However, compared to the point estimation there is still much work to be done for statistical inference of high-dimensional models. Existing work on the inference problems predominantly focuses on individual

coordinates of β_*. Early work typically imposes conditions that guarantee consistent variable selection (see Fan and Li (2001), Zou (2006), and Zhao and Yu (2006)) or develops methods that lead to conservative inferential guarantees (e.g. Bühlmann (2013)). However, recent work focusses on asymptotically accurate inference without relying on the variable selection consistency. Current advances in this domain are, however, restricted to the ultra-sparse case, where $\|\beta_*\|_0 = o(\sqrt{n}/\log p)$; see Zhang and Zhang (2014), Belloni, Chernozhukov, and Hansen (2014), Geer, Bühlmann, Ritov, and Dezeure (2014), Javanmard and Montanari (2014a), Ning and Liu (2014), Javanmard and Montanari (2015), Mitra and Zhang (2014), Bühlmann and Geer (2015), Belloni, Chernozhukov, and Kato (2015), and Chernozhukov, Hansen, and Spindler (2015). Under such sparsity condition, the expected length of the confidence intervals for individual coordinates is of the order $O(1/\sqrt{n})$ (van de Geer and Jankova 2016). Cai and Guo (2015) study the length of the confidence intervals allowing for $\|\beta_*\|_0 = o(n/\log p)$ and discover that lack of explicit knowledge of $\|\beta_*\|_0$ can fundamentally limit the efficiency of confidence intervals.

However, there is little reason to believe that the sparsity of β_* needs to hold in practice (Hall, Jin, and Miller 2014; Ward 2009; Jin and Ke 2014; Pritchard 2001). Unfortunately, there is almost no work on estimating or testing the true sparsity level of the underlying parameter. Hence, the theory of hypothesis testing under general sparsity structures is still a very challenging and important open problem. In particular, progress is very much required when $\|\beta_*\|_0$ is allowed to grow faster than $n/\log p$ and perhaps even larger than the sample size n. There are several articles showing that the regularized procedures have non-vanishing estimation errors in such settings (Donoho and Johnstone 1994; Raskutti, Wainwright, and Yu 2011; Cai and Guo 2016). However, is it still possible to develop a general methodology for testing β_* in this case? Can one construct valid inference procedures that do not require knowledge of $\|\beta_*\|_0$?

In the proposed inference procedure, we handle the high-dimensional, possibly non-sparse model parameters and/or non-sparse loadings, by developing a new methodology for testing. The proposed methodology is centered around a construction of augmented and synthesized features that are driven by a specific form of

the null hypothesis. Compared with the previous approaches of de-biasing (Zhang and Zhang 2014; Javanmard and Montanari 2014a; Geer, Bühlmann, Ritov, and Dezeure 2014; Mitra and Zhang 2014), scoring (Ning and Liu 2014; Chernozhukov, Hansen, and Spindler 2015), double-selection (Belloni, Chernozhukov, and Hansen 2014; Belloni, Chernozhukov, and Kato 2015), our new approach has two major distinctive features:

- We do not rely on a $l_1$norm consistent estimation of the unknown model parameters. In high-dimensional models with the lack of sparsity in the parameters, this may no longer be possible. Instead, we propose to reformulate the original parametric null hypothesis into a moment condition that can be successfully estimated even without sparsity in the model. This moment condition is different from the score equations employed for estimation as those are not estimable in non-sparse high-dimensional models.

- We advocate for a study and exploration of the correlation between feature vectors (and not the model parameters); this proves to be a valuable tool that overcomes the limit of estimation. Namely, we propose that the features be split and projected onto the loading vector a of the hypothesis (3.1.2), thereby fully utilizing the null hypothesis structure. This "decoupling" scheme allows for a successful estimation of the moment condition even without sparsity assumption. As a result the developed method provides a rich alternative to the classical Wald or Score principles.

3.1.2 Notation and organization of the article

We briefly describe notations used in the article. We use $\to^d$ to denote convergence in distribution and $\mathcal{N}(0,1)$ to denote the standard normal distribution with its cumulative distribution function denoted by $\Phi(\cdot)$. The (multivariate) normal distribution with mean (vector) μ and variance (matrix) Σ is denoted by $\mathcal{N}(\mu,\Sigma)$. We use $^\top$ to denote the transpose of (a vector or matrix) and denote by I_p the $p \times p$ identity matrix. For a vector $a = (a_1, \cdots, a_p)^\top \in \mathbb{R}^p$, its l_0 norm is the cardinality of $\text{supp}(a) = \{i \mid a_i \neq 0\}$ and $\|a\|_\infty = \max\{|a_1|, \cdots, |a_p|\}$; $\|a\|_1$

and $\|a\|_2$ denote the l_1 and l_2 norm of a, respectively. In this case, a_{-i} denotes the vector a with its ith coordinate removed. For two sequences of positive constants a_n and b_n, we use $a_n \asymp b_n$ to denote that $a_n/b_n = O(1)$ and $b_n/a_n = O(1)$. For two real numbers a_1 and a_2, $a_1 \vee a_2$ and $a_1 \wedge a_2$ denote $\max\{a_1, a_2\}$ and $\min\{a_1, a_2\}$, respectively.

The rest of this article is organized as follows. Section 3.2 introduces the main methodology under known Σ_X and establishes theoretical properties of the proposed test. Section 3.3 extends the proposed methodology to the case of the unknown Σ_X and provides theoretical results. Section 3.4 contains examples illustrating new methods that the proposed methodology brings to the literature on high-dimensional inference. Section 3.5 contains detailed numerical experiments on a number of dense high-dimensional linear models, including sparse and dense loadings a. In Section 3.5.1, we demonstrate the excellent finite-sample performance of the proposed methods through Monte Carlo simulations; in Section 3.5.2, we illustrate our method via a real data study. Appendix C contains complete details of the theoretical derivations.

3.2 Testing $H_0 : a^\top \beta_* = g_0$ with prior knowledge of Σ_X

In this section we promote a unified approach to a wide class of decision problems. Our main building block (which we believe is important in its own right) is a construction, named *restructured regression* allowing, under weak assumptions, to build tests for hypotheses on $a^\top \beta_*$, where β_* and/or a can be non-sparse. Considering the potential failure of sparsity in many practical problems, we strongly believe that our approach permits a diverse spectrum of applications. In this section our focus is to introduce the method with known Σ_X (an assumption relaxed in the next section).

Throughout the paper, we denote $\Omega_X = \Sigma_X^{-1}$. In the sequel, given the feature

vector $x_i \in \mathbb{R}^p$ and loading vector $a \in \mathbb{R}^p$, we consider the following decomposition

$$x_i = az_i + w_i, \tag{3.2.1}$$

with a scalar

$$z_i = \left(\frac{\Omega_X a}{a^\top \Omega_X a} \right)^\top x_i$$

and a p-dimensional vector

$$w_i = \left[I_p - \frac{aa^\top \Omega_X}{a^\top \Omega_X a} \right] x_i.$$

Observe that az_i can be viewed as the projection of x_i onto the vector a – taking into account Ω_X, hence extracting information in x_i regarding the null hypothesis. Notice that the model (3.1.1) and decomposition (3.2.1) imply

$$y_i = z_i \cdot (a^\top \beta_*) + w_i^\top \beta_* + \varepsilon_i, \tag{3.2.2}$$

referred to as *restructured regression*. The proposed construction gives rise to the method of *feature customization*. Given covariate vector x_i and the loading vector a representing the structure of the null hypothesis, we create the synthesized features $\tilde{x}_i := (z_i, w_i^\top)^\top$ so that the regression coefficient for z_i in the restructured regression (3.2.2) is the quantity under testing.

Remark 3.2.1. The synthesized features are not only an artifact of our new methodology but also admit intuitive interpretations. Consider the case where Σ_X is known to be I_p. The synthesized features z_i and w_i represent the relevant and the irrelevant information with respect to the null, respectively. To see this, suppose that the true distribution of the data is known. With the population expectations, we can identify the parameters in the restructured regression (3.2.2): $E(z_i y_i) = Ez_i^2 (a^\top \beta_*)$ and $Ew_i y_i = Ew_i w_i^\top \beta_*$. Notice that the latter equation contains no information regarding $a^\top \beta_*$ because it can be shown that a is orthogonal to columns in $Ew_i w_i^\top$. In other words, knowing $Ew_i w_i^\top \beta_*$ does not lead to knowing $a^\top \beta_*$. Therefore, $a^\top \beta_*$ is identified with the distribution of (y_i, z_i) and w_i does not contain information about the null hypothesis.

It is not hard to verify that, by the construction of the transformed features, $E[w_i z_i] = 0$. It follows that $E\left[z_i(y_i - z_i g_0)\right] = E\left[z_i\left(\varepsilon_i + w_i^\top \beta_* + z_i(a^\top \beta_* - g_0)\right)\right] = E\left[z_i^2(a^\top \beta_* - g_0)\right]$. Observe that the last expression is 0 if and only if the null hypothesis (3.1.2) holds. As a result, testing H_0 in (3.1.2) is equivalent to testing the following moment condition:

$$H_0 : E\left[z_1(y_1 - z_1 g_0)\right] = 0. \tag{3.2.3}$$

To test the above condition, we propose a studentized test statistic, $T_n(g_0)$, taking the form

$$T_n(g_0) := \frac{n^{-1/2}\sum_{i=1}^n l_i(g_0)}{\sqrt{n^{-1}\sum_{i=1}^n l_i(g_0)^2}}, \tag{3.2.4}$$

with $l_i(g_0) = z_i(y_i - z_i g_0)$. For a test of H_0 with nominal size $\alpha \in (0,1)$, we reject H_0 if

$$|T_n(g_0)| > \Phi(1 - \alpha/2).$$

The methodology proposed above is novel in a number of aspects. Unlike Wald or Score or Likelihood principles, centered around a consistent estimator of β_*, our methodology allows for extremely fast implementation and does not estimate the unknown parameter β_*. The novel methodology consists of two-stages. At the first stage, our procedure establishes a data-driven feature decomposition based on the structure of the null hypothesis directly. At the second stage, only "a moment condition" of the restructured regression is tested. It is critical to observe that restructured regression by itself is not sufficient to guarantee valid inference. The novel properties of the proposed method are based on the built-in, i.e., designed orthogonality of the synthesized features z_i and w_i. As such it enables us to construct a test statistic that does not contain the unknown parameter β_*, thereby allowing our methodology to handle dense (and thus possibly non-estimable) β_*. Moreover, no assumption is imposed on the loadings a either. As we will see in the next section, these properties under known Σ_X propagate to the case of the unknown Σ_X and underline all further developments.

Assumption 4. *Let the following hold: (i) there exists a positive constant C such*

that $E|z_i\sigma_z^{-1}|^8 \leq C$, $E\varepsilon_i^8 \leq C$ *and* $E|w_i^\top \beta_*|^8 < C$ *with* $C < \infty$. *Moreover, (ii) there exists a constant* $c \in (0, \infty)$, *such that* $\sigma_\varepsilon \geq c$. *Lastly, (iii) there exist constants* $D_1, D_2 > 0$ *such that the eigenvalues of* Σ_X *lie in* $[D_1, D_2]$.

The stated conditions in Assumption 4 are very weak and intuitive. Assumption 4(i) requires components in the restructured regression (3.2.2) to have bounded eighth moments. Assumption 4(ii) rules out the noiseless regression setting in the original model (3.1.1). Assumption 4(iii) is very weak in that it only imposes well-designed covariance matrix of the features x_i (see Bickel, Ritov, and Tsybakov (2009)).

Notice that Assumption 4 does not require any condition regarding the sparsity of β_*. Even in the case of sparse a, existing work, such as the debiasing method, heavily relies on the sparsity of β_*. Results regarding dense a are very limited even for sparse β_*. Cai and Guo (2015) impose the condition of $\max_{j\in\text{supp}(a)} |a_j| / \min_{j\in\text{supp}(a)} |a_j| = O(1)$; however, such a condition is quite hard to satisfy if a is drawn from a continuous distribution whose support contains zero. In contrast, our results do not require any condition on a and, hence, bridge the gap in the existing literature on high-dimensional inference.

Theorem 3.2.1. *Consider the model in (3.1.1) and the definition of* z_i *and* w_i *as in (3.2.1). Suppose that Assumption 4 holds. Under* H_0 *in (3.1.2), we have that (1) the test statistic* T_n, (3.2.4), *satisfies* $T_n(g_0) \to^d \mathcal{N}(0,1)$ *as* $n, p \to \infty$ *and that (2)*

$$\lim_{n,p\to\infty} P\Big(|T_n(g_0)| > \Phi^{-1}(1-\alpha/2)\Big) = \alpha.$$

Theorem 3.2.1 gives an asymptotic approximation for the null distribution of the test statistic $T_n(g_0)$ under general sparsity structure. The result of Theorem 3.2.1 has two striking features. The first is that it holds, no matter the size or sparsity of the loading vector a. The second is that the proposed test guarantees Type I error control when $p \geq n$ and $p, n \to \infty$ no matter of the sparsity of β_* and without the knowledge of the noise level σ_ε; in particular, it allows $\|\beta_*\|_0 = p$. Therefore, our test is fully adaptive, in the sense that its validity does not depend on in the sparse/dense level of either the model parameter β_* or the hypothesis

loading a. We also show that our test can detect deviations from the null that are larger than $O(\|a\|_2/\sqrt{n})$ while allowing β_* to be non-sparse and $p \geq n$.

Theorem 3.2.2. *Under the conditions of Theorem 3.2.1, suppose that* $a^\top \beta_* = g_0 + h_n$ *and* $\sqrt{n}|h_n|/\|a\|_2 \to \infty$. *Then, for any* $\alpha \in (0,1)$.

$$\lim_{n,p\to\infty} P\Big(|T_n(g_0)| > \Phi^{-1}(1-\alpha/2)\Big) = 1.$$

Remark 3.2.2. Theorem 3.2.2 also suggests that we can expect the length of the confidence interval for $a^\top \beta_*$ (obtained by inverting the proposed test) to be of the order of $O(\|a\|_2/\sqrt{n})$ regardless of the sparsity of β_* or a. To the best of our knowledge, it is the first result to explicitly allow non-sparse and simultaneously high-dimensional parameters β^* or vector loadings a. It is also closely connected with the existing results for the case of sparse parameters β^*. Cai and Guo (2015), state that under Gaussianity and sparsity in both β_* and a together with known Σ_X and σ_ε, the optimal expected length of confidence intervals for $a^\top \beta_*$ is of the order $O(\|a\|_2/\sqrt{n})$ (see Theorem 7 therein). Observe that our procedure achieves the same optimality without the knowledge of σ_ε and allowing dense vectors β_*.

We do not formally claim that this is the optimal rate for dense β_*, but we can consider an obvious benchmark. Let $\bar{\beta}$ be an estimator that attains an efficiency similar to (ordinary least square) OLS in low dimensions, i.e., $\bar{\beta}$ is distributed as $\mathcal{N}(\beta_*, \Omega_X \sigma_\varepsilon^2/n)$. Then $a^\top \bar{\beta}$ follows $\mathcal{N}(a^\top \beta_*, a^\top \Omega_X a \sigma_\varepsilon^2/n)$ distribution. Since Ω_X has eigenvalues bounded away from infinity, the standard deviation of $a^\top \bar{\beta}$ is of the order $\|a\|_2/\sqrt{n}$. Such an estimator might not be feasible in practice, but could serve as a benchmark for dense β_*. A rigorous study of the efficiency issue is likely to yield results that are quite different from current literature since existing results, e.g., Cai and Guo (2015), do not naturally extend to dense problems. For example, consider the case of $\|a\|_0 = \|\beta_*\|_0 = p$, naively extending Theorem 8 of Cai and Guo (2015) would conclude that the minimax expected length of a confidence interval for $a^\top \beta_*$ is of the order $\|a\|_\infty p\sqrt{(\log p)/n}$; however, this rate is larger than the rate $\|a\|_2/\sqrt{n}$, which is bounded above by $\|a\|_\infty \sqrt{p/n}$. Lastly, according to Theorem 3.2.2 our proposed test achieves the same rate at the benchmark $\bar{\beta}$.

3.3 Testing $H_0 : a^\top \beta_* = g_0$ without prior knowledge of Σ_X

The approach proposed in this section tackles the high-dimensional inference problem in a very general setting. The focus is the more realistic scenario in which the covariance matrix Σ_X and the variance of the model (3.1.1) are both unknown. We synthesize new features, create a new reference model and explore the correlations therein in order to design a suitable inferential procedure that is stable without sparsity assumption.

3.3.1 Feature synthetization and restructured regression

In order to design inference when Σ_X unknown, we take on a new perspective and build upon the methodology of Section 3.2. Consider feature synthetization of Section 3.2 where Σ_X is naively treated as I_p,

$$z_i = \Big(\frac{a}{a^\top a}\Big)^\top x_i \in \mathbb{R} \quad \text{and} \quad w_i = \big(I_p - aa^\top/(a^\top a)\big)\, x_i \in \mathbb{R}^p. \tag{3.3.1}$$

Although the decomposition $x_i = az_i + w_i$ still holds, features z_i and w_i might be correlated (because $\Sigma_X \neq I_p$). If such correlation is estimated successfully, we can use certain decoupling method to eliminate the impact of dense parameters while allowing exponentially growing dimensions.

The first challenge is that directly estimating the correlation between z_i and w_i (as defined) is not achievable (as the restricted eigenvalue (RE) condition Bickel, Ritov, and Tsybakov (2009) on $W = (w_1, \cdots, w_n)^\top$ is violated). To address this problem, we propose to *stabilize* the feature vector w_i and define *stabilized* features $\tilde{w}_i$. We stabilize the features in such a way that the RE condition on the stabilized design $\tilde{W} = (\tilde{w}_1, \cdots, \tilde{w}_n)^\top$ is satisfied with high probability. Since $I_p - aa^\top/(a^\top a)$ is a projection matrix, we can find $U_a \in \mathbb{R}^{p\times(p-1)}$ an orthogonal matrix such that

$$U_a^\top U_a = I_{p-1} \qquad \text{and} \qquad I_p - aa^\top/(a^\top a) = U_a U_a^\top.$$

Then

$$W\beta_* = X(I_p - aa^\top/(a^\top a))\beta_* = XU_aU_a^\top\beta_* = \tilde{W}\pi_*,$$

where

$$\tilde{W} = WU_a \qquad \text{and} \qquad \pi_* = U_a^\top\beta_*.$$

Since $y_i = z_i \cdot (a^\top\beta_*) + w_i^\top\beta_* + \varepsilon_i$, we have the *stabilized model*

$$y_i = z_i \cdot (a^\top\beta_*) + \tilde{w}_i^\top\pi_* + \varepsilon_i. \tag{3.3.2}$$

The model is balanced in the sense that $E\tilde{W}^\top\tilde{W}/n = U_a^\top\Sigma_X U_a \in \mathbb{R}^{(p-1)\times(p-1)}$ with eigenvalues bounded away from zero and infinity. Therefore, RE condition on $\tilde{W}$ holds under weak conditions; see Rudelson and Zhou (2013).

Remark 3.3.1. The synthesized feature $w_i \in \mathbb{R}^p$ is consolidated into $\tilde{w}_i \in \mathbb{R}^{p-1}$, in that $\tilde{w}_i$ has a smaller dimensionality and can be used to recover w_i via $w_i = U_a\tilde{w}_i$. In this sense, $\tilde{w}_i$ contains all the information in w_i. As an example, consider the case with a being the first column of I_p. In this case, it is not hard to verify that $z_i = x_{i,1}$, $w_i = (0, x_{i,2}, \cdots, x_{i,p})^\top \in \mathbb{R}^p$, $U_a = (0, I_{p-1})^\top \in \mathbb{R}^{p\times(p-1)}$ and thus $\tilde{w}_i = U_a^\top w_i = (x_{i,2}, \cdots, x_{i,p})^\top \in \mathbb{R}^{p-1}$.

We now introduce an additional model to account for the dependence between the *synthesized feature* z_i and the *stabilized feature* $\tilde{w}_i$:

$$z_i = \tilde{w}_i^\top\gamma_* + u_i, \tag{3.3.3}$$

where $\gamma_* \in \mathbb{R}^{p-1}$ is an unknown parameter and u_i is independent of $\tilde{w}_i$ with $Eu_i = 0$ and $Eu_i^2 = \sigma_u^2$.

In this article, we will assume that γ_* is sparse, in order to decouple the dependence between z_i and $\tilde{w}_i$ with the unknown Σ_X. In fact, sparse γ_* is a generalization of the sparsity condition on the precision matrix Ω_X, a regularity condition typically imposed in the literature; see Geer, Bühlmann, Ritov, and Dezeure (2014), Belloni, Chernozhukov, and Hansen (2014) and Belloni, Chernozhukov, and Kato (2015) and Ning and Liu (2014). Recall the example in Remark 3.3.1. Since $x_{i,1} = z_i = \tilde{w}_i^\top\gamma_* + u_i = x_{i,-1}^\top\gamma_* + u_i$, it is not hard to show that the first row of

Ω_X is $(\sigma_u^{-2}, -\sigma_u^{-2}\gamma_*^\top)$. Hence, the sparsity of γ_* is equivalent to the sparsity in the first row of Ω_X. The sparsity of γ_* can be justified for dense a as well. Consider the case of $\Sigma_X = cI_p$ for some $c > 0$; a prototypical model in compressive sensing corresponds to $c = 1$ (Nickl and Geer 2013). In this case, one can easily show that z_i and $\tilde{w}_i$ are uncorrelated, meaning that $\gamma_* = 0$ for any a. The synthesized features also admit intuitive interpretations in this case: the feature z_i contains useful information in testing the null hypothesis $a^\top\beta_* = g_0$, while the consolidated $\tilde{w}_i$ contain information not useful for inference.

Now, we are ready to construct the moment condition of interest. Observe that under H_0 in (3.1.2), $y_i - z_i g_0 - \tilde{w}_i^\top \pi_* = \varepsilon_i$ is uncorrelated with $z_i - \tilde{w}_i^\top\gamma_* = u_i$. If H_0 is false, then $y_i - z_i g_0 - \tilde{w}_i^\top \pi_* = \varepsilon_i + z_i(\theta_* - g_0) = \varepsilon_i + \tilde{w}_i^\top\gamma_*(\theta_* - g_0) + u_i(\theta_* - g_0)$ has non-zero correlation with $u_i = z_i - \tilde{w}_i^\top\gamma_*$. Hence, the initial null hypothesis, (3.1.2) is equivalent to the following null hypothesis

$$H_0 : E\Big[\big(z_1 - \tilde{w}_1^\top\gamma_*\big)\big(y_1 - z_1 g_0 - \tilde{w}_1^\top\pi_*\big)\Big] = 0. \tag{3.3.4}$$

Directly testing this moment condition is not feasible, due to the unknown values of parameters γ_* and π_*. As a result, we first provide estimates for these unknown parameters and consider the test statistic given by the studentized statistics.

We make a few remarks about the above proposed methodology. As mentioned above, the existing literature on high-dimensional inference adopts the approach of relying on an (almost) unbiased estimate of the model parameter to distinguish the null and alternative hypotheses. The existing methods largely differ by the means of constructing the unbiased estimate and/or its asymptotic variance. Many use an approximation of a one-step Newton method (Zhang and Zhang 2014; Geer, Bühlmann, Ritov, and Dezeure 2014; Javanmard and Montanari 2014a) to achieve consistency in estimation of possibly all p parameters. In order to test $a^\top\beta_*$ in this framework, one need to show that the debiased estimator for β_* can be used to construct an asymptotically unbiased and normal estimator for $a^\top\beta_*$; to the best of our knowledge, a formal theoretical justification is yet to be established even under sparse β_*. Other than the debiasing technique, some proposals center around Neyman's score orthogonalization ideas (Belloni, Chernozhukov, and Hansen 2014;

Belloni, Chernozhukov, and Kato 2015; Chernozhukov, Hansen, and Spindler 2015; Ning and Liu 2014). It is worth pointing out that such a method requires a clear separation of parameter under testing and the nuisance parameter. In the original problem, the model parameter is β_* and the quantity under testing is $a^\top \beta_*$; hence, it is not clear how to define the nuisance parameter since the $a^\top \beta_*$ is not just one entry (or a subset) of the parameter vector β_*. Lastly, the work of Cai and Guo (2015) propose a minimax optimal test that allows for dense loadings vector a, however in the dense case it provides a conservative error bounds and requires the knowledge of the sparsity size s.

However, our proposal deviates from the above methodologies in a few aspects. Firstly, we design a test statistic irrespective of a consistency of high-dimensional estimators for the model parameter; hence, any refitting or one-step approximations are unnecessary. Secondly, we aim to orthogonalize design features (rather than model parameters) by directly taking into account the structure of the null hypothesis (represented by a and g_0). In this way we achieve full adaptivity to the hypothesis testing problem of interest. Thirdly, we reformulate the original parametric hypothesis into a moment condition of which we provide adaptive estimators. The moment condition itself is not a simple first-order optimality identification (related to Z-estimators), but rather a moment that utilizes the special feature orthogonalization and fusion. Hence, even in setting where the existing work applies, our proposed method provides an alternative. However, apart from existing work, our proposed method applies much more broadly.

3.3.2 Adaptive estimation of the unknown quantities

In this subsection, we start with a brief introduction of the Dantzig selector, which is the basis of our estimators. Then we introduce the intuition and steps of our estimator as well as implementation details.

Dantzig selector review

Numerous studies have been conducted in regards to the consistent estimation of high-dimensional parameters in linear models. The canonical examples of

successful estimators represent Lasso and Dantzig selector, defined as $\hat{\beta}_l$ and $\hat{\beta}_d$ below,

$$\hat{\beta}_l = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda_l \|\beta\|_1 \right\}, \qquad \begin{aligned} \hat{\beta}_d &= \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \\ s.t &\left\| n^{-1} X^\top (Y - X\beta) \right\|_\infty \leq \lambda_d. \end{aligned} \tag{3.3.5}$$

Although Lasso and Dantzig selector are defined in different times, Bickel, Ritov, and Tsybakov (2009) established equivalence between the two estimators under the conditions of moderate design correlations and model sparsity, $\|\beta_*\|_0 \ll n$. Between these two estimator, the Dantzig selector, $\hat{\beta}_d$, offers easy implementation through linear programming techniques. Moreover, the constraint in the Dantzig selector can be interpreted as a relaxation of the least squares normal equations, $X^\top Y = X^\top X\beta$. However, the performance of both estimators is tightly connected to the choice of their respective tuning parameters λ_l and λ_d, i.e. the size of such relaxation. Several empirical and theoretical studies emphasized that tuning parameters should be chosen proportionally to the noise standard deviation σ_ε, i.e. $\lambda_d = \lambda_d(\sigma_\varepsilon) = \sigma_\varepsilon\sqrt{(\log p)/n}$. In such settings one can guarantee $\|\hat{\beta}_l - \beta_*\|_1 = O(\|\beta_*\|_0\sqrt{(\log p)/n})$. Unfortunately, in most applications, the variance of the noise is unavailable. It is therefore vital to design statistical procedures that estimate unknown parameters together with the size of model variance in a joint fashion. This topic received special attention, cf. Giraud, Huet, and Verzelen (2012) and the references therein. Most popular σ-adaptive procedures, the square-root Lasso (Belloni, Chernozhukov, and Wang 2011), the scaled Lasso (Sun and Zhang 2012) and the self-tuned Dantzig selector (Gautier and Tsybakov 2013; Belloni, Chernozhukov, and Hansen 2016) can be seen as maximum a posteriori estimators with a particular choice of prior distribution. However they do not provide estimates that are reasonable in non-sparse and high-dimensional models – after all in such settings it is impossible to consistently estimate the model parameters (see for more details Cai and Guo (2016) and Raskutti, Wainwright, and Yu (2011)). The aim of the present section is to present an alternative to these methods, which are closely related, but presents some advantages in terms of implementation and a more transparent theoretical

analysis in not necessarily sparse models; the main benefit is that our estimates are well controlled in certain sense.

Modified Dantzig selector: adaptive to signal-to-noise ratio

We start with the estimator for π_*, a parameter that is high-dimensional and yet not necessarily sparse. We extend the Dantzig selector above to conform to the testing problem that we have to perform. We begin by splitting the tuning parameter into a constant independent of the variance of the noise and introduce a parameter ρ, a square root of the noise to response ratio as an unknown in the optimization problem. At the population level, ρ is intended to represent $\sigma_\varepsilon/\sqrt{E(y_1 - z_1 g_0)^2}$ and ρ_0 is a lower bound for this ratio. One might attempt to use scaled Lasso by Sun and Zhang (2012) or self-tuning dantzig selector proposed by Gautier and Tsybakov (2013), but for non-sparse π_*, these methods cannot ensure that the estimated noise variance is bounded away from zero whenever the vector π_* is a dense vector (a case of special interest here).

For $Z = (z_1, \cdots, z_n)^\top$ and $Y = (y_1, \cdots, y_n)^\top$ defined in (3.3.1), we introduce the following version of Dantzig selector of π_*

$$\begin{array}{rcrcl} (\hat{\pi}, \hat{\rho}) & = & \underset{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}}{\arg\min} \|\pi\|_1 & & \\ & s.t & \left\|\tilde{W}^\top (Y - Zg_0 - \tilde{W}\pi)\right\|_\infty & \leq & \eta\ \rho\ \sqrt{n}\|Y - Zg_0\|_2 \\ & & (Y - Zg_0)^\top \left(Y - Zg_0 - \tilde{W}\pi\right) & \geq & \rho_0\ \rho\ \|Y - Zg_0\|_2^2/2 \\ & & \rho & \in & [\rho_0, 1], \end{array} \tag{3.3.6}$$

where $\eta \asymp \sqrt{n^{-1}\log p}$ and $\rho_0 \in (0,1)$ are scale-free tuning parameters.

The estimator (3.3.6) is different from (3.3.5) in two ways. First, the estimator (3.3.6) simultaneously estimates π_* and ρ. We introduce a ρ_0 the lower bound for ρ as a tuning parameter. Second, the estimator (3.3.6) has an additional constraint, which essentially serves as an upper bound for ρ. The intuition of this bound is the following. When π is replaced by the true π_* and the null hypothesis holds, this constraint (scaled by $1/n$) becomes $\pi_*^\top \tilde{W}^\top \varepsilon/n + \varepsilon^\top \varepsilon/n \geq \rho_0 \rho \|\tilde{W}\pi_* + \varepsilon\|_2^2/n$. By the law of large numbers, this means that $o_P(1) + \sigma_\varepsilon^2 \geq \rho_0 \rho E(y_1 - z_1 g_0)^2$,

which is satisfied if $\rho = \sigma_\varepsilon / \sqrt{E(y_1 - z_1 g_0)^2}$ and $\rho > \rho_0$.

The vector $\varepsilon = Y - Zg_0 - \tilde{W}\pi_*$ is a residual vector of the stabilized model (3.3.2) under the null hypothesis H_0. The first constraint on the residual vector imposes that for each i, much like the Dantzig selector, $\hat{\beta}_l$, maximal correlation $\|\tilde{W}^\top \varepsilon / n\|_\infty$ is not larger than the noise level $\eta\sigma_\varepsilon$. Yet, in contrast to $\hat{\beta}_l$, our estimator treats ρ as an unknown quantity and estimates it simultaneously with π_*. Moreover, we introduce the second constraint to stabilize estimation of the moment of interest (3.3.4) in the presence of non-sparse vectors π_*. Under the null hypothesis, this constraint prevents choice of ρ that is too large; namely, it constraints $\rho \leq C \left(Y - Zg_0\right)^\top \varepsilon / \left\|Y - Zg_0\right\|_2^2$ for a finite constant $C > 0$. In sparse settings, this additional constraint is redundant, so we remove it from our estimator of γ_* defined below (a vector that is assumed to be sparse). Hence, we consider the following estimator,$\hat{\gamma}$

$$\begin{array}{rcl} \hat{\gamma} & = & \underset{\gamma \in \mathbb{R}^{p-1}}{\arg\min} \ \|\gamma\|_1 \\ s.t & \left\| n^{-1} \tilde{W}^\top (Z - \tilde{W}\gamma) \right\|_\infty & \leq \ \lambda n^{-1/2} \|Z\|_2 \end{array} \tag{3.3.7}$$

where $\lambda \asymp \sqrt{n^{-1} \log p}$ is a scale-free tuning parameter and $n^{-1/2}\|Z\|_2$ serves as an upper bound of the unknown σ_u in the model (3.3.3). It is worth pointing out that the defined estimators change with a change in the hypothesis testing problem (3.1.2) through the new, synthesized and stabilized feature vectors $\tilde{W}$ and Z together with g_0. We present a few examples in Section 4.

Implementation

The optimization problem in (3.3.6), a generalization of the Dantzig selector (Candes and Tao 2007), can be recast as a linear program; the computational burden of our method is comparable to the Dantzig selector. Define scalars $d_1 = \rho_0 \|Y - Zg_0\|_2^2/2$, $d_2 = \|Y - Zg_0\|_2^2$, vectors $D_1 = \tilde{W}^\top (Y - Zg_0) \in \mathbb{R}^{p-1}$ and $D_2 = \sqrt{n}\eta \|Y - Zg_0\|_2 \mathbf{1}_{p-1}$ and matrix $D_3 = \tilde{W}^\top \tilde{W} \in \mathbb{R}^{(p-1)\times(p-1)}$.

Then, (3.3.6) is equivalent to the following linear program

$$\begin{array}{lllll} \min_{(c,\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}^{p-1}\times\mathbb{R}} & \mathbf{1}_{p-1}^{\top}c & & & \\ s.t. & -c \leq & \pi & \leq c & \\ & \rho_0 \leq & \rho & \leq 1 & \\ & & d_1\rho + D_1^{\top}\pi & \leq d_2 & \\ & -D_2\rho \leq & D_1 - D_3\pi & \leq D_2\rho, & \end{array} \tag{3.3.8}$$

where the optimization variables are $c \in \mathbb{R}^{p-1}$, $\pi \in \mathbb{R}^{p-1}$ and $\rho \in \mathbb{R}$. For application purposes we propose to choose the following choices of the tuning parameters: $\rho_0 = 0.01$ and $\eta = \sqrt{2\log(p)/n}$. They are universal choices and we show in simulations that they provide good results.

3.3.3 Test Statistic

With defined estimators of γ_* and π_*, we are ready to define a sample analog of the moment condition 3.3.4. Under our proposed method, a test of nominal size $\alpha \in (0,1)$ rejects H_0 in (3.1.2) if $|S_n| > \Phi^{-1}(1-\alpha/2)$, where

$$S_n = \sqrt{n}\frac{(Z - \tilde{W}\hat{\gamma})^{\top}(Y - Zg_0 - \tilde{W}\hat{\pi})}{\|Z - \tilde{W}\hat{\gamma}\|_2\|Y - Zg_0 - \tilde{W}\hat{\pi}\|_2}. \tag{3.3.9}$$

Other estimators of the first moment (3.3.4) are certainly possible, however we focus and analyze the natural case above; we leave future efficiency studies for future work since it is not apparent that any other choice is preferred. Moreover, the self-normalizing statistic above is directly dependent on the hypothesis of interest and is a function of synthesized features. Compared with the existing approaches where the normalization factor is a consistent estimator of the asymptotic variance, our self-normalized approach adopts an inconsistent estimator as the normalization factor, which in a sense corresponds to "inefficient Studentizing" (cf. Shao (2010)). However, we establish that the asymptotic distribution of the resulting statistic is pivotal and its percentiles can be obtained from the normal distribution.

In constructing estimates of γ_* and π_*, we do not impose any assumption regarding the sparsity of π_* or β_*. Notice that, except for the case of sparse a, it is

in general unreasonable to expect sparsity in π_*, even if β_* is sparse. Although we use estimates for both γ_* and π_* denoted by $\hat{\gamma}$ and $\hat{\pi}$, respectively, we only require l_1 consistency properties for $\hat{\gamma}$; in fact, $\hat{\pi}$ only serves to satisfy our decoupling argument in the proof and does not need to be consistent. We now briefly explain this point. The constraints imposed in the estimator (3.3.6) guarantee that for the test statistic S_n, the term $n^{-1/2}(Z - \tilde{W}\hat{\gamma})^\top(Y - Zg_0 - \tilde{W}\hat{\pi})$ can be approximated by a product of two independent terms, i.e. $n^{-1/2}(Z - \tilde{W}\gamma_*)^\top(Y - Zg_0 - \tilde{W}\hat{\pi})$. Then, the only requirement needed is to guarantee that the second term in the last expression does not grow to fast (it does not need to converge to zero) which in turn is provided by the constraints of the optimization problem (3.3.6).

3.3.4 Theoretical properties

In deriving the theoretical properties of our test, we impose the following assumption.

Assumption 5. *Let (i) x_i and ε_i have Gaussian distributions, $\mathcal{N}(0, \Sigma_X)$ and $\mathcal{N}(0, \sigma_\varepsilon^2)$, respectively. Moreover, assume (ii) that there exist constants $c_1, c_2 > 0$, such that σ_ε and the eigenvalues of Σ_X lie in $[c_1, c_2]$. Lastly, let (iii) there exist constants $c_3, c_4 \in (0, 1)$, such that $\sigma_u^2/\sigma_z^2 \geq c_3$ and $\sigma_\varepsilon^2/\sigma_y^2 \geq c_4$.*

Assumption 5(i) is only imposed to simplify the proof. In high-dimensional literature Gaussian design is a very common assumption (e.g. Javanmard and Montanari (2014b) and Cai and Guo (2015)). The same results, at the expense of more complicated proofs, can be derived for sub-Gaussian designs and errors. Assumption 5(ii) is very standard in high-dimensional literature (see Bickel, Ritov, and Tsybakov 2009; Ning and Liu 2014; Geer, Bühlmann, Ritov, and Dezeure 2014 for more details).

Assumption 5(iii) imposes nondegeneracy of signal-to-noise ratios for models (3.1.1) and (3.3.3). Since $\|a\|_2$ is allowed to tend to infinity, $\sigma_z^2 = a^\top \Sigma_X a/(a^\top a)^2$ can tend to zero and thus it is too restrictive to assume that σ_u is bounded away from zero. Hence, Assumption 5(iii) is a relaxation, as it only rules out the uninteresting case of asymptotic noiselessness.

Remark 3.3.2. The sparsity condition is imposed on neither a nor β_*. Theorem 3.3.1 below says that we can conduct valid inference of a non-sparse linear combination of a non-sparse high-dimensional parameter without knowing Σ_X. To the best of our knowledge, this is the first result that allows for such generality.

Theorem 3.3.1. *Let Assumption 5 hold. Consider estimators* (3.3.6) *and* (3.4.2) *with suitable choice of tuning parameters:* $\eta, \lambda \asymp \sqrt{n^{-1}\log p}$, $\rho_0^{-1} = O(1)$ *and* $\rho_0 \leq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{-1/2}$. *Suppose that* $\|\gamma_*\|_0 = o(\sqrt{n}/\log p)$. *Then, under* H_0 *in (3.1.2), optimization problems* (3.3.6) *and* (3.4.2) *are feasible with probability approaching one and*

$$\lim_{n,p\to\infty} P\left(|S_n| > \Phi^{-1}(1-\alpha/2)\right) = \alpha \qquad \forall \alpha \in (0,1),$$

where S_n *is defined in Equation* (3.3.9).

Theorem 3.3.1 establishes that the proposed test is asymptotically exact regardless of how sparse the model parameter or the loading vector are. In that sense, the result is unique in the existing literature as it covers cases of β sparse and a sparse (SS), β sparse and a dense (SD) , β dense and a sparse (DS) and especially β dense and a dense (DD). The (SS) case appears in a number of existing works (see Belloni, Chernozhukov, and Hansen (2014), Geer, Bühlmann, Ritov, and Dezeure (2014), Javanmard and Montanari (2014b), and Ning and Liu (2014)), case (SD) appears in Cai and Guo (2015). Whenever (SS) case holds, our result above matches the above mentioned work see Theorem 3.3.2. In the special setting of (SD) our result generalizes the one of Cai and Guo (2015) as Theorem 3.3.1 does not impose any restriction on the size of the loading vector a. The last two cases of (DS) and (DD) present an extremely challenging cases in which inference based on estimation (much like Wald or Rao or Likelihood principles) fails due to the inherit limit of detection – work of Cai and Guo (2016) provides details of impossibility of estimation in such settings. However, despite these challenges our method is able to provide asymptotically valid inference as we have developed inference based on a specifically designed moment condition (and not a parameter estimation alone).

The result in Theorem 3.3.1 is based on the assumption that $\hat{\pi}_*$ is a possibly

inconsistent estimator of the parameter vector π_*, i.e. the full model is dense with all non-zero entries. In the following, we will show that if the model is a sparse model, the proposed test (3.3.9) maintains strong power properties. To facilitate the mathematical derivations, we consider the local alternatives of the form

$$H_{1,n}: \ a^\top \beta_* = g_0 + n^{-1/2}(a^\top \Omega_X a)^{1/2}\sigma_\varepsilon d, \tag{3.3.10}$$

where $d \in \mathbb{R}$ is a fixed constant. The following result shows that the proposed test achieves certain optimality in detecting alternatives $H_{1,n}$.

Theorem 3.3.2. *Consider z_i and w_i defined in (3.3.1). Let Assumption 5 hold and consider the choice of tuning parameters, as in Theorem 3.3.1. Suppose that $\|\gamma_*\|_0 \vee \|\beta_*\|_0 \vee \|a\|_0 = o(\sqrt{n}/\log p)$. Then, under $H_{1,n}$ in (3.3.10), optimization problems* (3.3.6) *and* (3.4.2) *are feasible with probability approaching one and*

$$\lim_{n,p\to\infty} P\left(|S_n| > \Phi^{-1}(1-\alpha/2)\right) = \Psi_\alpha(d) \qquad \forall \alpha \in (0,1),$$

where $\Psi_\alpha(d) := \Phi\left(-\Phi^{-1}(1-\alpha/2) + d\right) + \Phi\left(-\Phi^{-1}(1-\alpha/2) - d\right)$.

To better understand the optimality of the result above, consider the estimator (possibly infeasible) discussed at the end of Section 3.2: let $\bar{\beta}$ denote an estimator satisfying $\sqrt{n}(\bar{\beta} - \beta_*) \sim \mathcal{N}(0, \Omega_X \sigma_\varepsilon^2)$. Notice that, for the low-dimensional components of β_*, $\bar{\beta}$ achieves semi-parametric efficiency; see Robinson (1988). Therefore, for sparse a, $a^\top \bar{\beta}$ is a semi-parametrically efficient estimator for $a^\top \beta_*$. Notice that $\sqrt{n}(a^\top \bar{\beta} - a^\top \beta_*) \sim \mathcal{N}(0, a^\top \Omega_X a \sigma_\varepsilon^2)$. Based on such efficient estimator, one might consider an "oracle" test: for a test of nominal size α, reject the null $H_0: \ a^\top \beta_* = g_0$ if and only if

$$\frac{\sqrt{n}|a^\top \bar{\beta} - g_0|}{(a^\top \Omega_X a)^{1/2}\sigma_\varepsilon} > \Phi^{-1}(1-\alpha/2).$$

It is easy to verify that the power of this "oracle" test of nominal size α against the local alternatives $H_{1,n}$ (3.3.10) is asymptotically equal to $\Psi_\alpha(d)$. Therefore, Theorem 3.3.2 says that our test asymptotically achieves the same power as the "oracle" test under sparse a and β_*, i.e. it is as efficient as the "oracle" test.

Moreover, in light of recent inferential results in the high-dimensional sparse

models, the rate of Theorem 4 can also be shown to be optimal. As existing results apply only to the case of $a = e_j$ for a coordinate vector e_j, $1 \leq j \leq p$, we discuss the relations of our work in this specific settings. We note that the tests based on VBRD and BCH are asymptotically equivalent to this "oracle" test and hence have the same asymptotic local power; the power of Wald or Score inferential methods (see Theorem 2.2 in Geer, Bühlmann, Ritov, and Dezeure (2014), Theorem 1 in Belloni, Chernozhukov, and Hansen (2014) or Theorem 4.7 in Ning and Liu (2014)) and that of Javanmard and Montanari (2014b) (see Theorem 2.3 therein) is asymptotically equal to and converges to $\Psi_\alpha(d)$, respectively. This in turn, implies that the proposed method is semi- parametrically efficient and asymptotically minimax. For vectors a that have more than one non-zero coordinate, we can only compare our work with that of Cai and Guo (2015), where we observe that the result of Theorems 1 and 3 therein matches those of Theorem 4 covering the case of extremely sparse beta and potentially dense vectors a. However, observe that the confidence intervals developed therein require specific knowledge of the sparsity of the parameter β_*, $\|\beta_*\|_0$, a quantity rarely known in practice. Unlike their method, our method can be directly implemented without the knowledge of the sparsity of β_* and yet achieves the same optimality guarantees.

3.4 Applications to non-sparse high-dimensional models

This section is devoted to three concrete applications of the general methodological results developed in Sections 3.2 and 3.3 – hence, showcasing the wide impact of the developed theories.

3.4.1 Testing pairwise homogeneity

The previous section deals with situations in which each coordinate of the parameters is allowed to vary independently and any subset of the coordinates can be non-zero simultaneously. This condition will not be satisfied if we are interested

in testing pairwise homogeneity in the linear model (group effect), that is, if we are interested in testing the hypothesis

$$H_0 : \beta_{*,k} = \beta_{*,j}$$

for $k, j \in \{1, 2, \ldots, p\}$ while also allowing β to be a dense and high-dimensional vector. To the best of our knowledge, such tests were not designed in the existing literature. The proposed methodology easily extends to this case, where the loading vector a takes the form $a = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^\top$, with the location of the 1's at the j-th and k-th coordinate, respectively. Without loss of generality, we assume that $k = 1$ and $j = 2$. Then it is not hard to show that $z_i = (x_{i,1} - x_{i,2})/2$ and $\tilde{w}_i = ((x_{i,1} + x_{i,2})/\sqrt{2}, x_{i,3}, \cdots, x_{i,p})^\top \in \mathbb{R}^{p-1}$. The proposed methodology for this problem simplifies, then, to finding $\hat{\pi}$ and $\hat{\rho}$ that satisfy

$$\begin{array}{rcl} (\hat{\pi}, \hat{\rho}) & = & \underset{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}_+}{\arg\min} \|\pi\|_1 \\ & s.t & \tilde{W} = [(X_1 + X_2)/\sqrt{2}, X_3, \cdots, X_p] \\ & & \|\tilde{W}^\top (Y - \tilde{W}\pi)\|_\infty \leq \eta\rho\sqrt{n}\|Y\|_2 \\ & & Y^\top \left(Y - \tilde{W}\pi\right) \geq \rho_0 \ \rho \ \|Y\|_2^2/2 \\ & & \rho \in [\rho_0, 1] \end{array} \tag{3.4.1}$$

and $\hat{\gamma}$ that satisfies

$$\begin{array}{rcl} \hat{\gamma} & = & \underset{\gamma\in\mathbb{R}^{p-1}}{\arg\min} \|\gamma\|_1 \\ & s.t & \tilde{W} = [(X_1 + X_2)/\sqrt{2}, X_3, \cdots, X_p] \\ & & \|\tilde{W}^\top (X_1 - X_2 - 2\tilde{W}\gamma)\|_\infty \leq \lambda\sqrt{n}\|X_1 - X_2\|_2 \quad , \end{array} \tag{3.4.2}$$

for $\lambda, \eta \asymp \sqrt{n^{-1}\log p}$.

Consequently, we reject $H_0 : \beta_{*,1} = \beta_{*,2}$ if $|S_n| > \Phi^{-1}(1 - \alpha/2)$, where

$$S_n = \sqrt{n}\frac{\left(X_1 - X_2 - 2\tilde{W}\hat{\gamma}\right)^\top \left(Y - \tilde{W}\hat{\pi}\right)}{\left\|X_1 - X_2 - 2\tilde{W}\hat{\gamma}\right\|_2 \left\|Y - \tilde{W}\hat{\pi}\right\|_2}. \tag{3.4.3}$$

3.4.2 Inference of conditional mean

Our methodology can also be used for the inference regarding the average value of the response i.e. regarding the conditional mean of the regression model. Suppose that the object of interest is $E(y_i \mid \zeta_i)$, where $y_i \in \mathbb{R}$ and $\zeta_i \in \mathbb{R}^k$. For a given value $d \in \mathbb{R}^k$ and $g_0 \in \mathbb{R}$, the focus is to test

$$H_0 : \ E(y_i \mid \zeta_i = d) = g_0.$$

Assuming that for some given dictionary of transformations of $\{\phi_j(\cdot)\}_{j=1}^p$, the conditional mean function admits the representation: $E(y_i \mid \zeta_i) = \sum_{j=1}^p \beta_{*,j}\phi_j(\zeta_i)$ for some vector $\beta_* = (\beta_{*,1}, \cdots, \beta_{*,k})^\top \in \mathbb{R}^p$. Then the conditional mean model can be written as

$$y_i = x_i^\top \beta_* + \varepsilon_i, \tag{3.4.4}$$

where $x_i = (\phi_1(\zeta_i), \cdots, \phi_p(\zeta_i))^\top \in \mathbb{R}^p$ and $E(\varepsilon_i \mid x_i) = 0$. In turn, the confidence intervals for the regression mean can be designed simply by inverting the test statistics

$$S_n = \sqrt{n}\frac{(Z - \tilde{W}\hat{\gamma})^\top (Y - Zg_0 - \tilde{W}\hat{\pi})}{\|Z - \tilde{W}\hat{\gamma}\|_2 \|Y - Zg_0 - \tilde{W}\hat{\pi}\|_2}$$

designed for the inference problem

$$H_0 : \ a^\top \beta_* = g_0,$$

where $a = (\phi_1(d), \cdots, \phi_p(d))^\top \in \mathbb{R}^p$ and $U_a U_a^\top = \left(I_p - aa^\top / \sum_{j=1}^p \phi_j^2(d)\right)$ with

$$z_i = \frac{\sum_{j=1}^p \phi_j(d)\phi_j(\zeta_i)}{\sum_{j=1}^p \phi_j^2(d)}, \qquad \text{and} \qquad \tilde{w}_{ij} = \sum_{l=1}^p \{U_a\}_{lj}\phi_l(\zeta_i), \qquad 1 \le j \le p-1.$$

Notice that we do not assume that the vector β_* is sparse and we allow for $p \gg n$. Therefore, representing the conditional mean function in terms of a large number of transformations of ζ_i, while simultaneously allowing all to be non-zero, does not lose much in generality. Additionally, it is worth mentioning that inference for such models has not been addressed in the existing literature: most of the existing work

is strictly focused around sparse or sparse additive models. With the general model considered here, one can consider tests regarding treatment effects (when viewed as the conditional mean) and allow for fully dense models and loading vectors, i.e. the treatment being a dense combination of many variables. Existing work, such as Belloni, Chernozhukov, and Hansen (2014), only allows the treatment to be a single variable.

3.4.3 Decomposition of conditional mean

In practice, the researcher might be interested in how much a certain group of features contribute to the conditional mean. Let $\mathcal{G} \subseteq \{1, ..., p\}$. The goal is to conduct inference on linear functionals of $\{\beta_{*,j}\}_{j\in\mathcal{G}}$, i.e., $\sum_{j\in\mathcal{G}} c_j\beta_{*,j}$ for some known $\{c_j\}_{j\in\mathcal{G}}$.

For example, consider the notations from Section 3.4.2. Let $\zeta_i = (\zeta_{i,1}, ..., \zeta_{i,k})^\top$ and suppose that one is interested in the impact of $\zeta_{i,1}$ on the conditional mean for $\zeta = d$. This is equivalent to quantifying $\sum_{j\in\mathcal{G}_1} \phi_j(d)\beta_{*,j}$, where the set contains all the indexes j such that the first entry of ζ_i has non-zero effect on $\phi_j(\zeta_i)$, i.e., $\mathcal{G}_1 = \{j : \phi_j(\zeta) \text{ is not constant in } \zeta_1\}$. If $\phi_j(\cdot)$'s are transformations of individual entries of $\{\zeta_{i,j}\}_{j=1}^k$, then $\mathcal{G}_1$ corresponds to transformations of $\zeta_{i,1}$. For another example, suppose that all the p features are genes. The domain scientist (biologist, doctor, geneticist, etc) might be interested in how much a group of genes contributes to the expected value of the response variable.

Without loss of generality, we assume that $\mathcal{G} = \{1, ..., H\}$ and $c = (c_1, ..., c_H)^\top \in \mathbb{R}^H$. Let $U_c \in \mathbb{R}^{H\times(H-1)}$ satisfy $I_H - cc^\top/(c^\top c) = U_cU_c^\top$ and $U_c^\top U_c = I_{H-1}$. Then the synthesized features can be constructed by $z_i = \|c\|_2^{-2}\sum_{j=1}^H c_jx_{i,j}$ and $\tilde{w}_i = \left(\sum_{l=1}^H (U_c)_{l,1}x_{i,l}, \cdots, \sum_{l=1}^H (U_c)_{l,H-1}x_{i,l}, x_{i,H}, \cdots, x_{i,p}\right)^\top \in \mathbb{R}^{p-1}$, where $(U_c)_{l,j}$ denotes the (l, j) entry of the matrix U_c. For example, whenever $H = 3$ and $c_j = 1$ for all $j = 1, 2, 3$, then

$$U_c = \begin{pmatrix} -\sqrt{3/2} & -1/\sqrt{2} \\ 0 & \sqrt{2} \\ \sqrt{3/2} & -1/\sqrt{2} \end{pmatrix}$$

and the procedure for testing $\beta_{*,1}+\beta_{*,2}+\beta_{*,3}=g_0$ would be as follows. We define

$$\begin{aligned}(\hat{\pi},\hat{\rho}) \quad &= \quad \underset{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}_+}{\arg\min} \|\pi\|_1 \\ &s.t \quad \tilde{W}=\left[\sqrt{\tfrac{3}{2}}(X_3-X_1),\ -\tfrac{1}{\sqrt{2}}(X_1-2X_2+X_3),\ X_4,\cdots,X_p\right] \\ &\quad \|\tilde{W}^\top[Y-(X_1+X_2+X_3)g_0/3-\tilde{W}\pi]\|_\infty \\ &\quad\quad \leq \eta\rho\sqrt{n}\|Y-(X_1+X_2+X_3)g_0/3\|_2 \\ &\quad (Y-(X_1+X_2+X_3)g_0/3)^\top\left(Y-(X_1+X_2+X_3)g_0/3-\tilde{W}\pi\right) \\ &\quad\quad \geq \rho_0\ \rho\ \|Y-(X_1+X_2+X_3)g_0/3\|_2^2/2 \\ &\quad \rho\in[\rho_0,1]\end{aligned} \tag{3.4.5}$$

and $\hat{\gamma}$ that satisfies

$$\begin{aligned}\hat{\gamma} \quad &= \quad \underset{\gamma\in\mathbb{R}^{p-1}}{\arg\min}\ \|\gamma\|_1 \\ &s.t \quad \tilde{W}=\left[\sqrt{\tfrac{3}{2}}(X_3-X_1),\ -\tfrac{1}{\sqrt{2}}(X_1-2X_2+X_3),\ X_4,\cdots,X_p\right] \\ &\quad \|\tilde{W}^\top\left((X_1+X_2+X_3)g_0-3\tilde{W}\gamma\right)\|_\infty \leq \lambda\sqrt{n}g_0\|X_1+X_2+X_3\|_2 \quad ,\end{aligned} \tag{3.4.6}$$

for $\lambda,\eta\asymp\sqrt{n^{-1}\log p}$.

For a test of nominal size α, we reject $H_0:\beta_{*,1}+\beta_{*,2}+\beta_{*,3}=g_0$ if $|S_n|>\Phi^{-1}(1-\alpha/2)$, where

$$S_n=\sqrt{n}\frac{\left((X_1+X_2+X_3)g_0-3\tilde{W}\hat{\gamma}\right)^\top\left(Y-(X_1+X_2+X_3)g_0/3-\tilde{W}\hat{\pi}\right)}{\left\|(X_1+X_2+X_3)g_0-3\tilde{W}\hat{\gamma}\right\|_2\left\|Y-(X_1+X_2+X_3)g_0/3-\tilde{W}\hat{\pi}\right\|_2}. \tag{3.4.7}$$

3.5 Numerical results

In this section we study the finite sample performance of the proposed methodology for both known Σ_X and unknown Σ_X. We explicitly consider dense loadings a and dense parameter vectors β_* as well as more common sparse settings.

3.5.1 Monte Carlo experiments

Consider the model (3.1.1) with the model error following standard normal distribution. In all the simulations, we set $n = 100$ and $p = 500$ and the nominal size of all the tests is 5%. The rejection probabilities are based on 500 repetitions. The null hypothesis we test is $H_0 : \ a^\top \beta_* = g_0$, where $g_0 = a^\top \beta_* + h$ and h is allowed to vary in order to capture both Type I and Type II error rates.

Setup

We consider in total four regimes on the structure of the model and the null hypothesis – sparse and dense regimes for β_* as well as sparse and dense regimes for the loading vector a.

(i) In the *Sparse parameter regime* we consider the parameter structure with $\beta_* = (0.8, 0.8, 0, ..., 0)^\top$.

(ii) In the *Dense parameter regime* we consider the parameter structure with $\beta_* = \frac{3}{\sqrt{p}}(1, 1, ..., 1)^\top$.

(iii) In the *Sparse loading regime* we consider the loading vector $a = (0, 1, 0, ..., 0)^\top$.

(iv) In the *Dense loading regime* we consider the loading vector $a = (1, 1, ..., 1)^\top$.

Observe that (iii) is an extreme sparse-loading case. We consider this special case in order to compare existing inferential methods, like VBRD and BCH. However, our method can be implement for various number of non-zero elements, whereas the existing one cannot.

We present results for three different designs settings including sparse, dense, Gaussian and non-Gaussian settings.

Example 1. Here we consider the standard Toeplitz design where the rows of X are drawn as an i.i.d random draws from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma_X)$, with covariance matrix $(\Sigma_X)_{i,j} = 0.4^{|i-j|}$.

Example 2. In this case we consider a non-sparse design matrix with equal correlations among the features. Namely, rows of X are i.i.d draws from the

multivariate Gaussian distribution $\mathcal{N}(0, \Sigma_X)$, where $(\Sigma_X)_{i,j}$ is 1 for $i = j$ and is 0.4 for $i \neq j$. Observe that this case is particularly hard for most inferential methods as all features are interdependent and Ω_X is not sparse.

Example 3. In this example we consider a highly non-Gaussian design that also has strong dependence structure. We consider the setting of Fan and Song (2010). We repeat the details here for the convenience of the reader. Let x be a typical row of X. For $j \in \{1, ..., 15\}$, $x_j = (\xi + c\xi_j)/\sqrt{1 + c^2}$, where ξ and $\{\xi_j\}_{j=1}^{15}$ are i.i.d $\mathcal{N}(0, 1)$ and c is chosen such that $\text{corr}(x_1, x_2) = 0.4$. For $j \in \{16, ..., [p/3]\}$, x_j is i.i.d $\mathcal{N}(0, 1)$. For $j \in \{[p/3] + 1, ..., [2p/3]\}$, x_j is i.i.d from a double exponential distributions with location parameter zero and scale parameter one. For $j \in \{[2p/3] + 1, ..., p\}$, x_j is i.i.d from the half-half mixture of $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 0.5)$. Observe that in this case 2/3 of the features follow non-Gaussian distributions. Thus, in this case it is extremely difficult to even obtain consistent estimation of the model parameters.

Implementation details

We compare the proposed tests with VBRD and BCH; methods proposed in Cai and Guo (2015) contain constants whose values could be very conservative in finite samples. Our tests with known and unknown Σ_X are implemented as discussed in Sections 3.2 and 3.3, respectively.

The VBRD method is implemented for both dense and sparse loadings as follows. We first compute the debiased estimator $\hat{\beta}_{\text{debias}}$ and the nodewise Lasso estimator $\hat{\Omega}_{\text{Lasso}}$ for the precision matrix Σ_X as in VBRD. Then test is to reject H_0 if and only if

$$\sqrt{n}|a^\top \hat{\beta}_{\text{debias}} - g_0|/\sqrt{a^\top \hat{\Omega}_{\text{Lasso}} \hat{\Sigma}_X \hat{\Omega}_{\text{Lasso}}^\top a \sigma_\varepsilon^2} > \Phi^{-1}(1 - 0.05/2).$$

The BCH method is only implemented for the sparse loadings. We compute the generic post-double-selection estimator for the second entry of β as in Equation (2.8) of BCH and compute the standard error as in Theorem 2 therein. Then a usual t-test is conducted. It is not clear how BCH can be extended to handle any

loading vector a different from an extremely sparse case (see (iii) above): first, for any other loading structure it is not defined how to gather selected features of what would be multiple simultaneous equations; second, naively extending the original BCH to the problem of dense a ($\|a\|_0 = p$) means running an OLS regression of the response against all the features, which is not feasible for $p > n$.

Results

We start with the size properties of competing tests. For this purpose, we examine the distributions of the test statistics under the null hypothesis by comparing empirical distributions of the tests with the theoretical benchmark of standard normal random variable. For simplicity of presentation, we only consider the Toeplitz design. For the testing problem with sparse β_* and sparse a, our tests, VBRD and BCH exihibit the validity guaranteed by the theory; in Figure 3.1, the histograms of the test statistics are close to $\mathcal{N}(0,1)$ with large p-values of the Kolmogorov-Smirnov (KS) tests. For all the other problems, our tests outperform existing methods. As shown in Figure 3.2, the histogram of VBRD test visually is still close to the standard normal distribution but the KS test suggests discernible discrepancies between the two distributions. In Figure 3.3, we see that lack of sparsity in β_* causes serious problems in Type I error for both VBRD and BCH. Inference under dense β_* and dense a turns out to be the most challenging problem for existing methods; in Figure 3.4, we see quite noticeable difference between the histogram of VBRD test and $\mathcal{N}(0,1)$. In contrast, the distribution of the test statistics of the proposed methods closely match $\mathcal{N}(0,1)$ in all the scenarios, as established in Theorems 3.2.1 and 3.3.1. The Type I errors, reported in Table 3.1, confirm the above findings: existing methods can suffer greatly from lack of sparsity in β_* and/or a in terms of validity – observed Type I error of BCH or VBRD can easily reach 40%.

We also contrast the power properties of the proposed tests with respect to the existing methods. Results are collected in Figures 3.5, 3.6 and 3.7, where we plot the power curves of competing methods for design Examples 1, 2 and 3 described above with hypothesis setting of (i)-(iv). The overall message is clear from

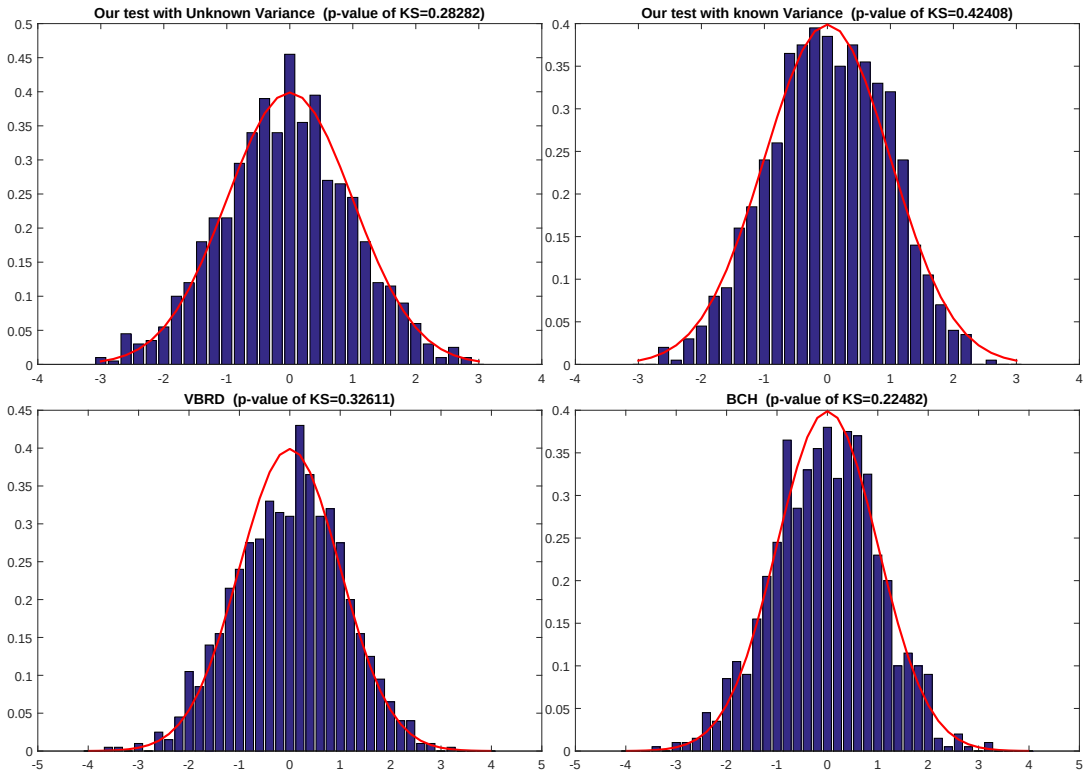


Figure 3.1: Distribution of the test statistics under the null hypothesis $H_0 : \beta_{*,2} = 0.8$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example we consider sparse β and sparse a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov-Smirnov test statistics in the subtitles.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

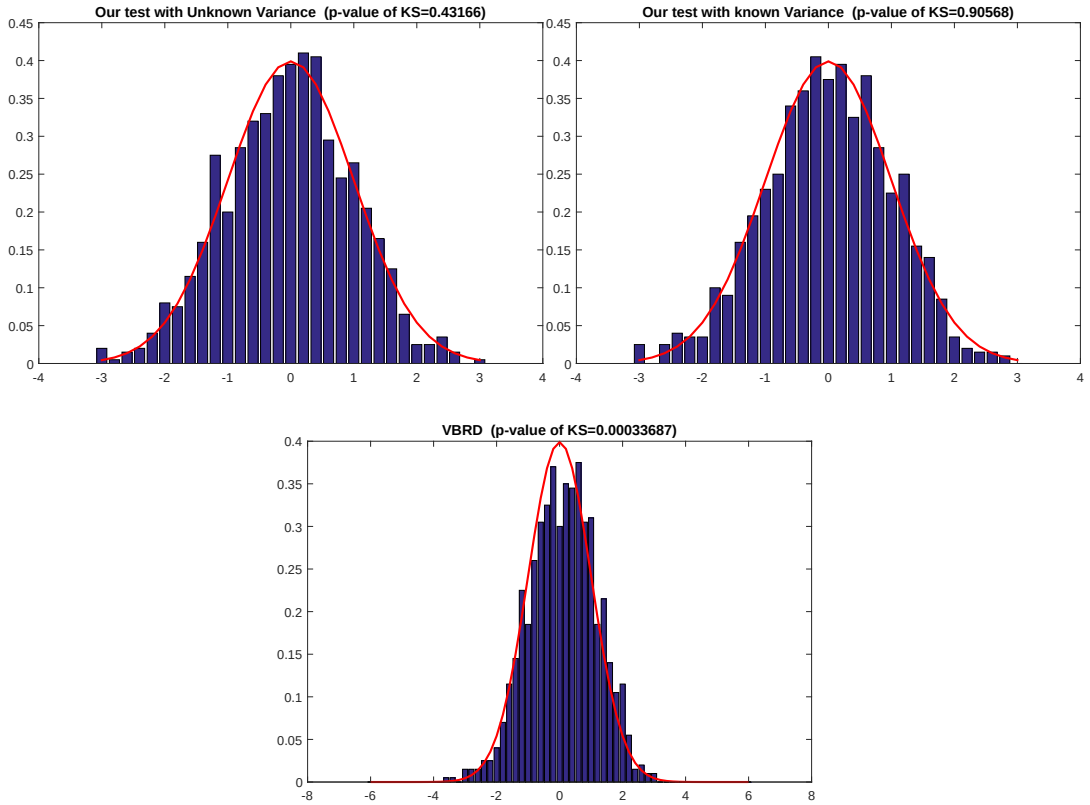


Figure 3.2: Distribution of the test statistics under the null hypothesis $H_0 : \sum_{j=1}^{p} a_j \beta_{*,j} = 1.6$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example we consider sparse β and dense a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov-Smirnov test statistics in the subtitles.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

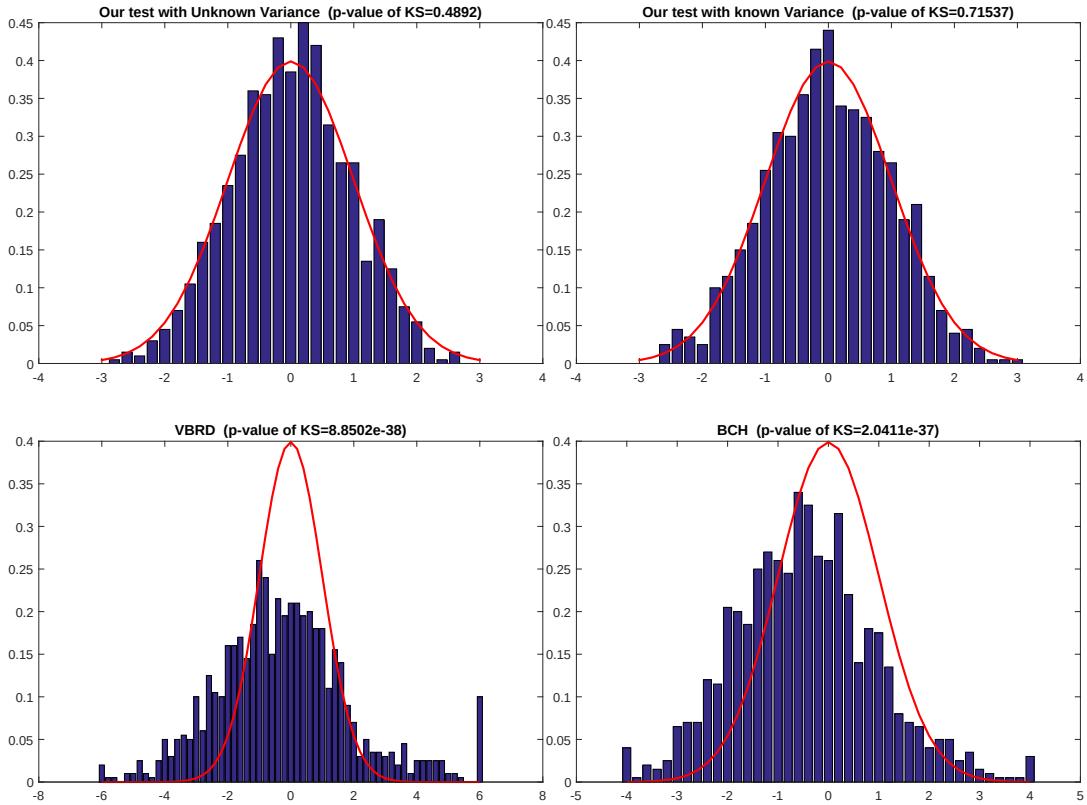


Figure 3.3: Distribution of the test statistics under the null hypothesis $H_0 : \beta_{*,2} = 3/\sqrt{p}$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example we consider dense β and sparse a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov-Smirnov test statistics in the subtitles.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

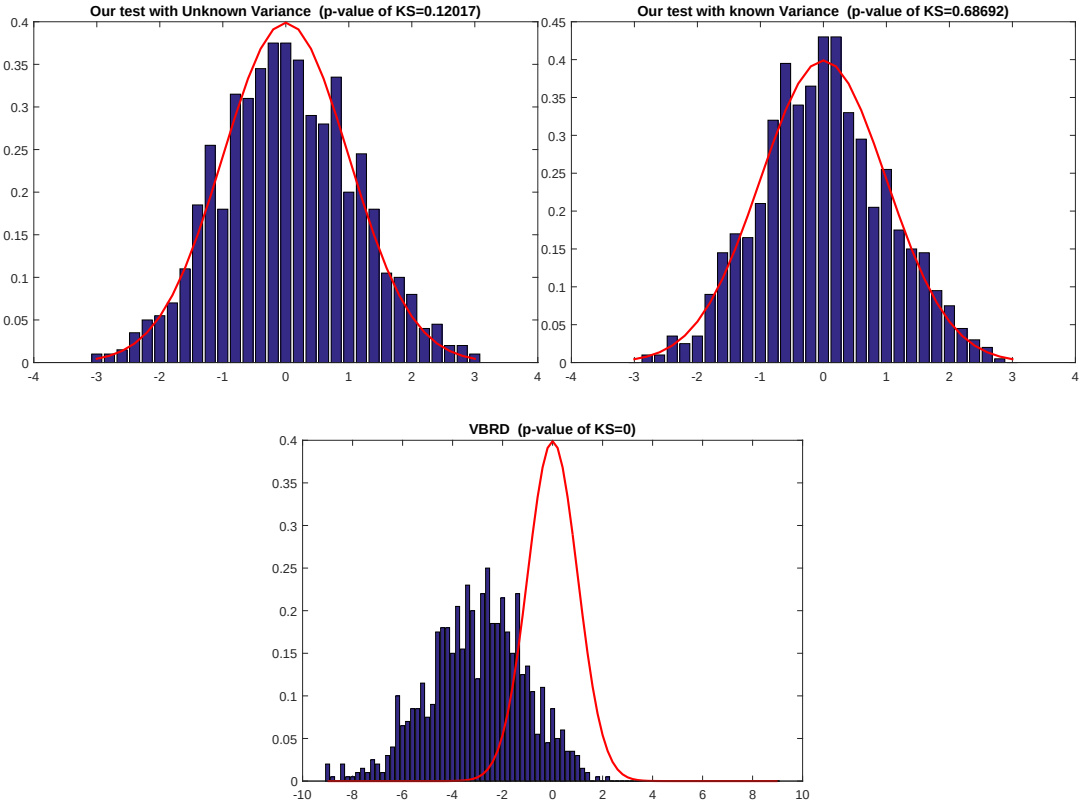


Figure 3.4: Distribution of the test statistics under the null hypothesis $H_0 : \sum_{j=1}^{p} \beta_{*,j} = 3\sqrt{p}$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example we consider dense β and dense a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov-Smirnov test statistics in the subtitles.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

Table 3.1: Type I errors over 500 repetitions of the 5% level proposed tests together with VBRD and BCH. In the table, NA symbol indicates that the method cannot be implemented "as is".

	Type I Error			
Hypothesis Setting	Unknown Σ_X	Known Σ_X	VBRD	BCH
Sparse β and Sparse a	7.4%	5.6%	8.2%	6.6%
Sparse β and Dense a	4.4%	4.8%	7.4%	NA
Dense β and Sparse a	3.6%	4.4%	33.4%	27.2%
Dense β and Dense a	5.6%	3.0%	67.2%	NA

these figures: our tests and existing methods are quite similar for sparse β_* and sparse a, whereas our tests behave nominally for other problems with preserving both low Type I error rates and Type II error rates. The biggest advantages are seen for dense vectors β_* with other methods behaving in a manner close to random guessing. In addition to the advantages in Type I error, our methods also display certain power advantages. In the case of equal-correlation setting we observe that our methods consistently reach faster power than BCH method even in the case of all sparse setting. Observe that the precision matrix in this setting is not sparse and our methods are still well-behaved. In the case of dense models, VBRD method completely breaks down with Type I or Type II error being close to 1. For non-Gaussian design we see that VBRD may not be a nominal test any more regardless of the model sparsity. BCH behaves more stably in this case but fails to apply for the hypothesis settings (ii) and (iv) as described at the beginning of the Section. In conclusion, we observe that our methods are stable across vastly different designs and model setting whereas existing methods fail to control either Type I error rate or Type II error rate. Hence the proposed methodology offers a robust and more widely applicable alternative to the existing inferential procedures, achieving better error control in difficult setting and not loosing much in the simples cases.

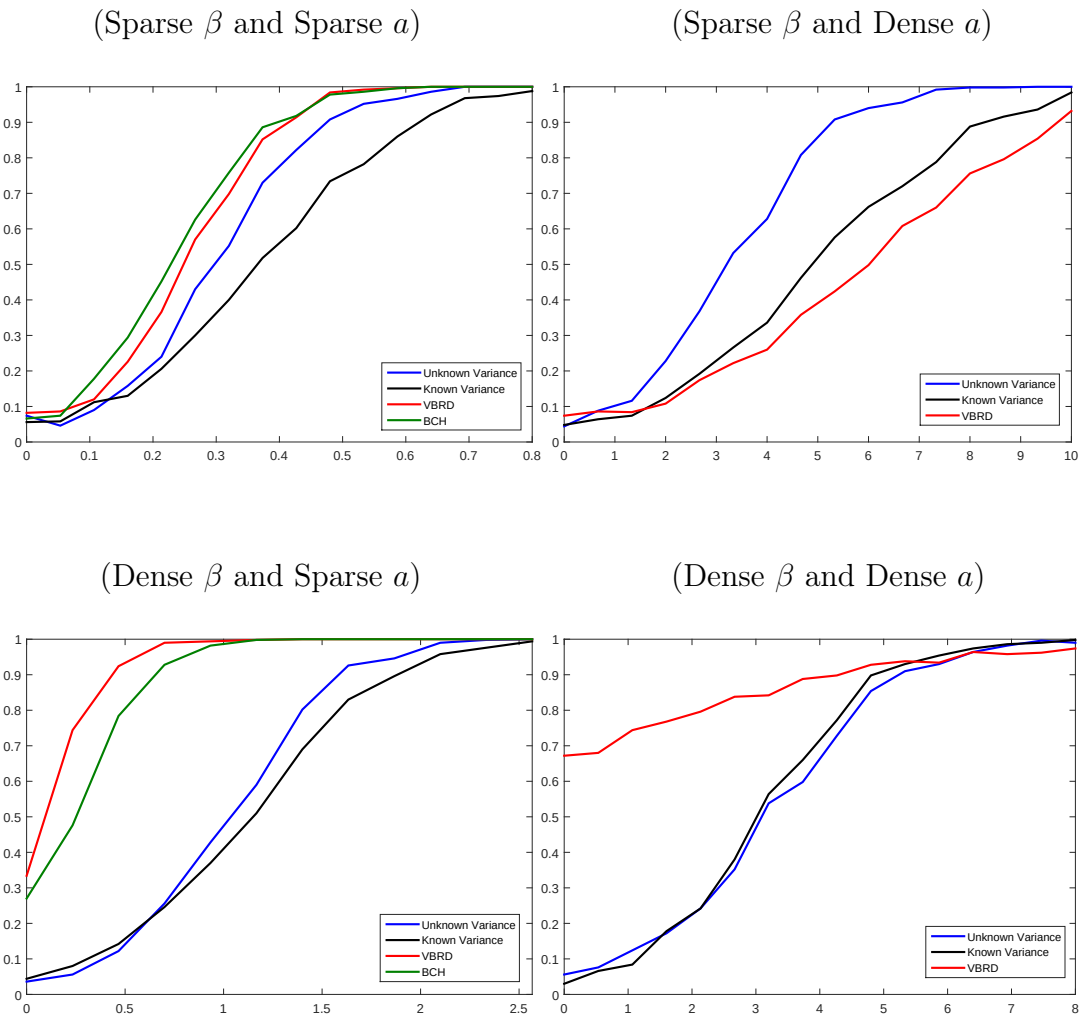


Figure 3.5: Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings.

Design settings follows Example 1 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axes. The y-axes contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type-I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Geer, Bühlmann, Ritov, and Dezeure (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective paper it wasn't included in the graph.

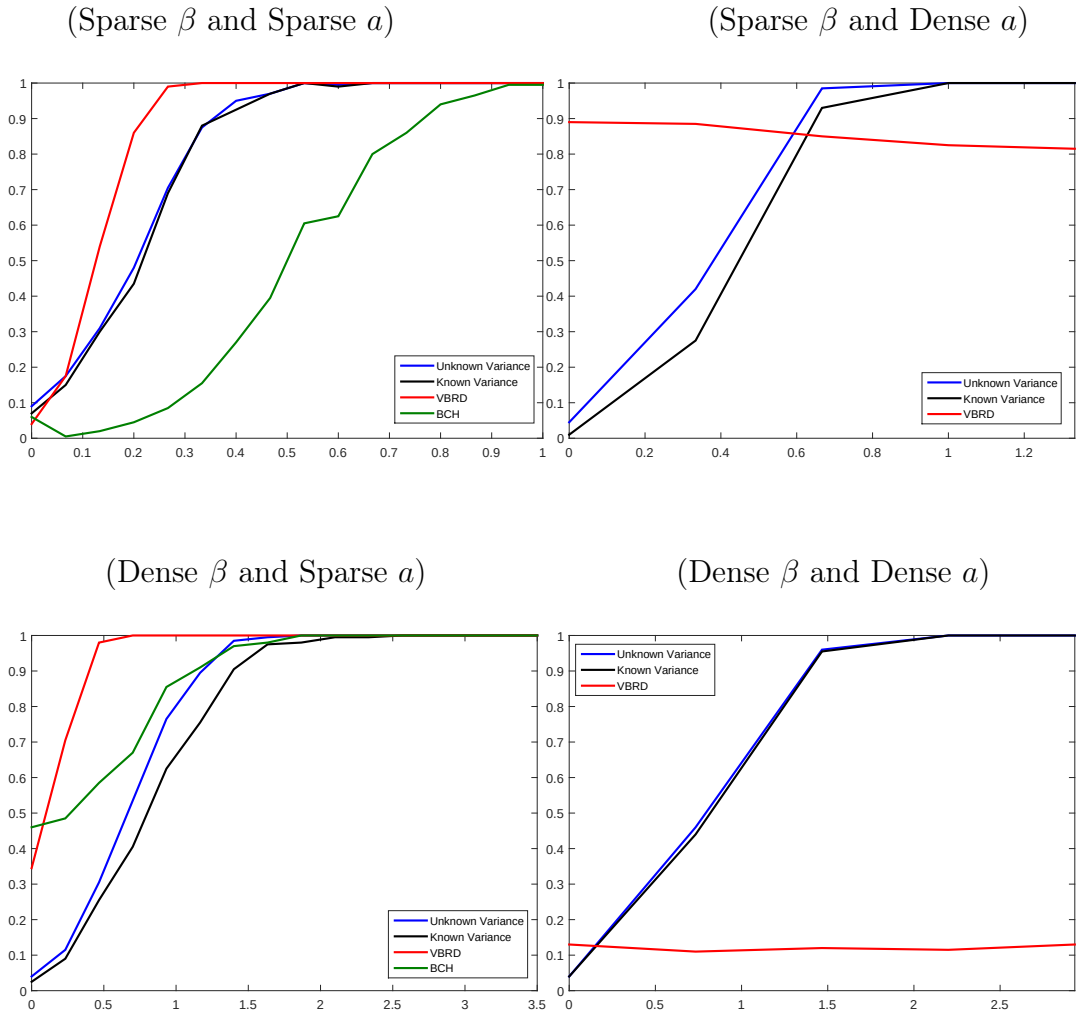


Figure 3.6: Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings. Design settings follows Example 2 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axes. The y-axes contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type-I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Geer, Bühlmann, Ritov, and Dezeure (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective paper it wasn't included in the graph.

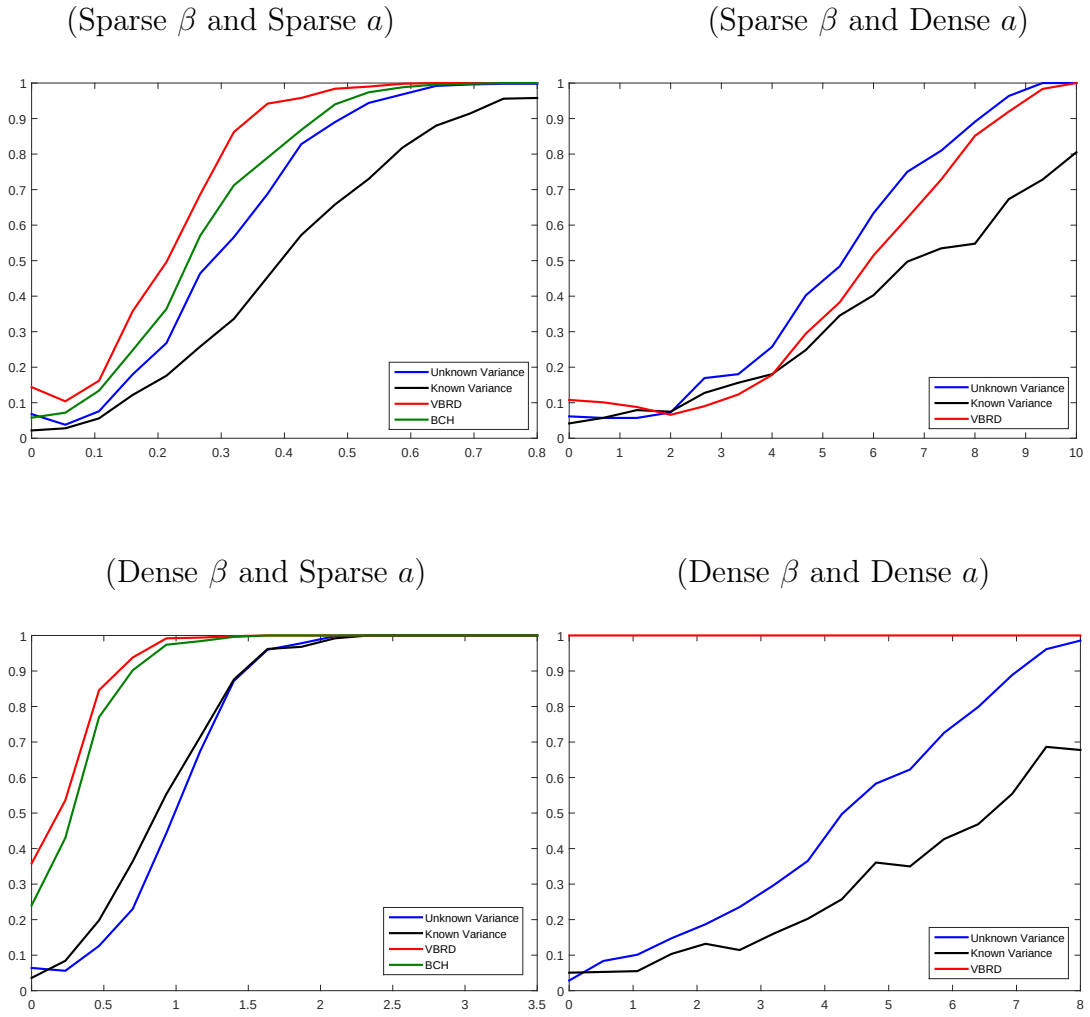


Figure 3.7: Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings. Design settings follows Example 3 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axes. The y-axes contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type-I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Geer, Bühlmann, Ritov, and Dezeure (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively.

Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective paper it wasn't included in the graph.

3.5.2 Real data example: equity risk premia

We apply the methods developed in Section 3.3 to inference of equity risk premia during different states of the economy. Some studies have found that the risk premia of stock market returns have different predictability, depending on whether the macroeconomy is in recession or expansion; see Rapach, Strauss, and Zhou (2010), Henkel, Martin, and Nardari (2011) and Dangl and Halling (2012). One common explanation for this is time variation in risk premia; see Henkel, Martin, and Nardari (2011). It is plausible that the stock market is riskier in recessions than in expansions and thus a higher expected return is demanded by investors, implying that the expected stock returns can be predicted by the state of the macroeconomy. In this section, we revisit this argument by directly conducting inference on the expected return of the stock market conditional on a large number of macroeconomic variables.

Let y_t be the excess return of the U.S stock market observed at time t and $x_{t-1} \in \mathbb{R}^p$ be a large number of macroeconomic variables observed at time $t-1$. Let $s_t \in \{0, 1\}$ denote the NBER recession indicator; $s_t = 1$ means that the economy is in recession at time t. We would like to conduct inference on $E(y_t \mid x_{t-1})$ for the two different values of s_{t-1}. Formally, we wish to construct confidence intervals for the following quantities: (a) $E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 1]$, (b) $E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 0]$ and (c) $E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 1] - E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 0]$.

We impose a linear model on the risk premia: $E(y_t \mid x_{t-1}) = x_{t-1}^\top \beta_*$ for some unknown $\beta_* \in \mathbb{R}^p$. Hence, the quantities of interest are: $a_1^\top \beta_*$, $a_0^\top \beta_*$ and $(a_1 - a_0)^\top \beta_*$, where $a_j = E(x_{t-1} \mid s_{t-1} = j)$. The macroeconomic variables we use are from the dataset constructed by McCracken and Ng (2015). We also include the squared, cubed and fourth power of these variables, leading to $p = 440$ (after removing variables with more than 30 missing observations). It is possible that $\beta_* \in \mathbb{R}^p$ is not a sparse vector because many macroeconomic variables might be relevant and each might only explain a tiny fraction of the equity risk premia. Therefore, the methods proposed in this article are particularly useful because they do not assume the sparsity of β_*.

Remark 3.5.1. There have been numerous attempts to include information from

Table 3.2: 95% confidence intervals for equity risk premia

The values are reported in annualized percentage, i.e., 2.79 means 2.79%.

	Lower bound	Upper bound
Risk premia in expansion $a_0^\top \beta_*$:	2.79	10.94
Risk premia in recession $a_1^\top \beta_*$:	6.32	36.92
Risk premia difference $(a_1 - a_0)^\top \beta_*$:	5.13	38.30

many macroeconomic variables in estimating the equity risk premium. Rapach, Strauss, and Zhou (2010) use the model combination approach by taking the simple average of 14 univariate linear models. Although this approach manages to reduce the variance in the predictions, it only produces a single point prediction and does not deliver a confidence interval. Moreover, under the specification of $E(y_t \mid x_{t-1}) = x_{t-1}^\top \beta_*$, we should not expect the simple average of predictions by individual components of x_{t-1} to be close to $x_{t-1}^\top \beta_*$, especially with highly correlated regressors. Another popular approach is to use factor models. This method is widely used in macroeconomics for predictions; see Stock and Watson (2002a), Stock and Watson (2002b) and McCracken and Ng (2015). The idea is to extract a few principal components (PC's) from x_t and to predict y_t using these PC's. Although the PC's account for a large variation in x_{t-1}, they are not hard-wired to have high predictive power for y_t unless we assume that the PC's capture the factors that drive y_t. In some sense, this factor approach only uses information in x_{t-1} that is relevant for predicting variations among different components of x_{t-1}; by contrast, the methodology we propose in this article allows us to use all the information in x_{t-1}.

Our dataset has 659 monthly observations starting from 1960. We use the first 20 years ($n = 240$) to train the data and the last $659 - n$ months to compute $a_j = \sum_{t=n+1}^{659} x_t \mathbf{1}\{s_t = j\} / \sum_{t=n+1}^{659} \mathbf{1}\{s_t = j\}$. In other words, we investigate the equity risk premia between 1980 and 2014. We conduct inference on the average equity risk premia in different states of the macroeconomy. The 95% confidence intervals for $a_1^\top \beta_*$, $a_0^\top \beta_*$ and $(a_1 - a_0)^\top \beta_*$ are reported in Table 3.2.

The confidence intervals in Table 3.2 are very informative for our purpose.

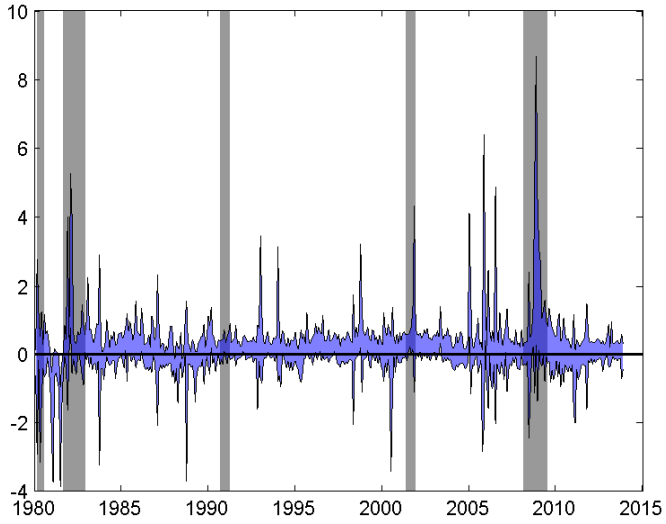


Figure 3.8: 95% confidence interval for the risk premia at each time period (the blue band) with the grey shades representing the NBER recession periods.

The results presented in Table 3.2, imply that the risk premia in recessions are higher than in expansions and that the magnitude of difference is economically meaningful. These results are consistent with existing literature; see Table 1 of Henkel, Martin, and Nardari (2011). Figure 3.8 plots the confidence intervals for $E(y_t \mid x_{t-1})$ at each t. This figure is consistent with the hypothesis that, during the Recessions (e.g., in the early 80's or around 2008), the risk premia went up substantially.

3.6 Discussions

In this article, we develop new methodology for testing hypotheses on $a^\top \beta_*$, where a is given and β_* is the regression parameter of a high-dimensional linear model. Under the proposed methodology, a new restructured regression and with features that are synthesized and augmented, is constructed based on a and is used to obtain moment conditions that are equivalent to the null hypothesis. Estimators proposed are tailored to the problem at hand and solve constrained high-dimensional optimization problems. The two proposed methods deal with the scenario with

known Σ_X and the scenario with unknown Σ_X, respectively. The first can be used when a prior information about correlation among the features exists; a case of independent features, whereas the second applies more broadly to many scientific examples where feature correlations need to be estimated. To solve a high-dimensional inference problem, there exists at least one competing choice. It is based on the "debiasing" principles of Zhang and Zhang (2014). However, the principles laid out therein only apply to strictly sparse linear models. Therefore, we fulfill an important gap in the existing literature by developing methodology that allows fully non-sparse linear models.

Restructuring the model according to the hypothesis under testing allows for the high-dimensional a and β_* that are not necessarily sparse. The synthesized features are customized based on the null hypothesis and are close to being orthogonal. We note that this customization is the key, since the orthogonality per se is not useful. Techniques that only induce feature orthogonality, such as pre-conditioning by Jia and Rohe (2012) and DECO by Wang, Dunson, and Leng (2016), still cannot be used to test $H_0 : \ a^\top \beta_* = g_0$ when a and β_* are dense.

Observe that we have proposed two different inferential methods. However, it is not necessarily true that the method proposed in Section 2 dominates the one proposed in Section 3 in terms of power. The main difference between the method is in the definition of the moment condition. The method assuming knowledge of Σ_X avoids estimation of β_* and hence is extremely easy to implement; however, when β_* is sparse (and thus easy to estimate), not using information on β_* can cause some loss of power. The method proposed in Section 2 essentially treats $w_i^\top \beta_*$ as the error term. In contrast, the method proposed in Section 3 computes an estimate for $\tilde{w}_i^\top \pi_*$ (which in spirit corresponds to $w_i^\top \beta_*$); when the model turns out to be sparse, the method without knowledge of Σ_X can essentially "remove" $w_i^\top \beta_*$ from the error term, thereby achieving better power. For dense models, this reasoning does not apply and thus it is not clear which one should be more powerful.

To conclude the article, we would like to discuss here valuable topics for future research. The proposed methodology can be used to conduct inference of conditional distributions of the response, whenever the distribution function of ε,

$Q(\cdot)$ is known or is consistently estimated. Specific example includes construction of prediction intervals for high-dimensional linear models – a topic of extreme importance. For $F_{Y|X}(y,x) = P(y_{n+1} \leq y \mid x_{n+1} = x)$ $F_{Y|X}$ can be parametrized as $F_{Y|X}(y,x;\beta_*,Q) = Q(y - x^\top \beta_*)$. For a given x, we can obtain a confidence set for $x^\top\beta_*$: $\hat{I}(1-\alpha,x)$ such that $P(x^\top\beta_* \in \hat{I}(1-\alpha,x)) \to 1-\alpha$, by inverting the tests proposed in this article. This leads to a natural confidence set for the $F_{Y|X}(y,x)$: $P(F_{Y|X}(\cdot,x) \in \hat{S}(1-\alpha,x)) \to 1-\alpha$, where

$$\hat{S}(1-\alpha,x) = \{Q(\cdot - c) \mid c \in \hat{I}(1-\alpha,x)\}.$$

If we restrict the model parameters to be sparse, then we can consistently estimate ε_i (and thus $Q(\cdot)$) and consequently form valid prediction intervals – a topic of specific importance for practitioners. However, when the model is allowed to be non-sparse and high-dimensional, the question of construction of prediction intervals hasn't been answered and needs special considerations. Additionally, under this setup, the proposed methods also lead to an inference method for (possibly nonlinear) functionals of the conditional distribution of y_{n+1} given x_{n+1}. For example, suppose that one is interested in $H(u,x) = \inf\{y \in \mathbb{R} \mid F_{Y|X}(y,x) \geq u\}$. Following the above proposal, we can simply take

$$\hat{\mathcal{H}}(u,x,\alpha) = \{\inf\{y \in \mathbb{R} \mid Q(y-c) \geq u\} \mid c \in \hat{I}(1-\alpha,x)\}$$

as a confidence set for $H(u,x)$.

3.7 Acknowledgements

Chapter 3, in full, is joint work with Jelena Bradic and has been submitted for publication of the material as it may appear in Zhu, Yinchu; Bradic, Jelena, Journal of the American Statistical Association, 2017. The dissertation author was the primary investigator and author of this paper.

Appendix A

Proofs and examples for Chapter 1

In Appendix A.1, we provide a simple example illustrating the difficulties arising from the cross-sectional dependence in the error terms. The rest of the appendix contains proofs for the technical results in the main text. The theoretical results in the main text are proved in Appendix A.2. In Appendix A.3, we provide useful technical tools that are used in Appendix A.2.

We introduce some notations that will be used extensively for the rest of the paper. We denote $\max\{a, b\}$ and $\min\{a, b\}$ by $a \vee b$ and $a \wedge b$, respectively. For a matrix A, we define $\|A\|_\infty = \|\text{vec}A\|_\infty$. For a positive integer q, we define $[q] = \{1, \ldots, q\}$. For a set A, $|A|$ denotes the cardinality (number of elements) of A. We will repeatedly use the notation $O_P(\log^{O(1)} n)$ to denote a term of order $O_P((\log n)^r)$ for some constant $0 < r < \infty$. Finally, "wpa1" denotes the phrase "with probability approaching one".

A.1 An example of difficulties due to cross-sectional dependence

Suppose that $y_{i,t} = x'_{i,t}\beta_t + \varepsilon_{i,t}$ with $\varepsilon_{i,t} = u_{i,t} + L'_i F_t$. Assume that, for each t, $\{(L_i, x_{i,t})\}_{i=1}^n$ is independent of $\{(F_t, u_{i,t})\}_{i=1}^n$, $\mathbb{E}F_t = 0$ and $\mathbb{E}u_{i,t} = 0$. Therefore, strict exogeneity holds: $\mathbb{E}(\varepsilon_{i,t} \mid \{x_{i,t}\}_{i=1}^n) = 0$.

However, the OLS estimator for each t might not be consistent for β_t. To

see this, let $\hat{\beta}_{OLS,t} = (\sum_{i=1}^{n} x_{i,t}x_{i,t}')^{-1}(\sum_{i=1}^{n} x_{i,t}y_{i,t})$. The estimation error takes the form

$$\hat{\beta}_{OLS,t} - \beta_t = \left[n^{-1} \sum_{i=1}^{n} x_{i,t}x_{i,t}' \right]^{-1} \left[n^{-1} \sum_{i=1}^{n} x_{i,t}u_{i,t} + \left(n^{-1} \sum_{i=1}^{n} x_{i,t}L_i' \right) F_t \right]$$

The problem is that $n^{-1} \sum_{i=1}^{n} x_{i,t}L_i'$ need not be close to zero since the sequence $\{L_{\alpha,i}d_{i,t-1}\}_{i=1}^{n}$ might not have weak dependence across i and $\mathbb{E}x_{i,t}L_i'$ could be non-zero.

A.2 Proofs of theoretical results in the main text

We provide the proof of Theorem 1.3.1 in Appendix A.2.1. Appendix A.2.2 contains proofs of Theorems 1.3.3, 1.3.4 and 1.3.5, as well as Theorem 1.3.2 and Corollary 1.3.1. Other results, including Theorems 1.3.6, 1.3.7 and 1.4.1, are proved in Appendix A.2.3. In Appendix A.2.4, we show Lemma A.2.16, which establishes strong mixing properties for the process described in Example 1.2.1. We recall some definitions used in the main text as well as introducing some new definitions

that will be used in the rest of this section:

$$\begin{cases} \Sigma_t & = n^{-1}\sum_{i=1}^n \mathbb{E} v_{t,i} v_{t,i}' \\ \bar{v}_{i,t} & = \Sigma_t^{-1} v_{t,i} \\ G_{i,t} & = \bar{v}_{i,t} u_{i,t} \\ \hat{\Sigma}_t & = n^{-1}\hat{v}_t'\hat{v}_t = n^{-1}\sum_{i=1}^n \hat{v}_{i,t}\hat{v}_{i,t}' \\ \hat{\bar{v}}_{i,t} & = \hat{\Sigma}_t^{-1}\hat{v}_{i,t} \\ \hat{G}_{i,t} & = \hat{\bar{v}}_{i,t}\hat{u}_{i,t} \\ G_i & = (G_{i,1}', \ldots, G_{i,T}')' \\ \hat{G}_i & = (\hat{G}_{i,1}', \ldots, \hat{G}_{i,T}')' \\ \Omega & = n^{-1}\sum_{i=1}^n \mathbb{E} J G_i G_i' J' \\ \hat{\Omega} & = n^{-1}\sum_{i=1}^n J\hat{G}_i\hat{G}_i' J' \\ u_t & = (u_{1,t}, \ldots, u_{n,t})' \in \mathbb{R}^n \\ v_t & = (v_{1,t}, \ldots, v_{n,t})' \in \mathbb{R}^{n\times k} \\ \hat{v}_t & = (\hat{v}_{1,t}, \ldots, \hat{v}_{n,t})' \in \mathbb{R}^{n\times k} \\ \alpha_t & = (\alpha_{1,t}, \ldots, \alpha_{n,t})' \in \mathbb{R}^n \qquad \text{with } \alpha_{i,t} = L_{\alpha,i}' F_{\alpha,t} \\ \hat{\alpha}_t & = \hat{L}_\alpha \hat{F}_{\alpha,t} \\ \hat{u}_t & = y_t - X_t\hat{\beta}_t - \hat{\alpha}_t \\ D_{n,t} & = n^{-1/2}\hat{\Sigma}_t^{-1}\hat{v}_t'(\alpha_t - \hat{\alpha}_t) + n^{-1/2}\left(\hat{\Sigma}_t^{-1}\hat{v}_t' - \Sigma_t^{-1}v_t'\right)u_t \\ D_n & = (D_{n,1}', \ldots, D_{n,T}')' \end{cases} \tag{A.2.1}$$

A.2.1 Proof of Theorem 1.3.1

Lemma A.2.1. *Under Assumption 1, the following hold:*

(1) $\|L_Q\|_\infty$, $\|L_\alpha\|_\infty$, $\|F_Q\|_\infty$, $\|F_\alpha\|_\infty$, $\|u\|_\infty$, $\|v\|_\infty$, $\max_{(i,t)\in[n]\times[T]}\|\bar{v}_{i,t}\|$, $\max_{i,t}\|\bar{v}_{i,t}u_{i,t}\|$, *and* $\max_{(i,t)\in[n]\times[T]}\|x_{i,t}\|$ *are* $O_P(\log^{O(1)} n)$.

(2) both $\|u\|$ *and* $\|v\|$ *are* $O_P(\sqrt{n\log n})$.

Proof. **Proof of part (1).** The first six claims hold by Lemma A.3.7 and the

exponential-type tails in Assumption 1.

To bound $\max_{i,t} \|\bar{v}_{i,t}\|$, notice that the $\|\cdot\|_1$-norm of rows of Σ_t^{-1} are bounded by some constants due to Assumption 1. Therefore, by Lemma A.3.3(1), entries of $\bar{v}_{i,t}$ have exponential-type tails with parameters that depend only on the constants in Assumption 1. Thus, Lemma A.3.7 implies $\max_{(i,t)\in[n]\times[T]} \|\bar{v}_{i,t}\| \leq \sqrt{k} \max_{(i,t)\in[n]\times[T]} \|\bar{v}_{i,t}\|_\infty = \sqrt{k} O_P(\log^{O(1)} |[n] \times [T]|) = O_P(\log^{O(1)} n)$.

To see the bound for $\max_{i,t} \|\bar{v}_{i,t} u_{i,t}\|$, notice that Lemma A.3.3(3) implies the exponential-type tail for entries of $\bar{v}_{i,t} u_{i,t}$. Then the bound follows by Lemma A.3.7.

To see the last claim of part (1), notice that $x_{i,t} = L'_{Q,i} F_{Q,t} + v_{i,t}$. Since entries of $L_{Q,i}$, $F_{Q,t}$ and $v_{i,t}$ have exponential-type tails, it follows, by Lemma A.3.3, that entries of $x_{i,t}$ also have exponential-type tails with parameters that only depend on the constants in Assumption 1. Thus, the bound for $\max_{(i,t)\in[n]\times[T]} \|x_{i,t}\|$ follows by Lemma A.3.7. We have proved part (1).

Proof of part (2). We apply the random matrix theory. By Theorem 5.48 and Remark 5.49 in Vershynin (2010),

$$\mathbb{E}\|u\| \leq C^{1/2} n^{1/2} + \bar{C} \sqrt{m \log (n \wedge T)}, \tag{A.2.2}$$

where $\bar{C}$ is an absolute constant and $m := \mathbb{E} \max_i \|\underline{u}_i\|^2$, where $\underline{u}_i = (u_{i,1}, \ldots, u_{i,T})' \in \mathbb{R}^T$. Let $s_i^2 = \mathbb{E}\|\underline{u}_i\|^2$.

By Lemma A.3.3(3)-(4), there exists a constant $b_* > 0$ such that $u_{i,t}^2 - \mathbb{E}u_{i,t}^2$ has an exponential-type tail with parameter (b_*, γ_1), where $\gamma_1 = \gamma_*/2$. Let $\gamma_2 = \min\{\gamma_{**}, 1/2\}$. Then $\alpha_n(t) \leq b_2 \exp(-t^{\gamma_2})$ and $\gamma_2 < 1$. Hence, $\gamma = (\gamma_1^{-1} + \gamma_2^{-1})^{-1} < \gamma_2 < 1$. By Theorem 1 in Merlevède, Peligrad, and Rio (2011), there exist positive constants $C_1, \ldots, C_5 > 0$ depending only on b_*, b_2, γ and γ_2 such that $\forall x > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\|\underline{u}_i\|^2 - s_i^2\right| > a_n x\right) \\
&= \mathbb{P}\left(\left|\sum_{t=1}^{T} \left(u_{i,t}^2 - \mathbb{E}u_{i,t}^2\right)\right| > a_n x\right)
\end{aligned}$$

$$\leq T\exp\left(-C_1 a_n^{\gamma}x^{\gamma}\right)+\exp\left(-\frac{C_2 a_n^2 x^2}{1+C_3 T}\right)$$
$$+\exp\left[-\frac{C_4 a_n^2 x^2}{T}\exp\left(C_5(a_n x)^{\gamma/(1-\gamma)}(\log a_n x)^{-\gamma}\right)\right],$$

where $a_n = d_*\sqrt{T\log n}$ and d_* is a constant to be determined. The union bound implies that $\forall x > 0$,

$$\mathbb{P}\left(\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| > a_n x\right)$$
$$\leq \sum_{i=1}^{n}\mathbb{P}\left(\left|\|\underline{u}_i\|^2 - s_i^2\right| > a_n x\right)$$
$$\leq nT\exp\left(-C_1 a_n^{\gamma}x^{\gamma}\right)+n\exp\left(-\frac{C_2 a_n^2 x^2}{1+C_3 T}\right)$$
$$+n\exp\left[-\frac{C_4 a_n^2 x^2}{T}\exp\left(C_5(a_n x)^{\gamma/(1-\gamma)}(\log a_n x)^{-\gamma}\right)\right].$$

Thus, by elementary computations, we can choose large constants $a_*, b_*, d_* > 0$ such that $\forall x \geq a_*$

$$\mathbb{P}\left(\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| / \left(d_*\sqrt{T\log n}\right) > x\right) = \mathbb{P}\left(\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| > a_n x\right)$$
$$\leq b_*\exp\left(-x^{\gamma}\right). \quad \text{(A.2.3)}$$

Therefore,

$$\mathbb{E}\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| / \left(d_*\sqrt{T\log n}\right)$$
$$\overset{(i)}{=} \int_0^{\infty}\mathbb{P}\left(\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| / \left(d_*\sqrt{n\log n}\right) > x\right)dx$$
$$\leq a_* + \int_{a_*}^{\infty}\mathbb{P}\left(\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| / \left(d_*\sqrt{n\log n}\right) > x\right)dx$$
$$\overset{(ii)}{\leq} a_* + b_*\int_{a_*}^{\infty}\exp\left(-x^{\gamma}\right)dx$$
$$= O(1),$$

where (i) follows by the identity $\mathbb{E}X = \int_0^{\infty}\mathbb{P}(X > x)dx$ for any non-negative

random variable X and (ii) holds by (A.2.3). The above display implies that

$$m := \mathbb{E}\max_i \|\underline{u}_i\|^2 \leq \mathbb{E}\max_i \left|\|\underline{u}_i\|^2 - s_i^2\right| + \max_i s_i^2 = \sqrt{T\log n}O(1) + Cn$$
$$= O(n \vee \sqrt{T\log n}) \overset{(i)}{=} O(n),$$

where (i) holds by $T \asymp n^\xi$ with $\xi \in (6/7, 2)$. The above display and (A.2.2) implies that $\mathbb{E}\|u\| = O(\sqrt{n\log n})$ and thus $\|u\| = O_P(\sqrt{n\log n})$. This proves part (2) for $\|u\|$. The result for $\|v\|$ follows by an analogous argument. The proof is complete. □

Lemma A.2.2. *Let Assumption 1 hold. Then the following hold:*
(1) $\max_{(i,j,t)\in[n]\times[k]\times[T]} \sum_{s=1}^T |\mathbb{E}(v_{i,t,j}u_{i,s})| = O(1)$*;*
(2) $\max_{(i,t,j_1,j_2)\in[n]\times[T]\times[k]\times[k]} \sum_{s=1}^T |\mathbb{E}(v_{i,t,j_1}v_{i,s,j_2})| = O(1)$*;*
(3) $\max_{(i,s)} \sum_{t=1}^T |\mathbb{E}(G'_{i,t}G_{i,s})| = O(1)$*.*

Proof. We first show part (1). By the exponential-type tails, Lemma A.3.3(2) implies that there exists a constant $C > 0$ such that $\max_{i,t,j} \mathbb{E}|v_{i,t,j}|^4 \leq C$ and $\max_{i,s} \mathbb{E}|u_{i,s}|^4 \leq C$. By Corollary 16.2.4 of Athreya and Lahiri (2006), we have that, $\forall i,t,j,s$, $|\mathbb{E}(v_{i,t,j}u_{i,s})| \leq 4\left[2\alpha_{mixing}(|t-s|)\right]^{1/2} C^2$. Therefore,

$$\max_{i,j,t}\sum_{s=1}^T |\mathbb{E}(v_{i,t,j}u_{i,s})| \leq 4\sqrt{2}C^2 \max_t \sum_{s=1}^T \sqrt{\alpha_{mixing}(|t-s|)}$$
$$\overset{(i)}{\leq} 4\sqrt{2}C^2c_* \max_t \sum_{s=1}^T \exp[-|t-s|^{\gamma_{**}}] \leq 8\sqrt{2}C^2c_* \sum_{\tau=1}^\infty \exp[-\tau^{\gamma_{**}}],$$

where (i) holds by Assumption 1. Since $\sum_{\tau=1}^\infty \exp[-\tau^{\gamma_{**}}] < \infty$, the part (1) follows.

Notice that for i,t,s,j_1,j_2, Corollary 16.2.4 of Athreya and Lahiri (2006) still implies that $|\mathbb{E}(v_{i,t,j_1}v_{i,s,j_2})| \leq 4\left[2\alpha_{mixing}(|t-s|)\right]^{1/2} C^2$. Part (2) follows by the same argument.

To see part (3), let $\bar{v}_{i,t,j}$ denote the j-th component of $\bar{v}_{i,t}$ (defined in (A.2.1)). Since each row of Σ_t^{-1} is bounded in $\|\cdot\|_1$-norm and entries of $v_{i,t}$ have exponential-type tails, it follows, by Lemma A.3.3(1), that $\bar{v}_{i,t,j}$ has an exponential-type tail. Using Lemma A.3.3(3), we have that $\bar{v}_{i,t,j}u_{i,t}$ has an exponential-type

tail. Then by the same argument as in the proof of part (1), we have that $\max_{(i,s,j)\in[n]\times[T]\times[k]} \sum_{t=1}^{T} |\mathbb{E}\bar{v}_{i,t,j}u_{i,t}\bar{v}_{i,s,j}u_{i,s}| = O(1)$. Since k is fixed, part (3) follows by $G'_{i,t}G_{i,s} = \sum_{j=1}^{k} \bar{v}_{i,t,j}u_{i,t}\bar{v}_{i,s,j}u_{i,s}$. The proof is complete. □

Lemma A.2.3. *Under Assumption 1, the following hold:*
(1) $\max_{i\in[n]} \| \sum_{t=1}^{T} v'_{i,t}F'_{Q,t}\| = O_P(T^{1/2}\log^{O(1)} n)$.
(2) $\max_{i\in[n]} \| \sum_{t=1}^{T} u_{i,t}F'_{\alpha,t}\| = O_P(T^{1/2}\log^{O(1)} n)$.
(3) $\max_{t\in[T]} \|L'_Q v_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(4) $\max_{t\in[T]} \|L'_\alpha v_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(5) $\max_{t} \|L'_Q u_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(6) $\max_{t} \|L'_\alpha u_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(7) $\max_{t\in[T]} \|v'_t u_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(8) $\max_{t\in[T]} \|n^{-1}v'_t v_t - \Sigma_t\| = O_P(n^{-1/2}\log^{O(1)} n)$.
(9) $\max_{s,t\in[T]} \| \sum_{i=1}^{n}(v_{i,t}x'_{i,s} - \mathbb{E}v_{i,t}x'_{i,s})\| = O_P(n^{1/2}\log^{O(1)} n)$.
(10) $\max_{t\in[T]} \|v'_t\alpha_t\| = O_P(n^{1/2}\log^{O(1)} n)$.
(11) $\max_{t\in[T]} \|v'_t u F_\alpha\| = O_P\left([\sqrt{nT}+n]\log^{O(1)} n\right)$.
(12) $\max_{t\in[T]} \|u'_t v F_Q\| = O_P\left([\sqrt{nT}+n]\log^{O(1)} n\right)$.

Proof. **Proof of part (1).** Let $j = (j_1, j_2) \in J := [n]\times[r_Q]$ and $\mathcal{F}_n$ the σ-algebra generated by F_Q. Since $\|F_Q\|_\infty = O_P(\log^{O(1)} n)$ (Lemma A.2.1), there exists a constant $c_0 > 0$ such that $\mathbb{P}(\mathcal{A}_n) \to 1$, where $\mathcal{A}_n = \{\|F_Q\|_\infty \le h_n\}$ with $h_n = \log^{c_0} n$.

Define $e_{t,j} = v'_{j_1,t}F'_{Q,t}\tau_{j_2}h_n^{-1}\mathbf{1}\{\|F_{Q,t}\|_\infty \le h_n\}$, where τ_{j_2} is the j_2th column of I_{r_Q}. Notice that $|e_{t,j}| \le |v_{j_1,t}|$. By Assumption 1, $\forall z > 0$, $\mathbb{P}(|e_{t,j}| > z \mid \mathcal{F}_n) \le \mathbb{P}(|v_{j_1,t}| > z \mid \mathcal{F}_n) \le \exp[1-(z/b_*)^{\gamma_*}]$ a.s. Since $\{(e_{t,j})_{j\in J}\}_{t=1}^{T}$ is strong mixing with mixing coefficient $\alpha_{mixing}(\cdot)$ defined in Assumption 1 and v is independent of $\mathcal{F}_n$, $\{(e_{t,j})_{j\in J}\}_{t=1}^{T}$ is strong mixing in the sense of Lemma A.3.8. It follows, by Lemma A.3.8, that there exist constants $c, r > 0$ depending only on the constants in Assumption 1 such that

$$\mathbb{P}\left(\left|T^{-1/2}\sum_{t=1}^{T} e_{t,j}\right| > z \,\middle|\, \mathcal{F}_n\right) \le \exp[1-(z/c)^r] \quad a.s. \qquad \forall j \in J,\ \forall z > 0.$$

This exponential-type tail condition and Lemma A.3.7 imply that

$\max_{j\in J}|T^{-1/2}\sum_{t=1}^{T}e_{t,j}| = O_P(\log^{O(1)} n)$. Therefore,

$$\max_{i\in[n]}\left\|\sum_{t=1}^{n}v_{i,t}'F_{Q,t}'\right\|\mathbf{1}_{\mathcal{A}_n} \le h_n r_Q^{1/2}T^{1/2}\max_{j\in J}\left|T^{-1/2}\sum_{t=1}^{T}e_{t,j}\right| \overset{(i)}{=} O_P(T^{1/2}\log^{O(1)} n),$$

where (i) holds by $h_n = O(\log^{O(1)} n)$. Since $\mathbb{P}(\mathcal{A}_n)\to 1$, part (1) follows.

Parts (2)-(7) follow by analogous arguments.

Proof of part (8). Notice that $n^{-1}v_t'v_t - \Sigma_t = n^{-1}\sum_{i=1}^{n}[v_{i,t}v_{i,t}' - \mathbb{E}v_{i,t}v_{i,t}']$. By Lemma A.3.3(3), there exist constants $c, r > 0$ such that each entry of $v_{i,t}v_{i,t}'$ has an exponential-type tail with parameter (c, r) for all i, t. Then part (8) follows by Lemma A.3.6.

Part (9) follows by an analogous argument.

Part (10) follows by $\max_{t\in[T]}\|v_t'\alpha_t\| \le \max_{t\in[T]}\|v_t'L_\alpha\|\max_{t\in[T]}\|F_{\alpha,t}\|$, together with part (4) and Lemma A.2.1.

Proof of part (11). Let $j = (j_1, j_2) \in J := [k]\times[r_\alpha]$. Let $\mathcal{F}_n$ be σ-algebra generated by F_α. As before, since $\|F_\alpha\|_\infty = O_P(\log^{O(1)} n)$ (Lemma A.2.1), there exists a constant $c_0 > 0$ such that $\mathbb{P}(\mathcal{A}_n)\to 1$, where $\mathcal{A}_n = \{\|F_\alpha\|_\infty \le h_n\}$ with $h_n = \log^{c_0} n$.

For $i \in [n]$ and $j = (j_1, j_2) \in J$, define $d_{i,j_2} = T^{-1/2}\sum_{s=1}^{T}u_{i,s}F_{\alpha,s,j_2}h_n^{-1}\mathbf{1}\{|F_{\alpha,s,j_2}| \le h_n\}$. Notice that $[v_t'uF_\alpha]_{j_1,j_2}\mathbf{1}_{\mathcal{A}_n} = T^{1/2}h_n\sum_{i=1}^{n}v_{i,t,j_1}d_{i,j_2}$. Notice that $\forall z > 0$,

$$\mathbb{P}(|u_{i,s}F_{\alpha,s,j_2}h_n^{-1}\mathbf{1}\{|F_{\alpha,s,j_2}|\le h_n\}| > z \mid \mathcal{F}_n) \le \mathbb{P}(|u_{i,s}| > z \mid \mathcal{F}_n) \le \exp[-(z/b_*)^{\gamma_*}].$$

Since u and F_α are independent, the sequence $\{u_{i,s}F_{\alpha,s,j_2}h_n^{-1}\mathbf{1}\{|F_{\alpha,s,j_2}|\}\}_{s=1}^{T}$ is strong mixing conditional on $\mathcal{F}_n$ in the sense of Lemma A.3.8. It follows, by Lemma A.3.8, that there exist constants $c_1, r_1 > 0$ such that $\mathbb{P}(|d_{i,j_2}| > z \mid \mathcal{F}_n) \le \exp[-(z/c_1)^{r_1}]$, $\forall z > 0$ for all i, j_2.

Since v_t is independent of $\mathcal{F}_n$, it follows, by the exponential-type tails of entries in v_t and Lemma A.3.3(3), that there exist constants $c_2, r_2 > 0$ such that $\mathbb{P}(|v_{i,t,j_1}d_{i,j_2}| > z \mid \mathcal{F}_n) \le \exp[-(z/c_2)^{r_2}]$ $\forall z > 0$. Thus, Lemma A.3.6 implies that

$$\max_{(t,j)\in[T]\times J}\left|\sum_{i=1}^{n}[v_{i,t,j_1}d_{i,j_2}-\mathbb{E}(v_{i,t,j_1}d_{i,j_2})]\right| = O_P(\sqrt{n\log|[T]\times J|}) = O_P(\sqrt{n\log n}). \quad \text{(A.2.4)}$$

Therefore,

$$\begin{aligned}
&\max_t \|u_t' v F_Q\|\mathbf{1}_{\mathcal{A}_n} \\
&\overset{(i)}{\leq} T^{1/2}\sqrt{kr_\alpha}h_n \max_{j,t}\left|\sum_{i=1}^{n} v_{i,t,j_1}d_{i,j_2}\right| \\
&\leq T^{1/2}\sqrt{kr_\alpha}h_n\left(\max_{j,t}\left|\sum_{i=1}^{n}[v_{i,t,j_1}d_{i,j_2}-\mathbb{E}(v_{i,t,j_1}d_{i,j_2})]\right| + \max_{j,t}\sum_{i=1}^{n}|\mathbb{E}(v_{i,t,j_1}d_{i,j_2})|\right) \\
&\overset{(ii)}{=} T^{1/2}O_P(\log^{O(1)}n)\left(O_P(\sqrt{n\log n}) + \max_{j,t}\sum_{i=1}^{n}|\mathbb{E}(v_{i,t,j_1}d_{i,j_2})|\right), \qquad \text{(A.2.5)}
\end{aligned}$$

where (i) follows by $[v_t' u F_\alpha]_{j_1,j_2}\mathbf{1}_{\mathcal{A}_n} = T^{1/2}h_n\sum_{i=1}^{n} v_{i,t,j_1}d_{i,j_2}$, (ii) follows by (A.2.4) and $h_n = O_P(\log^{O(1)}n)$. Notice that

$$\begin{aligned}
&\max_{j,t}\sum_{i=1}^{n}|\mathbb{E}(v_{i,t,j_1}d_{i,j_2})| \\
&\leq \max_{j,t}\sum_{i=1}^{n}\mathbb{E}|\mathbb{E}(v_{i,t,j_1}d_{i,j_2}\mid\mathcal{F}_n)| \\
&\overset{(i)}{\leq} \max_{j,t}T^{-1/2}\sum_{i=1}^{n}\mathbb{E}\left\{\sum_{s=1}^{T}|\mathbb{E}(v_{i,t,j_1}u_{i,s})|\cdot|F_{\alpha,s,j_2}|h_n^{-1}\mathbf{1}\{|F_{\alpha,s,j_2}|\leq h_n\}\right\} \\
&\leq nT^{-1/2}\max_{j,t,i}\sum_{s=1}^{T}|\mathbb{E}(u_{i,t,j_1}u_{i,s})| \\
&\overset{(ii)}{=} O(nT^{-1/2}),
\end{aligned}$$

where (i) holds by the definition of d_{i,j_2} and the independence between $\mathcal{F}_n$ and $v_{i,t,j_1}u_{i,s}$ and (ii) holds by Lemma A.2.2. The above display and (A.2.5) imply that $\max_t\|u_t' v F_Q\|\mathbf{1}_{\mathcal{A}_n} = O_P\left([\sqrt{nT}+n]\log^{O(1)}n\right)$. Since $\mathbb{P}(\mathcal{A}_n)\to 1$, part (11) follows.

Part (12) follows by an analogous argument as part (11). The proof

is complete. □

Lemma A.2.4. *Under Assumption 1,*

$$\max_{1\le i\le n, 1\le t\le T} \|\hat{v}_{i,t} - v_{i,t}\| = O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right]\log^{O(1)} n\right).$$

Proof. First notice that $\|\hat{v}_{i,t} - v_{i,t}\| = \|\hat{L}'_{Q,i}\hat{F}_{Q,t} - L'_{Q,i}F_{Q,t}\| = \|\tau_i'(\hat{L}_Q\hat{F}_{Q,t} - L_QF_{Q,t})\|$. We apply Lemma A.3.11 with $L = L_Q$, $F = F_Q$, $e = v$ and $a = \tau_i$, where τ_i is the ith column of I_n. By Lemmas A.2.1 and A.2.3(4), we have $\max_t \|F_{Q,t}\| = O_P(\log^{O(1)} n)$, $\|v\| = O_P(\sqrt{n}\log^{O(1)} n)$ and $\max_t \|L'_Q v_t\| = O_P(\sqrt{n}\log^{O(1)} n)$. Therefore, Lemma A.3.11(4) and $T \asymp n^\xi$ imply that

$$\begin{aligned}\max_{i,t} \|\tau_i'(\hat{L}_Q\hat{F}_{Q,t} - L_QF_{Q,t})\| \;\le\; & O_P(n^{-\xi}\log^{O(1)} n)\max_t \left\|\sum_{t=1}^T v'_{i,t}F'_{Q,t}\right\| \\ & + O_P\left(\left[n^{-\xi/2} + n^{-1/2}\right]\log^{O(1)} n\right)\max_i \|L_{Q,i}\| \\ & + O_P\left(\left[n^{1/2-\xi} + n^{-\xi/2}\right]\log^{O(1)} n\right)\max_i \|\tau_i\|.\end{aligned}$$

Notice that $\max_i \|\tau_i\| = 1$ and $\max_i \|L_{Q,i}\| \le \sqrt{r_Q}\|L_Q\|_\infty = O_P(\log^{O(1)} n)$ (due to Lemma A.2.1). Thus, by the above display and Lemma A.2.3(1), we have

$$\max_{i,t} \|\tau_i'(\hat{L}_Q\hat{F}_{Q,t} - L_QF_{Q,t})\| = O_P\left(\left[n^{-1/2} + n^{1/2-\xi} + n^{-\xi/2}\right]\log^{O(1)} n\right).$$

Since $\max\left\{n^{-1/2},\ n^{1/2-\xi},\ n^{-\xi/2}\right\} \le 2n^{-1/2} + 2n^{1/2-\xi}$, the desired result follows. □

Lemma A.2.5. *Under Assumption 1, both* $\max_t \|\hat{\Sigma}_t - \Sigma_t\|$ *and* $\max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\|$ *are* $O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right]\log^{O(1)} n\right)$.

Proof. First notice that

$$\begin{aligned}&\max_t \|\hat{\Sigma}_t - \Sigma_t\| \\ &\le \max_t \|n^{-1}(\hat{v}_t'\hat{v}_t - v_t'v_t)\| + \max_t \|n^{-1}v_t'v_t - \Sigma_t\| \\ &\overset{(i)}{\le} \max_t \|n^{-1}(\hat{v}_t - v_t)'\hat{v}_t\| + \max_t \|n^{-1}v_t'(\hat{v}_t - v_t)\| + O_P(n^{-1/2}\log^{O(1)} n)\end{aligned}$$

$$\leq n^{-1} \max_t \|\hat{v}_t - v_t\| \max_t \|\hat{v}_t\| + n^{-1} \max_t \|v_t\| \max_t \|\hat{v}_t - v_t\| + O_P(n^{-1/2} \log^{O(1)} n), \tag{A.2.6}$$

where (i) holds by Lemma A.2.3(8). By Lemma A.2.4, we have that

$$\max_t \|\hat{v}_t - v_t\| \leq n^{1/2} \max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| = O_P\left(\left[1 + n^{1-\xi}\right] \log^{O(1)} n\right). \tag{A.2.7}$$

By Lemma A.2.1(2), $\max_t \|v_t\| \leq \|v\| = O_P(n^{1/2} \log^{O(1)} n)$. Since $\xi > 1/2$ (Assumption 1), it follows that

$$\max_t \|\hat{v}_t\| \leq \max_t \|v_t\| + \max_t \|\hat{v}_t - v_t\| = O_P\left(n^{1/2} \log^{O(1)} n\right). \tag{A.2.8}$$

Now we combine (A.2.6) with (A.2.7) and (A.2.8) and obtain

$$\max_t \|\hat{\Sigma}_t - \Sigma_t\| = O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right] \log^{O(1)} n\right) = o_P(1).$$

Notice that $\|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\| = \|\hat{\Sigma}_t^{-1}(\Sigma_t - \hat{\Sigma}_t)\Sigma_t^{-1}\| \leq \|\hat{\Sigma}_t^{-1}\| \|\Sigma_t - \hat{\Sigma}_t\| \|\Sigma_t^{-1}\| = \|\Sigma_t - \hat{\Sigma}_t\| / (s_{\min}(\hat{\Sigma}_t) s_{\min}(\Sigma_t))$. By Lemma A.3.10(1), $s_k(\hat{\Sigma}_t) + s_1(\Sigma_t - \hat{\Sigma}_t) \geq s_k(\Sigma_t)$. It follows that

$$\max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\| \leq \frac{\max_t \|\hat{\Sigma}_t - \Sigma_t\|}{\min_t s_{\min}(\Sigma_t) \left(\min_t s_{\min}(\Sigma_t) - \max_t \|\hat{\Sigma}_t - \Sigma_t\|\right)} = O_P(1) \max_t \|\hat{\Sigma}_t - \Sigma_t\|.$$

The proof is complete. □

Proof of Theorem 1.3.1. Let $X = U_X S_X V_X'$ be an SVD, where $U_X \in \mathbb{R}^{n \times n}$ and $V_X \in \mathbb{R}^{kT \times kT}$ are orthogonal matrices and $S_X = \begin{bmatrix} S_{X,1} & 0 \\ 0 & S_{X,2} \end{bmatrix} \in \mathbb{R}^{n \times kT}$ with $S_{X,1} \in \mathbb{R}^{r_Q \times r_Q}$. By the definition of $\hat{Q}$ and $\hat{v}$, we have

$$\hat{Q} = U_X \begin{bmatrix} S_{X,1} & 0 \\ 0 & 0 \end{bmatrix} V_X' \qquad \text{and} \qquad \hat{v} = U_X \begin{bmatrix} 0 & 0 \\ 0 & S_{X,2} \end{bmatrix} V_X'.$$

Therefore, $\hat{v}'\hat{Q} = 0$. This means that $\hat{v}_t'\hat{Q}_t = 0 \ \forall 1 \leq t \leq T$. Since $Y_t = \alpha_t + X_t\beta_t + u_t$ and $X_t = \hat{Q}_t + \hat{v}_t$, it follows that

$$\begin{aligned}\hat{\beta}_t - \beta_t = (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'Y_t - \beta_t &= (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'(\alpha_t + u_t) + (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'\hat{Q}_t \\ &\overset{(i)}{=} (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'(\alpha_t + u_t) = n^{-1}\hat{\Sigma}_t^{-1}\hat{v}_t'(\alpha_t + u_t), \quad \text{(A.2.9)}\end{aligned}$$

where (i) holds by $\hat{v}_t'\hat{Q}_t = 0$. By Lemma A.2.3(7) and (10), $\max_t \|v_t'u_t\| = O_P(n^{1/2}\log^{O(1)} n)$ and $\max_t \|v_t'\alpha_t\| = O_P(n^{1/2}\log^{O(1)} n)$. Hence,

$$\max_t \|v_t'(\alpha_t + u_t)\| \leq \max_t \|v_t'\alpha_t\| + \max_t \|v_t'u_t\| = O_P(n^{1/2}\log^{O(1)} n). \quad \text{(A.2.10)}$$

Notice that

$$\begin{aligned}\max_t \|(\hat{v}_t - v_t)'(\alpha_t + u_t)\| &\leq \max_t \|\hat{v}_t - v_t\| \max_t \|\alpha_t + u_t\| \\ &\leq n^{1/2}\max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| \max_t \|\alpha_t + u_t\| \\ &\leq n^{1/2}\max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| \left[\max_t \|L_\alpha F_{\alpha,t}\| + \max_t \|u_t\|\right] \\ &\overset{(i)}{\leq} n^{1/2}\max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| O_P(\sqrt{n}\log^{O(1)} n) \\ &\overset{(ii)}{=} n^{1/2} O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right]\log^{O(1)} n\right) O_P(\sqrt{n}\log^{O(1)} n) \\ &= O_P\left(\left[\sqrt{n} + n^{3/2-\xi}\right]\log^{O(1)} n\right), \quad \text{(A.2.11)}\end{aligned}$$

where (i) follows by $\max_t \|L_\alpha F_{\alpha,t}\| \leq \sqrt{nr_\alpha}\|L_\alpha\|_\infty \|F_\alpha\|_\infty = O_P(\sqrt{n}\log^{O(1)} n)$ (due to Lemma A.2.1) and $\max_t \|u_t\| \leq \|u\| = O_P(\sqrt{n}\log^{O(1)} n)$ (due to Lemma A.2.1(2)) and (ii) follows by Lemma A.2.4. Therefore, we obtain that

$$\begin{aligned}&\max_t \|\hat{\beta}_t - \beta_t\| \\ \overset{(i)}{=} \ & n^{-1}\max_t \left\|\hat{\Sigma}_t^{-1}\left[v_t'(\alpha_t + u_t) + (\hat{v}_t - v_t)'(\alpha_t + u_t)\right]\right\| \\ \leq \ & n^{-1}\max_t \|\hat{\Sigma}_t^{-1}\| \left(\max_t \|v_t'(\alpha_t + u_t)\| + \max_t \|(\hat{v}_t - v_t)'(\alpha_t + u_t)\|\right) \\ \leq \ & n^{-1}\left(\max_t \|\Sigma_t^{-1}\| + \max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\|\right)\end{aligned}$$

$$\times \left(\max_t \|v_t'(\alpha_t + u_t)\| + \max_t \|(\hat{v}_t - v_t)'(\alpha_t + u_t)\| \right)$$
$$\overset{(ii)}{=} O_P\left(\left[n^{-1/2} + n^{1/2-\xi} \right] \log^{O(1)} n \right),$$

where (i) holds by (A.2.9) and (ii) holds by (A.2.10), (A.2.11) and Lemma A.2.5. The desired result follows by $\max_t \|\hat{\beta}_t - \beta_t\|_\infty \leq \max_t \|\hat{\beta}_t - \beta_t\|$. □

A.2.2 Proofs for Theorems 1.3.3, 1.3.4 and 1.3.5 and Corollary 1.3.1

Lemma A.2.6. *Let $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_T) \in \mathbb{R}^{n \times T}$ with $\tilde{u}_t = (\tilde{u}_{1,t}, \ldots, \tilde{u}_{n,t})'$ and $\tilde{u}_{i,t} = u_{i,t} + x_{i.t}'(\beta_t - \hat{\beta}_t)$. Under Assumption 1, we have*

(1) $\max_t \|X_t(\hat{\beta}_t - \beta_t)\| = O_P\left(\left[1 + n^{1-\xi} \right] \log^{O(1)} n \right)$ *and* $\|\tilde{u}\| = O_P\left([n^{\xi/2} + n^{1-\xi/2}] \log^{O(1)} n \right)$.

(2) $\max_t \|v_t'(\tilde{u} - u)F_\alpha\| = O_P\left([n + n^{\xi}] \log^{O(1)} n \right)$.

(3) $\max_t \|u_t'(\hat{v}_t - v_t)\| = O_P\left([1 + n^{1-\xi}] \log^{O(1)} n \right)$.

Proof. **Proof for part (1).** Notice that

$$\max_t \|X_t(\hat{\beta}_t - \beta_t)\| \leq \sqrt{n} \max_{i,t} |x_{i,t}'(\hat{\beta}_t - \beta_t)|$$
$$\leq \sqrt{n} \max_{i,t} \|x_{i,t}\| \max_t \|\hat{\beta}_t - \beta_t\| \overset{(i)}{=} O_P\left(\left[1 + n^{1-\xi} \right] \log^{O(1)} n \right), \quad \text{(A.2.12)}$$

where (i) holds by Lemma A.2.1 and Theorem 1.3.1. The definition of $\tilde{u}$ implies that

$$\|\tilde{u}\| \leq \|\tilde{u} - u\| + \|u\| \leq \sqrt{T} \max_t \|X_t(\hat{\beta}_t - \beta_t)\| + \|u\|$$
$$\overset{(i)}{=} O_P\left([n^{\xi/2} + n^{1-\xi/2} + n^{1/2}] \log^{O(1)} n \right),$$

where (i) follows by (A.2.12), Lemma A.2.1(2) and $T \asymp n^\xi$.

Notice that $n^{\xi/2} + n^{1-\xi/2} \geq 2\sqrt{n^{\xi/2} n^{1-\xi/2}} = 2n^{1/2} > n^{1/2}$. Thus,

$\max\{n^{\xi/2},\ n^{1-\xi/2},\ n^{1/2}\} \le n^{\xi/2} + n^{1-\xi/2}$. Thus, the stated bound for $\|\tilde{u}\|$ follows.

Proof for part (2). The definition of $\tilde{u}$ implies that

$$
\begin{aligned}
& \max_t \|v_t'(\tilde{u}-u)F_\alpha\| \qquad\qquad (A.2.13)\\
= \ & \max_t \left\| \sum_{s=1}^{T} (\sum_{i=1}^{n} v_{i,t}x_{i,s}')(\hat{\beta}_s - \beta_s)F_{\alpha,s}' \right\| \\
\le \ & \max_t \sum_{s=1}^{T} \left\| (\sum_{i=1}^{n} v_{i,t}x_{i,s}')(\hat{\beta}_s - \beta_s)F_{\alpha,s}' \right\| \\
\le \ & \left[\max_t \sum_{s=1}^{T} \left\| \sum_{i=1}^{n} v_{i,t}x_{i,s}' \right\| \right] \max_s \left\| (\hat{\beta}_s - \beta_s)F_{\alpha,s}' \right\| \\
\le \ & \left[\max_t \sum_{s=1}^{T} \left\| \sum_{i=1}^{n} (v_{i,t}x_{i,s}' - \mathbb{E}v_{i,t}x_{i,s}') \right\| + \max_t \sum_{s=1}^{T} \left\| \sum_{i=1}^{n} \mathbb{E}v_{i,t}x_{i,s}' \right\| \right] \\
& \times \max_s \left\| (\hat{\beta}_s - \beta_s)F_{\alpha,s}' \right\|.
\end{aligned}
$$

By Lemma A.3.3(3) and (4) (applied entry-wise), there exist constants $b, \gamma > 0$ such that $\forall i, t, s$, each entry of $v_{i,t}x_{i,s}' - \mathbb{E}v_{i,t}x_{i,s}'$ has an exponential-type tail with parameter (b, γ). Hence,

$$
\begin{aligned}
\max_t \sum_{s=1}^{T} \left\| \sum_{i=1}^{n} (v_{i,t}x_{i,s}' - \mathbb{E}v_{i,t}x_{i,s}') \right\| &\le T \max_{s,t} \left\| \sum_{i=1}^{n} (v_{i,t}x_{i,s}' - \mathbb{E}v_{i,t}x_{i,s}') \right\| \\
&\overset{(i)}{=} O_P(n^{1/2+\xi} \log^{O(1)} n), \qquad (A.2.14)
\end{aligned}
$$

where (i) follows by Lemma A.2.3(9) and $T \asymp n^{\xi}$. Notice that

$$
\begin{aligned}
\max_t \sum_{s=1}^{T} \left\| \sum_{i=1}^{n} \mathbb{E}v_{i,t}x_{i,s}' \right\| &\le \max_t \sum_{i=1}^{n} \sum_{s=1}^{T} \left\| \mathbb{E}v_{i,t}x_{i,s}' \right\| \\
&\overset{(i)}{\le} \max_t \sum_{i=1}^{n} \sum_{s=1}^{T} \left\| \mathbb{E}v_{i,t}v_{i,s}' \right\| + \max_t \sum_{i=1}^{n} \sum_{s=1}^{T} \left\| \mathbb{E}v_{i,t}Q_{i,s}' \right\| \\
&\le \max_{i,t} n \sum_{s=1}^{T} \left\| \mathbb{E}v_{i,t}v_{i,s}' \right\| + \max_{i,t} n \sum_{s=1}^{T} \left\| \mathbb{E}v_{i,t}Q_{i,s}' \right\|
\end{aligned}
$$

$$\overset{(ii)}{=} O(n), \tag{A.2.15}$$

where (i) holds by $x_{i,s} = Q_{i,s} + v_{i,s}$ and (ii) follows by $\max_{i,t} \sum_{s=1}^{T} \|\mathbb{E} v_{i,t} v'_{i,s}\| \leq \sqrt{k} \max_{i,t} \sum_{s=1}^{T} \|\mathbb{E} v_{i,t} v'_{i,s}\|_\infty = O(1)$ (due to Lemma A.2.2) and $\mathbb{E} v_{i,t} Q'_{i,s} = 0$ (by the independence between $v_{i,t}$ and $Q_{i,s}$). Also observe that

$$\max_s \left\| (\hat{\beta}_s - \beta_s) F'_{\alpha,s} \right\| \leq \sqrt{r_\alpha} \|F_\alpha\|_\infty \max_s \|\hat{\beta}_s - \beta_s\| \overset{(i)}{\leq} O_P\left([n^{-1/2} + n^{1/2-\xi}] \log^{O(1)} n \right), \tag{A.2.16}$$

where (i) holds by Theorem 1.3.1 and Lemma A.2.1.

Combining (A.2.13) with (A.2.14), (A.2.15) and (A.2.16), we obtain $\max_t \|v'_t(\tilde{u} - u) F_\alpha\| = O_P([n^\xi + n + n^{3/2-\xi}] \log^{O(1)} n)$. Since $3/2 - \xi < 1$ (as $\xi > 6/7$), part (2) follows.

Proof of part (3). Notice that $\max_t \|u'_t(\hat{v}_t - v_t)\| = \max_t \|u'_t(\hat{L}_Q \hat{F}_{Q,t} - L_Q F_{Q,t})\|$. We apply Lemma A.3.11(4) with $L = L_Q$, $F = F_Q$ and $e = v$ (as well as $a = u_t$ in Lemma A.3.11(4)). By Lemmas A.2.1 and A.2.3(3), we have $\|v\| = O_P(\sqrt{n} \log^{O(1)} n)$, $\max_t \|F_{Q,t}\| = O_P(\log^{O(1)} n)$ and $\max_t \|L'_Q v_t\| = O_P(n^{1/2} \log^{O(1)} n)$. Therefore, Lemma A.3.11(4) and $T \asymp n^\xi$ imply that

$$\begin{aligned} \max_t \|u'_t(\hat{L}_Q \hat{F}_{Q,t} - L_Q F_{Q,t})\| \leq\ & O_P(n^{-\xi} \log^{O(1)} n) \max_t \|F'_Q v' u_t\| \\ & + O_P\left([n^{-\xi/2} + n^{-1/2}] \log^{O(1)} n \right) \max_t \|L'_Q u_t\| \\ & + O_P\left([n^{1/2-\xi} + n^{-\xi/2}] \log^{O(1)} n \right) \max_t \|u_t\|. \end{aligned}$$

By Lemmas A.2.3(5) and (9), $\max_t \|L'_Q u_t\| = O_P(n^{1/2} \log^{O(1)} n)$ and $\max_t \|u'_t v F_Q\| = O_P(n \log^{O(1)} n)$. Since $\max_t \|u_t\| \leq \|u\| = O_P(n^{1/2} \log^{O(1)} n)$ (due to Lemma A.2.1(2)), we have

$$\max_t \|u'_t(\hat{L}_Q \hat{F}_{Q,t} - L_Q F_{Q,t})\| = O_P\left([1 + n^{1-\xi} + n^{(1-\xi)/2}] \log^{O(1)} n \right).$$

Since $1 + n^{1-\xi} \geq 2\sqrt{1 \cdot n^{1-\xi}} = 2n^{(1-\xi)/2} > n^{(1-\xi)/2}$, we have $\max\{1,\ n^{1-\xi},\ n^{(1-\xi)/2}\} \leq 1 + n^{1-\xi}$ and thus part (3) follows. □

Lemma A.2.7. *Under Assumption 1,*

(1) $\max_t \|v_t'(\hat{\alpha}_t - \alpha_t)\| = O_P\left(\left[n^{5(1-\xi)/2} + n^{(\xi-1)/2}\right]\log^{O(1)} n\right)$.
(2) $\max_{i,t} |\hat{\alpha}_{i,t} - \alpha_{i,t}| = O_P\left(\left[n^{\xi/2-1} + n^{2-5\xi/2}\right]\log^{O(1)} n\right)$.

Proof. **Proof for part (1).** Notice that $\max_t \|v_t'(\hat{\alpha}_t - \alpha_t)\| = \max_t \|v_t'(\hat{L}_\alpha \hat{F}_{\alpha,t} - L_\alpha F_{\alpha,t})\|$. Recall that $y_{i,t} - x_{i,t}'\hat{\beta}_t = \alpha_{i,t} + \tilde{u}_{i,t}$, where $\tilde{u}_{i,t}$ is defined in Lemma A.2.6. We apply Lemma A.3.11(4) with $L = L_\alpha$, $F = F_\alpha$ and $e = \tilde{u}$ (as well as $a = v_t$ for Lemma A.3.11(4)). By Lemmas A.2.6(1) and A.2.1, we have $\|\tilde{u}\| = O_P([n^{\xi/2} + n^{1-\xi/2}]\log^{O(1)} n)$ and $\max_t \|F_{\alpha,t}\| = O_P(\log^{O(1)} n)$. It follows, by Lemma A.3.11(4), $T \asymp n^\xi$ and straight-forward computations, that

$$\begin{aligned}
&\max_t \|v_t'(\hat{L}_\alpha \hat{F}_{\alpha,t} - L_\alpha F_{\alpha,t})\| \\
\leq\ & O_P\left(n^{-\xi}\log^{O(1)} n\right)\max_t \|F_\alpha'\tilde{u}'v_t\| \\
&+O_P(\log^{O(1)} n)\left[n^{-1}M + n^{-1/2} + n^{1/2-\xi} + n^{\xi/2-1} + n^{1-3\xi/2}\right]\max_t \|L_\alpha' v_t\| \\
&+O_P(\log^{O(1)} n)\left[n^{1-3\xi/2} + n^{\xi/2-1} + n^{2-5\xi/2} + (n^{-1} + n^{-\xi})M\right]\max_t \|v_t\|, \qquad \text{(A.2.17)}
\end{aligned}$$

where $M = \max_t \|L_\alpha'\tilde{u}_t\|$. Since $\tilde{u}_t - u_t = X_t(\beta_t - \hat{\beta}_t)$, we have that

$$M \leq \max_t \|L_\alpha' u_t\| + \|L_\alpha\| \max_t \|X_t(\hat{\beta}_t - \beta_t)\| \overset{(i)}{\leq} O_P\left(\left[n^{1/2} + n^{3/2-\xi}\right]\log^{O(1)} n\right), \qquad \text{(A.2.18)}$$

where (i) holds by $\|L_\alpha\| \leq \sqrt{nr_\alpha}\|L_\alpha\|_\infty$ and Lemmas A.2.3(6), A.2.1(1) and A.2.6(1). Notice that

$$\max_t \|F_\alpha'\tilde{u}'v_t\| \leq \max_t \|F_\alpha' u' v_t\| + \max_t \|F_\alpha'(\tilde{u} - u)'v_t\| \overset{(i)}{=} O_P\left(\left[n + n^\xi\right]\log^{O(1)} n\right), \qquad \text{(A.2.19)}$$

where (i) holds by Lemmas A.2.3(11) and A.2.6(2), together with $T \asymp n^\xi$ and $n^{(1+\xi)/2} \leq n + n^\xi$.

Now we combine (A.2.17) with (A.2.18), (A.2.19), $\max_t \|L_\alpha' v_t\| = O_P(\sqrt{n}\log^{O(1)} n)$ (Lemma A.2.3(4)) and $\max_t \|v_t\| \leq \|v\| = O_P(\sqrt{n}\log^{O(1)} n)$ (Lemma A.2.1(2)). After some tedious computations, this yields

$$\max_t \|v_t'(\hat{L}_\alpha \hat{F}_{\alpha,t} - L_\alpha F_{\alpha,t})\| = O_P\left(\left[1 + a_n^{-1/2} + a_n + a_n^{3/2} + a_n^2 + a_n^{5/2}\right]\log^{O(1)} n\right),$$

where $a_n = n^{1-\xi}$. By Lemma A.3.9, $1+a_n^{-1/2}+a_n+a_n^{3/2}+a_n^2+a_n^{5/2} \leq 6(a_n^{-1/2}+a_n^{5/2})$ for $a_n > 0$. Thus, part (1) follows.

Proof for part (2). The argument is similar to the one in part (1). We apply Lemma A.3.11(4) with $L = L_\alpha$, $F = F_\alpha$ and $e = \tilde{u}$ (as well as $a = \tau_i$ in Lemma A.3.11(4)), where τ_i is the ith column of I_n. Recall, from the proof of part (1), that $\|\tilde{u}\| = O_P([n^{\xi/2} + n^{1-\xi/2}]\log^{O(1)} n)$ and $\max_t \|F_{\alpha,t}\| = O_P(\log^{O(1)} n)$. Notice that $\max_i \|\tau_i\| = 1$ and $\max_i \|L'_\alpha \tau_i\| = O_P(\log^{O(1)} n)$ (due to A.2.1). Thus, Lemma A.3.11(4) and $T \asymp n^\xi$ imply that

$$
\begin{aligned}
&\max_{i,t} |\hat{\alpha}_{i,t} - \alpha_{i,t}| \\
&= \max_{i,t} |\tau_i'(\hat{L}_\alpha \hat{F}_t - L_\alpha F_t)| \\
&\leq O_P(n^{-\xi}\log^{O(1)} n) \max_i \|F'_\alpha \tilde{u}' \tau_i\| \\
&\quad + O_P(\log^{O(1)} n)\Big[n^{-1}M + n^{-1/2} + n^{1/2-\xi} + n^{\xi/2-1} + n^{1-3\xi/2}\Big] \\
&\quad + O_P(\log^{O(1)} n)\Big[n^{1-3\xi/2} + n^{\xi/2-1} + n^{2-5\xi/2} + (n^{-1}+n^{-\xi})M\Big]. \qquad \text{(A.2.20)}
\end{aligned}
$$

Notice that

$$
\begin{aligned}
\max_i \|\tau_i' \tilde{u} F_\alpha\| &\leq \max_i \|\tau_i' u F_\alpha\| + \max_i \|\tau_i'(\tilde{u}-u)F_\alpha\| \\
&\overset{(i)}{=} O_P(T^{1/2}\log^{O(1)} n) + \max_i \left\| \sum_{t=1}^T x'_{i,t}(\hat{\beta}_t - \beta_t) F'_{\alpha,t} \right\| \\
&\leq O_P(T^{1/2}\log^{O(1)} n) + T \max_{i,t} \|x_{i,t}\| \cdot \max_t \|\hat{\beta}_t - \beta_t\| \cdot \sqrt{r_\alpha}\|F_\alpha\|_\infty \\
&\overset{(ii)}{=} O_P\left([n^{\xi/2} + n^{1/2} + n^{\xi-1/2}]\log^{O(1)} n\right), \qquad \text{(A.2.21)}
\end{aligned}
$$

where (i) follows by Lemma A.2.3(2) and (ii) follows by $T \asymp n^\xi$, Lemma A.2.1(1) and Theorem 1.3.1. Hence, We combine (A.2.20) with (A.2.21) and (A.2.18). After straight-forward (but tedious) computations, this yields

$$
\max_{i,t} |\hat{\alpha}_{i,t} - \alpha_{i,t}| = O_P\left(n^{-1/2}[1 + a_n^{-1/2} + a_n^{1/2} + a_n + a_n^{3/2} + a_n^2 + a_n^{5/2}]\log^{O(1)} n\right),
$$

where $a_n = n^{1-\xi}$. By Lemma A.3.9, $1 + a_n^{-1/2} + a_n^{1/2} + a_n + a_n^{3/2} + a_n^2 + a_n^{5/2} \leq$

$7(a_n^{-1/2} + a_n^{5/2})$ for $a_n > 0$. Thus, part (2) follows. The proof is complete. □

Lemma A.2.8. *Let $\hat{\bar{v}}_{i,t}$, $\bar{v}_{i,t}$ and $D_{n,t}$ be defined in* (A.2.1)*. Suppose that Assumption 1 holds. Then*
(1) $\max_t \|\hat{\Sigma}_t^{-1}\| = O_P(1)$ *and* $\max_{i,t} \|\hat{\bar{v}}_{i,t} - \bar{v}_{i,t}\| = O_P\left([n^{-1/2} + n^{1/2-\xi}] \log^{O(1)} n\right)$,
(2) $\|D_n\|_\infty = \max_t \|D_{n,t}\|_\infty = O_P\left([n^{\xi/2-1} + n^{3-7\xi/2}] \log^{O(1)} n\right)$.

Proof. **Proof for part (1).** By Lemma A.2.5,

$$\begin{aligned}\max_t \|\hat{\Sigma}_t^{-1}\| &\le \max_t \|\Sigma_t\| + \max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\| \\ &= O(1) + O_P\left(\left[n^{-1/2} + n^{1/2-\xi}\right] \log^{O(1)} n\right) \overset{(i)}{=} O_P(1), \end{aligned} \tag{A.2.22}$$

where (i) holds by $\xi > 1/2$. Notice that

$$\begin{aligned}&\max_{i,t} \|\hat{\Sigma}_t^{-1}\hat{v}_{i,t} - \Sigma_t^{-1}v_{i,t}\| \\ &\le \max_{i,t} \|\hat{\Sigma}_t^{-1}(\hat{v}_{i,t} - v_{i,t})\| + \max_{i,t} \|(\hat{\Sigma}_t^{-1} - \Sigma_t^{-1})v_{i,t}\| \\ &\le \max_t \|\hat{\Sigma}_t^{-1}\| \max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| + \max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\| \max_{i,t} \|v_{i,t}\| \\ &\overset{(i)}{=} O_P\left([n^{-1/2} + n^{1/2-\xi}] \log^{O(1)} n\right), \end{aligned}$$

where (i) holds by the bounds for $\max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\|$ and for $\max_t \|\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\|$ (Lemmas A.2.4 and A.2.5), together with $\max_{i,t} \|v_{i,t}\| \le \sqrt{k}\|v\|_\infty = O_P(\log^{O(1)} n)$ (Lemma A.2.1). This proves part (1).

Proof for part (2). By the definition of $D_{n,t}$ in (A.2.1), we have the following decomposition

$$D_{n,t} = n^{-1/2}\left(\hat{\Sigma}_t^{-1}\hat{v}_t' - \Sigma_t^{-1}v_t'\right)u_t + n^{-1/2}\hat{\Sigma}_t^{-1}v_t'(\alpha_t - \hat{\alpha}_t) + n^{-1/2}\hat{\Sigma}_t^{-1}(\hat{v}_t - v_t)'(\alpha_t - \hat{\alpha}_t). \tag{A.2.23}$$

Now we derive bounds for each of these three terms. Let $a_n = n^{1-\xi}$. For the first term, notice that

$$\begin{aligned}&\max_t \left\| n^{-1/2}\left(\hat{\Sigma}_t^{-1}\hat{v}_t' - \Sigma_t^{-1}v_t'\right)u_t \right\| \\ = \ &\max_t \left\| n^{-1/2}\hat{\Sigma}_t^{-1}(\hat{v}_t - v_t)'u_t \right\| + \max_t \left\| n^{-1/2}\left(\hat{\Sigma}_t^{-1} - \Sigma_t^{-1}\right)v_t'u_t \right\| \end{aligned}$$

$$\begin{aligned} &\leq \max_t \left\| n^{-1/2}\hat{\Sigma}_t^{-1} \right\| \max_t \|(\hat{v}_t - v_t)' u_t\| + \max_t \left\| n^{-1/2} \left(\hat{\Sigma}_t^{-1} - \Sigma_t^{-1} \right) v_t' u_t \right\| \\ &\overset{(i)}{=} O_P\left(n^{-1/2}[1 + a_n] \log^{O(1)} n \right) + \max_t \left\| n^{-1/2} \left(\hat{\Sigma}_t^{-1} - \Sigma_t^{-1} \right) v_t' u_t \right\|, \quad \text{(A.2.24)} \end{aligned}$$

where (i) follows by (A.2.22) and Lemma A.2.6(3). Also notice that

$$\begin{aligned} \max_t \left\| n^{-1/2} \left(\hat{\Sigma}_t^{-1} - \Sigma_t^{-1} \right) v_t' u_t \right\| &\leq n^{-1/2} \max_t \left\| \hat{\Sigma}_t^{-1} - \Sigma_t^{-1} \right\| \max_t \|v_t' u_t\| \\ &\overset{(i)}{=} O_P\left(n^{-1/2}[1 + a_n] \log^{O(1)} n \right), \quad \text{(A.2.25)} \end{aligned}$$

where (i) follows by Lemmas A.2.3(7) and A.2.5. We combine (A.2.24) and (A.2.25) and obtain that

$$\max_t \left\| n^{-1/2} \left(\hat{\Sigma}_t^{-1} \hat{v}_t' - \Sigma_t^{-1} v_t' \right) u_t \right\| = O_P\left(n^{-1/2}[1 + a_n] \log^{O(1)} n \right). \quad \text{(A.2.26)}$$

To bound the second term in (A.2.23), observe that

$$\begin{aligned} \max_t \|n^{-1/2}\hat{\Sigma}_t^{-1} v_t'(\alpha_t - \hat{\alpha}_t)\| &\leq n^{-1/2} \max_t \|\hat{\Sigma}_t^{-1}\| \max_t \|v_t'(\hat{\alpha}_t - \alpha_t)\| \\ &\overset{(i)}{=} O_P\left([n^{2-5\xi/2} + n^{\xi/2-1}] \log^{O(1)} n \right) \\ &= O_P\left(n^{-1/2}[a_n^{5/2} + a_n^{-1/2}] \log^{O(1)} n \right), \quad \text{(A.2.27)} \end{aligned}$$

where (i) holds by (A.2.22) and Lemma A.2.7. To bound the third term in (A.2.23), we have that

$$\begin{aligned} &\max_t \|n^{-1/2}\hat{\Sigma}_t^{-1}(\hat{v}_t - v_t)'(\alpha_t - \hat{\alpha}_t)\| \\ &\leq n^{-1/2} \max_t \|\hat{\Sigma}_t^{-1}\| \max_t \|\hat{v}_t - v_t\| \max_t \|\hat{\alpha}_t - \alpha_t\| \\ &\leq n^{-1/2} \max_t \|\hat{\Sigma}_t^{-1}\| n^{1/2} \max_{i,t} \|\hat{v}_{i,t} - v_{i,t}\| n^{1/2} \max_{i,t} |\hat{\alpha}_{i,t} - \alpha_{i,t}| \\ &\overset{(i)}{=} O_P\left(n^{-1/2}[a_n^{-1/2} + a_n^{1/2} + a_n^{5/2} + a_n^{7/2}] \log^{O(1)} n \right), \quad \text{(A.2.28)} \end{aligned}$$

where (i) holds by (A.2.22) and Lemmas A.2.4 and A.2.7(2), together with $a_n = n^{1-\xi}$.

Now we combine (A.2.23) with (A.2.26), (A.2.27) and (A.2.28) and obtain

$$\max_t \|D_{n,t}\|_\infty = O_P\left(n^{-1/2}[1 + a_n^{-1/2} + a_n + a_n^{1/2} + a_n^{5/2} + a_n^{7/2}]\log^{O(1)} n\right).$$

By Lemma A.3.9, $1 + a_n^{-1/2} + a_n + a_n^{1/2} + a_n^{5/2} + a_n^{7/2} \le 6(a_n^{-1/2} + a_n^{7/2})$. Part (2) follows. □

Lemma A.2.9. *Recall $\hat{u}_{i,t}$ defined in Algorithm 2 and $\hat{\bar{v}}_{i,t}$ and $\bar{v}_{i,t}$ defined in* (A.2.1). *Under Assumption 1,* $\max_{1\le i\le n, 1\le t\le T} \|\hat{\bar{v}}_{i,t}\hat{u}_{i,t} - \bar{v}_{i,t}u_{i,t}\| = O_P\left([n^{2-5\xi/2} + n^{\xi/2-1}]\log^{O(1)} n\right)$.

Proof. Let $a_n = n^{1-\xi}$. Notice that

$$\begin{aligned}
\max_{i,t}|\hat{u}_{i,t} - u_{i,t}| &= \max_{i,t}|y_{i,t} - x_{i,t}'\hat{\beta}_t - \hat{\alpha}_{i,t} - u_{i,t}| \\
&\overset{(i)}{=} \max_{i,t}|\alpha_{i,t} + \tilde{u}_{i,t} - \hat{\alpha}_{i,t} - u_{i,t}| \\
&\le \max_{i,t}|\tilde{u}_{i,t} - u_{i,t}| + \max_{i,t}|\hat{\alpha}_{i,t} - \alpha_{i,t}| \\
&\le \max_{i,t}\|x_{i,t}\| \max\|\hat{\beta}_t - \beta_t\| + \max_{i,t}|\hat{\alpha}_{i,t} - \alpha_{i,t}| \\
&\overset{(ii)}{=} O_P\left(n^{-1/2}[1 + a_n + a_n^{-1/2} + a_n^{5/2}]\log^{O(1)} n\right) \\
&\overset{(iii)}{=} O_P\left(n^{-1/2}[a_n^{-1/2} + a_n^{5/2}]\log^{O(1)} n\right), \qquad \text{(A.2.29)}
\end{aligned}$$

where (i) holds by $y_{i,t} = \alpha_{i,t} + x_{i,t}'\beta_t + u_{i,t}$ and the definition of $\tilde{u}_{i,t}$ in Lemma A.2.6, (ii) holds by Lemma A.2.1, Theorem 1.3.1 and Lemma A.2.7(2) and (iii) holds by noticing that $1 + a_n + a_n^{-1/2} + a_n^{5/2} \le 4(a_n^{-1/2} + a_n^{5/2})$ (due to Lemma A.3.9). Notice that

$$\begin{aligned}
&\max_{i,t}\|\hat{\bar{v}}_{i,t}\hat{u}_{i,t} - \bar{v}_{i,t}u_{i,t}\| \\
\le&\max_{i,t}\|\hat{\bar{v}}_{i,t} - \bar{v}_{i,t}\| \max_{i,t}|\hat{u}_{i,t}| + \max_{i,t}\|\bar{v}_{i,t}\| \max_{i,t}|\hat{u}_{i,t} - u_{i,t}| \\
\le&\max_{i,t}\|\hat{\bar{v}}_{i,t} - \bar{v}_{i,t}\|\left(\|u\|_\infty + \max_{i,t}|\hat{u}_{i,t} - u_{i,t}|\right) + \max_{i,t}\|\bar{v}_{i,t}\| \max_{i,t}|\hat{u}_{i,t} - u_{i,t}| \\
\overset{(i)}{=}&O_P\left(\left\{n^{-1/2}[1 + a_n + a_n^{5/2} + a_n^{-1/2}] + n^{-1}[a_n^{-1/2} + a_n^{1/2} + a_n^{5/2} + a_n^{7/2}]\right\}\log^{O(1)} n\right) \\
\overset{(ii)}{=}&O_P\left(\left\{n^{-1/2}[a_n^{5/2} + a_n^{-1/2}] + n^{-1}[a_n^{-1/2} + a_n^{7/2}]\right\}\log^{O(1)} n\right),
\end{aligned}$$

where (i) follows by (A.2.29) and Lemmas A.2.8(1) and A.2.1 and (ii) follows by $1+a_n+a_n^{5/2}+a_n^{-1/2} \leq 4(a_n^{-1/2}+a_n^{5/2})$ and $a_n^{-1/2}+a_n^{1/2}+a_n^{5/2}+a_n^{7/2} \leq 4(a_n^{-1/2}+a_n^{7/2})$ (due to Lemma A.3.9). Plugging in $a_n = n^{1-\xi}$, we obtain

$$\begin{aligned}\max_{i,t} \|\hat{\bar{v}}_{i,t}\hat{u}_{i,t} - \bar{v}_{i,t}u_{i,t}\| &= O_P\left([n^{2-5\xi/2} + n^{\xi/2-1} + n^{\xi/2-3/2} + n^{5/2-7\xi/2}]\log^{O(1)} n\right) \\ &\overset{(i)}{=} O_P\left([n^{2-5\xi/2} + n^{\xi/2-1}]\log^{O(1)} n\right),\end{aligned}$$

where (i) holds by $n^{\xi/2-3/2} = o(n^{\xi/2-1/2})$ and $n^{5/2-7\xi/2} = o(n^{2-5\xi/2})$ (since $\xi \in (6/7, 2)$). □

Lemma A.2.10. *Recall Ω and $\hat{\Omega}$ defined in* (A.2.1)*. Let Assumptions 1 and 2 hold. Then*

$$\sup_{x\in\mathbb{R}} \left|\Phi(x, \hat{\Omega}) - \Phi(x, \Omega)\right| = o_P(1).$$

Proof. **Step 1: derive the exponential-type tails for $(J_j'G_i)(J_k'G_i)$.** By Assumption 1, there is a constant $M > 0$ such that $\forall t \in [T]$, each row of Σ_t^{-1} is bounded (in $\|\cdot\|_1$) by M. It follows, by Lemma A.3.3(1) and Assumption 1, that there exist constants $b > 0$ depending only on M, k and γ_* such that $\forall (i,t) \in [n] \times [T]$, each entry of $\bar{v}_{i,t}$ has an exponential-type tail with parameter (b, γ_*). By Lemma A.3.3(3), there exists $b_G > 0$ such that $\forall (i,t) \in [n] \times [T]$, each entry of $G_{i,t} = \bar{v}_{i,t}u_{i,t}$ has an exponential-type tails with parameters $(b_G, \gamma_*/2)$.

By Assumption 2 and Lemma A.3.3(1), $J_j'G_i$ has an exponential-type tail with parameter $(c_n, \gamma_*/2)$, where $c_n = b_G A_1 \log^{2/\gamma_*}(m_J + 2)$. Then Lemma A.3.3(3) implies that for $j, k \in [m_J]$, $(J_j'G_i)(J_k'G_i)$ has an exponential-type tail with parameter $(C_n, \gamma_*/4)$, where $C_n = 2^{4/\gamma_*}c_n^2$. Hence, $(J_j'G_i)(J_k'G_i)C_n^{-1}$ has an exponential-type tail with parameter $(1, \gamma_*/4)$.

Step 2: show the desired result by bounding $\|\hat{\Omega} - \Omega\|_\infty$. Since $\{(J_j'G_i)(J_k'G_i)\}_{i=1}^n$ is independent across i, it follows, by Lemma A.3.6, that

$$\max_{1\leq j,k\leq m_J} \left| n^{-1}\sum_{i=1}^n \left[(J_j'G_i)(J_k'G_i) - \mathbb{E}(J_j'G_i)(J_k'G_i)\right]/C_n \right| = O_P\left(\sqrt{n^{-1}\log(m_J^2)}\right).$$

Let $\tilde{\Omega} = n^{-1}\sum_{i=1}^n G_i G_i'$. The above display implies that

$$\begin{aligned}\|\tilde{\Omega} - \Omega\|_\infty &= \max_{1\le j,k\le m_J}\left|n^{-1}\sum_{i=1}^n\left[(J_j'G_i)(J_k'G_i) - \mathbb{E}(J_j'G_i)(J_k'G_i)\right]\right| \\ &= C_n O_P\left(\sqrt{n^{-1}\log(m_J^2)}\right) \overset{(i)}{=} O_P\left(n^{-1/2}\log^{O(1)} n\right), \quad \text{(A.2.30)}\end{aligned}$$

where (i) holds by the definition of C_n. Notice that

$$\begin{aligned}&\left\|n^{-1}\sum_{i=1}^n(\hat{G}_i\hat{G}_i - G_iG_i')\right\|_\infty \\ \le\ & \max_i\|\hat{G}_i\hat{G}_i' - G_iG_i'\|_\infty \\ \le\ & \max_i\left(\|\hat{G}_i\|_\infty\|\hat{G}_i - G_i\|_\infty + \|G_i\|_\infty\|\hat{G}_i - G_i\|_\infty\right) \\ \overset{(i)}{\le}\ & \max_i\left(2\|G_i\|_\infty\|\hat{G}_i - G_i\|_\infty + \|\hat{G}_i - G_i\|_\infty^2\right) \\ \le\ & \left(2\max_{i,t}\|\bar{v}_{i,t}u_{i,t}\|\max_{i,t}\|\hat{\bar{v}}_{i,t}\hat{u}_{i,t} - \bar{v}_{i,t}u_{i,t}\|_\infty + \max_{i,t}\|\hat{\bar{v}}_{i,t}\hat{u}_{i,t} - \bar{v}_{i,t}u_{i,t}\|_\infty^2\right) \\ \overset{(ii)}{\le}\ & O_P\left([n^{2-5\xi/2} + n^{\xi/2-1}]\log^{O(1)} n\right), \quad \text{(A.2.31)}\end{aligned}$$

where (i) holds by $\|\hat{G}_i\|_\infty \le \|G_i\|_\infty + \|\hat{G}_i - G_i\|_\infty$ and (ii) follows by Lemmas A.2.1(1) and A.2.9. Therefore,

$$\begin{aligned}&\|\hat{\Omega} - \tilde{\Omega}\|_\infty = \max_{1\le j,k\le m_J}\left|J_j'\left[n^{-1}\sum_{i=1}^n(\hat{G}_i\hat{G}_i - G_iG_i')\right]J_k\right| \\ &\overset{(i)}{\le}\max_{1\le j\le m_J}\|J_j\|_1^2\left\|n^{-1}\sum_{i=1}^n(\hat{G}_i\hat{G}_i - G_iG_i')\right\|_\infty \overset{(ii)}{=} O_P\left([n^{2-5\xi/2} + n^{\xi/2-1}]\log^{O(1)} n\right), \quad \text{(A.2.32)}\end{aligned}$$

where (i) follows by Holder's inequality and (ii) follows by $\max_{1\le j\le m_J}\|J_j\|_1^2 \le A_1$ and (A.2.31). We combine (A.2.30) and (A.2.32) and obtain

$$\|\hat{\Omega} - \Omega\|_\infty = O_P\left([n^{2-5\xi/2} + n^{\xi/2-1} + n^{-1/2}]\log^{O(1)} n\right). \quad \text{(A.2.33)}$$

By Assumption 2, the diagonal entries of Ω are bounded away from zero

and infinity. Therefore, Lemma A.3.5 implies that

$$\sup_{x\in\mathbb{R}}\left|\Phi(x,\hat{\Omega})-\Phi(x,\Omega)\right|\leq M\Delta^{1/3}\left(1\vee\log(2m_J/\Delta)\right)^{2/3}, \tag{A.2.34}$$

where $M>0$ is a constant and $\Delta=\|\hat{\Omega}-\Omega\|_\infty$. The desired result follows by (A.2.34), together with (A.2.33) and $T\asymp n^\xi$ with $\xi\in(6/7,2)$. □

Lemma A.2.11. *Recall Ω defined in* (A.2.1)*. Let Assumptions 1 and 2 hold. Then*

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|n^{-1/2}\sum_{i=1}^n JG_i\right\|_\infty\leq x\right)-\Phi(x,\Omega)\right|=0.$$

Proof. For $j\in[m_J]$ and $i\in[n]$, define $W_{i,j}=J_j'G_i$ and denote $W_i=JG_i=(W_{i,1},\ldots,W_{i,m_J})'\in\mathbb{R}^{m_J}$. By Assumption 2, $\min_{1\leq j\leq m_R}n^{-1}\sum_{i=1}^n\mathbb{E}W_{i,j}^2\geq c_1$.

As argued at the beginning of the proof of Lemma A.2.10, $W_{i,j}$ has an exponential-type tail with parameter (d_n,γ_1), where $d_n=c_0A_1\log^{1/\gamma_1}(m_J+2)$, $\gamma_1=\gamma_*/2$ and $c_0>0$ is a constant. Define $B_n=C_1n^{l/q}d_n$, where $q=4(l+1)$ and $C_1>0$ is a constant to be chosen later. Then by Lemma A.3.3(2), we have

$$\begin{cases}n^{-1}\sum_{i=1}^n\mathbb{E}|W_{i,j}|^3/B_n\leq C_{\gamma_1,3}d_n^3B_n^{-1}=O\left(n^{-l/(4l+4)}\log^{O(1)}n\right)=o(1)\\ n^{-1}\sum_{i=1}^n\mathbb{E}W_{i,j}^4/B_n^2\leq C_{\gamma_1,4}d_n^4B_n^{-2}=O\left(n^{-l/(2l+2)}\log^{O(1)}n\right)=o(1)\\ \max_{1\leq i\leq n}\mathbb{E}\max_{1\leq j\leq m_J}|W_{i,j}/B_n|^q\leq C_{\gamma_1,q}m_Jd_n^qB_n^{-q}=O(1),\end{cases}$$

where $C_{\gamma_1,3}$, $C_{\gamma_1,4}$ and $C_{\gamma_1,q}$ are constants depending only on γ_1 and q. Therefore, we can choose a constant $C_1>0$ such that

$$\begin{cases}n^{-1}\sum_{i=1}^n\mathbb{E}|W_{i,j}|^3\leq B_n & \forall 1\leq j\leq m_J\\ n^{-1}\sum_{i=1}^n\mathbb{E}W_{i,j}^4\leq B_n^2 & \forall 1\leq j\leq m_J\\ \mathbb{E}\max_{1\leq j\leq m_J}|W_{i,j}/B_n|^q\leq 2 & \forall 1\leq i\leq n.\end{cases}$$

Notice that $\{z\in\mathbb{R}^{m_J}\mid\|z\|_\infty\leq x\}$ is a rectangle in $\mathbb{R}^{m_J}$. It follows, by Proposition 2.1 of Chernozhukov, Chetverikov, and Kato (2014) (applied to

$\{W_i\}_{i=1}^n$), that there exists a constant $C > 0$ depending only on c_1 and q such that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|n^{-1/2}\sum_{i=1}^n JG_i\right\|_\infty \le x\right) - \Phi(x,\Omega)\right| \\
\le\ & C\left\{\left(n^{-1}B_n^2\log^7(m_J n)\right)^{1/6} + \left(n^{2/q-1}B_n^2\log^3(m_J n)\right)^{1/3}\right\} \\
\overset{(i)}{=}\ & O\left(\left[n^{-\frac{l+2}{12(l+1)}} + n^{-\frac{1}{6(l+1)}}\right]\log^{O(1)} n\right) = o(1),
\end{aligned}$$

where (i) holds by the definition of B_n. The proof is complete.

□

Proof of Theorem 1.3.3. Since $\tilde{\beta}_t = \hat{\beta}_t - (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'(\alpha_t + u_t)$, we have that

$$\begin{aligned}
\tilde{\beta}_t - \beta_t &= \hat{\beta}_t - \beta_t - (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'\hat{\alpha}_t \\
&\overset{(i)}{=} (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'(\alpha_t - \hat{\alpha}_t) + (\hat{v}_t'\hat{v}_t)^{-1}\hat{v}_t'u_t \overset{(ii)}{=} n^{-1/2}D_{n,t} + n^{-1}\sum_{i=1}^n G_{i,t},
\end{aligned}$$

where (i) follows by (A.2.9) in the proof of Theorem 1.3.1 and (ii) follows by the definition of $D_{n,t}$ and $G_{i,t}$ in (A.2.1). For the rest of the proof, recall G_i, D_n, Ω and $\hat{\Omega}$ defined in (A.2.1). The above display means that

$$\sqrt{n}J(\tilde{\beta} - \beta) = JD_n + S_n^{JG}, \tag{A.2.35}$$

where $S_n^{JG} = n^{-1/2}\sum_{i=1}^n JG_i$. Define

$$\varepsilon = n^{-\kappa_*/2} \qquad \text{with} \qquad \kappa_* = \min\left\{1 - \frac{\xi}{2},\ \frac{7\xi}{2} - 3\right\}. \tag{A.2.36}$$

Notice that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta} - \beta)\right\|_\infty \le x\right) - \Phi(x,\Omega)\right| \\
\le &\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta} - \beta)\right\|_\infty \le x\right) - \mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \le x\right)\right| \\
&+ \sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \le x\right) - \Phi(x,\Omega)\right|
\end{aligned}$$

$$\begin{aligned}
&\overset{(i)}{\leq}\mathbb{P}\left(\|JD_n\|_\infty > \varepsilon\right) + \sup_{x\in\mathbb{R}} \mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \in (x-\varepsilon, x+\varepsilon]\right) \\
&\quad + \sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \leq x\right) - \Phi(x,\Omega)\right|, \qquad \text{(A.2.37)}
\end{aligned}$$

where (i) follows by (A.2.35) and Lemma A.3.1. Also notice that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \in (x-\varepsilon, x+\varepsilon]\right) - \left[\Phi(x+\varepsilon,\Omega) - \Phi(x-\varepsilon,\Omega)\right]\right| \\
&= \left|\left[\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \leq x+\varepsilon\right) - \Phi(x+\varepsilon,\Omega)\right] - \left[\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \leq x-\varepsilon\right) - \Phi(x-\varepsilon,\Omega)\right]\right| \\
&\leq 2\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \leq t\right) - \Phi(t,\Omega)\right|. \qquad \text{(A.2.38)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta}-\beta)\right\|_\infty \leq x\right) - \Phi(x,\Omega)\right| \\
&\overset{(i)}{\leq}\mathbb{P}\left(\|JD_n\|_\infty > \varepsilon\right) + \sup_{x\in\mathbb{R}} \left[\Phi(x+\varepsilon,\Omega) - \Phi(x-\varepsilon,\Omega)\right] \\
&\quad + 3\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\left\|S_n^{JG}\right\|_\infty \leq x\right) - \Phi(x,\Omega)\right| \\
&\overset{(ii)}{=}\mathbb{P}\left(\|JD_n\|_\infty > \varepsilon\right) + \sup_{x\in\mathbb{R}} \left[\Phi(x+\varepsilon,\Omega) - \Phi(x-\varepsilon,\Omega)\right] + o(1) \\
&\overset{(iii)}{\leq}\mathbb{P}\left(\|JD_n\|_\infty > \varepsilon\right) + C_0\varepsilon\sqrt{\log m_J} + o(1) \\
&\overset{(iv)}{\leq}\mathbb{P}(A_1\|D_n\|_\infty > \varepsilon) + C_0\varepsilon\sqrt{\log m_J} + o(1), \qquad \text{(A.2.39)}
\end{aligned}$$

where $C_0 > 0$ is a constant depending only on the constants in Assumption 2; in the above display, (i) follows by (A.2.37) and (A.2.38), (ii) follows by Lemma A.2.11, (iii) holds by Lemma A.3.4 and (iv) follows by Holder's inequality $\|JD_n\|_\infty \leq \max_{1\leq j\leq m_J}\|J_j\|_1\|D_n\|_\infty$ and Assumption 2.

By Lemma A.2.8 and $T \asymp n^\xi$ with $\xi \in (6/7, 2)$ (Assumption 1), we have $\|D_n\|_\infty = O_P(n^{-\kappa_*})$ and $\kappa_* > 0$. Therefore, (A.2.39) and (A.2.36) imply that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta}-\beta)\right\|_\infty \leq x\right) - \Phi(x,\Omega)\right| \\
&\qquad \leq \mathbb{P}\left(A_1 O_P(n^{-\kappa_*}) > n^{-\kappa_*/2}\right) + C_0 n^{-\kappa_*/2}\sqrt{\log m_J} + o(1) = o(1).
\end{aligned}$$

By Lemma A.2.10, $\sup_{x\in\mathbb{R}}|\Phi(x,\hat{\Omega})-\Phi(x,\Omega)|=o_P(1)$. Hence, the above display implies that

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta}-\beta)\right\|_\infty\le x\right)-\Phi(x,\hat{\Omega})\right|=o_P(1). \tag{A.2.40}$$

Fix an arbitrary constant $\delta>0$. Then Lemma A.3.2 and (A.2.40) imply that

$$\begin{aligned}&\limsup_{n\to\infty}\sup_{\eta\in(0,1)}\left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta}-\beta)\right\|_\infty>\Phi^{-1}(1-\eta,\hat{\Omega})\right)-\eta\right|\\&\qquad\le\delta+\limsup_{n\to\infty}\mathbb{P}\left\{\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\left\|\sqrt{n}J(\tilde{\beta}-\beta)\right\|_\infty\le x\right)-\Phi(x,\hat{\Omega})\right|>\delta\right\}=\delta.\end{aligned}$$

Since $\delta>0$ is arbitrary, the desired result follows. □

Proof of Corollary 1.3.1. Notice that $J\beta\in\mathcal{C}_{1-\eta}(J)$ if and only if $\sqrt{n}\|J(\tilde{\beta}-\beta)\|_\infty\le\Phi^{-1}(1-\eta,\hat{\Omega})$. It follows, by Theorem 1.3.3, that

$$\begin{aligned}&\mathbb{P}\left(\sqrt{n}\|J(\tilde{\beta}-\beta)\|_\infty\le\Phi^{-1}(1-\eta,\hat{\Omega})\right)\\&\qquad=1-\mathbb{P}\left(\sqrt{n}\|J(\tilde{\beta}-\beta)\|_\infty>\Phi^{-1}(1-\eta,\hat{\Omega})\right)\to1-\eta.\end{aligned}$$

The proof is complete. □

Proof of Theorem 1.3.4. If $l=0$, the result clearly holds since $\Phi^{-1}(1-\eta,\hat{\Omega})=O_P(1)$ for $m_J=O(1)$. Now we assume $l>1$ and thus $m_J\to\infty$. Let $\zeta\sim N(0,\Omega)$ with $\zeta=(\zeta_1,\ldots,\zeta_{m_J})'\in\mathbb{R}^{m_J}$. At the beginning of the proof of Lemma A.2.10, we have showed that the entries of $G_{i,t}$ have exponential-type tails. Thus, $\|\Sigma_G\|_\infty\le K_1$ for some constant K_1, where $\Sigma_G=n^{-1}\sum_{i=1}^n\mathbb{E}G_iG_i'$. Since $\Omega_{j,j}=J_j'\Sigma_GJ_j$ and $\|J_j\|_1\le A_1$, Holder's inequality implies that $\Omega_{j,j}\le\|J_j\|_1^2\|\Sigma_G\|_\infty\le A_1^2K_1$.

Since $\zeta_j\sim N(0,\Omega_{j,j})$, there exists a constant $K_2>0$ such that ζ_j has an exponential-type tail with parameter $(K_2,2)$. Then,

$$\begin{aligned}&\Phi\left(2K_2\sqrt{\log m_J},\Omega\right)\\&=1-\mathbb{P}\left(\|\zeta\|_\infty>2K_2\sqrt{\log m_J}\right)\end{aligned}$$

$$\geq 1 - \sum_{j=1}^{m_J} \mathbb{P}\left(|\zeta_j| > 2K_2\sqrt{\log m_J}\right)$$
$$\overset{(i)}{\geq} 1 - m_J \exp\left[1 - (2K_2\sqrt{\log m_J}/K_2)^2\right] = 1 - e/m_J \to 1.$$

where (i) holds by the exponential-type tails of ζ_j. By Lemma A.2.10, $\Phi\left(2K_2\sqrt{\log m_J}, \hat{\Omega}\right) = 1 + o_P(1)$ and thus $2K_2\sqrt{\log m_J} = \Phi^{-1}(1 + o_P(1), \hat{\Omega})$. Since $\Phi^{-1}(1 + o_P(1), \hat{\Omega}) \geq \Phi^{-1}(1 - \eta, \hat{\Omega})$ with probability approaching one, we have that

$$\mathbb{P}\left(\Phi^{-1}(1 - \eta, \hat{\Omega}) \leq 2K_2\sqrt{\log m_J}\right) \to 1.$$

Since $m_J = O(n^l)$ (which means that $\log m_J = O(\log n)$, the desired result follows. □

Proof of Theorem 1.3.5. This is Lemma A.2.7(2). □

Proof of Theorem 1.3.2. The result follows by combining Theorems 1.3.3 and 1.3.4 using $J = I_{kT}$. □

A.2.3 Proof of Theorems 1.3.6, 1.3.7 and 1.4.1

The following result is useful for proving Theorem 1.3.6.

Lemma A.2.12. *Consider matrices $W, e \in \mathbb{R}^{n_1 \times n_2}$. Suppose that $s_{r+1}(W) = 0$ for some $r \geq 1$. Let $\mu > 0$ and define $\hat{r} = \max\{j \mid s_j(W + e) \geq \mu\}$. Then $\hat{r} \neq r$ implies that either $s_r(W) < 2\mu$ or $s_1(e) \geq \mu$.*

Proof. We proceed by contradiction. Suppose that $\hat{r} \neq r$, $s_r(W) \geq 2\mu$ and $s_1(e) < \mu$. We discuss two cases separately: case (A) with $\hat{r} > r$ and case (B) with $\hat{r} < r$.

We first consider case (A). By definition, $s_{\hat{r}}(W + e) \geq \mu$. Since $\hat{r} > r$, we have $\hat{r} \geq r + 1$ and thus $s_{r+1}(W + e) \geq \mu$. By Lemma A.3.10(1), we have $s_{r+1}(W) + s_1(e) \geq s_{r+1}(W + e)$ and thus $s_{r+1}(W) + s_1(e) \geq \mu$. Therefore, the assumption of $s_{r+1}(W) = 0$ implies that $s_1(e) \geq \mu$, contradicting $s_1(e) < \mu$.

We now consider case (B). By definition $s_{\hat{r}+1}(W + e) < \mu$. Since $\hat{r} < r$, we have $\hat{r} + 1 \leq r$ and thus $s_r(W + e) < \mu$. Hence, Lemma A.3.10(1) implies

that $s_r(W) \leq s_r(W+e) + s_1(-e) < \mu + s_1(e)$. Since $s_1(e) < \mu$, we have that $s_r(W) < 2\mu$, contradicting $s_r(W) \geq 2\mu$.

Therefore, it is impossible that these three conditions hold simultaneously: $\hat{r} \neq r$, $s_r(W) \geq 2\mu$ and $s_1(e) < \mu$. Hence, $\hat{r} \neq r$ implies that at least one of the other two conditions does not hold. □

Proof of Theorem 1.3.6. **Proof of part (1).** Since $X = L_Q F_Q' + v$, Lemma A.2.12 implies that it suffices to verify

(1a) $\mathbb{P}[s_{r_Q}(L_Q F_Q') < 2\mu_n] \to 0$;

(1b) $\mathbb{P}[\|v\| \geq \mu_n] \to 0$.

Notice that (1b) follows by Lemma A.2.1 and $\sqrt{n}\log^{O(1)} n = o(\mu_n)$. By Assumption 1, both $s_1(F_Q/\sqrt{T})$ and $s_{r_Q}(L_Q/\sqrt{n})$ are bounded away from zero. It follows, by Lemma A.3.10(2), that there exists a constant $b_0 > 0$ such that $\mathbb{P}[s_{r_Q}(L_Q F_Q'/\sqrt{nT}) > b_0] \to 1$. Since $\sqrt{nT}/\mu_n \to \infty$, condition (1a) follows. We have proved part (1).

Proof of part (2). Recall $\tilde{u}_t = u_t + X_t(\beta_t - \hat{\beta}_t)$ (defined in Lemma A.2.6). Then $y_t - X_t\hat{\beta}_t = \alpha_t + \tilde{u}_t$. By Lemma A.2.12, it suffices to verify

(2a) $\mathbb{P}[s_{r_\alpha}(L_\alpha F_\alpha') < 2\tilde{\mu}_n] \to 0$;

(2b) $\mathbb{P}\left(\|\tilde{u}\| \geq \tilde{\mu}_n\right) \to 0$.

The same argument as in showing (1a) (with (L_Q, F_Q, r_Q) replaced by $(L_\alpha, F_\alpha, r_\alpha)$) proves (2a). Lemma A.2.6(1) and $T \asymp n^\xi$imply that $\|\tilde{u}\| = O_P([\sqrt{T} + n/\sqrt{T}]\log^{O(1)} n)$. Since $[\sqrt{T} + n/\sqrt{T}]\log^{O(1)} n = o(\tilde{\mu}_n)$, condition (2b) follows. We have proved part (2). □

The following results are useful for proving Theorem 1.3.7.

Lemma A.2.13. *Let Assumption 1 hold. Then for any fixed positive integer r, there exists a constant $c > 0$ such that $\mathbb{P}\left(s_r(u) > \sqrt{n}c\right) \to 1$. Moreover, $\mathbb{P}(s_r(\tilde{u}) > \sqrt{n}c/2) \to 1$, where $\tilde{u}$ is defined in Lemma A.2.6.*

Proof. Notice that there exist a constant $c_1 > 0$ and a sequence $\{t_j\}_{j=1}^r \subset [T]$ such that $t_1 < t_2 < \ldots < t_r$ and $d_n \geq c_1 T$, where $d_n = \min_{1\leq j\leq r-1}(t_{j+1} - t_j)$. Let $A \in \mathbb{R}^{n\times r}$ be defined as $A_{i,j} = u_{i,t_j}$ for $(i,j) \in [n] \times [r]$. Since A is a matrix consisting of r columns of u, Lemma A.3.10(4) implies

$$s_r(u) \geq s_r(A). \tag{A.2.41}$$

Let $\Sigma_A = n^{-1}\mathbb{E}A'A \in \mathbb{R}^{r\times r}$. We only need to show the lower bound for $s_r(A)$. We proceed in two steps. First, we show that singular values of Σ_A are bounded below; then we show the desired result. Finally, we shall show the result for $s_r(\tilde{u})$.

Step 1: derive lower bound for $s_r(\Sigma_A)$. Fix arbitrary $j_1, j_2 \in [r]$ with $j_1 \neq j_2$. Notice that $\Sigma_{A,j_1,j_2} = n^{-1}\sum_{i=1}^n \mathbb{E}u_{i,t_{j_1}}u_{i,t_{j_2}}$. By Lemma A.3.3(2) and the exponential-type tails of $u_{i,t}$'s, there exists a constant $c_1 > 0$ such that $\mathbb{E}|u_{i,t}|^4 \leq c_1$. It follows, by Corollary 16.2.4 of Athreya and Lahiri (2006), that

$$\begin{aligned}|\Sigma_{A,j_1,j_2}| &\leq \max_i |\mathbb{E}(u_{i,t_{j_1}}u_{i,t_{j_2}})| \leq 4c_1^2\sqrt{2\alpha_{mixing}(|t_{j_1} - t_{j_2}|)} \\ &\leq 4c_1^2\sqrt{c_*}\exp[-|t_{j_1} - t_{j_2}|^{\gamma_{**}}/2] \leq 4c_1^2\sqrt{c_*}\exp[-d_n^{\gamma_{**}}/2] = o(1).\end{aligned}$$

Let $\tilde{\Sigma}_A \in \mathbb{R}^{r\times r}$ be the diagonal matrix such that $\tilde{\Sigma}_{A,j,j} = \Sigma_{A,j,j}$ for $j \in [r]$. Then the above display implies that $\|\tilde{\Sigma}_A - \Sigma_A\| = o(1)$.

For $j \in [r]$, $\tilde{\Sigma}_{A,j,j} = \Sigma_{A,j,j} = n^{-1}\sum_{i=1}^n \mathbb{E}u_{i,t_j}^2$. By Assumption 1, there exists a constant $c_2 > 0$ with $\tilde{\Sigma}_{A,j,j} \geq c_2$ for $j \in [r]$. It follows, by Lemma A.3.10(1), that $s_r(\Sigma_A) + s_1(\tilde{\Sigma}_A - \Sigma_A) \geq s_r(\tilde{\Sigma}_A) \geq c_2$. Since $\|\tilde{\Sigma}_A - \Sigma_A\| = o(1)$, we have

$$s_r(\Sigma_A) \geq c_2/2. \tag{A.2.42}$$

Step 2: show the desired result for $s_r(u)$. By the law of large numbers, we have that for any $j_1, j_2 \in [r]$, $n^{-1}\sum_{i=1}^n \left(u_{i,t_{j_1}}u_{i,t_{j_2}} - \mathbb{E}u_{i,t_{j_1}}u_{i,t_{j_2}}\right) = o_P(1)$. Since r is fixed, this means that $\|n^{-1}A'A - \Sigma_A\| = o_P(1)$. By Lemma A.3.10(1), we have

$$s_r(n^{-1}A'A) + s_1(\Sigma_A - n^{-1}A'A) \geq s_r(\Sigma_A) \overset{(i)}{\geq} c_2/2,$$

where (i) holds by (A.2.42). Since $\|\Sigma_A - n^{-1}A'A\| = o_P(1)$, we have that $s_r(n^{-1}A'A) \geq c_2/2 - o_P(1)$. By $s_r(A) = \sqrt{s_r(A'A)}$ and (A.2.41), the desired result for $s_r(u)$ holds with $c = c_2/3$.

Step 3: show the desired result for $s_r(\tilde{u})$. Let $B \in \mathbb{R}^{n\times r}$ with $B_{i,j} = \tilde{u}_{i,t_j} - u_{i,j}$. By the definition of $\tilde{u}_{i,t}$,

$$\max_{i,t} |\tilde{u}_{i,t} - u_{i,t}| \leq \max_{i,t} \|x_{i,t}\| \max_t \|\hat{\beta}_t - \beta_t\| \overset{(i)}{=} O_P\left([n^{-1/2} + n^{1/2-\xi}] \log^{O(1)} n\right) = o_P(1),$$

where (i) holds by Lemma A.2.1(1) and Theorem 1.3.1, together with $\xi > 6/7$. Since $\|B\| \leq \sqrt{nr} \max_{i,t} |\tilde{u}_{i,t} - u_{i,t}|$, we have that $\|B\| = o_P(\sqrt{n})$.

Notice that $A + B$ is a matrix consisting of r columns of $\tilde{u}$. Hence, Lemma A.3.10(4) implies $s_r(\tilde{u}) \geq s_r(A + B)$. By Lemma A.3.10(1), $s_r(A + B) + s_1(-B) \geq s_r(A)$. It follows, by $\|B\| = o_P(\sqrt{n})$, that $s_r(A + B)/\sqrt{n} \geq s_r(A)/\sqrt{n} - o_P(1)$. Since $s_r(A)/\sqrt{n} \geq \sqrt{c_2/3}$ with probability approaching one (due to Step 2), the desired result for $s_r(\tilde{u})$ follows. □

Lemma A.2.14. *Let Assumption 1 hold. Then for any fixed positive integer r, there exists a constant $c > 0$ such that $\mathbb{P}(s_r(v) > \sqrt{n}c) \to 1$.*

Proof. The proof is the same as that of Lemma A.2.13 with u replaced by v. □

Proof of Theorem 1.3.7. **Step 1: show consistency of $\hat{r}_Q^{SV}$.** It suffices to show the following:

(1a) $\max_{r_Q+1\leq r\leq r_{\max}} [s_r(X)/s_{r+1}(X)] = O_P(\sqrt{\log n})$;

(1b) $\max_{1\leq r\leq r_Q-1} [s_r(X)/s_{r+1}(X)] = O_P(1)$;

(1c) $\mathbb{P}(s_{r_Q}(X)/s_{r_Q+1}(X) > T^{1/3}) \to 1$.

We first show condition (1a). Lemma A.3.10(1) implies that, for $r > r_Q$,

$$s_r(X) = s_r(L_Q F_Q' + v) \leq s_r(L_Q F_Q') + s_1(v) \overset{(i)}{=} 0 + \|v\|, \tag{A.2.43}$$

where (i) holds by $\text{rank}L_Q = r_Q$. Lemma A.3.10(1) also implies that, for $r > r_Q$,

$$s_{r+1}(X) + s_1(-L_Q F_Q') \geq s_{r+1}(X - L_Q F_Q') = s_{r+1}(v). \tag{A.2.44}$$

By (A.2.43) and (A.2.44), we have

$$\max_{r_Q+1\le r\le r_{\max}} \frac{s_r(X)}{s_{r+1}(X)} \le \frac{\|v\|}{s_{r_{\max}+1}(v)} \overset{(i)}{=} O_P(\sqrt{\log n}),$$

where (i) holds by Lemmas A.2.1(2) and A.2.14. This proves condition (1a).

Since $\|v\| = O_P(\sqrt{n\log n})$ (Lemma A.2.1(2)), $\|X - L_Q F_Q'\|/\sqrt{nT} = \|v\|/\sqrt{nT} = o_P(1)$. By Assumption 1, the largest r_Q singular values of $L_Q F_Q'/\sqrt{nT}$ are bounded away from zero and infinity. Therefore, there exist constants $c_1, c_2 > 0$ such that the largest r_Q singular values of $X/\sqrt{nT}$ lie in $[c_1, c_2]$ with probability approaching one. This proves condition (1b).

Since $\mathbb{P}[s_{r_Q}(X)/\sqrt{nT} \ge c_1] \to 1$ and $s_{r_Q+1}(X) \le \|v\| = O_P(\sqrt{n\log n})$ (due to (A.2.43) and Lemma A.2.1(2)), we have that $\mathbb{P}[s_{r_Q}(X)/s_{r_Q+1}(X) \ge c_3\sqrt{T/\log n}] \to 1$ for some constant $c_3 > 0$. Condition (1c) follows. We have proved the consistency of $\hat{r}_Q^{SV}$.

Step 2: show consistency of $\hat{r}_\alpha^{SV}$. The argument is similar to Step 1. Notice that $y_t - X_t\hat{\beta}_t = \alpha_t + \tilde{u}_t$, where $\tilde{u}$ is defined in Lemma A.2.6. Recall that $\alpha = L_\alpha F_\alpha'$ with rank$\alpha = r_\alpha$. We shall verify the following conditions:

(2a) $\max_{r_\alpha+1\le r\le r_{\max}} [s_r(\alpha+\tilde{u})/s_{r+1}(\alpha+\tilde{u})] = O_P\left([n^{(\xi-1)/2} + n^{(1-\xi)/2}]\log^{O(1)} n\right)$;

(2b) $\max_{1\le r\le r_\alpha-1} [s_r(\alpha+\tilde{u})/s_{r+1}(\alpha+\tilde{u})] = O_P(1)$;

(2c) $\mathbb{P}\left[s_{r_\alpha}(\alpha+\tilde{u})/s_{r_\alpha+1}(\alpha+\tilde{u}) > M_1 n^{(1+\xi)/2}[n^{\xi/2} + n^{1-\xi/2}]^{-1}/\log^{M_2} n\right] \to 1$ for constants $M_1, M_2 > 0$.

Notice that the above three conditions imply the desired result because for $\xi \in (6/7, 2)$,

$$\frac{n^{(1+\xi)/2}[n^{\xi/2} + n^{1-\xi/2}]^{-1}/\log^{M_2} n}{[n^{(\xi-1)/2} + n^{(1-\xi)/2}]\log^{O(1)} n} \to \infty \qquad \text{and} \qquad [n^{(\xi-1)/2} + n^{(1-\xi)/2}]\log^{O(1)} n \to \infty.$$

Similar to (A.2.43) and (A.2.44), we have that, for $r > r_\alpha$, $s_r(\alpha+\tilde{u}) \le \|\tilde{u}\|$ and $s_{r+1}(\alpha+\tilde{u}) \ge s_{r+1}(\tilde{u})$. Therefore,

$$\max_{r_\alpha+1\le r\le r_{\max}} \frac{s_r(\alpha+\tilde{u})}{s_{r+1}(\alpha+\tilde{u})} \le \frac{\|\tilde{u}\|}{s_{r_{\max}+1}(\tilde{u})} \overset{(i)}{=} O_P\left([n^{(\xi-1)/2} + n^{(1-\xi)/2}]\log^{O(1)} n\right),$$

where (i) holds by Lemmas A.2.6(1) and A.2.13. This proves condition (2a).

Since $\|\tilde{u}\|/\sqrt{nT} = o_P(1)$ (due to Lemma A.2.6(1) and $T \asymp n^{\xi}$), condition (2b) follows from the same argument as the proof of condition (1b), except that (L_Q, F_Q, r_Q) is replaced with $(L_\alpha, F_\alpha, r_\alpha)$.

Similar to the proof of condition (1c), we notice that $\mathbb{P}[s_{r_\alpha}(\alpha + \tilde{u})/\sqrt{nT} \geq c_3] \to 1$ for some constant $c_3 > 0$ and $s_{r_\alpha+1}(\alpha + \tilde{u}) \leq \|\tilde{u}\| \overset{(i)}{=} O_P([n^{\xi/2} + n^{1-\xi/2}] \log^{O(1)} n)$ (with (i) due to Lemma A.2.6(1)). Hence, condition (2c) follows by $T \asymp n^{\xi}$. We have proved the consistency of $\hat{r}_\alpha^{SV}$. □

Proof of Theorem 1.4.1. To avoid complicated notations involving j_0, we prove the result for the full vector β_t (rather than $\beta_{j_0,t}$), i.e.,

$$\sqrt{T}(\tilde{\Theta} - \hat{\Theta}) = o_P(1), \tag{A.2.45}$$

where $\hat{\Theta} = (\sum_{t=1}^T \beta_t z_t')(\sum_{t=1}^T z_t z_t')^{-1}$ and $\tilde{\Theta} = (\sum_{t=1}^T \tilde{\beta}_t z_t')(\sum_{t=1}^T z_t z_t')^{-1}$. The result stated in Theorem 1.4.1 corresponds to the j_0-th row of the above display.

By (A.2.35) in the proof of Theorem 1.3.3 (with $J = I_{kT}$), we have $\tilde{\beta}_t - \beta_t = n^{-1/2} D_{n,t} + n^{-1/2}\delta_t$, where $\delta_t = n^{-1/2} \sum_{i=1}^n G_{i,t}$ and $D_{n,t}$ and $G_{i,t}$ are defined in (A.2.1). Thus, the definitions of $\tilde{\Theta}$ and $\hat{\Theta}$ imply that

$$\begin{aligned}
\sqrt{T}(\tilde{\Theta} - \hat{\Theta}) &= \left(T^{-1/2} \sum_{t=1}^T (\tilde{\beta}_t - \beta_t) z_t' \right) \left(T^{-1} \sum_{t=1}^T z_t z_t' \right)^{-1} \\
&\overset{(i)}{=} \left(\frac{1}{\sqrt{nT}} \sum_{t=1}^T D_{n,t} z_t' \right) O_P(1) + \left(\frac{1}{\sqrt{nT}} \sum_{t=1}^T \delta_t z_t' \right) O_P(1),
\end{aligned} \tag{A.2.46}$$

where (i) holds by $(T^{-1} \sum_{t=1}^T z_t z_t')^{-1} = O_P(1)$. The rest of the proof proceeds in two steps in which we bound the two terms in (A.2.46).

Step 1: show $\frac{1}{\sqrt{nT}} \sum_{t=1}^T D_{n,t} z_t' = o_P(1)$. Notice that

$$\begin{aligned}
\left\| \frac{1}{\sqrt{nT}} \sum_{t=1}^T D_{n,t} z_t' \right\| &\leq \frac{1}{\sqrt{nT}} \sum_{t=1}^T \|D_{n,t}\| \|z_t\| \\
&\overset{(i)}{\leq} \|D_n\|_\infty \sqrt{\frac{k}{nT}} \sum_{t=1}^T \|z_t\|
\end{aligned}$$

$$\overset{(ii)}{=} \|D_n\|_\infty \sqrt{\frac{Tk}{n}} O_P(1)$$
$$\overset{(iii)}{=} O_P\left([n^{\xi/2-1} + n^{3-7\xi/2}] \log^{O(1)} n \right) \sqrt{n^{\xi-1}} O_P(1) \overset{(iv)}{=} o_P(1), \tag{A.2.47}$$

where (i) follows by Holder's inequality and $\|D_{n,t}\| \leq \sqrt{k}\|D_n\|_\infty$, (ii) follows by $T^{-1}\sum_{t=1}^T \|z_t\| = O_P(1)$ (due to $T^{-1}\sum_{t=1}^T \mathbb{E}\|z_t\| \leq \max_t \mathbb{E}\|z_t\| = O(1)$), (iii) follows by $T \asymp n^\xi$ and Lemma A.2.8 and (iv) holds by $6/7 < \xi < 3/2$.

Step 2: show $\frac{1}{\sqrt{nT}}\sum_{t=1}^T \delta_t z_t' = o_P(1)$. Therefore, by Lemma A.3.6 (applied with $\mathcal{F}_n$ being the trivial σ-algebra), $\max_{1\leq t\leq T} \|\delta_t\|_\infty = O_P(\sqrt{\log(kT)})$. Let $r_n = \sum_{t=1}^T z_t \otimes \delta_t$. Then

$$\mathbb{E}\|r_n\|^2 = \sum_{s,t=1}^T \mathbb{E}\left[(z_t' \otimes \delta_t')(z_s \otimes \delta_s)\right] = \sum_{s,t=1}^T \mathbb{E}\left[(z_t' z_s)(\delta_t'\delta_s)\right] \overset{(i)}{=} \sum_{s,t=1}^T \mathbb{E}(z_t' z_s)\mathbb{E}(\delta_t'\delta_s)$$
$$\leq T \max_{s,t} |\mathbb{E}(z_t' z_s)| \max_s \sum_{t=1}^T |\mathbb{E}(\delta_t'\delta_s)| \overset{(ii)}{=} O(T) \max_s \sum_{t=1}^T |\mathbb{E}(\delta_t'\delta_s)|, \tag{A.2.48}$$

where (i) holds by the independence between $\{z_t\}_{t=1}^T$ and (u, v) and (ii) holds by $\max_{s,t} |\mathbb{E}(z_t' z_s)| \leq \max_t \mathbb{E}\|z_t\|^2 = O(1)$. Notice that

$$\max_s \sum_{t=1}^T |\mathbb{E}(\delta_t'\delta_s)| = \max_s \sum_{t=1}^T \left| n^{-1} \sum_{i,j=1}^n \mathbb{E} G_{i,t}' G_{j,s} \right| \overset{(i)}{=} \max_s \sum_{t=1}^T \left| n^{-1} \sum_{i=1}^n \mathbb{E} G_{i,t}' G_{i,s} \right|$$
$$\leq \max_s n^{-1} \sum_{i=1}^n \sum_{t=1}^T |\mathbb{E} G_{i,t}' G_{i,s}| \leq \max_{(i,s)\in[n]\times[s]} \sum_{t=1}^T |\mathbb{E} G_{i,t}' G_{i,s}| \overset{(ii)}{=} O(1), \tag{A.2.49}$$

where (i) follows by the independence of $\{(G_{i,s}, G_{i,t})\}_{i=1}^n$ across i and (ii) holds by Lemma A.2.2(3). It follows, by (A.2.48) and (A.2.49), that $\mathbb{E}\|r_n\|^2 = O(T)$ and thus

$$\operatorname{vec}\left(\frac{1}{\sqrt{nT}} \sum_{t=1}^T \delta_t z_t' \right) = \frac{1}{\sqrt{nT}} r_n = \frac{1}{\sqrt{nT}} O_P(\sqrt{T}) = o_P(1). \tag{A.2.50}$$

Hence, we obtain (A.2.45) by combining (A.2.46) with (A.2.47) and (A.2.50). The proof is complete. □

The following lemma is needed to prove Lemma A.2.16.

A.2.4 Strong mixing with geometric decay for Example 1.2.1

Lemma A.2.15. *Let* $\varphi(x,y;r)$ *be the p.d.f of* $N\left(0,\begin{bmatrix}1 & r\\ r & 1\end{bmatrix}\right)$ *for* $|r| < 1$. *Suppose that* $\forall t,s \geq 0$, $\sigma_t,\sigma_s \geq 1$ *and* $|r_{t,s}| \leq C\exp(-|t-s|)$ *for some constant* $C > 0$. *Then there exist constants* $\tau, M > 0$ *such that* $\sup_{t,s,|t-s|>\tau}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\int_0^1 \varphi(i/\sigma_t, j/\sigma_s; hr_{t,s}/(\sigma_t\sigma_s))dh \leq M$.

Proof. Using the formula for the p.d.f of bivariate Gaussian random vectors, we have that

$$\varphi(x,y;r) = \frac{1}{2\pi\sqrt{1-r^2}}\exp\left[-\frac{x^2-2rxy+y^2}{2(1-r^2)}\right].$$

Choose $\tau > 0$ such that $|r_{t,s}|/(\sigma_t\sigma_s) \leq 1/2$ for $|t-s| \geq \tau$. Hence, for $0 \leq h \leq 1$,

$$\varphi\left(x,y;hr_{t,s}/(\sigma_t\sigma_s)\right) \leq \frac{1}{2\pi\sqrt{3/4}}\exp\left[-\frac{x^2-xy+y^2}{3/2}\right]$$
$$\leq \frac{1}{\sqrt{3}\pi}\exp\left[-\frac{2}{3}\left((x^2+y^2)/2+(x^2+y^2-xy)/2\right)\right] \leq \frac{1}{\sqrt{3}\pi}\exp\left[-\frac{1}{3}\left(x^2+y^2\right)\right].$$

It follows that for $|t-s| \geq \tau$,

$$\begin{aligned}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\int_0^1 \varphi(i/\sigma_t, j/\sigma_s; hr_{t,s}/(\sigma_t\sigma_s))dh &\leq \frac{1}{\sqrt{3}\pi}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\exp\left[-\frac{1}{3}\left(i^2+j^2\right)\right]\\ &< \frac{1}{\sqrt{3}\pi}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\exp\left[-\frac{1}{3}\left(i+j\right)\right]\\ &= \frac{1}{\sqrt{3}\pi}\left(\sum_{i=0}^{\infty}\exp\left(-i/3\right)\right)^2 < \infty.\end{aligned}$$

Therefore, the desired result holds with $M = \left(\sum_{i=0}^{\infty}\exp\left(-i/3\right)\right)^2/\left(\sqrt{3}\pi\right)$. □

Lemma A.2.16. *Under the setup of Example 1.2.1, there exist a constant $M > 0$ such that $\forall t \geq 1$, $\alpha_{mixing}(t) \leq Mc^t$.*

Proof. Recall the notations in Example 1.2.1. By Theorem 2.2 of Piterbarg (2012), it suffices to verify that

(i) For some constants $\tau_0, K_1 > 0$, we have that $\forall \tau \geq \tau_0$

$$\sup_{t,s,|t-s|>\tau} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \int_0^1 \varphi\left(i/\sigma_t, j/\sigma_s; hr(t,s)/(\sigma_t\sigma_s)\right) dh \leq K_1$$

(ii) For some constant $K_2 \in (0,1)$, we have that $\forall \tau \geq \tau_0$

$$\sup_{v} \sum_{t=v+\tau}^{\infty} \sum_{s=0}^{v} \frac{|r(t,s)|}{\sigma_t\sigma_s} \leq K_2^{|t-s|}.$$

Notice that $\sigma_t^2 = \sum_{j=0}^{\infty} \gamma_{t-1,j}^2 \geq \gamma_{t-1,0}^2 = 1$. For $t > s$,

$$\begin{aligned} |r(t,s)| = \left| \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} \gamma_{t-1,j_1}\gamma_{s-1,j_2} \mathbb{E} u_{t-1-j_1} u_{s-1-j_2} \right| &\overset{(i)}{=} \left| \sum_{j=0}^{\infty} \gamma_{t-1,j+t-s}\gamma_{s-1,j} \right| \\ &\leq \sum_{j=0}^{\infty} c^{2j+t-s} = Dc^{t-s} \quad \text{for} \quad D = \frac{1}{1-c^2}, \quad \text{(A.2.51)} \end{aligned}$$

where (i) follows by u_t being i.i.d $N(0,1)$. Hence, claim (i) holds by Lemma A.2.15.

By (A.2.51), we have that for any $v \geq 0$,

$$\begin{aligned} &\sum_{t=v+\tau}^{\infty} \sum_{s=0}^{v} \frac{|r(t,s)|}{\sigma_t\sigma_s} \\ &\leq D \sum_{t=v+\tau}^{\infty} \sum_{s=0}^{v} c^{t-s} = \frac{D}{1-c} \sum_{t=v+\tau}^{\infty} c^{t-v}(1-c^{v+1}) \overset{(i)}{\leq} \frac{D}{1-c} \sum_{t=v+\tau}^{\infty} c^{t-v} = \frac{D}{(1-c)^2} c^{\tau}, \end{aligned}$$

where (i) holds by $c \in (0,1)$. Claim (ii) follows. The proof is complete. □

A.3 Useful technical tools

A.3.1 Useful results on probability theory

Lemma A.3.1. *Let X and Y be two random vectors. Then $\forall t, \varepsilon > 0$, we have*

$$|\mathbb{P}(\|X\|_\infty \le t) - \mathbb{P}(\|Y\|_\infty \le t)| \le \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty \in (t-\varepsilon, t+\varepsilon]).$$

Proof. By the triangular inequality,

$$\begin{aligned}\mathbb{P}(\|X\|_\infty > t) \le& \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty > t - \varepsilon) \\ =& \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty > t) + \mathbb{P}(\|Y\|_\infty \in (t-\varepsilon, t]).\end{aligned} \tag{A.3.1}$$

On the other hand, also by the triangular inequality,

$$\begin{aligned}\mathbb{P}(\|X\|_\infty > t) \ge& \mathbb{P}(\|Y\|_\infty > t + \varepsilon) - \mathbb{P}(\|X - Y\|_\infty > \varepsilon) \\ =& \mathbb{P}(\|Y\|_\infty > t) - \mathbb{P}(\|Y\|_\infty \in (t, t+\varepsilon]) - \mathbb{P}(\|X - Y\|_\infty > \varepsilon).\end{aligned} \tag{A.3.2}$$

It follows, by (A.3.1) and (A.3.2), that

$$|\mathbb{P}(\|X\|_\infty > t) - \mathbb{P}(\|Y\|_\infty > t)| \le \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty \in (t-\varepsilon, t+\varepsilon]).$$

The desired result follows by $|\mathbb{P}(\|X\|_\infty > t) - \mathbb{P}(\|Y\|_\infty > t)| = |\mathbb{P}(\|X\|_\infty \le t) - \mathbb{P}(\|Y\|_\infty \le t)|$. □

Lemma A.3.2. *Let X and Y be two random vectors. Define $F_X(x) = \mathbb{P}(\|X\|_\infty \le x)$ and $F_Y(x) = \mathbb{P}(\|Y\|_\infty \le x)$. Then $\forall \varepsilon > 0$,*

$$\sup_{\alpha \in (0,1)} \left|\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha)\right) - \alpha\right| \le \varepsilon + \mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > \varepsilon\right).$$

Proof. Fix $\alpha \in (0,1)$ and notice that

$$\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha)\right)$$

$$
\begin{aligned}
=&\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \le \varepsilon\right)\\
&+\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right)\\
\overset{(i)}{\le}&\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \le \varepsilon\right)\\
&+\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right)\\
\le&\mathbb{P}\left(\|X\|_\infty > F_X^{-1}(1-\alpha-\varepsilon)\right)+\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right)\\
\overset{(ii)}{=}&\alpha+\varepsilon+\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right),
\end{aligned}
\tag{A.3.3}
$$

where (i) follows from Lemma A.1(ii) in Romano and Shaikh (2012) (if $\sup_{x\in\mathbb{R}}[F_Y(x)-F_X(x)] \le \varepsilon$ then $F_X^{-1}(1-\alpha-\varepsilon) \le F_Y^{-1}(1-\alpha)$) and (ii) follows by the definition of $F_X(\cdot)$. Also notice that

$$
\begin{aligned}
&\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha)\right)\\
\ge\ &\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \le \varepsilon\right)\\
\overset{(i)}{\ge}\ &\mathbb{P}\left(\|X\|_\infty > F_X^{-1}(1-\alpha+\varepsilon) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \le \varepsilon\right)\\
\overset{(ii)}{\ge}\ &\mathbb{P}\left(\|X\|_\infty > F_X^{-1}(1-\alpha+\varepsilon)\right)-\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right)\\
\overset{(iii)}{=}\ &\alpha-\varepsilon-\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right)
\end{aligned}
\tag{A.3.4}
$$

where (i) follows from Lemma A.1(ii) in Romano and Shaikh (2012) (if $\sup_{x\in\mathbb{R}}[F_X(x)-F_Y(x)] \le \varepsilon$ then $F_Y^{-1}(1-\alpha) \le F_X^{-1}(1-\alpha+\varepsilon)$), (ii) follows by the elementary inequality that for any two events A and B, $\mathbb{P}(A\bigcap B)+\mathbb{P}(B^c) \ge \mathbb{P}(A\bigcap B)+\mathbb{P}(A\bigcap B^c) = \mathbb{P}(A)$ or equivalently $\mathbb{P}(A\bigcap B) \ge \mathbb{P}(A)-\mathbb{P}(B^c)$, and (iii) follows by the definition of $F_X(\cdot)$. The desired result follows by (A.3.3) and (A.3.4). □

Lemma A.3.3. *The following hold.*

(1) Let $Z \in \mathbb{R}^{m_Z}$ be a random vector whose jth entry is denoted by Z_j. Suppose

that there exist constants $b, \gamma > 0$ such that $\forall j \in [m_Z]$, Z_j has an exponential-type tail with parameter (b, γ). Then for any nonrandom vector $a \in \mathbb{R}^{m_Z}$, $a'Z$ has an exponential-type tail with parameter $\Big(b\|a\|_1 \log^{1/\gamma}(\|a\|_0 + 2), \gamma\Big)$.

(2) Let $\{Z_j\}_{j=1}^{m_Z}$ be a sequence of random variables. Suppose that constants $b, \gamma > 0$ satisfy that $\forall j \in [m_Z]$, Z_j has an exponential-type tail with parameter (b, γ). Let $q > 0$ be any nonrandom number. Then there exists a constant $C_{\gamma,q} > 0$ depending only on γ and q such that $\mathbb{E} \max_{1 \le j \le m_Z} |Z_j|^q \le C_{\gamma,q} m_Z b^q$ and $\mathbb{E}|Z_j|^q \le C_{\gamma,q} b^q$ $\forall j \in [m_Z]$.

(3) Let Z_1 and Z_2 be two random variables having exponential-type tails with parameters (b_1, γ_1) and (b_2, γ_2), respectively. Then $\forall \gamma \in (0, \gamma_1\gamma_2(\gamma_1 + \gamma_2)^{-1})$, $Z_1 Z_2$ has an exponential-type tail with parameter $\big(2^{1/\gamma_0} b_1 b_2, \gamma_0\big)$, where $\gamma_0 = \gamma_1\gamma_2(\gamma_1 + \gamma_2)^{-1}$

(4) Let X have an exponential-type tail with parameter (b_X, γ_X). Then $\forall a \in \mathbb{R}$, $X - a$ has an exponential-type tail with parameter $(b_X + |a|, \gamma_X)$.

Proof. **Proof of part (1).** Let $A_0 := \{i \mid a_i \neq 0\}$. Then by Holder's inequality and the union bound, $\mathbb{P}(|a'Z| > x) \le \mathbb{P}(\|a\|_1 \max_{i \in A_0} |Z_i| > x) \le \sum_{i \in A_0} \mathbb{P}(\|a\|_1 |Z_i| > x) \le \|a\|_0 \exp\big[1 - (xb^{-1}\|a\|_1^{-1})^\gamma\big]$. If $\|a\|_0 = 1$, then the result follows by $b\|a\|_1 < b\|a\|_1 \log^{1/\gamma}(3)$. For $\|a\|_0 > 1$, we let $c = b\|a\|_1 \log^{1/\gamma} \|a\|_0 < b\|a\|_1 \log^{1/\gamma}(\|a\|_0 + 2)$. For $x \le c$, $\mathbb{P}(|a'Z| > x) \le 1 \le \exp(1 - (x/c)^\gamma)$. Since $\mathbb{P}(|a'Z| > x) \le \|a\|_0 \exp\big[1 - (xb^{-1}\|a\|_1^{-1})^\gamma\big]$, it suffices to show that $\forall x > c$, $\log \|a\|_0 - (xb^{-1}\|a\|_1^{-1})^\gamma \le 1 - (xc^{-1})^\gamma$. This is to say that $x^\gamma \ge (\log \|a\|_0 - 1)/((b\|a\|_1)^{-\gamma} - c^{-\gamma})$ $\forall x > c$. By simple computations, one can show that $c^\gamma = (\log \|a\|_0 - 1)/((b\|a\|_1)^{-\gamma} - c^{-\gamma})$. Part (1) follows.

Proof of part (2). Notice that, by the union bound, $\mathbb{P}(\max_{1 \le j \le m_Z} |Z_j| > x) \le \sum_{j=1}^{m_Z} \mathbb{P}(|Z_j| > x) \le m_Z \exp[1 - (x/b)^\gamma]$. Then

$$\mathbb{E} \max_{1 \le j \le m_Z} |Z_j|^q \overset{(i)}{=} \int_0^\infty \mathbb{P}\left(\max_{1 \le j \le m_Z} |Z_j|^q > x\right) dx = \int_0^\infty \mathbb{P}\left(\max_{1 \le j \le m_Z} |Z_j| > x^{1/q}\right) dx$$
$$\overset{(ii)}{\le} m_Z \int_0^\infty \exp\Big[1 - \big(x^{1/q}/b\big)^\gamma\Big] dx$$

$$\overset{(iii)}{=} m_Z b^q \left(q\gamma^{-1} \int_0^\infty e^{1-z} z^{q/\gamma-1} dz \right),$$

where (i) follows by the identity $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)dx$ for any non-negative random variable X, (ii) follows by $\mathbb{P}(\max_{1\le j\le m_Z} |Z_j| > x) \le m_Z \exp\left[1 - (x/b)^\gamma\right]$ and (iii) follows by a change of variable $z = \left(x^{1/q}/b\right)^\gamma$. The bound for $\mathbb{E}\max_{1\le j\le m_Z} |Z_j|^q$ follows with $C_{\gamma,q} = q\gamma^{-1}\int_0^\infty e^{1-z} z^{q/\gamma-1} dz$. The bound for $\mathbb{E}|Z_j|^q$ follows by the same reasoning with $\max_{1\le j\le m_Z} |Z_j|$ replaced by $|Z_j|$. This completes the proof for part (2).

Proof of part (3). The proof of Lemma A.2 of Fan, Liao, and Mincheva (2011) implies that $\forall\gamma \in (0, \gamma_0)$, $Z_1 Z_2$ has an exponential-type tail with parameter (b_3, γ) $\forall b_3 > b_0 \max\{(\gamma/\gamma_0)^{1/\gamma_0}, (1 + \log 2)^{1/\gamma_0}\}$, where $b_0 = b_1 b_2$. Let $b_* = 2^{1/\gamma_0} b_1 b_2$. It is easy to check that $b_* > b_0 \max\{(\gamma/\gamma_0)^{1/\gamma_0}, (1 + \log 2)^{1/\gamma_0}\}$. Thus, $Z_1 Z_2$ has an exponential-type tail with parameter (b_*, γ) $\forall\gamma \in (0, \gamma_0)$. In other words, for any $x > 0$, $\mathbb{P}(|Z_1 Z_2| > x) \le \exp[-(x/b_*)^\gamma]$ $\forall\gamma \in (0, \gamma_0)$. We take the infimum of the right-hand side over γ and obtain for any $x > 0$, $\mathbb{P}(|Z_1 Z_2| > x) \le \exp[-(x/b_*)^{\gamma_0}]$. Part (3) follows.

Proof of part (4). Let $c = b_X + |a|$. Notice that $\mathbb{P}(|X - a| > t) \le \mathbb{P}(|X| + |a| > t) = \mathbb{P}(|X| > t - |a|)$. For $t \in (0, c]$, $\mathbb{P}(|X| > t - |a|) \le 1 \le \exp\left[1 - (t/c)^{\gamma_X}\right]$. For $t > c$, $t - |a| > 0$ and $\mathbb{P}(|X| > t - |a|) \le \exp\left[1 - ((t - |a|)/b_x)^{\gamma_X}\right]$. It is easy to check that $(t - |a|)/b_x \ge t/c$ $\forall t > c$. Part (4) follows. □

Lemma A.3.4. *Let $\Sigma \in \mathbb{R}^{p\times p}$ be a positive-semi-definite matrix. Suppose that there exists a constant $b > 0$ such that $\min_{1\le j\le p} \Sigma_{j,j} \ge b$. Then there exists a constant $C_b > 0$ depending only on b such that $\forall\varepsilon > 0$,*

$$\sup_{x\in\mathbb{R}} |\Phi(x - \varepsilon, \Sigma) - \Phi(x + \varepsilon, \Sigma)| \le C_b \varepsilon \sqrt{\log p}.$$

Proof. Let $Y \sim N(0, \Sigma)$ with its jth component denoted by Y_j. By Nazarov's anti-concentration inequality (Lemma A.1 in Chernozhukov, Chetverikov, and Kato (2014)), there exists a constant $C_b' > 0$ depending only on b such that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\left(\max_{1\le j\le p} Y_j \in (x - \varepsilon, x + \varepsilon] \right) \quad \le \quad 2C_b' \varepsilon \sqrt{\log p}$$

$$\sup_{x\in\mathbb{R}} \mathbb{P}\left(\max_{1\le j\le p}(-Y_j)\in(x-\varepsilon,x+\varepsilon]\right) \;\le\; 2C_b'\varepsilon\sqrt{\log p}.$$

Therefore,

$$\begin{aligned}\sup_{x\in\mathbb{R}}|\Phi(x-\varepsilon,\Sigma)-\Phi(x+\varepsilon,\Sigma)| &= \sup_{x\in\mathbb{R}}\mathbb{P}\left(\|Y\|_\infty\in(x-\varepsilon,x+\varepsilon]\right)\\ &\overset{(i)}{\le} \sup_{x\in\mathbb{R}}\mathbb{P}\left(\max_{1\le j\le p}Y_j\in(x-\varepsilon,x+\varepsilon]\right)\\ &\quad+\sup_{x\in\mathbb{R}}\mathbb{P}\left(\max_{1\le j\le p}(-Y_j)\in(x-\varepsilon,x+\varepsilon]\right)\\ &\le 4C_b'\varepsilon\sqrt{\log p},\end{aligned}$$

where (i) holds by $\|Y\|_\infty\in\{\max_{1\le j\le p}Y_j,\ \max_{1\le j\le p}(-Y_j)\}$. The proof is complete. □

Lemma A.3.5. *Let Σ_A and Σ_B be $p\times p$ positive semi-definite matrices. Define $\Delta=\max_{1\le j,k\le p}|\Sigma_{A,j,k}-\Sigma_{B,j,k}|$. Suppose that there exist constants $c,C>0$ such that $c\le\Sigma_{A,j,j}\le C$ for $1\le j\le p$. Then there exists a constant $K>0$ depending only on c and C such that*

$$\sup_{x\in\mathbb{R}}|\Phi(x,\Sigma_A)-\Phi(x,\Sigma_B)|\le C\Delta^{1/3}\left(1\vee\log(2p/\Delta)\right)^{2/3}.$$

Proof. Consider random vectors $X\sim N(0,\Sigma_A)$ and $Y\sim N(0,\Sigma_B)$. Define $\bar{X}=(X',-X')'$ and $\bar{Y}=(Y',-Y')'$. Notice that $\bar{X}\sim N(0,\bar{\Sigma}_A)$ and $\bar{Y}\sim N(0,\bar{\Sigma}_B)$, where $\bar{\Sigma}_A=D\otimes\Sigma_A$, $\bar{\Sigma}_B=D\otimes\Sigma_B$ and $D=\begin{pmatrix}1&-1\\-1&1\end{pmatrix}$.

The definition of $\bar{\Sigma}_A$ and $\bar{\Sigma}_B$ also implies that (1) $\max_{1\le j,k\le 2p}|\bar{\Sigma}_{A,j,k}-\bar{\Sigma}_{B,j,k}|=\max_{1\le j,k\le p}|\Sigma_{A,j,k}-\Sigma_{B,j,k}|=\Delta$ and (2) the diagonal entries of $\bar{\Sigma}_A$ lie in $[c,C]$. It follows, by Lemma 3.1 of Chernozhukov, Chetverikov, and Kato (2013), that there exists a constant $M>0$ depending only on c and C such that

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\max_{1\le j\le 2p}\bar{X}_j\le x\right)-\mathbb{P}\left(\max_{1\le j\le 2p}\bar{Y}_j\le x\right)\right|\le M\Delta^{1/3}\left(1\vee\log(2p/\Delta)\right)^{2/3}.$$

We obtain the desired result by noticing that $\|X\|_\infty=\max_{1\le j\le 2p}\bar{X}_j$ and

$\|Y\|_\infty = \max_{1\le j\le 2p} \bar{Y}_j$. □

Lemma A.3.6. *Let $\{u_{i,j}\}_{(i,j)\in[n]\times J}$ be an array of random variables and $\mathcal{F}_n$ be a σ-algebra. Suppose the following hold:*
(i) Condition on $\mathcal{F}_n$, u_i is independent across i, where $u_i = \{u_{i,j} \mid j \in J\}$.
(ii) There exist constants $b, \gamma > 0$ such that $\forall (i,j) \in [n] \times J$ and $\forall x > 0$, $\mathbb{P}(|u_{i,j}| > x \mid \mathcal{F}_n) \le \exp(1 - (x/b)^\gamma)$ a.s.
(iii) $\forall 0 < c < \infty$, $n^{-c} \log |J| \to 0$, where $|J|$ denotes the cardinality of J.

Then

$$\max_{j\in J} \left| \sum_{i=1}^n [u_{i,j} - \mathbb{E}(u_{i,j} \mid \mathcal{F}_n)] \right| = O_P(\sqrt{n \log |J|}).$$

Proof. Let $\tilde{u}_{i,j} = u_{i,j} - \mathbb{E}(u_{i,j} \mid \mathcal{F}_n)$. By Lemma A.3.3(2) and (4) applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$, we have that there exists a constant $b_1 > 0$ depending only on b and γ such that $\forall z > 0$ $\forall (i,j) \in [n] \times J$, $\mathbb{P}(|\tilde{u}_{i,j}| > x \mid \mathcal{F}_n) \le \exp(1 - (x/b_1)^\gamma)$ a.s. Due to the conditional independence, the strong mixing coefficients are always zero.

Then by Theorem 1 in Merlevède, Peligrad, and Rio (2011) (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), there exist positive constants C_1, C_2, C_3, C_4, C_5 and r depending only on bM_ε and γ such that $r < 1$ and $\forall z > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{i=1}^n \tilde{u}_{i,j} \right| > z\sqrt{n \log |J|} \mid \mathcal{F}_n \right) \\
&\le n \exp\left[-C_1 \left(z\sqrt{n \log |J|} \right)^r \right] + \exp\left[-\frac{C_2 n z^2 \log |J|}{1 + nC_3} \right] \\
&\quad + \exp\left\{ -C_4 z^2 \log |J| \exp\left[C_5 \log^{-r}\left(z\sqrt{n \log |J|} \right) \left(z\sqrt{n \log |J|} \right)^{r/(1-r)} \right] \right\} \ a.s.
\end{aligned}$$

Then, by the union bound, we have that

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in J} \left| \sum_{i=1}^n \tilde{u}_{i,j} \right| > z\sqrt{n \log |J|} \mid \mathcal{F}_n \right) \\
&\le \sum_{j\in J} \mathbb{P}\left(\max_{j\in J} \left| \sum_{i=1}^n \tilde{u}_{i,j} \right| > z\sqrt{n \log |J|} \mid \mathcal{F}_n \right) \\
&\le |J| n \exp\left[-C_1 \left(z\sqrt{n \log |J|} \right)^r \right] + |J| \exp\left[-\frac{C_2 n z^2 \log |J|}{1 + nC_3} \right]
\end{aligned}$$

$$+ |J| \exp\left\{-C_4 z^2 \log|J| \exp\left[C_5 \log^{-r}\left(z\sqrt{n\log|J|}\right)\left(z\sqrt{n\log|J|}\right)^{r/(1-r)}\right]\right\} \ a.s.$$

By assumption (iv), the first and third terms in the above display go to zero for any $z > 0$. Hence, $\forall \varepsilon > 0$, we can choose a large constant $z_* > 0$ such that

$$\mathbb{P}\left(\max_{j\in J}\left|\sum_{i=1}^{n}\tilde{u}_{i,j}\right| > z_*\sqrt{n\log|J|} \mid \mathcal{F}_n\right) \leq \varepsilon \ a.s. \tag{A.3.5}$$

Hence, we have proved the result since, for an arbitrary $\varepsilon > 0$, we can find $z_* > 0$ such that the above equation holds. The result follows by the law of iterated expectations. □

Lemma A.3.7. *Let $\{W_j\}_{j\in J}$ be random variables. If there exist constant $b, \gamma > 0$ such that $\forall j \in J$, W_j has an exponential-type tail with parameter (b, γ), then $\max_{j\in J}|W_j| = O_P(\log^{1/\gamma}|J|)$, where $|J|$ is the cardinality of J.*

Proof. By the union bound, we have

$$\begin{aligned}
\mathbb{P}\left(\max_{j\in J}|W_j| > (\log|J|)^{1/\gamma} x\right) &\leq \sum_{j\in J}\mathbb{P}\left(|W_j| > (\log|J|)^{1/\gamma} x\right) \\
&\leq |J|\exp\left[1 - \left((\log|J|)^{1/\gamma} x/b\right)^{\gamma}\right] \\
&= \exp\left[1 + (1 - (x/b)^{\gamma})\log|J|\right].
\end{aligned}$$

Hence, for any $\varepsilon > 0$, one can choose large enough x such that the right-hand side of the above display is smaller than ε. The result follows. □

Lemma A.3.8. *Let $\mathcal{F}_n$ be a σ-algebra and $\{W_t\}_{t=1}^{T}$ be random variables with $\mathbb{E}(W_t \mid \mathcal{F}_n) = 0$. Suppose that the following hold:*
(i) There exist constants $\gamma_1, b_1 > 0$ such that $\forall t \in [T]$ and $\forall z > 0$, $\mathbb{P}(|W_t| > z \mid \mathcal{F}_n) \leq \exp[1 - (z/b_1)^{\gamma_1}]$ a.s.
(ii) There exist constants $\gamma_2, b_2 > 0$ such that $\alpha_n(t \mid \mathcal{F}_n) \leq b_2\exp(-t^{\gamma_2})$ a.s, where

$$\alpha_n(t \mid \mathcal{F}_n) := \sup\Big\{\Big|\mathbb{P}(A \mid \mathcal{F}_n)\mathbb{P}(B \mid \mathcal{F}_n) - \mathbb{P}(A\bigcap B \mid \mathcal{F}_n)\Big| :$$
$$A \in \sigma\left(\{(W_s, \ldots, W_s) \mid s \leq \tau\}\right),\ B \in \sigma\left(\{(W_s, \ldots, W_s) \mid s \geq \tau + t\}\right) \text{ and } \tau \in \mathbb{Z}\Big\}.$$

Then $\forall z > 0$, $\mathbb{P}(|T^{-1/2}\sum_{t=1}^{T} W_t| > z \mid \mathcal{F}_n) \leq \exp[1-(z/b)^{\gamma}]$ *a.s., where* $b, \gamma > 0$ *are constants depending only on* γ_1, γ_2, b_1 *and* b_2.

Proof. Let $\gamma_3 = \min\{\gamma_2, 1/2\}$. Notice that $\alpha_n(t \mid \mathcal{F}_n) \leq b_2 \exp(-t^{\gamma_3})$ and $\gamma_3 < 1$. Thus, $\gamma := (\gamma_1^{-1} + \gamma_3^{-1})^{-1} < 1$. Hence, by Theorem 1 in Merlevède, Peligrad, and Rio (2011) (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), there exist constants $C_1, C_2, C_3, C_4, C_5 > 0$ depending only on γ, γ_3, b_1 and b_2, such that $\forall z > 0$,

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} W_t\right| > zT^{1/2} \mid \mathcal{F}_n\right) \leq \underbrace{T\exp\left(-C_1 T^{\gamma/2} z^{\gamma}\right)}_{J_{1,T}(z)} + \underbrace{\exp\left(-\frac{C_2 z^2 T}{1+C_3 T}\right)}_{J_{2,T}(z)} + \underbrace{\exp\left[-C_4 z^2 \exp\left(C_5 \frac{(T^{1/2}z)^{\gamma/(1-\gamma)}}{[\log(T^{1/2}z)]^{\gamma}}\right)\right]}_{J_{3,T}(z)} \quad a.s.$$

It is not hard to see that one can choose a large enough constant $K > 0$ such that $\forall z \geq K$, $J_{1,T}(z) \leq \exp(-C_1 z^{\gamma})$, $J_{3,T}(z) \leq J_{1,T}(z)$ and $J_{2,T}(z) \leq \exp(-C_6 z^2)$, where $C_6 = C_2/(1+C_3)$. Hence, $\forall z \geq K$, $J_{1,T}(z) + J_{2,T}(z) + J_{3,T}(z) \leq 2\exp(-C_1 z^{\gamma}) + \exp(-C_6 z^2)$. Since $\gamma < 1$, we have that $\forall z \geq K$,

$$\mathbb{P}\left(T^{-1/2}\left|\sum_{t=1}^{T} W_t\right| > z \mid \mathcal{F}_n\right) \leq 3\exp\left(-C_7 z^{\gamma}\right) \ a.s, \tag{A.3.6}$$

where $C_7 = \min\{C_1, C_6\}$. Let $b := \max\left\{K, (C_7^{-1}\log 3)^{1/\gamma}\right\}$.

For $z \in (0, b]$, $\exp[1-(z/b)^{\gamma}] \geq 1 \geq \mathbb{P}(T^{-1/2}|\sum_{t=1}^{T} W_t| > z \mid \mathcal{F}_n)$. It is easy to verify that, $\forall z > b$, $3\exp(-C_7 z^{\gamma}) \leq \exp[1-(z/b)^{\gamma}]$. It follows, by (A.3.6), that $\forall z > b$, $\mathbb{P}(T^{-1/2}|\sum_{t=1}^{T} W_t| > z \mid \mathcal{F}_n) \leq \exp[1-(z/b)^{\gamma}]$. The proof is complete. □

Lemma A.3.9. *Let* $x > 0$ *and* $\{b_j\}_{j=1}^{q} \subset \mathbb{R}$. *Then* $\sum_{j=1}^{q} x^{b_j} \leq q(x^{b_{\min}} + x^{b_{\max}})$, *where* $b_{\min} = \min_{1\leq j\leq q} b_j$ *and* $b_{\max} = \max_{1\leq j\leq q} b_j$.

Proof. We discuss two cases: (A) $x \in (0, 1]$ and (B) $x > 1$. In Case (A), $x^{b_{\min}} \geq x^{b_j}$ for $1 \leq j \leq q$ and thus $\sum_{j=1}^{q} x^{b_j} \leq q x^{b_{\min}} \leq q(x^{b_{\min}} + x^{b_{\max}})$. In Case (B), $x^{b_{\max}} \geq x^{b_j}$ for $1 \leq j \leq q$ and thus $\sum_{j=1}^{q} x^{b_j} \leq q x^{b_{\max}} \leq q(x^{b_{\min}} + x^{b_{\max}})$. The proof is complete. □

A.3.2 Useful results on PCA

Lemma A.3.10. *The following hold.*
(1) Let $A, B \in \mathbb{R}^{n_1 \times n_2}$ be two matrices. If $i + j - 1 \leq \min\{n_1, n_2\}$, then $s_{i+j-1}(A+B) \leq s_i(A) + s_j(B)$, where $s_j(\cdot)$ denotes the jth largest singular value.
(2) Let $A \in \mathbb{R}^{n_1 \times n_0}$ and $B \in \mathbb{R}^{n_0 \times n_2}$. If $1 \leq i \leq n_0$, then $s_i(AB) \geq s_i(A)s_{n_0-i+1}(B)$.
(3) Let $A, B \in \mathbb{R}^{n_1 \times n_2}$ be two matrices. If $\text{rank}B \leq r$ and $1 \leq j \leq \min\{n_1, n_2\} - r$, then $s_j(A) \geq s_{j+r}(A+B) \geq s_{2r+j}(A)$.
(4) Let $A \in \mathbb{R}^{n_1 \times n_2}$. Let $B \in \mathbb{R}^{n_1 \times m}$ be the matrix consisting of the first m columns of A with $m \leq n_2$. Then for $j \in [m \wedge n_1]$, $s_j(B) \leq s_j(A)$.

Proof. Part (1) and (4) are Fact 6(b) and Fact 3, respectively, in Chapter 17.4 of Hogben (2006). Part (2) follows by Lemma 3 of Wang and Xi (1997). Part (3) follows by applying part (1): $s_j(A) = s_j(A) + s_{r+1}(B) \geq s_{j+r}(A+B)$ and $s_{j+r}(A+B) = s_{j+r}(A+B) + s_{r+1}(-B) \geq s_{2r+j}(A)$. □

Lemma A.3.11. *Let $W = LF' + e$ with $L \in \mathbb{R}^{n \times r}$ and $F \in \mathbb{R}^{T \times r}$. Let $W = \hat{U}\hat{\Sigma}\hat{V}'$ be an SVD and $\hat{U}_1 \in \mathbb{R}^{n \times r}$ the first r columns of $\hat{U}$. Define $\hat{L} = \sqrt{n}\hat{U}_1$ and $\hat{F}_t = n^{-1}\hat{L}'W_t$, where $W = (W_1, \ldots, W_T)$, $e = (e_1, \ldots, e_T)$ and $F = (F_1, \ldots, F_T)'$. Suppose that the following hold:*
(i) $\|e\| = o_P(\sqrt{nT})$ and $T \asymp n^\kappa$ for a constant $\kappa > 0$
(ii) There exist $0 < m_1 \leq m_2 < \infty$ such that all the eigenvalues of $\Sigma_F := T^{-1}F'F$ and $\Sigma_L := n^{-1}L'L$ belong to $[m_1, m_2]$ wpa1.
Then the following hold:

(1) $\hat{L}\hat{F}_t - LF_t = (n^{-1}\hat{L}\hat{L}' - I)LF_t + n^{-1}\hat{L}\hat{L}'e_t$.

(2) $\hat{L} - LH = \Delta_L$, where $H = F'FL'\hat{L}\hat{\Omega}_1^{-2}(nT)^{-1}$, $\Delta_L = (nT)^{-1}(LF'e' + eW')\hat{L}\hat{\Omega}_1^{-2}$ and $\hat{\Omega}_1 = \hat{\Sigma}_1(nT)^{-1/2}$ and $\hat{\Sigma}_1$ is the upper-left $r \times r$ submatrix of $\hat{\Sigma}$.

(3) $\|\hat{\Omega}_1^{-2}\| = O_P(1)$, $\|\Delta_L\| = O_P(\|e\|/\sqrt{T})$, $\|H\| = O_P(1)$ and $\|HH' - \Sigma_L^{-1}\| = O_P(\|e\|/\sqrt{nT})$.

(4) There exists a random variable $A_* = O_P(1)$ that does not depend on t such that, with probability one, $\forall a \in \mathbb{R}^n$,

$$\begin{aligned}\left|a'(\hat{L}\hat{F}_t - LF_t)\right| \le\ & n^{-\kappa} M_1 \|F' e' a\| A_* \\ & + \left[n^{-(1+\kappa)/2}\|e\| M_1 + n^{-1} M_2 + n^{-1-\kappa/2}\|e\|^2\right] A_* \|L' a\| \\ & + \left[n^{-1/2-\kappa}\|e\|^2 M_1 + n^{-(1+\kappa)}\|e\|^3 + n^{-1-\kappa/2}\|e\| M_2\right] A_* \|a\|,\end{aligned}$$

where $M_1 = \|F\|_\infty$ and $M_2 = \max_t \|L' e_t\|$.

Proof. **Proof for part (1).** Since $\hat{F}_t = n^{-1}\hat{L}' W_t = n^{-1}\hat{L}'(LF_t + e_t)$, we have $\hat{L}\hat{F}_t = n^{-1}\hat{L}\hat{L}'(LF_t + e_t)$ and thus $\hat{L}\hat{F}_t - LF_t = (n^{-1}\hat{L}\hat{L}' - I)LF_t + n^{-1}\hat{L}\hat{L}' e_t$. Part (1) follows.

Proof for part (2). By the definition of $\hat{L}$, we have $WW'\hat{L} = \hat{L}\hat{\Sigma}_1^2$ and thus $\hat{L} = WW'\hat{L}\hat{\Sigma}_1^{-2}$. We obtain part (2) by noticing that

$$\begin{aligned}WW'\hat{L}\hat{\Sigma}_1^{-2} &= (LF' + e)W'\hat{L}\hat{\Sigma}_1^{-2} = LF'W'\hat{L}\hat{\Sigma}_1^{-2} + eW'\hat{L}\hat{\Sigma}_1^{-2} \\ &= LF'(FL' + e')\hat{L}\hat{\Sigma}_1^{-2} + eW'\hat{L}\hat{\Sigma}_1^{-2} = LH + (LF'e' + eW')\hat{L}\hat{\Sigma}_1^{-2}.\end{aligned}$$

Proof for part (3). Notice that by Lemma A.3.10(2), $s_r(LF') \ge s_1(L)s_r(F)$. Thus, by assumption (ii), it follows that there exists $b > 0$ such that $\mathbb{P}\left((nT)^{-1/2}s_r(LF') > b\right) \to 1$. By Lemma A.3.10(1), $s_r(W) + \|e\| = s_r(LF' + e) + s_1(-e) \ge s_r(LF')$. Thus,

$$\begin{aligned}\mathbb{P}\left((nT)^{-1/2}s_r(W) + (nT)^{-1/2}\|e\| \ge b\right) &= \mathbb{P}\left(s_r(W) + \|e\| > \sqrt{nT}b\right) \\ &\ge \mathbb{P}\left(s_r(LF') > \sqrt{nT}b\right) \to 1.\end{aligned}$$

Since $\|e\|/\sqrt{nT} = o_P(1)$, $\mathbb{P}\left(s_r(W)/\sqrt{nT} > b/2\right) \to 1$. Notice that $\|\hat{\Omega}_1^{-2}\| = nTs_r^{-2}(W)$. Therefore, $\|\hat{\Omega}_1^{-2}\|$ is bounded above by $4/b^2$ with probability approaching one. In other words, $\|\hat{\Omega}_1^{-2}\| = O_P(1)$.

The definition of Δ_L (in part (2)) implies that

$$\|\Delta_L\| \le (nT)^{-1}\left[\|L\|\|F\|\|e\| + \|e\|\|LF' + e\|\right]\|\hat{L}\|\|\hat{\Omega}_1^{-2}\|$$

$$\leq (nT)^{-1}\Big[2\|L\|\|F\|\|e\| + \|e\|^2\Big]\|\hat{L}\|\|\hat{\Omega}_1^{-2}\| \overset{(i)}{=} O_P(T^{-1/2}\|e\|),$$

where (i) follows by $\|L\| = O_P(n^{1/2})$, $\|F\| = O_P(T^{1/2})$, $\|\hat{L}\| = n^{1/2}$, $\|\hat{\Omega}_1^{-2}\| = O_P(1)$ and $\|e\|/\sqrt{nT} = o_P(1)$. Notice that

$$\|H\| = \|F'F\|\|L\|\|\hat{L}\|\|\hat{\Omega}_1^{-2}\|/(nT) = O_P(T)O_P(n^{1/2})n^{1/2}O_P(1)/(nT) = O_P(1).$$

Observe that

$$\begin{aligned} I_r = n^{-1}\hat{L}'\hat{L} &= n^{-1}(LH + \Delta_L)'(LH + \Delta_L) \\ &= H'\Sigma_L H + n^{-1}H'L'\Delta_L + n^{-1}\Delta_L'LH + n^{-1}\Delta_L'\Delta_L. \quad \text{(A.3.7)} \end{aligned}$$

Also observe that

$$\begin{cases} \|H'L'\Delta_L\| \leq \|H\| \cdot \|L\| \cdot \|\Delta_L\| = O_P(\|e\|\sqrt{n/T}) \\ \|\Delta_L'\Delta_L\| \leq \|\Delta_L\|^2 = O_P(\|e\|^2/T) \overset{(i)}{=} o_P(\|e\|\sqrt{n/T}), \end{cases} \quad \text{(A.3.8)}$$

where (i) holds by $\|e\| = o_P(\sqrt{nT})$. Then (A.3.7) and (A.3.8) imply $H'\Sigma_L H + O_P(\|e\|/\sqrt{nT}) = I_r$. By $O_P(\|e\|/\sqrt{nT}) = o_P(1)$, it follows that $I_r - (H'\Sigma_L H)^{-1} = O_P(\|e\|/\sqrt{nT})$ and thus

$$\begin{aligned} \|HH' - \Sigma_L^{-1}\| &= \|H(I_r - (H'\Sigma_L H)^{-1})H'\| \\ &\leq \|H\| \cdot \|I_r - (H'\Sigma_L H)^{-1}\| \cdot \|H\| = O_P(\|e\|/\sqrt{nT}). \end{aligned}$$

Proof for part (4). Let $A_{n,1} = \|n^{-1}L'L\|\|(HH' - \Sigma_L^{-1})\|$, $A_{n,2} = n^{-1}\|L\|\|\Delta_L\|^2$, $A_{n,3} = n^{-1}\|L\|\|\Delta_L\|\|H\|$, $A_{n,4} = \|n^{-1}L'L\|\|H\|$, $A_{n,5} = (nT)^{-1}\|\hat{\Omega}_1^{-2}\|\|\hat{L}\|$, $A_{n,6} = A_{n,1} + A_{n,3} + A_{n,4}A_{n,5}\|e\|\|F\|$, $A_{n,7} = A_{n,2} + A_{n,4}A_{n,5}\|e\|^2$ and $A_{n,8} = A_{n,4}A_{n,5}\|L\|$. Notice that

$$|a'(\hat{L}\hat{F}_t - LF_t)| \overset{(i)}{\leq} \|L'(n^{-1}\hat{L}\hat{L}' - I)a\|\|F_t\| + n^{-1}\|\hat{L}'a\|\|\hat{L}'e_t\|, \quad \text{(A.3.9)}$$

where (i) holds by part (1). Also notice that

$$\begin{aligned}
&\|L'(n^{-1}\hat{L}\hat{L}' - I)a\| \\
&= \|n^{-1}L'(\hat{L}\hat{L}' - L\Sigma_L^{-1}L')a\| \\
&\overset{(i)}{=} \|n^{-1}L'\left(L(HH' - \Sigma_L^{-1})L' + \Delta_L\Delta_L' + \Delta_L H'L' + LH\Delta_L'\right)a\| \\
&\overset{(ii)}{\leq} A_{n,1}\|L'a\| + A_{n,2}\|a\| + A_{n,3}\|L'a\| + A_{n,4}\|\Delta_L'a\| \\
&\overset{(iii)}{\leq} (A_{n,1} + A_{n,3})\|L'a\| + A_{n,2}\|a\| + A_{n,4}A_{n,5}\left(\|We'a\| + \|e\|\|F\|\|L'a\|\right) \\
&\overset{(iv)}{\leq} (A_{n,1} + A_{n,3})\|L'a\| + A_{n,2}\|a\| \\
&\quad + A_{n,4}A_{n,5}\left(\|L\|\|F'e'a\| + \|e\|^2\|a\| + \|e\|\|F\|\|L'a\|\right) \\
&= A_{n,6}\|L'a\| + A_{n,7}\|a\| + A_{n,8}\|F'e'a\|, \qquad \text{(A.3.10)}
\end{aligned}$$

where (i) follows by $\hat{L} = LH + \Delta_L$, (ii) follows by the triangular inequality and the submultiplicativity of $\|\cdot\|$, (iii) follows by the definition of Δ_L (part (2)) and (iv) follows by $W = LF' + e$. Then,

$$\begin{aligned}
&|a'(\hat{L}\hat{F}_t - LF_t)| \\
&\overset{(i)}{\leq} \sqrt{r}\Big[A_{n,6}\|L'a\| + A_{n,7}\|a\| + A_{n,8}\|F'e'a\|\Big]M_1 + n^{-1}\|\hat{L}'a\|\|\hat{L}'e_t\| \\
&\overset{(ii)}{\leq} \sqrt{r}\Big[A_{n,6}\|L'a\| + A_{n,7}\|a\| + A_{n,8}\|F'e'a\|\Big]M_1 \\
&\quad + n^{-1}\left(\|H\|\|L'a\| + \|\Delta_L\|\|a\|\right)\left(\|H\|M_2 + \|\Delta_L\|\|e_t\|\right) \\
&\overset{(iii)}{\leq} \underbrace{\sqrt{r}A_{n,8}}_{J_1}\|F'e'a\|M_1 + \underbrace{\Big[\sqrt{r}A_{n,6}M_1 + n^{-1}\|H\|^2M_2 + n^{-1}\|H\|\|\Delta_L\|\|e\|\Big]}_{J_2}\|L'a\| \\
&\quad + \underbrace{\Big[\sqrt{r}A_{n,7}M_1 + n^{-1}\|\Delta_L\|^2\|e\| + n^{-1}\|H\|\|\Delta_L\|M_2\Big]}_{J_3}\|a\|, \qquad \text{(A.3.11)}
\end{aligned}$$

where (i) is due to (A.3.9), (A.3.10) and $\|F_t\| \leq \sqrt{r}M_1$, (ii) follows by $\hat{L} = LH + \Delta_L$ and (iii) follows by $\|e_t\| \leq \|e\|$.

By simple computations using part (3) and $T \asymp n^{\kappa}$, we have that

$$\begin{cases} J_1 = O_P(n^{-\kappa}) \\ J_2 = O_P\left(n^{-(1+\kappa)/2}\|e\|M_1 + n^{-1}M_2 + n^{-1-\kappa/2}\|e\|^2\right) \\ J_3 = O_P\left(n^{-1/2-\kappa}\|e\|^2 M_1 + n^{-(1+\kappa)}\|e\|^3 + n^{-1-\kappa/2}\|e\|M_2\right). \end{cases} \tag{A.3.12}$$

Notice that J_1, J_2 and J_3 do not depend on a. Therefore, part (4) follows by (A.3.11) and (A.3.12). The proof is complete. □

Appendix B

Proofs and examples for Chapter 2

B.1 Approximate bootstrap

In this section, we present the key theoretical tool we use, whose proof is presented after the auxiliary lemmas. The proofs for results in Sections 2.2 and 2.3 are contained in Appendix B.2. Appendix B.3 contains technical tools used in the proof.

Proposition B.1.1. *Let $\mathcal{F}_n$ be a σ-algebra and $\{\Upsilon_i\}_{i=1}^n$ a sequence of zero-mean random vectors in $\mathbb{R}^p$ such that Υ_i, conditional on $\mathcal{F}_n$, is independent across i. Let $\hat{S}_n$ and $\hat{\Upsilon}_i$ be random vectors in $\mathbb{R}^p$. Suppose that the following hold:*

(i) There exist constants $q_1, q_2 > 0$ such that $\|\hat{S}_n - S_n^{\Upsilon}\|_\infty = O_{\mathbb{P}}(n^{-q_1})$ and $\max_{1\le j\le p} n^{-1}\sum_{i=1}^n(\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2 = O_{\mathbb{P}}(n^{-q_2})$, where $S_n^{\Upsilon} = n^{-1/2}\sum_{i=1}^n \Upsilon_i$ and $\hat{\Upsilon}_{i,j}$ and Υ_i denote the jth component of $\hat{\Upsilon}_i$ and Υ_i, respectively.

(ii) There exist a constant $r > 2$ and an $\mathcal{F}_n$-measurable positive random variable B_n such that, almost surely, $n^{-1}\sum_{i=1}^n \mathbb{E}(|\Upsilon_{i,j}|^3 \mid \mathcal{F}_n) \le B_n$, $n^{-1}\sum_{i=1}^n \mathbb{E}(|\Upsilon_{i,j}|^4 \mid \mathcal{F}_n) \le B_n^2$ and $\mathbb{E}(\max_{1\le j\le p}|\Upsilon_{i,j}|^r \mid \mathcal{F}_n) \le 2B_n^r$.

(iii) There exists a constant $b > 0$ such that $\mathbb{P}\left(\min_{1\le j\le p} n^{-1}\sum_{i=1}^n \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n) > b\right) \to 1$, $\max_{1\le j\le p}|n^{-1}\sum_{i=1}^n[\Upsilon_{i,j}^2 - \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n)]| = o_{\mathbb{P}}(1)$ and $\max_{1\le j\le p}|n^{-1}\sum_{i=1}^n \Upsilon_{i,j}| = o_{\mathbb{P}}(1)$.

(iv) $n^{2/r-1}B_n^2\log^3(p\vee n)=o_{\mathbb{P}}(1)$, $n^{-1}B_n^2\log^7(p\vee n)=o_{\mathbb{P}}(1)$, $n^{q_2}/\log^4 p\to\infty$ *and* $n^{2q_1}/\log p\to\infty$.

Then

$$\limsup_{n\to\infty}\sup_{\eta\in(0,1)}\left|\mathbb{P}\left[\|\hat{S}_n\|_\infty>\mathcal{Q}\left(1-\eta,\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right)\right]-\eta\right|=0,$$

where $\mathcal{G}_n$ *is the* σ*-algebra generated by* $\mathcal{F}_n$, $\{\Upsilon_i\}_{i=1}^n$ *and* $\{\hat{\Upsilon}_i\}_{i=1}^n$, $\mathcal{Q}\left(1-\eta,\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right)=\inf\left\{x\in\mathbb{R}\mid\mathbb{P}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty>x\mid\mathcal{G}_n\right)\le\alpha\right\}$, $\tilde{S}_n^{\hat{\Upsilon}}=n^{-1/2}\sum_{i=1}^n(\hat{\Upsilon}_i-\bar{\hat{\Upsilon}})\xi_i$, $\bar{\hat{\Upsilon}}=n^{-1}\sum_{i=1}^n\hat{\Upsilon}_i\in\mathbb{R}^p$ *and* $\{\xi_i\}_{i=1}^n$ *is a sequence of independent* $N(0,1)$ *random variables independent of* $\mathcal{G}_n$.

Lemma B.1.1 (Chernozhukov, Chetverikov, and Kato (2014))**.** *Consider the setup in the statement of Proposition B.1.1. Let the assumptions of Proposition B.1.1 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|S_n^{\Upsilon}\|_\infty\le x\mid\mathcal{F}_n\right)-\mathbb{P}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty\le x\mid\mathcal{G}_n\right)\right|=o_{\mathbb{P}}(1).$$

where $\tilde{S}_n^{\Upsilon}=n^{-1/2}\sum_{i=1}^n(\Upsilon_i-\bar{\Upsilon})\xi_i$ *with* $\{\xi_i\}_{i=1}^n$ *and* $\bar{\Upsilon}=(\bar{\Upsilon}_1,\cdots,\bar{\Upsilon}_p)'\in\mathbb{R}^p$ *defined in the statement of Proposition B.1.1.*

Proof. For notational simplicity, we denote $\mathbb{P}(\cdot\mid\mathcal{F}_n)$ and $\mathbb{P}(\cdot\mid\mathcal{G}_n)$ by $\mathbb{P}_{|\mathcal{F}_n}(\cdot)$ and $\mathbb{P}_{|\mathcal{G}_n}(\cdot)$, respectively. Let $\{\Phi_i\}_{i=1}^n$ be a sequence of random elements in $\mathbb{R}^p$ such that conditional on $\mathcal{F}_n$, $\{\Phi_i\}_{i=1}^n$ is independent across i and $\Phi_i\mid\mathcal{F}_n$ is Gaussian with mean zero and variance $\mathbb{E}(\Upsilon_i\Upsilon_i'\mid\mathcal{F}_n)$. Notice that for any $x\in\mathbb{R}$, $\{a\in\mathbb{R}^p\mid\|a\|_\infty\le x\}$ is a rectangle in $\mathbb{R}^p$. By Proposition 2.1 of Chernozhukov, Chetverikov, and Kato (2014) applied to the conditional probability measure $\mathbb{P}_{|\mathcal{F}_n}(\cdot)$, we have

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Upsilon}\|_\infty\le x\right)-\mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Phi}\|_\infty\le x\right)\right|\le C_1(D_{n,1}+D_{n,2})\quad a.s,\quad\text{(B.1.1)}$$

where $C_1>0$ is a constant depending only on b, $S_n^{\Phi}=n^{-1/2}\sum_{i=1}^n\Phi_i$, $D_{n,1}=(n^{-1}B_n^2\log^7(pn))^{1/6}$ and $D_{n,2}=(n^{2r^{-1}-1}B_n^2\log^3(pn))^{1/3}$.

Applying Corollary 4.2 of Chernozhukov, Chetverikov, and Kato (2014) to the conditional probability measure $\mathbb{P}(\cdot\mid\mathcal{F}_n)$, we obtain that, for $\alpha_n=$

$\min\{\exp(-1), n^{1/r-1/2}B_n \log^{3/2}(pn)\}$,

$$\mathbb{P}_{|\mathcal{F}_n}\left[\sup_{x\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Phi}\|_\infty \le x\right) - \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty \le x\right)\right| > C_2(\tilde{D}_{n,1}+\tilde{D}_{n,2})\right] \le \alpha_n \quad a.s, \tag{B.1.2}$$

where $C_2 > 0$ is a constant depending only on b, $\tilde{D}_{n,1} = (n^{-1}B_n^2\log^5(pn)\log^2(\alpha_n^{-1}))^{1/6}$ and $\tilde{D}_{n,2} = (\alpha_n^{-2}n^{2/q-1}B_n^2\log^3(pn))^{1/3}$. Straightforward computations show that α_n, $\tilde{D}_{n,1}$, $\tilde{D}_{n,2}$, $D_{n,1}$ and $D_{n,2}$ are $o_{\mathbb{P}}(1)$. Thus, by (B.1.1) and (B.1.2), we have

$$\begin{cases} \sup_{x\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Upsilon}\|_\infty \le x\right) - \mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Phi}\|_\infty \le x\right)\right| = o_{\mathbb{P}}(1) \\ \sup_{x\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{F}_n}\left(\|S_n^{\Phi}\|_\infty \le x\right) - \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty \le x\right)\right| = o_{\mathbb{P}}(1). \end{cases}$$

The desired result follows. □

Lemma B.1.2. *Let the assumptions of Proposition B.1.1 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty \le x \mid \mathcal{G}_n\right) - \mathbb{P}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty \le x \mid \mathcal{G}_n\right)\right| = o_{\mathbb{P}}(1),$$

where $\tilde{S}_n^{\Upsilon}$ is defined in Lemma B.1.1.

Proof. For notational simplicity, we denote $\mathbb{P}(\cdot \mid \mathcal{G}_n)$ by $\mathbb{P}_{|\mathcal{G}_n}(\cdot)$. Define $\varepsilon_n = n^{-q_2/4}\log^{-1/2} p$. Let $b > 0$ be a constant such that $\mathbb{P}(\min_{1\le j\le p} n^{-1}\sum_{i=1}^n \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n) > b) \to 1$. Since both $\max_{1\le j\le p}|n^{-1}\sum_{i=1}^n[\Upsilon_{i,j}^2 - \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n)]|$ and $\max_{1\le j\le p}|n^{-1}\sum_{i=1}^n \Upsilon_{i,j}|$ are $o_{\mathbb{P}}(1)$, we have $\mathbb{P}(\mathcal{J}_n) \to 1$, where $\mathcal{J}_n = \{\min_{j\in J} n^{-1}\sum_{i=1}^n(\Upsilon_{i,j} - \bar{\Upsilon}_j)^2 > b/2\}$ and $\bar{\Upsilon}_j = n^{-1}\sum_{i=1}^n \Upsilon_{i,j}$. By Lemma B.3.2,

$$\begin{aligned} &\sup_{x\in\mathbb{R}}\left|\mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty > x\right) - \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty > x\right)\right| \\ &\quad \le \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon} - \tilde{S}_n^{\hat{\Upsilon}}\|_\infty > \varepsilon_n\right) + \sup_{x\in\mathbb{R}}\mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty \in (x-\varepsilon_n, x+\varepsilon_n]\right). \end{aligned} \tag{B.1.3}$$

Notice that conditional on $\mathcal{G}_n$, $\tilde{S}_n^{\Upsilon}$ is a zero-mean Gaussian vector whose jth entry has variance of $n^{-1}\sum_{i=1}^n(\Upsilon_{i,j} - \bar{\Upsilon}_j)^2$. Hence, by Lemma B.3.4, there exists a

constant $C_b > 0$ depending only on b such that

$$\sup_{x\in\mathbb{R}} \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty \in (x-\varepsilon_n, x+\varepsilon_n]\right) \leq C_b\varepsilon_n\sqrt{\log p} + \mathbf{1}_{\mathcal{J}_n^c}. \tag{B.1.4}$$

Also notice that conditional on $\mathcal{G}_n$, $\tilde{S}_n^{\Upsilon} - \tilde{S}_n^{\hat{\Upsilon}}$ is a zero-mean Gaussian vector whose jth entry has variance equal to

$$n^{-1}\sum_{i=1}^n \left[(\hat{\Upsilon}_{i,j} - \bar{\hat{\Upsilon}}_j) - (\Upsilon_{i,j} - \bar{\Upsilon}_j)\right]^2 = n^{-1}\sum_{i=1}^n (\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2 - (\bar{\hat{\Upsilon}}_j - \bar{\Upsilon}_j)^2$$
$$\leq n^{-1}\sum_{i=1}^n (\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2.$$

Observe that for any Gaussian random variable $Z \sim N(0,\sigma^2)$ and $x > 0$, $\mathbb{P}(|Z| > x) \leq C\exp(-C\sigma^{-2}x^2)$ for some universal constant $C > 0$. This elementary fact implies that

$$\mathbb{P}_{|\mathcal{G}_n}(\|\tilde{S}_n^{\Upsilon} - \tilde{S}_n^{\hat{\Upsilon}}\|_\infty > \varepsilon_n) \leq \sum_{j=1}^p \mathbb{P}_{|\mathcal{G}_n}(|\tilde{S}_{n,j}^{\Upsilon} - \tilde{S}_{n,j}^{\hat{\Upsilon}}| > \varepsilon_n) \leq pC\exp(-C\varepsilon_n^2\sigma_{n,*}^{-2}), \tag{B.1.5}$$

where $\sigma_{n,*}^2 = \max_{1\leq j\leq p} n^{-1}\sum_{i=1}^n (\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2$. Combining (B.1.3), (B.1.4) and (B.1.5), we have

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty > x\right) - \mathbb{P}_{|\mathcal{G}_n}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty > x\right)\right|$$
$$\leq C_b\varepsilon_n\sqrt{\log p} + pC\exp(-C\varepsilon_n^2\sigma_{n,*}^{-2}) + \mathbf{1}_{\mathcal{J}_n^c}.$$

By assumption, $\sigma_{n,*}^2 = O_{\mathbb{P}}(n^{-q_2})$. Thus, $\varepsilon_n^2\sigma_{n,*}^{-2}/\log p = n^{q_2/2}/\log^2 p \to \infty$ and $p\exp(-C\varepsilon_n^2\sigma_{n,*}^{-2}) = o_{\mathbb{P}}(1)$. Notice that $\varepsilon_n\sqrt{\log p} = n^{-q_2/4} = o(1)$. The desired result follows from the above display, together with these observations and $\mathbb{P}(\mathcal{J}) \to 1$. □

Proof of Proposition B.1.1. We use the notations in the statement of Lemmas B.1.1 and B.1.2. For $x \in \mathbb{R}$, let $Q_n(x) = \mathbb{P}\left(\|S_n^{\Upsilon}\|_\infty > x \mid \mathcal{F}_n\right)$, $\tilde{Q}_n(x) = \mathbb{P}\left(\|\tilde{S}_n^{\Upsilon}\|_\infty > x \mid \mathcal{G}_n\right)$ and $\hat{Q}_n(x) = \mathbb{P}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty > x \mid \mathcal{G}_n\right)$. Define $a_{n,1} =$

$\sup_{x\in\mathbb{R}}|Q_n(x)-\tilde{Q}_n(x)|$ and $a_{n,2}=\sup_{x\in\mathbb{R}}|\tilde{Q}_n(x)-\hat{Q}_n(x)|$. Let $b>0$ be a constant such that $\mathbb{P}(\min_{1\le j\le p} n^{-1}\sum_{i=1}^n \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n) > b) \to 1$. As argued at the beginning of the proof of Lemma B.1.2, $\mathbb{P}(\mathcal{J}_n)\to 1$, where $\mathcal{J}_n := \{\min_{j\in J} n^{-1}\sum_{i=1}^n(\Upsilon_{i,j}-\bar{\Upsilon}_j)^2 \ge b/2\}$ and $\bar{\Upsilon}_j = n^{-1}\sum_{i=1}^n \Upsilon_{i,j}$.

Define the event $\mathcal{J}_n := \{\min_{j\in J} n^{-1}\sum_{i=1}^n(\Upsilon_{i,j}-\bar{\Upsilon}_{n,j})^2 \ge b\}$, where $b>0$ is a constant satisfying $\mathbb{P}(\mathcal{J}_n)\to 1$. Define $s_n = n^{-q_1/2}\log^{-1/4} p$. Notice that, $\forall x\in\mathbb{R}$,

$$\begin{aligned}
&\left|\mathbb{P}\left(\|S_n^\Upsilon\|_\infty \in (x-s_n, x+s_n] \mid \mathcal{F}_n\right) - \mathbb{P}\left(\|\tilde{S}_n^\Upsilon\|_\infty \in (x-s_n, x+s_n] \mid \mathcal{G}_n\right)\right| \\
= &\left|[Q_n(x-s_n)-Q_n(x+s_n)] - \left[\tilde{Q}_n(x-s_n)-\tilde{Q}_n(x+s_n)\right]\right| \\
\le &\left|Q_n(x-s_n)-\tilde{Q}_n(x-s_n)\right| + \left|Q_n(x+s_n)-\tilde{Q}_n(x+s_n)\right| \le 2a_{n,1}. \qquad \text{(B.1.6)}
\end{aligned}$$

Let $\Delta_n = \hat{S}_n - S_n^\Upsilon$. We have

$$\begin{aligned}
&\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - \tilde{Q}_n(x)\right| \\
&\le \left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - Q_n(x)\right| + a_{n,1} \\
&\overset{(i)}{\le} \mathbb{P}\left(\|\Delta_n\|_\infty > s_n \mid \mathcal{F}_n\right) + \mathbb{P}\left(\|S_n^\Upsilon\|_\infty \in (x-s_n, x+s_n] \mid \mathcal{F}_n\right) + a_{n,1} \\
&\overset{(ii)}{\le} \mathbb{P}\left(\|\Delta_n\|_\infty > s_n \mid \mathcal{F}_n\right) + \mathbb{P}\left(\|\tilde{S}_n^\Upsilon\|_\infty \in (x-s_n, x+s_n] \mid \mathcal{G}_n\right) + 3a_{n,1}, \qquad \text{(B.1.7)}
\end{aligned}$$

where (i) follows by Lemma B.3.2 and (ii) follows by (B.1.6). Notice that conditional on $\mathcal{G}_n$, $\tilde{S}_n^\Upsilon$ is a zero-mean Gaussian vector in $\mathbb{R}^p$ whose jth component has variance equal to $n^{-1}\sum_{i=1}^n(\Upsilon_{i,j}-\bar{\Upsilon}_j)^2$. By Lemma B.3.4, there exists a constant $C_b>0$ depending only on b such that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\left(\|\tilde{S}_n^\Upsilon\|_\infty \in (x-s_n, x+s_n] \mid \mathcal{G}_n\right) \le s_n C_b\sqrt{\log p} + \mathbf{1}_{\mathcal{J}_n^c} \quad a.s. \qquad \text{(B.1.8)}$$

Therefore,

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - \hat{Q}_n(x)\right| \\
\le &\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - \tilde{Q}_n(x)\right| + \sup_{x\in\mathbb{R}}|\tilde{Q}_n(x)-\hat{Q}_n(x)| \\
\overset{(i)}{\le} &\mathbb{P}\left(\|\Delta_n\|_\infty > s_n \mid \mathcal{F}_n\right) + s_n C_b\sqrt{\log p} + \mathbf{1}_{\mathcal{J}_n^c} + 3a_{n,1} + a_{n,2}
\end{aligned}$$

$$\overset{(ii)}{=} \mathbb{P}\left(s_n^{-1}\|\Delta_n\|_\infty > 1 \mid \mathcal{F}_n\right) + s_n C_b\sqrt{\log p} + o_{\mathbb{P}}(1) \overset{(iii)}{=} o_{\mathbb{P}}(1), \qquad \text{(B.1.9)}$$

where (i) follows by (B.1.7) and (B.1.8) and the definition of $a_{n,2}$, (ii) follows by $\mathbb{P}(\mathcal{J}_n^c) = o(1)$, $a_{n,1} = o_{\mathbb{P}}(1)$ (by Lemma B.1.1) and $a_{n,2} = o_{\mathbb{P}}(1)$ (by Lemma B.1.2) and (iii) holds by the assumptions: $\|\Delta_n\|_\infty = O_{\mathbb{P}}(n^{-q_1})$, $s_n n^{q_1} = (n^{q_1}\log^{-1/2})^{1/2} \to \infty$ and $s_n\sqrt{\log p} = (n^{q_1}\log^{-1/2})^{-1/2} = o(1)$. Notice that $\forall\delta > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\eta\in(0,1)}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > \mathcal{Q}\left(1-\eta, \|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right) \mid \mathcal{F}_n\right) - \eta\right|\right] \\
&= \mathbb{E}\left[\sup_{\eta\in(0,1)}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > \hat{Q}_n^{-1}(\eta) \mid \mathcal{F}_n\right) - \eta\right|\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\delta + \mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - \hat{Q}_n(x)\right| > \delta \middle| \mathcal{F}_n\right)\right] \\
&= \delta + \mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > x \mid \mathcal{F}_n\right) - \hat{Q}_n(x)\right| > \delta\right) \overset{(ii)}{\leq} \delta + o(1) \qquad \text{(B.1.10)}
\end{aligned}$$

where (i) follows by Lemma B.3.3 and (ii) follows by (B.1.9). Since δ is arbitrary, (B.1.10) implies

$$\mathbb{E}\left[\sup_{\eta\in(0,1)}\left|\mathbb{P}\left(\|\hat{S}_n\|_\infty > \mathcal{Q}\left(1-\eta, \|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right) \mid \mathcal{F}_n\right) - \eta\right|\right] = o(1).$$

The desired result follows by noticing that $\sup_\eta |\mathbb{E}(Z_n(\eta))| \leq \mathbb{E}\sup_\eta |Z_n(\eta)|$, where $Z_n(\eta) = \mathbb{P}\left(\|\hat{S}_n\|_\infty > \mathcal{Q}\left(1-\eta, \|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right) \mid \mathcal{F}_n\right) - \eta$. □

B.2 Proof of results in Sections 2.2 and 2.3

B.2.1 Proof of Lemma 2.2.1.

Proof of Lemma 2.2.1. We first show that $r + p_C = p_Y + p_W$, which is equivalent to the following claims:

(i) $r + p_C \geq p_Y + p_W$.

(ii) $r + p_C \leq p_Y + p_W$.

To see Claim (i), recall that X and F are both assumed to have full column rank. Hence, $r = \text{rank}[F, X] = \text{rank}[M_X F, X] = p_W + \text{rank}M_X F$. Thus, $\text{rank}M_X F = r - p_W$. It follows, by the rank-nullity theorem, that there exist a matrix $Q_1 \in \mathbb{R}^{p_Y \times (p_Y - (r-p_W))}$ such that $M_X F Q_1 = 0$ and $\text{rank}Q_1 = p_Y - (r - p_W) = p_Y + p_W - r$. Since $M_X F Q_1 = [I - X(X'X)^{-1}X']FQ_1$, we have $FQ_1 = XQ_2$, where $Q_2 = (X'X)^{-1}X'FQ_1$. Since F has full column rank, $\text{rank}FQ_1 = \text{rank}Q_1$. Similarly, $\text{rank}XQ_2 = \text{rank}Q_2$. It follows, by $FQ_1 = XQ_2$, that $p_Y + p_W - r = \text{rank}Q_1 = \text{rank}FQ_1 = \text{rank}XQ_2 = \text{rank}Q_2$. By Definition 2.2.1, $p_Y + p_W - r \in \mathcal{C}_{F,X}$. Hence, $p_Y + p_W - r \leq p_C = \max_{k \in \mathcal{C}_{F,X}} k$. This proves Claim (i).

To see Claim (ii), notice that, by Definition 2.2.1, there exists matrices $R_1 \in \mathbb{R}^{p_Y \times p_C}$ and $R_2 \in \mathbb{R}^{p_W \times p_C}$ such that $\text{rank}R_1 = \text{rank}R_2 = p_C$ and $FR_1 = XR_2$. Let $R_{1,C} \in \mathbb{R}^{p_Y \times (p_Y - k)}$ and $R_{2,C} \in \mathbb{R}^{p_W \times (p_W - k)}$ such that matrices $\bar{R}_1 = [R_1, R_{1,C}] \in \mathbb{R}^{p_Y \times p_Y}$ and $\bar{R}_2 = [R_2, R_{2,C}] \in \mathbb{R}^{p_W \times p_W}$ satisfy that $\text{rank}\bar{R}_1 = p_Y$ and $\text{rank}\bar{R}_2 = p_W$. Since $\text{Blockdiag}(\bar{R}_1, \bar{R}_2)$ has full row rank of $p_W + p_Y$, $r = \text{rank}[F, X] = \text{rank}([F, X]\text{Blockdiag}(\bar{R}_1, \bar{R}_2))$. Notice that $[F, X]\text{Blockdiag}(\bar{R}_1, \bar{R}_2) = [FR_1, FR_{1,C}, XR_2, XR_{2,C}]$. Since removing the redudant columns $FR_1 = XR_2$, we have $r = \text{rank}[FR_{1,C}, XR_2, XR_{2,C}] \in \mathbb{R}^{T \times (p_Y + p_W - p_C)}$. Since the rank of a matrix cannot exceed the number of columns, $r \leq p_Y + p_W - p_C$. This proves Claim (ii).

Combing Claims (i) and (ii) yields $r + p_C = p_Y + p_W$. The "if" part in Lemma 2.2.1 follows. To see the "only if" part, it remains to show that $M_Z[F, X] = 0$. Since M_Z represents projection onto the space orthogonal to those spanned by columns in Z. Since columns in Z and those in $[F, X]$ span the same space. We have $M_Z = M_{[F,X]}$. Thus, $M_Z[F, X] = M_{[F,X]}[F, X] = 0$. The proof is complete. □

In the rest of this section, we prove Theorems 2.3.1 and 2.3.2. The proof of these theorems is presented in Appendix B.2.3 after we derive auxiliary results in Appendix B.2.2. We adopt the following notations.

Recall the quantities defined in (2.2.6). Let $\Omega_{Y,(k)} = S_{Y,(k)}/\sqrt{nT}$, where $S_{Y,(k)}$ is the upper-left $k \times k$ matrix of S_Y and $LF' = U_Y S_Y V_Y'$ is an SVD. Similarly, let $\Omega_{W,(p)} = \Sigma_{W,(k)}/\sqrt{nT}$, where $S_{W,(p)}$ is the upper-left $p \times p$ matrix of Σ_W and $RX' = U_W \Sigma_W V_W'$ is an SVD. Recall the SVD's in Algorithm 3: $Y = \hat{U}_Y \hat{S}_Y \hat{V}_Y'$ and

$W = \hat{U}_W \hat{S}_W \hat{V}_W'$. Let $\hat{\Omega}_{Y,(k)} = \hat{S}_{Y,(k)}/\sqrt{nT}$ and $\hat{\Omega}_{W,(k)} = \hat{S}_{W,(k)}/\sqrt{nT}$, where $\hat{S}_{Y,(k)}$ and $\hat{S}_{W,(k)}$ are the upper-left $k \times k$ matrix of $\hat{S}_Y$ and the upper-left $p \times p$ matrix of $\hat{S}_W$, respectively. Also define $\tilde{L} = \sqrt{n} U_{Y,(k)} \in \mathbb{R}^{n \times k}$ and $\tilde{H}_L = (L'L)^{-1} L' \tilde{L}$.

Let $s_j(\cdot)$ denote the jth largest singular value counting multiplicity. For any $j \in [q]$, $\iota_{j,q}$ denotes the jth column of I_q; when there is no ambiguity, we write ι_j instead of $\iota_{j,q}$. We also use the notation $\log^{O(1)} n$ to denote a term $O(\log^c n)$ for some constant $c \in (0, \infty)$.

We define $\Xi(b, r)$ to be the set of random variables with exponential-type tails with parameter (b, r); $\Xi(b, r, p_1, p_2)$ denotes the set of $p_1 \times p_2$ random matrices whose entries belong to $\Xi(b, r)$. We also introduce similar notations for random variables whose conditional distributions have exponential-type tails. For constants $b, r > 0$ and a σ-algebra, we define $\Xi(b, r, \mathcal{F}) = \{\zeta \mid \forall d > 0 \ \mathbb{P}(|\zeta| > d \mid \mathcal{F}) \leq \exp[1 - (d/b)^\gamma] \ a.s\}$. Unless stated otherwise, all the constants in the rest of the paper depend only on β, γ, κ and ρ in Assumption 3.

B.2.2 Preliminary results for Theorems 2.3.1 and 2.3.2

Lemma B.2.1. *Let A and B be matrices of dimension $k_1 \times k_2$ and $k_2 \times k_3$, respectively. Then*
(1) $\|AB\|_\infty \leq \sqrt{k_2} \|A\|_\infty \|B\|$ and $\|AB\|_\infty \leq \sqrt{k_2} \|B\|_\infty \|A\|$.
(2) $\|AB\|_\infty \leq k_2 \|A\|_\infty \|B\|_\infty$.

Proof. For part (1), let $A = (a_1, \cdots, a_{k_1})'$ with $a_i \in \mathbb{R}^{k_2}$. Then $\|AB\|_\infty = \max_{1 \leq i \leq k_1} \|B' a_i\|_\infty \leq \max_{1 \leq i \leq k_1} \|a_i\| \|B\| \leq \sqrt{k_2} \|A\|_\infty \|B\|$. The proof for $\|AB\|_\infty \leq \sqrt{k_2} \|B\|_\infty \|A\|$ is analogous. For part (2), let $B = (b_1, \cdots, b_{k_3})$ with $b_i \in \mathbb{R}^{k_2}$. Then $\|AB\|_\infty = \max_{1 \leq i \leq k_1, 1 \leq j \leq k_3} |a_i' b_j| \leq \max_{1 \leq i \leq k_1, 1 \leq j \leq k_3} \|a_i\| \|b_i\| \leq k_2 \max_{1 \leq i \leq k_1, 1 \leq j \leq k_3} \|a_i\|_\infty \|b_i\|_\infty \leq k_2 \|A\|_\infty \|B\|_\infty$. □

Lemma B.2.2. *Let Assumption 3 hold. Then the following hold. (1) $\|Z\|_\infty$, $\|F\|_\infty$, $\|L\|_\infty$, $\|\Lambda\|_\infty$, $\|e\|_\infty$ and $\|v\|_\infty$ are $O_{\mathbb{P}}(\log^{O(1)} n)$. (2) $\|Z\|$, $\|F\|$, $\|L\|$, $\|\Lambda\|$, $\|\hat{L}\|$, $\|\hat{\Lambda}\|$ and $\|\tilde{L}\|$ are $O_{\mathbb{P}}(n^{1/2})$. (3) $\|e\| = O_{\mathbb{P}}(\sqrt{n \log n})$ and $\|v\| = O_{\mathbb{P}}(\sqrt{n \log n})$.*

Proof. Part (1) follows by the exponential-type tail condition together with Lemma B.3.6. Part (2) follows by Assumption 3 and definitions of $\hat{L}$, $\hat{\Lambda}$ and $\tilde{L}$. Now we

prove part (3). By Theorem 5.48 and Remark 5.49 in Vershynin (2010), there exists a universal constant $\bar{C} > 0$ such that

$$\mathbb{E}\|e\| \leq \sqrt{n_Y \max_{i\in[n_Y]} s_i^2} + \bar{C}\sqrt{\mathbb{E}\max_i \|e_i\|^2 \log(n_Y \wedge T)}, \tag{B.2.1}$$

where $s_i^2 = \mathbb{E}\|e_i\|^2 = \sum_{t=1}^T \mathbb{E}e_{i,t}^2$. By Assumption 3 and Lemma B.3.5, $\max_{i\in[n_Y]} s_i^2 = O(n)$.

By Lemma B.3.5(3)-(4), there exists a constant $b > 0$ such that $e_{i,t}^2 - \mathbb{E}e_{i,t}^2 \in \Xi(b, \gamma_1/3)$. Let $a_n = c_*\sqrt{n \log n}$, where $c_* > 0$ is a constant to be determined. By Theorem 1 in Merlevède, Peligrad, and Rio (2011), there exists a constant $C > 0$ such that $\forall x > 0$,

$$\begin{aligned}
\mathbb{P}\left(\max_{i\in[n_Y]} \left|\|e_i\|^2 - s_i^2\right| > a_n x\right) &\leq \sum_{i=1}^{n_Y} \mathbb{P}\left(\left|\|e_i\|^2 - s_i^2\right| > a_n x\right) \\
&= \sum_{i=1}^{n_Y} \mathbb{P}\left(\left|\sum_{t=1}^T \left(e_{i,t}^2 - \mathbb{E}e_{i,t}^2\right)\right| > a_n x\right) \\
&\leq n_Y T \exp\left(-C a_n^\gamma x^\gamma\right) + n_Y \exp\left(-\frac{C a_n^2 x^2}{1 + CT}\right) \\
&\quad + n_Y \exp\left[-\frac{C a_n^2 x^2}{T} \exp\left(C(a_n x)^{\gamma/(1-\gamma)} (\log a_n x)^{-\gamma}\right)\right].
\end{aligned}$$

Thus, by elementary computations, we can choose large constants $c_*, a_* > 0$ and small constants $b_* > 0$ such that $\forall x \geq a_*$

$$\mathbb{P}\left(\max_i \left|\|e_i\|^2 - s_i^2\right| / \left(c_*\sqrt{n \log n}\right) > x\right) \leq \exp\left(-b_* x^2\right). \tag{B.2.2}$$

By (B.2.2) and the identity $\mathbb{E}|\zeta| = \int_0^\infty \mathbb{P}(|\zeta| > z)dz$ for any random variable ζ, we have

$$\begin{aligned}
&\mathbb{E}\max_i \left|\|e_i\|^2 - s_i^2\right| / \left(c_*\sqrt{n_Y \log n_Y}\right) \\
&= \int_0^\infty \mathbb{P}\left(\max_i \left|\|e_i\|^2 - s_i^2\right| / \left(c_*\sqrt{n_Y \log n_Y}\right) > x\right) dx \\
&\leq a_* + \int_{a_*}^\infty \mathbb{P}\left(\max_i \left|\|e_i\|^2 - s_i^2\right| / \left(c_*\sqrt{n \log n}\right) > x\right) dx
\end{aligned}$$

$$\leq a_* + \int_{a_*}^{\infty} \exp\left(-b_* x^2\right) dx = O(1).$$

It follows that $\mathbb{E}\max_i \|e_i\|^2 \leq \mathbb{E}\max_i |\|e_i\|^2 - s_i^2| + \max_i s_i^2 = O(\sqrt{n\log n}) + O(n) = O(n)$. Thus, by (B.2.1) and Markov's inequality, $\|e\| = O_{\mathbb{P}}(\sqrt{n\log n})$. An analogous argument yields $\|v\| = O_{\mathbb{P}}(\sqrt{n\log n})$. □

Lemma B.2.3. *Let Assumption 3 hold. Then*
(1) $\|e'L\|_\infty = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$ *and* $\|v'\Lambda\|_\infty = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$
(2) $\|F'e'L\|_\infty$, $\|Z'e'L\|_\infty$ *and* $\|Z'v'\Lambda\|_\infty$ *are* $O_{\mathbb{P}}(n\log^{O(1)} n)$.
(3) $\|e'eF\|_\infty = O_{\mathbb{P}}(n\log^{O(1)} n)$
(4) $\|eF\|_\infty = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$ *and* $\|v\Lambda\|_\infty = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$.

Proof. We show these results by repeatedly applying Lemmas B.3.8 and B.3.7.

Proof of part (1). For $j = (j_1, j_2) \in J = [T] \times [p_Y]$, we define $\mu_{i,j} = e_{i,j_1} L_{i,j_2}$, where e_{i,j_1} is the j_1th entry of e_i and L_{i,j_2} is the j_2th entry of L_i. Let $\zeta_n = \|L\|_\infty$ and $\mathcal{F}_n$ be the σ-algebra generated by L. Since $e_{i,j_1} \in \Xi(\beta,\gamma)$ (by Assumption 3), we have $\mu_{i,j}/\zeta_n \in \Xi(\beta,\gamma,\mathcal{F}_n)$. Notice that $|J| = p_Y T = O(n)$. Hence, by Lemmas B.2.2 and B.3.8, we have $\|e'L\|_\infty = \zeta_n \max_{j\in J} |\sum_{i=1}^{n_Y} \mu_{i,j}/\zeta_n| = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$. The result for $\|v'\Lambda\|_\infty$ follows by an analogous argument.

Proof of part (2). For $j = (j_1, j_2) \in J = [p_Y] \times [p_Y]$, we define $\mu_{i,j} = d_{i,j_1} L_{i,j_2} \|L\|_\infty^{-1}$ and $d_{i,j_1} = T^{-1/2}\sum_{t=1}^{T} h_{i,t,j_1}$ with $h_{i,t,j_1} = e_{i,t}F_{t,j_1}\|F\|_\infty^{-1}$, where F_{t,j_1} is the j_1th entry of F_t. Let $\mathcal{F}_n$ be the σ-algebra generated by F and L.

By Assumption 3, $h_{i,t,j_1} \in \Xi(\beta,\gamma,\mathcal{F}_n)$. By the independence between $e_{i,t}$ and (F, L), the mixing condition in Assumption 3 implies the mixing condition in the statement of Lemma B.3.7, by which it follows that there exist constants $b_1, r_1 > 0$ such that $d_{i,j_1} \in \Xi(b_1, r_1, \mathcal{F}_n)$. Since $|\mu_{i,j}| \leq |d_{i,j_1}|$, $\mu_{i,j} \in \Xi(b_1, r_1, \mathcal{F}_n)$. Since $\mu_{i,j}$ is, conditional on $\mathcal{F}_n$, independent across i, we can apply Lemma B.3.8 and obtain that $\max_{j\in J} |\sum_{i=1}^{n_Y} \mu_{i,j}| = O_{\mathbb{P}}(n)$. This, together with Lemma B.2.2, implies $\|F'e'L\|_\infty \leq T^{1/2}\max_{j\in J} |\sum_{i=1}^{n_Y} \mu_{i,j}| \|F\|_\infty \|L\|_\infty = O_{\mathbb{P}}(n\log^{O(1)} n)$. The results for $\|Z'e'L\|_\infty$ and $\|Z'v'\Lambda\|_\infty$ follow by analogous arguments.

Proof of part (3). For $j = (j_1, j_2) \in J = [T] \times [p_Y]$, we define $\mu_{i,j} = e_{i,j_1} d_{i,j_2}$ and $d_{i,j_2} = T^{-1/2}\sum_{t=1}^{T} e_{i,t}F_{t,j_2}\|F\|_\infty^{-1}$. Let $\mathcal{F}_n$ be the σ-algebra generated by F.

Similar to previous arguments, we have $e_{i,t}F_{t,j_2}\|F\|_\infty^{-1} \in \Xi(b_2, r_2, \mathcal{F}_n)$ for some constants $b_2, r_2 > 0$. By Lemma B.3.7, $d_{i,j_2} \in \Xi(b_3, r_3, \mathcal{F}_n)$ for some constants $b_3, r_3 > 0$. By Lemma B.3.5 (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), it follows that $\mu_{i,j} \in \Xi(b_3, r_3, \mathcal{F}_n)$, where $b_4, r_4 > 0$ are constants. By Lemma B.3.8, $T^{-1/2}\|F\|_\infty^{-1}\|e'eF - \mathbb{E}(e'e)F\|_\infty = \max_{j\in J} |\sum_{i=1}^{n_Y}[\mu_{i,j} - \mathbb{E}(\mu_{i,j} \mid \mathcal{F}_n)]| = O_\mathbb{P}(n\log^{1/2} n)$. By Lemma B.2.2, $\|e'eF - \mathbb{E}(e'e)F\|_\infty = O_\mathbb{P}(n^{-1}\log^{O(1)} n)$. By Holder's inequality,

$$\begin{aligned}
\|\mathbb{E}(e'e)F\|_\infty &= \max_{j\in J}\left\|\sum_{i=1}^{n_Y}\sum_{s=1}^{T}(\mathbb{E}e_{i,j_1}e_{i,s})F_{s,j_2}\right\|_\infty \\
&\le \|F\|_\infty \max_{j\in J}\sum_{s=1}^{T}\sum_{i=1}^{n_Y}|\mathbb{E}e_{i,j_1}e_{i,s}| \\
&\le \|F\|_\infty \max_{(i,t)\in[n_Y]\times[T]} n_Y\sum_{s=1}^{T}|\mathbb{E}e_{i,t}e_{i,s}| \overset{(i)}{=} O_\mathbb{P}(n\log^{O(1)} n),
\end{aligned}$$

where (i) follows by Assumption 3 and Lemma B.2.2. Hence, $\|e'eF\|_\infty = O_\mathbb{P}(n\log^{O(1)} n)$.

Proof of part (4). By the law of iterated expectations and the proof of part (3), we have that $T^{-1/2}\sum_{t=1}^{T} e_{i,t}F_{t,j_1}\|F\|_\infty^{-1} \in \Xi(b_1, r_1)$, where $b_3, r_3 > 0$ are constants defined in the proof of part (3). Then by Lemmas B.2.2 and B.3.6,

$$\begin{aligned}
\|eF\|_\infty &= T^{1/2}\|F\|_\infty \max_{1\le i\le n_Y}\left|T^{-1/2}\sum_{t=1}^{T} e_{i,t}F_{t,j_1}\|F\|_\infty^{-1}\right| \\
&= T^{1/2}\|F\|_\infty O_\mathbb{P}(\log^{O(1)} n) = O_\mathbb{P}(n^{1/2}\log^{O(1)} n).
\end{aligned}$$

The result for $\|v\Lambda\|_\infty$ follows by an analogous argument. The proof is complete. □

Lemma B.2.4. *Under Assumption 3, the following hold:*
(1) $\hat{L} = LH_L + \Delta_L$, *where* $H_L = F'FL'\hat{L}\hat{\Omega}_1^{-2}(n_YT)^{-1}$ *and* $\Delta_L = (n_YT)^{-1}(LF'e' + eY')\hat{L}\hat{\Omega}_1^{-2}$.
(2) $\hat{F} = FH_F + \Delta_F$, *where* $H_F = L'\hat{L}/n_Y$ *and* $\Delta_F = n_Y^{-1}e'\hat{L}$.
(3) $\|\Omega_1\| = O_\mathbb{P}(1)$, $\|\Omega_1^{-1}\| = O_\mathbb{P}(1)$ *and* $\|\hat{\Omega}_1 - \Omega_1\| = O_\mathbb{P}(n^{-1/2}\log^{O(1)} n)$
(4) $\|\Delta_L\| = O_\mathbb{P}(\log^{O(1)} n)$, $\|H_L\| = O_\mathbb{P}(1)$, $\|H_LH_L' - \Sigma_L^{-1}\| = O_\mathbb{P}(n^{-1/2}\log^{O(1)} n)$,

$\|H_F^{-1}\| = O_{\mathbb{P}}(1)$ *and* $\|\Delta_F\| = O_{\mathbb{P}}(\log^{O(1)} n)$.

Proof. By the definition of $\hat{L}$, we have $YY'\hat{L} = \hat{L}\hat{\Sigma}_1^2$ and thus $\hat{L} = YY'\hat{L}\hat{\Sigma}_1^{-2}$. Notice that

$$\begin{aligned} YY'\hat{L}\hat{\Sigma}_1^{-2} = (LF' + e)Y'\hat{L}\hat{\Sigma}_1^{-2} &= LF'Y'\hat{L}\hat{\Sigma}_1^{-2} + eY'\hat{L}\hat{\Sigma}_1^{-2} \\ &= LF'(FL' + e')\hat{L}\hat{\Sigma}_1^{-2} + eY'\hat{L}\hat{\Sigma}_1^{-2} \\ &= LH_L + (LF'e' + eY')\hat{L}\hat{\Sigma}_1^{-2}. \end{aligned}$$

Part (1) follows. Part (2) follows by $\hat{F} = Y'\hat{L}/n_Y = (FL' + e')\hat{L}/n_Y$.

Notice that $\|\Omega_1\| = (n_Y T)^{-1/2}\|FL'\| \leq (n_Y T)^{-1/2}\|F\|\|L\| = O_{\mathbb{P}}(1)$. Notice that by Lemma B.3.1(2), $\sqrt{n_Y T}\Omega_{1,k} = s_k(FL') \geq s_1(F)s_k(L)$, where $\Omega_{1,i}$ is the ith entry on the diagonal of Ω_1. Thus, by Assumption 3, it follows that there exists $b > 0$ such that $\mathbb{P}(\Omega_{1,k} > b) \to 1$. Hence, $\|\Omega_1^{-1}\| = O_{\mathbb{P}}(1)$.

For any $1 \leq j \leq k$, $\sqrt{n_Y^{-1}T}\Omega_{1,j} + \|e\| = s_j(FL') + s_1(e) \geq s_j(FL' + e) = \hat{\Omega}_{1,j}\sqrt{n_Y T}$, where $\hat{\Omega}_{1,j}$ denote the jth entry on the diagonal of $\hat{\Omega}_1$. By Lemma B.2.2, $\Omega_{1,j} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \geq \hat{\Omega}_{1,j}$. Similarly, we use $s_j(Y) + s_1(-e) \geq s_j(Y - e)$ to obtain that $\hat{\Omega}_{1,j} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \geq \Omega_{1,j}$. Hence, $\|\hat{\Omega}_1 - \Omega_1\| = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. Part (3) follows.

It remains to show part (4). Notice that

$$\|\Delta_L\| \leq (n_Y T)^{-1}\left[\|L\|\|F\|\|e\| + \|e\|\left(\|LF' + e\|\right)\right]\|\hat{L}\|\|\hat{\Omega}_1^{-2}\| = O_{\mathbb{P}}(\log^{O(1)} n),$$

where the equality follows by Lemma B.2.2 and $\|\hat{\Omega}_1^{-2}\| = O_{\mathbb{P}}(1)$. Notice that $\|H_L\| = \|F'F\|\|L\|\|\hat{L}\|\|\hat{\Omega}_1^{-2}\|/(n_Y^{-1}T) = O_{\mathbb{P}}(T)O_{\mathbb{P}}(n_Y^{1/2})n_Y^{1/2}O_{\mathbb{P}}(1)/(n_Y T) = O_{\mathbb{P}}(1)$.

Notice that

$$I = n_Y^{-1}\hat{L}'\hat{L} = n_Y^{-1}(LH_L + \Delta_L)'(LH_L + \Delta_L) \overset{(i)}{=} H_L'\Sigma_L H_L + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),$$

where (i) follows by $\|H_L\| = O_{\mathbb{P}}(1)$, $\|\Delta_L\| = O_{\mathbb{P}}(\log^{O(1)} n)$ and $\|e\| = O_{\mathbb{P}}(n^{1/2}\log^{O(1)} n)$. Use singular value inequalities to say that the singular values of H_L is bounded away from zero and thus $\|H_F^{-1}\| = O_{\mathbb{P}}(1)$. Hence,

$I - (H_L'\Sigma_L H_L)^{-1} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ and

$$\begin{aligned}\|H_L H_L' - \Sigma_L^{-1}\| &= \|H_L(I - (H_L'\Sigma_L H_L)^{-1})H_L'\| \\ &\leq \|H_L\| \cdot \|I - (H_L'\Sigma_L H_L)^{-1}\| \cdot \|H_L\| = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).\end{aligned}$$

This, together with $H_F = \Sigma_L H_L + L'\Delta_L/n_Y = \Sigma_L H_L + o_{\mathbb{P}}(1)$, implies that $\|H_F^{-1}\| = O_{\mathbb{P}}(1)$. By Lemma B.2.2, $\|\Delta_F\| \leq n_Y^{-1}\|e\|\|\hat{L}\| = O_{\mathbb{P}}(\log^{O(1)} n)$. Part (4) follows. □

Lemma B.2.5. *Under Assumption 3, there exists a diagonal (possibly random) matrix $D_{n,F}$ whose diagonal entries take values in $\{-1, 1\}$ such that $H_L = (\tilde{H}_L'\Sigma_L)^{-1}D_{n,F} + O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$. Moreover, $H_F D_{n,F} - (\tilde{H}_L')^{-1} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$.*

Proof. We proceed in two steps. In the first step, we show that $n_Y^{-1}\hat{L}'\tilde{L}$ is approximately a diagonal matrix. In the second step, the square of this diagonal matrix is shown to be I_k.

Step 1: show that $n_Y^{-1}\hat{L}'\tilde{L}$ is approximately diagonal. Notice that by definition, we have $LF'FL'\tilde{L} = n_Y^{-1}T\tilde{L}\Omega_1^2$ and $\hat{L}'YY' = n_Y^{-1}T\hat{\Omega}_1^2\hat{L}'$. Thus, (a) $\hat{L}'LF'FL'\tilde{L} = n_Y^{-1}T\hat{L}'\tilde{L}\Omega_1^2$ and (b) $\hat{L}'YY'\tilde{L} = n_Y^{-1}T\hat{\Omega}_1^2\hat{L}'\tilde{L}$. Plugging $Y = FL' + e$ into (b) and using (a), we obtain

$$\hat{\Omega}_1^2\hat{L}'\tilde{L}n_Y^{-1} - n_Y^{-1}\hat{L}'\tilde{L}\Omega_1^2 = n_Y^{-2}T\hat{L}'(eFL' + LF'e' + ee')\tilde{L} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),$$

where the last equality follows by Lemma B.2.2. Thus, since $\hat{\Omega}_1^2 = \Omega_1^2 + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ (Lemma B.2.4), we have

$$n_Y^{-1}\hat{L}'\tilde{L}\Omega_1^2 + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) = n_Y^{-1}\Omega_1^2\hat{L}'\tilde{L}.$$

Let $A = n_Y^{-1}\hat{L}'\tilde{L}$ with its (i,j) entry denoted by $A_{i,j}$. Also let $\Omega_{1,j}^2$ denote the jth entry on the diagonal of Ω_1^2. Then the above equation implies that $\forall(i,j) \in [k] \times [k]$ with $i \neq j$, $A_{i,j}(\Omega_{1,j}^2 - \Omega_{1,i}^2) = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. By the distinctive singular value assumption, it follows that $A_{i,j} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ for

$i \neq j$. Hence, there exists a diagonal matrix $\tilde{D} = \text{diag}(\tilde{D}_1, \cdots, \tilde{D}_k)$ such that

$$n_Y^{-1}\hat{L}'\tilde{L} = \tilde{D} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n). \tag{B.2.3}$$

Step 2: show the square of the diagonal matrix is approximate I_{p_Y}. Observe

$$n_Y^{-1}H_L'L'\tilde{L} = n_Y^{-1}\hat{L}'\tilde{L} - n_Y^{-1}\Delta_L'\tilde{L} = \tilde{D} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),$$

where both equalities follow by Lemma B.2.4 since $LH_L = \hat{L} - \Delta_L$ and $\|\Delta_L'\tilde{L}\| \leq \|\Delta_L\| \cdot \|\tilde{L}\|$. Hence, $(n_Y^{-1}H_L'L'\tilde{L})'(n_Y^{-1}H_L'L'\tilde{L}) = \tilde{D}^2 + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$.

On the other hand, by Lemma B.2.4(3),

$$\begin{aligned}(n_Y^{-1}H_L'L'\tilde{L})'(n_Y^{-1}H_L'L'\tilde{L}) &= (n_Y^{-1}\tilde{L}'L)(H_LH_L')(n_Y^{-1}L'\tilde{L}) \\ &= (n_Y^{-1}\tilde{L}'L)\Sigma_L^{-1}(n_Y^{-1}L'\tilde{L}) + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).\end{aligned}$$

Thus,

$$n_Y^{-1}\tilde{L}'(n_Y^{-1}L\Sigma_L^{-1}L')\tilde{L} = \tilde{D}^2 + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \tag{B.2.4}$$

Let $M_L = I_{n_Y} - L(L'L)^{-1}L'$. Since $M_LL = 0$ and $\tilde{L} = L\tilde{H}_L$, we have $\tilde{L}'M_L\tilde{L} = 0$. By the definition of M_L, we have $\tilde{L}'\tilde{L} = n_Y^{-1}\tilde{L}'L\Sigma_L^{-1}L'\tilde{L}$ and thus $n_Y^{-1}\tilde{L}'(n_Y^{-1}L\Sigma_L^{-1}L')\tilde{L} = I_k$. By (B.2.4), $\tilde{D}^2 = I_k + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. Then there exists a diagonal matrix $D_{n,F}$ such that

$$D_{n,F}^2 = I_{p_Y} \qquad \text{and} \qquad \tilde{D} = D_{n,F} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n). \tag{B.2.5}$$

Step 3: show the desired result. It follows, by (B.2.3) and (B.2.5), that $n_Y^{-1}\hat{L}'\tilde{L} = D_{n,F} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. Using $\hat{L} = LH_L + \Delta_L$ and $\|\Delta_L\| = O_{\mathbb{P}}(\log^{O(1)} n)$ from Lemma B.2.4, we have $H_L'\Sigma_L\tilde{H}_L - D_{n,F} = n_Y^{-1}\Delta_L'\tilde{L} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. In other words, $H_L = (\tilde{H}_L'\Sigma_L)^{-1}D_{n,F} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. Notice that the diagonal entries of $D_{n,F}$ are either -1 or 1 (since $D_{n,F}^2 = I_{p_Y}$). The first claim follows.

It follows, by Lemma B.2.4, that

$$\begin{aligned} H_F D_{n,F} - (\tilde{H}'_L)^{-1} &= n_Y^{-1} L' \hat{L} D_{n,F} - (\tilde{H}'_L)^{-1} \\ &= n_Y^{-1} L'(LH_L + \Delta_L) D_{n,F} - (\tilde{H}'_L)^{-1} \\ &= \Sigma_L H_L D_{n,F} - (\tilde{H}'_L)^{-1} + n_Y^{-1} L' \Delta_L D_{n,F} \\ &\overset{(i)}{=} O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n), \end{aligned}$$

where (i) holds by $n_Y^{-1}\|L'\Delta_L D_{n,F}\| \leq n_Y^{-1}\|L\| \cdot \|\Delta_L\| \cdot \|D_{n,F}\| = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ (Lemma B.2.4) and the first claim: $\Sigma_L H_L D_{n,F} = (\tilde{H}'_L)^{-1} D_{n,F}^2 + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ with $D_{n,F}^2 = I_{p_Y}$. This proves the second claim. □

Lemma B.2.6. *Let Assumption 3 hold. Then (1)* $\|\Delta_L\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ *and (2)* $\|e'\Delta_L\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$*, where* Δ_L *is defined in Lemma B.2.4(1).*

Proof. To see part (1), notice that by the definition of Δ_L and $Y = LF' + e$, we have

$$\begin{aligned} \|\Delta_L\|_\infty &\leq (n_Y T)^{-1} \left[\|LF'e'\hat{L}\hat{\Omega}_1^{-2}\|_\infty + \|eFL'\hat{L}\hat{\Omega}_1^{-2}\|_\infty + \|ee'\hat{L}\hat{\Omega}_1^{-2}\|_\infty \right] \\ &\overset{(i)}{\leq} (n_Y T)^{-1} \Big[\|LF'e'LH_L\hat{\Omega}_1^{-2}\|_\infty + \|LF'e'\Delta_L\hat{\Omega}_1^{-2}\|_\infty + \|eFL'\hat{L}\hat{\Omega}_1^{-2}\|_\infty \\ &\qquad + \|ee'\hat{L}\hat{\Omega}_1^{-2}\|_\infty \Big] \\ &\overset{(ii)}{\leq} (n_Y T)^{-1} \Big[p_Y^2 \|L\|_\infty \|F'e'L\|_\infty \|H_L\hat{\Omega}_1^{-2}\|_\infty + \|LF'e'\Delta_L\hat{\Omega}_1^{-2}\| \\ &\qquad + \sqrt{p_Y}\|eF\|_\infty \|L'\hat{L}\hat{\Omega}_1^{-2}\| + \|ee'\hat{L}\hat{\Omega}_1^{-2}\| \Big] \\ &\overset{(iii)}{\leq} (n_Y T)^{-1} \Big[p_Y^2 \|L\|_\infty \|F'e'L\|_\infty \|H_L\hat{\Omega}_1^{-2}\|_\infty + \|L\|\|F\|\|e\|\|\Delta_L\|\|\hat{\Omega}_1^{-2}\| \\ &\qquad + \sqrt{p_Y}\|eF\|_\infty \|L\| \cdot \|\hat{L}\| \cdot \|\hat{\Omega}_1^{-2}\| + \|e\|^2 \|\hat{L}\| \cdot \|\hat{\Omega}_1^{-2}\| \Big] \\ &\overset{(iv)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n), \end{aligned}$$

where (i) and (ii) follow by and the elementary inequality $\|A\|_\infty \leq \|A\|$, Lemmas B.2.4 and B.2.1; (iii) follows by the sub-multiplicity of the spectral norm; finally, (iv) follows by Lemmas B.2.2, B.2.4 and B.2.3. We have proved part (1).

The argument for part (2) is similar. By the definition of Δ_L and $Y = LF' + e$,

we have

$$\begin{aligned}
\|e'\Delta_L\|_\infty &= \|(n_Y T)^{-1} e'(LF'e' + eFL' + ee')\hat{L}\hat{\Omega}_1^{-2}\|_\infty \\
&\leq (n_Y T)^{-1}\left[\|e'LF'e'\hat{L}\hat{\Omega}_1^{-2}\|_\infty + \|e'eFL'\hat{L}\hat{\Omega}_1^{-2}\|_\infty + \|e'ee'\hat{L}\hat{\Omega}_1^{-2}\|_\infty\right] \\
&\overset{(i)}{\leq} (n_Y T)^{-1}\Big[\sqrt{p_Y}\|e'L\|_\infty\|F'e'\hat{L}\hat{\Omega}_1^{-2}\| + \sqrt{p_Y}\|e'eF\|_\infty\|L'\hat{L}\hat{\Omega}_1^{-2}\| \\
&\qquad\qquad + \|e'ee'\hat{L}\hat{\Omega}_1^{-2}\|\Big] \\
&\overset{(ii)}{\leq} (n_Y T)^{-1}\Big[\sqrt{p_Y}\|e'L\|_\infty\|F\| \cdot \|e\| \cdot \|\hat{L}\| \cdot \|\hat{\Omega}_1^{-2}\| \\
&\qquad\qquad + \sqrt{p_Y}\|e'eF\|_\infty\|L\| \cdot \|\hat{L}\| \cdot \|\hat{\Omega}_1^{-2}\| + \|e\|^3\|\hat{L}\| \cdot \|\hat{\Omega}_1^{-2}\|\Big] \\
&\overset{(iii)}{=} O_{\mathbb{P}}(\log^{O(1)} n),
\end{aligned}$$

where (i) follows by Lemma B.2.1 and the elementary inequality $\|A\|_\infty \leq \|A\|$, (ii) holds by the sub-multiplicity of the spectral norm and (iii) follows by Lemmas B.2.2, B.2.4 and B.2.3. This proves part (2). The proof is complete. □

Lemma B.2.7. *Under Assumption 3, the following hold:*
(1) $\hat{\Lambda} = \Lambda H_\Lambda + \Delta_\Lambda$, *where* $\|H_\Lambda\| = O_{\mathbb{P}}(1)$, $\|\Delta_\Lambda\| = O_{\mathbb{P}}(\log^{O(1)} n)$, $\|\Delta_\Lambda\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ *and* $\|v'\Delta_\Lambda\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$
(2) $\hat{Z} = ZH_Z + \Delta_Z$ *and* $\|H_Z^{-1}\| = O_{\mathbb{P}}(1)$, *where* $H_Z = \Lambda'\hat{\Lambda}/n$ *and* $\Delta_Z = n^{-1}v'\hat{\Lambda}$.
(3) $\bar{Z} = Z + \bar{\Delta}_Z$, *where* $\bar{Z} = ZH_Z^{-1}$ *and* $\bar{\Delta}_Z = \Delta_Z H_Z^{-1}$.
(4) $\|\bar{\Delta}_Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty \leq O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$, $\|\bar{\Delta}_Z\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$, $\|\bar{\Delta}_Z\| = O_{\mathbb{P}}(\log^{O(1)} n)$, $\|Z'\bar{\Delta}_Z\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$, $\bar{\Sigma}_Z - \Sigma_Z = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$, *where* $\bar{\Sigma}_Z = \bar{Z}'\bar{Z}/T$.

Proof. Part (1) and part (2) follow by the same argument as in Lemmas B.2.4 and B.2.6, except that (L, F, e, p_Y) is replaced by (Λ, Z, v, r). Part (3) follows by part (2).

It remains to show part (4). By Lemma B.2.1,

$$\begin{aligned}
&\|\bar{\Delta}_Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty \\
&\leq \|\Delta_Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}H_Z\|_\infty\|H_Z^{-1}\|_\infty r \\
&\overset{(i)}{=} \|n^{-1}v'\hat{\Lambda} - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}H_Z\|_\infty\|H_Z^{-1}\|_\infty r
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{=} \|n^{-1}v'\Delta_\Lambda - n^{-2}v'\Lambda\Sigma_\Lambda^{-1}\Lambda'\Delta_\Lambda\|_\infty \|H_Z^{-1}\|_\infty r \\
&\leq \left(n^{-1}\|v'\Delta_\Lambda\|_\infty + n^{-2}r^2\|v'\Lambda\|_\infty\|\Sigma_\Lambda^{-1}\|_\infty\|\Lambda\|\|\Delta_\Lambda\|\right)\|H_Z^{-1}\|_\infty r \\
&\overset{(iii)}{=} O_{\mathbb{P}}(n^{-1}\log^{O(1)} n),
\end{aligned}
$$

where (i) and (ii) follow by the expressions for $\hat{\Lambda}$, H_Z and Δ_Z from parts (1)-(2) and (iii) follows by parts (1)-(2), together with Lemmas B.2.2 and B.2.3.

By Lemmas B.2.1 and B.2.3, $\|n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty \leq \sqrt{r}\|n^{-1}v'\Lambda\|_\infty\|\Sigma_\Lambda^{-1}\| = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$, we have $\|\bar{\Delta}_Z\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. Notice that $\|\bar{\Delta}_Z\| \leq \sqrt{Tr}\|\bar{\Delta}_Z\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$. By parts (1)-(2) and Lemma B.2.1, we have

$$
\begin{aligned}
\|Z'\bar{\Delta}_Z\|_\infty &= \|n^{-1}Z'v'(\Lambda H_\Lambda + \Delta_\Lambda)\|_\infty \\
&\leq \|n^{-1}Z'v'\Lambda H_\Lambda\|_\infty + \|n^{-1}Z'v'\Delta_\Lambda\| \\
&\leq n^{-1}\|Z'v'\Lambda\|_\infty\|H_\Lambda\|\sqrt{r} + n^{-1}\|Z\|\cdot\|v\|\cdot\|\Delta_\Lambda\| \\
&\overset{(i)}{=} O_{\mathbb{P}}(\log^{O(1)} n),
\end{aligned}
$$

where (i) holds by part (1) and Lemmas B.2.2 and B.2.3. Notice that

$$
\bar{\Sigma}_Z - \Sigma_Z = T^{-1}(Z+\bar{\Delta}_Z)'(Z+\bar{\Delta}_Z) - T^{-1}Z'Z = T^{-1}Z'\bar{\Delta}_Z + T^{-1}\bar{\Delta}_Z'Z + T^{-1}\bar{\Delta}_Z'\bar{\Delta}_Z.
$$

Since $\|\bar{\Delta}_Z'Z\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$ and $\|\bar{\Delta}_Z\| \leq \sqrt{rT}\|\bar{\Delta}_Z\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$, it follows that $\bar{\Sigma}_Z - \Sigma_Z = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$. We have proved all the claims in part (4). □

Lemma B.2.8. *Let Assumption 3 hold. Then* $\|(\Pi_{\hat{Z}} - \Pi_Z)Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$.

Proof. We adopt all the notations introduced in Lemma B.2.7. Notice that $\Pi_{\hat{Z}} = \Pi_{\bar{Z}}$ and

$$
\begin{aligned}
&\|(\Pi_{\bar{Z}} - \Pi_Z)Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty \\
&= \left\|T^{-1}\left(\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}' - Z\Sigma_Z^{-1}Z'\right)Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\right\|_\infty \\
&= \|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'Z - Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty
\end{aligned}
$$

$$\leq \|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'Z - Z - \bar{\Delta}_Z\|_\infty + \|\bar{\Delta}_Z - n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty$$
$$\overset{(i)}{=} \|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'Z - \bar{Z}\|_\infty + O_{\mathbb{P}}(n^{-1}\log^{O(1)} n), \tag{B.2.6}$$

where (i) holds by Lemma B.2.7(4) and $\bar{Z} = Z + \bar{\Delta}_Z$ (Lemma B.2.7(3)). Thus,

$$\begin{aligned}
\|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'Z - \bar{Z}\|_\infty &= \|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'(\bar{Z} - \bar{\Delta}_Z) - \bar{Z}\|_\infty \qquad \text{(B.2.7)}\\
&= \|T^{-1}\bar{Z}\bar{\Sigma}_Z^{-1}\bar{Z}'\bar{\Delta}_Z\|_\infty\\
&\overset{(i)}{\leq} T^{-1}\|\bar{Z}\|_\infty\|\bar{\Sigma}_Z^{-1}\|_\infty\|\bar{Z}'\bar{\Delta}_Z\|_\infty r^2\\
&\leq T^{-1}\left(\|Z\|_\infty + \|\bar{\Delta}_Z\|_\infty\right)\|\bar{\Sigma}_Z^{-1}\|_\infty\|\bar{Z}'\bar{\Delta}_Z\|_\infty r^2\\
&\overset{(ii)}{=} O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)\|\bar{Z}'\bar{\Delta}_Z\|_\infty\\
&\overset{(iii)}{\leq} O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)\left(\|Z'\bar{\Delta}_Z\|_\infty + \|\bar{\Delta}_Z'\bar{\Delta}_Z\|_\infty\right)\\
&\overset{(iv)}{\leq} O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)\left(\|Z'\bar{\Delta}_Z\|_\infty + T\|\bar{\Delta}_Z\|_\infty^2\right)\\
&\overset{(v)}{=} O_{\mathbb{P}}(n^{-1}\log^{O(1)} n),
\end{aligned}$$

where (i) holds by Lemma B.2.1, (ii) holds by Lemmas B.2.2 and B.2.7(4), (iii) holds by $\bar{Z} = Z + \bar{\Delta}_Z$, (iv) holds by Lemma B.2.1 and finally (v) follows by Lemma B.2.7(4). The desired result follows by (B.2.6) and (B.2.7). □

Lemma B.2.9. *Let Assumption 3 hold. Then* $\|(\Pi_{\hat{Z}} - \Pi_Z)\Delta_F\|_\infty = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$*, where* Δ_F *is defined in Lemma B.2.4.*

Proof. We adopt all the notations introduced in Lemma B.2.7. Notice that $\Pi_{\hat{Z}} = \Pi_{\bar{Z}}$ and

$$(\Pi_{\bar{Z}} - \Pi_Z)\Delta_F = \underbrace{T^{-1}\bar{\Delta}_Z\bar{\Sigma}_Z^{-1}\bar{Z}'\Delta_F}_{J_1} + \underbrace{T^{-1}Z[\bar{\Sigma}_Z^{-1} - \Sigma_Z^{-1}]\bar{Z}'\Delta_F}_{J_2} + \underbrace{T^{-1}Z\Sigma_Z^{-1}\bar{\Delta}_Z'\Delta_F}_{J_3}. \tag{B.2.8}$$

By Lemmas B.2.1, B.2.2 and B.2.4(4), that

$$\begin{aligned}
\|J_3\|_\infty \leq T^{-1}\|Z\|_\infty\|\Sigma_Z^{-1}\bar{\Delta}_Z'\Delta_F\|\sqrt{r} &\leq T^{-1}\|Z\|_\infty\|\Sigma_Z^{-1}\| \cdot \|\bar{\Delta}_Z\| \cdot \|\Delta_F\|\sqrt{r}\\
&= O_{\mathbb{P}}(n^{-1}\log^{O(1)} n).
\end{aligned}$$

Similarly, by Lemmas B.2.1, B.2.2, B.2.4 and B.2.7, we have

$$\|J_2\|_\infty \le T^{-1}\|Z\|_\infty \|\bar{\Sigma}_Z^{-1} - \Sigma_Z^{-1}\| \cdot \|\bar{Z}\| \cdot \|\Delta_F\|\sqrt{r} \overset{(i)}{=} O_{\mathbb{P}}(n^{-3/2}\log^{O(1)} n),$$

where in (i), we invoke $\|\bar{Z}\|_\infty \le \|Z\|_\infty + \|\bar{\Delta}_Z\|_\infty$ and $\bar{\Sigma}_Z^{-1} - \Sigma_Z^{-1} = -\bar{\Sigma}_Z^{-1}(\bar{\Sigma}_Z - \Sigma_Z)\Sigma_Z^{-1}$, together with bounds in Lemma B.2.7. Similar argument also yields

$$\|J_1\|_\infty \le T^{-1}\|\bar{\Delta}_Z\|_\infty \|\bar{\Sigma}_Z^{-1}\| \cdot \|\bar{Z}\| \cdot \|\Delta_F\|\sqrt{r} = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n).$$

The result follows by (B.2.8) together with the bounds for $\|J_1\|_\infty$, $\|J_2\|_\infty$ and $\|J_3\|_\infty$. □

Lemma B.2.10. *Under Assumption 3,* $\|\hat{e} - e\|_\infty = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$ *and* $\|\hat{v} - v\|_\infty = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n)$.

Proof. First notice that

$$\begin{aligned}
&\|\hat{L}\hat{L}' - L\Sigma_L^{-1}L'\|_\infty \\
&\overset{(i)}{=} \left\|(LH_L + \Delta_L)(H_L'L' + \Delta_L') - L\Sigma_L^{-1}L'\right\|_\infty \\
&\le \left\|L(H_LH_L' - \Sigma_L^{-1})L'\right\|_\infty + 2\left\|LH_L\Delta_L'\right\|_\infty + \left\|\Delta_L\Delta_L'\right\|_\infty \\
&\overset{(ii)}{\le} p_Y\|L\|_\infty\|H_LH_L' - \Sigma_L^{-1}\|\|L\|_\infty + 2p_Y\|L\|_\infty\|H_L\|\|\Delta_L\|_\infty + p_Y\|\Delta_L\|_\infty^2 \\
&\overset{(iii)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),
\end{aligned} \tag{B.2.9}$$

where (i) follows by Lemma B.2.4, (ii) follows by Lemma B.2.1 and (iii) follows by Lemmas B.2.4(4), B.2.6 and B.2.2. Moreover,

$$\begin{aligned}
&\|\hat{e} - e\| \\
&= \|\hat{L}\hat{F}' - LF'\|_\infty \\
&\overset{(i)}{=} \|(n_Y^{-1}\hat{L}\hat{L}' - I_{n_Y})LF' + n_Y^{-1}\hat{L}\hat{L}'e\|_\infty \\
&\overset{(ii)}{\le} \|n_Y^{-1}(\hat{L}\hat{L}' - L\Sigma_L^{-1}L')LF'\|_\infty + \|n_Y^{-1}(\hat{L}\hat{L}' - L\Sigma_L^{-1}L')e\|_\infty + \|n_Y^{-1}L\Sigma_L^{-1}L'e\|_\infty \\
&\overset{(iii)}{\le} n_Y^{-1}\|\hat{L}\hat{L}' - L\Sigma_L^{-1}L'\|_\infty\|L\|_\infty\|F\|_\infty n_Y p_Y + n_Y^{-1}\|\hat{L}\hat{L}' - L\Sigma_L^{-1}L'\|_\infty n_Y\|e\|_\infty \\
&\quad + n_Y^{-1}\|L\|_\infty\|\Sigma_L^{-1}\| \cdot \|L'e\|_\infty p_Y
\end{aligned}$$

$$\overset{(iv)}{\leq} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),$$

where (i) follows by the definition in (2.2.6) $\hat{F} = Y'\hat{L}/n_Y = (LF' + e)'\hat{L}/n_Y$, (ii) follows by the triangular inequality, (iii) follows by Lemma B.2.1 and (iv) follows by (B.2.9) and Lemmas B.2.2 and B.2.3(1). We have proved the result for $\|\hat{e} - e\|_\infty$.

The result for $\|\hat{v} - v\|_\infty$ follows by the same arguments (including in auxiliary lemmas), except that (F, L, e, p_Y) is replaced by (Z, Λ, v, r). □

B.2.3 Proof of Theorems 2.3.1 and 2.3.2

We introduce/recall the following definitions, which will be used in the rest of this section.

$$\begin{aligned}
n =& n_Y + n_W \qquad\qquad (B.2.10)\\
\mathcal{F}_n =& \sigma-\text{algebra generated by } L,\ F,\ X,\ R,\ \Lambda \text{ and } Z\\
\mathcal{G}_n =& \sigma-\text{algebra generated by } L,\ F,\ X,\ R,\ \Lambda,\ Z \text{ and } v\\
\{\xi_i\}_{i=1}^n =& \text{an i.i.d sequence of } N(0,1) \text{ independent of } \mathcal{G}_n\\
Q_F =& (Z'Z)^{-1}Z'F\\
Q_X =& (Z'Z)^{-1}Z'X\\
\hat{Q}_F =& (\hat{Z}'\hat{Z})^{-1}(\hat{Z}'\hat{F})\\
\hat{Q}_X =& (\hat{Z}'\hat{Z})^{-1}(\hat{Z}'\hat{X})\\
\Psi_{(F),i} =& \begin{cases} v_i\left[(n_Y^{-1}n)L_i'\Sigma_L^{-1}(\tilde{H}_L')^{-1} - \Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\right] & \text{for } 1 \leq i \leq n_Y\\ -v_i\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1} & \text{for } n_Y + 1 \leq i \leq n \end{cases}\\
\Psi_{(X),i} =& \begin{cases} v_i\left[(n_W^{-1}n)R_i'\Sigma_R^{-1}(\tilde{H}_R')^{-1} - \Lambda_i'\Sigma_\Lambda^{-1}Q_X(\tilde{H}_R')^{-1}\right] & \text{for } 1 \leq i \leq n_W\\ -v_i\Lambda_i'\Sigma_\Lambda^{-1}Q_X(\tilde{H}_R')^{-1} & \text{for } n_W + 1 \leq i \leq n \end{cases}\\
\Psi_i =& \begin{bmatrix}\Psi_{(F),i} & \Psi_{(X),i}\end{bmatrix} \in \mathbb{R}^{T\times(p_Y+p_W)}\\
\hat{\Psi}'_{(F),i} =& \begin{cases} \hat{v}_i\left[(n_Y^{-1}n)\hat{L}_i' - \hat{\Lambda}_i'\hat{Q}_F\right] & \text{for } 1 \leq i \leq n_Y\\ -\hat{v}_i\hat{\Lambda}_i'\hat{Q}_F & \text{for } n_Y + 1 \leq i \leq n \end{cases}
\end{aligned}$$

$$\hat{\Psi}'_{(X),i} = \begin{cases} -\hat{v}_i \hat{\Lambda}'_i \hat{Q}_X & \text{for } 1 \le i \le n_W \\ \hat{v}_i \left[(n_W^{-1} n) \hat{R}'_i - \hat{\Lambda}'_i \hat{Q}_X \right] & \text{for } n_W + 1 \le i \le n \end{cases}$$

$$S_n^{\Psi_{(F)}} = n^{-1/2} \sum_{i=1}^{n} \Psi_{(F),i}$$

$$S_n^{\Psi_{(X)}} = n^{-1/2} r \sum_{i=1}^{n} \Psi_{(X),i}$$

$$\hat{\Psi}_i = \begin{bmatrix} \hat{\Psi}_{(F),i} & \hat{\Psi}_{(X),i} \end{bmatrix} \in \mathbb{R}^{T \times (p_Y + p_W)}$$

$$S_n^{\Psi} = \begin{bmatrix} S_n^{\Psi_{(F)}} & S_n^{\Psi_{(X)}} \end{bmatrix}$$

$$\tilde{S}_n^{\Psi} = n^{-1/2} \sum_{i=1}^{n} (\Psi_i - \bar{\Psi}) \xi_i \quad \text{with } \bar{\Psi} = n^{-1} \sum_{i=1}^{n} \Psi_i$$

$$\tilde{S}_n^{\hat{\Psi}} = n^{-1/2} \sum_{i=1}^{n} (\hat{\Psi}_i - \bar{\hat{\Psi}}) \xi_i \quad \text{with } \bar{\hat{\Psi}} = n^{-1} \sum_{i=1}^{n} \hat{\Psi}_i$$

We also use the following partitions :

$$\tilde{S}_n^{\Psi} = \begin{bmatrix} \tilde{S}_n^{\Psi_{(F)}} & \tilde{S}_n^{\Psi_{(X)}} \end{bmatrix} \text{ with } \tilde{S}_n^{\Psi_{(F)}} \in \mathbb{R}^{T \times p_Y} \text{ and } \tilde{S}_n^{\Psi_{(X)}} \in \mathbb{R}^{T \times p_W}$$

$$\tilde{S}_n^{\hat{\Psi}} = \begin{bmatrix} \tilde{S}_n^{\hat{\Psi}_{(F)}} & \tilde{S}_n^{\hat{\Psi}_{(X)}} \end{bmatrix} \text{ with } \tilde{S}_n^{\hat{\Psi}_{(F)}} \in \mathbb{R}^{T \times p_Y} \text{ and } \tilde{S}_n^{\hat{\Psi}_{(X)}} \in \mathbb{R}^{T \times p_W}$$

$$J = [T] \times [p_Y + p_W]$$

$$\Psi_{i,j} = \iota'_{j_1,T} \Psi_i \iota_{j_2,k} \qquad \text{for } j = (j_1, j_2) \in J$$

$$\hat{\Psi}_{i,j} = \iota'_{j_1,T} \hat{\Psi}_i \iota_{j_2,k} \qquad \text{for } j = (j_1, j_2) \in J$$

$$S^{\Psi}_{n,j} = \iota'_{j_1,T} S^{\Psi}_{n,j} \iota_{j_2,k} \qquad \text{for } j = (j_1, j_2) \in J$$

Lemma B.2.11. *Consider the notations in (B.2.10). Let Assumption 3 and H_0 hold. Then*

$$\left\| \sqrt{n} M_{\hat{Z}} \hat{F} D_{n,F} - S_n^{\Psi_{(F)}} \right\|_{\infty} = O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n),$$

where $D_{n,F}$ is defined in Lemma B.2.5.

Proof. First notice that $S_n^{\Psi_{(F)}} = e' L \Sigma_L^{-1} (\tilde{H}'_L)^{-1} \sqrt{n}/n_Y - v' \Lambda \Sigma_\Lambda^{-1} Q_F (\tilde{H}'_L)^{-1}/\sqrt{n}$. By Lemma B.2.4, we have

$$\sqrt{n} M_{\hat{Z}} \hat{F} D_{n,F} = \sqrt{n} M_Z (\hat{F} - F H_F) D_{n,F} + \sqrt{n} (M_{\hat{Z}} - M_Z) \hat{F} D_{n,F}$$

$$= \sqrt{n} M_Z \Delta_F D_{n,F} + \sqrt{n}(\Pi_Z - \Pi_{\hat{Z}})(Z Q_F H_F D_{n,F} + \Delta_F D_{n,F}).$$

Hence,

$$\begin{aligned}
&\sqrt{n} M_{\hat{Z}} \hat{F} D_{n,F} - S_n^{\Psi_{(F)}} \\
&\quad = \underbrace{\sqrt{n}\left(M_Z \Delta_F D_{n,F} - n_Y^{-1} e' L \Sigma_L^{-1} (\tilde{H}_L')^{-1} \right)}_{J_1} + \underbrace{\sqrt{n}(\Pi_Z - \Pi_{\hat{Z}}) \Delta_F D_{n,F}}_{J_2} \\
&\qquad + \underbrace{\sqrt{n}\left((\Pi_Z - \Pi_{\hat{Z}}) Z Q_F H_F D_{n,F} + n^{-1} v' \Lambda \Sigma_\Lambda^{-1} Q_F (\tilde{H}_L')^{-1} \right)}_{J_3}. \quad \text{(B.2.11)}
\end{aligned}$$

We show that $\|J_1\|_\infty$, $\|J_2\|_\infty$ and $\|J_3\|_\infty$ are all $O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n)$. By Lemma B.2.9 and the fact that $\|(\Pi_Z - \Pi_{\hat{Z}}) \Delta_F D_{n,F}\|_\infty = \|(\Pi_Z - \Pi_{\hat{Z}}) \Delta_F\|_\infty$, we have

$$\|J_2\|_\infty = O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n). \quad \text{(B.2.12)}$$

The bound for $\|J_1\|_\infty$ is based on the following observation:

$$\begin{aligned}
\|J_1\|_\infty &\overset{(i)}{\leq} \sqrt{n} \|\Delta_F D_{n,F} - n_Y^{-1} e' L \Sigma_L^{-1} (\tilde{H}_L')^{-1}\|_\infty + \sqrt{n} \|\Pi_Z \Delta_F D_{n,F}\|_\infty \\
&\overset{(ii)}{\leq} \sqrt{n/n_Y} \|n_Y^{-1/2} e' \hat{L} D_{n,F} - n_Y^{-1/2} e' L \Sigma_L^{-1} (\tilde{H}_L')^{-1}\|_\infty + \sqrt{n} n_Y^{-1} \|\Pi_Z e' \hat{L}\|_\infty \\
&\overset{(iii)}{\leq} O(n^{-1/2}) \|e' L [H_L D_{n,F} - \Sigma_L^{-1} (\tilde{H}_L')^{-1}]\|_\infty + O(n^{-1/2}) \|e' \Delta_L D_{n,F}\|_\infty \\
&\qquad + O(n^{-1/2}) \|\Pi_Z e' \hat{L}\|_\infty \\
&\overset{(iv)}{\leq} O(n^{-1/2}) \|e' L\|_\infty \|H_L D_{n,F} - \Sigma_L^{-1} (\tilde{H}')^{-1}\|_\infty p_Y + O(n^{-1/2}) \|e' \Delta_L\|_\infty \\
&\qquad + O(n^{-1/2}) \|\Pi_Z e' \hat{L}\|_\infty \\
&\overset{(v)}{=} O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n) + n^{-1/2} \|\Pi_Z e' \hat{L}\|_\infty \\
&\overset{(vi)}{=} O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n) + n^{-1/2} \|T^{-1} Z \Sigma_Z^{-1} Z' e' (L H_L + \Delta_L)\|_\infty \\
&\overset{(vii)}{\leq} O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n) + n^{-1/2} T^{-1} \|Z\|_\infty \|\Sigma_Z^{-1}\|_\infty \|Z' e' L\|_\infty \|H_L\|_\infty p_Y^3 \\
&\qquad + n^{-1/2} T^{-1} \|Z\|_\infty \|\Sigma_Z^{-1}\|_\infty \|Z\| \cdot \|e' \Delta_L\|_\infty p_Y^2 \sqrt{T} \overset{(viii)}{=} O_{\mathbb{P}}(n^{-1/2} \log^{O(1)} n),
\end{aligned} \quad \text{(B.2.13)}$$

where (i) holds by $M_Z = I_T - \Pi_Z$ and the triangular inequality; (ii) and (iii) hold

by $\Delta_F = n_Y^{-1} e' \hat{L}$ and $\hat{L} = LH_L + \Delta_L$ (Lemma B.2.4) together with the fact that $D_{n,F}$ is diagonal with diagonal entries taking values in $\{-1, 1\}$ (Lemma B.2.5); (iv) follows by Lemmas B.2.1 and B.2.5; (v) follows by Lemmas B.2.3(1), B.2.6(2) and $H_L D_{n,F} - \Sigma_L^{-1}(\tilde{H}')^{-1} = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ (due to the first claim of Lemma B.2.5); (vi) follows by $\hat{L} = LH_L + \Delta_L$ (Lemma B.2.4); (vii) follows by Lemma B.2.1; finally, $(viii)$ follows by Lemmas B.2.2, B.2.3(2) and B.2.6(2). Also notice that

$$\begin{aligned}
\|J_3\|_\infty &\leq \sqrt{n}\|[(\Pi_Z - \Pi_{\hat{Z}})Z + n^{-1}v'\Lambda\Sigma_\Lambda^{-1}]Q_F H_F D_{n,F}\|_\infty \\
&\quad + n^{-1/2}\|v'\Lambda\Sigma_\Lambda^{-1}Q_F[(\tilde{H}_L')^{-1} - H_F D_{n,F}]\|_\infty \\
&\overset{(i)}{\leq} \sqrt{n}\|(\Pi_Z - \Pi_{\hat{Z}})Z + n^{-1}v'\Lambda\Sigma_\Lambda^{-1}\|_\infty \|Q_F H_F D_{n,F}\|_\infty p_Y \\
&\quad + n^{-1/2}\|v'\Lambda\|_\infty \|\Sigma_\Lambda^{-1}Q_F\|_\infty \|(\tilde{H}')^{-1} - H_F D_{n,F}\|_\infty p_Y^2 \\
&\overset{(ii)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n), \qquad \text{(B.2.14)}
\end{aligned}$$

where (i) follows by Lemma B.2.1 and (ii) follows by Lemmas B.2.3(1) and B.2.8, together with the second claim of Lemma B.2.5. The desired result follows by (B.2.11), together with (B.2.12), (B.2.13) and (B.2.14). □

Lemma B.2.12. *Let Assumption 3 and H_0 hold. Then there exists a diagonal matrix $D_{n,X}$ whose diagonal elements are either -1 or 1 such that*

$$\left\| \sqrt{n} M_{\hat{Z}} \hat{X} D_{n,X} - S_n^{\Psi_{(X)}} \right\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).$$

Proof. The proof is analogous to that of Lemma B.2.11. □

Lemma B.2.13. *Let Assumption 3 hold. Recall J, $\Psi_{i,j}$'s and $\mathcal{F}_n$ defined in (B.2.10) and let $\max_{i,j}$ denote $\max_{(i,j)\in[n]\times J}$. Then $\forall q > 2$, there exists an $\mathcal{F}_n$-measurable random variable $B_n = O_{\mathbb{P}}(n^{2/q}\log^{O(1)} n)$ such that almost surely (1) $\max_{i,j}\mathbb{E}(|\Psi_{i,j}|^3 \mid \mathcal{F}_n) \leq B_n$, (2) $\max_{i,j}\mathbb{E}(|\Psi_{i,j}|^4 \mid \mathcal{F}_n) \leq B_n^2$ and (3) $\mathbb{E}[\max_{i,j}|\Psi_{i,j}|^q \mid \mathcal{F}_n] \leq B_n^q$.*

Proof. Define

$$A := \ 1 + \|\Lambda\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty + \|\Lambda\Sigma_\Lambda^{-1}Q_X(\tilde{H}_R')^{-1}\|_\infty$$

$$+\|(n_Y^{-1}n)L\Sigma_L^{-1}(\tilde{H}_L')^{-1}\|_\infty + \|(n_W^{-1}n)R\Sigma_R^{-1}(\tilde{H}_R')^{-1}\|_\infty. \quad \text{(B.2.15)}$$

Notice that $A = O_{\mathbb{P}}(\log^{O(1)} n)$ by Lemma B.2.2. For $j = (j_1, j_2) \in J$ and $i \in [n]$, let $\mu_{i,j} = \Psi_{i,j}/A$. Notice that , $|\mu_{i,j}| \leq |v_{i,j_1}|$. By the exponential-type tail condition in Assumption 3, we have that $\forall (i,j) \in [n] \times J$ and $\mu_{i,j} \in \Xi(\beta, \gamma, \mathcal{F}_n)$. Hence, by Lemma B.3.5 (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), $\max_{i,j} \mathbb{E}(|\mu_{i,j}|^3 \mid \mathcal{F}_n) \leq C_1$, $\max_{i,j} \mathbb{E}(|\mu_{i,j}|^4 \mid \mathcal{F}_n) \leq C_1$ and $\mathbb{E}(\max_{i,j} |\mu_{i,j}|^q \mid \mathcal{F}_n) \leq C_1 nT$, where $C_1 > 0$ is a constant depending only on q and the constants in Assumption 3. The desired result holds with $B_n = C_1(nT)^{1/q}A^3$. □

Lemma B.2.14. *Let Assumption 3 hold. Recall $\Psi_{i,j}$, J and $\mathcal{F}_n$ defined in (B.2.10). Then*

(1) There exists a constant $b > 0$ such that $\mathbb{P}\left(\min_{j\in J} n^{-1}\sum_{i=1}^n \mathbb{E}[\Psi_{i,j}^2 \mid \mathcal{F}_n] \geq b\right) \to 1$.

(2) $\max_{j\in J} \left|n^{-1}\sum_{i=1}^n \left[\Psi_{i,j}^2 - \mathbb{E}(\Psi_{i,j}^2 \mid \mathcal{F}_n)\right]\right| = o_{\mathbb{P}}(1)$.

(3) $\max_{j\in J} |n^{-1}\sum_{i=1}^n \Psi_{i,j}| = o_{\mathbb{P}}(1)$.

Proof. Let $J_F = [T] \times [p_Y]$ and $J_X = J \backslash J_F$. Notice that $\forall j = (j_1, j_2) \in J_F$

$$\Psi_{i,j} = \Psi_{(F),i,j} = \begin{cases} v_{i,j_1}\left[(n_Y^{-1}n)L_i'\Sigma_L^{-1}(\tilde{H}_L')^{-1} - \Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\right]\iota_{j_2} & \forall i \leq n_Y \\ -v_{i,j_1}\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\iota_{j_2} & \forall i > n_Y \end{cases}$$

and

$$\mathbb{E}(\Psi_{i,j}^2 \mid \mathcal{F}_n) = \mathbb{E}(v_{i,j_1}^2)\left[\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\iota_{j_2}\right]^2 \qquad \text{for } n_Y + 1 \leq i \leq n. \quad \text{(B.2.16)}$$

We first show part (1). Let $\Lambda_{(2)} = (\Lambda_{n_Y+1}, \cdots, \Lambda_n)' \in \mathbb{R}^{n_W \times r}$. Notice that $\Lambda_{(2)}Z' = RX'$. Thus, $\Lambda_{(2)} = RX'Z(Z'Z)^{-1} = RQ_X$. Therefore,

$$\begin{aligned} \min_{j\in J_F} n^{-1}\sum_{i=1}^n \mathbb{E}[\Psi_{i,j}^2 \mid \mathcal{F}_n] &\geq \min_{j\in J_F} n^{-1}\sum_{i=n_Y+1}^n \mathbb{E}[\Psi_{(F),i,j}^2 \mid \mathcal{F}_n] \\ &\overset{(i)}{\geq} n^{-1}\kappa_1 \min_{1\leq d\leq p_Y} \sum_{i=n_Y+1}^n \left[\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\iota_{d,p_Y}\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= n^{-1}\kappa_1 \min_{1\le d\le p_Y} \left\| \Lambda_{(2)}\Sigma_\Lambda^{-1} Q_F (\tilde{H}_L')^{-1} \iota_{d,p_Y} \right\|_2^2 \\
&\overset{(ii)}{\ge} n^{-1}\kappa_1 \left[s_{p_Y}\left(RQ_X \Sigma_\Lambda^{-1} Q_F (\tilde{H}_L')^{-1} \right) \right]^2,
\end{aligned}
$$

where (i) follows by (B.2.16) and $\min_{i,t} \mathbb{E} v_{i,t}^2 \ge \kappa_1$ (by Assumption 3) and (ii) follows by $\Lambda_{(2)} = RQ_X$ and the fact that for matrix $A \in \mathbb{R}^{r_1 \times r_2}$ and vector $x \in \mathbb{R}^{r_2}$, $\|Ax\|_2 \ge s_{r_2}(A)\|x\|_2$. Since the singular values of Σ_R, Q_X, Σ_Λ, Q_F and $\tilde{H}_L$ are bounded away from infinity and zero, we have that the right-hand size of the above display is bounded away from zero. Similarly, we can show the same result for $\min_{j\in J_X} n^{-1}\sum_{i=1}^n \mathbb{E}[\Psi_{i,j}^2 \mid \mathcal{F}_n]$. We have proved part (1).

To show part (2), recall the random variable A defined in (B.2.15) in the proof of Lemma B.2.13. Let $\mu_{i,j} = \Psi_{i,j}^2/A^2$ for $(i,j) \in [n] \times J$. By the definition of $\Psi_{i,j}$ in (B.2.10), we have $|\mu_{i,j}| \le v_{i,j}^2$. By Lemma B.3.5 and Assumption 3, there exist a constant $C_1 > 0$ such that $v_{i,j}^2 \in \Xi(C_1, 2\gamma, \mathcal{F}_n)$. Thus, $\mu_{i,j} \in \Xi(C_1, 2\gamma, \mathcal{F}_n)$. Since $\{\mu_{i,j}\}_{j\in J}$ is independent across i conditional on $\mathcal{F}_n$, it follows, by Lemma B.3.8, that $\max_{j\in J} |\sum_{i=1}^n [\mu_{i,j} - \mathbb{E}(\mu_{i,j} \mid \mathcal{F}_n)]| = O_{\mathbb{P}}(\sqrt{n}\log^{O(1)} n)$. Therefore, part (2) follows by noticing that

$$
\begin{aligned}
\max_{j\in J} \left| n^{-1}\sum_{i=1}^n \left[\Psi_{i,j}^2 - \mathbb{E}(\Psi_{i,j}^2 \mid \mathcal{F}_n) \right] \right| &= A^2 \max_{j\in J} \left| n^{-1}\sum_{i=1}^n \left[\mu_{i,j} - \mathbb{E}(\mu_{i,j} \mid \mathcal{F}_n) \right] \right| \\
&= A^2 O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \overset{(i)}{=} o_{\mathbb{P}}(1),
\end{aligned}
$$

where (i) holds by $A = O_{\mathbb{P}}(\log^{O(1)} n)$ as argued in the proof of Lemma B.2.13.

To show part (3), we use a similar argument. Let $d_{i,j} = \Psi_{i,j}/A$. Then $|d_{i,j}| \le |v_{i,j}|$. Since $v_{i,j} \in \Xi(\beta, \gamma, \mathcal{F}_n)$ (Assumption 3), we have $d_{i,j} \in \Xi(\beta, \gamma, \mathcal{F}_n)$. Hence, by Lemma B.3.8 and $\mathbb{E}(d_{i,j} \mid \mathcal{F}_n)$, we have that $\max_{j\in J} |\sum_{i=1}^n d_{i,j}| = O_{\mathbb{P}}(\sqrt{n}\log^{O(1)} n)$. Then part (3) follows by

$$
\max_{j\in J} \left| n^{-1}\sum_{i=1}^n \Psi_{i,j} \right| = A \max_{j\in J} \left| n^{-1}\sum_{i=1}^n d_{i,j} \right| = A O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) = o_{\mathbb{P}}(1).
$$

The proof is complete. □

Lemma B.2.15. *Recall the definitions in (B.2.10) and $D_{n,F}$ in Lemma B.2.5.*

Under Assumption 3 and H_0, $\|\hat{\Lambda}\hat{Q}_F D_{n,F} - \Lambda\Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1}\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$.

Proof. We adopt all the notations in Lemma B.2.7. Notice that

$$\begin{aligned}
&\|\hat{\Lambda}\hat{Q}_F D_{n,F} - \Lambda\Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1}\|_\infty \\
&\overset{(i)}{\leq} \|\Lambda[H_\Lambda \hat{Q}_F D_{n,F} - \Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1}]\|_\infty + \|\Delta_\Lambda \hat{Q}_F D_{n,F}\|_\infty \\
&\overset{(ii)}{\leq} \|\Lambda\|_\infty \|H_\Lambda \hat{Q}_F D_{n,F} - \Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1}\|\sqrt{p} + \|\Delta_\Lambda\|_\infty \|\hat{Q}_F\|\sqrt{p} \\
&\overset{(iii)}{=} O_{\mathbb{P}}(\log^{O(1)} n)\|H_\Lambda \hat{Q}_F D_{n,F} - \Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1}\| + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n), \quad \text{(B.2.17)}
\end{aligned}$$

where (i) holds by Lemma B.2.7(1), (ii) holds by Lemma B.2.1 and (iii) holds by Lemmas B.2.2 and B.2.7.

By Lemmas B.2.7 and B.2.4, we have that under H_0,

$$\begin{cases}
T^{-1}\hat{Z}'\hat{Z} = H_Z'\bar{\Sigma}_Z H_Z = H_Z'\Sigma_Z H_Z + O_{\mathbb{P}}(n^{-1}\log^{O(1)} n) \\
T^{-1}\hat{Z}'\hat{F} = T^{-1}(H_Z'Z + \Delta_Z')(ZQ_F H_F + \Delta_F) = H_Z'\Sigma_Z Q_F H_F + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \\
H_Z = \Lambda'(\Lambda H_\Lambda + \Delta_\Lambda)/n = \Sigma_\Lambda H_\Lambda + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \\
H_F = L'(LH_L + \Delta_L)/n_Y = \Sigma_L H_L + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).
\end{cases}$$

Thus,

$$\hat{Q}_F = (T^{-1}\hat{Z}'\hat{Z})^{-1}(T^{-1}\hat{Z}'\hat{F}) = H_\Lambda^{-1}\Sigma_\Lambda^{-1} Q_F \Sigma_L H_L + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).$$

It follows, by Lemma B.2.5, that

$$\begin{aligned}
&H_\Lambda \hat{Q}_F D_{n,F} - \Sigma_\Lambda^{-1} Q_F(\tilde{H}_L')^{-1} \\
&= \Sigma_\Lambda^{-1} Q_F\left(\Sigma_L H_L D_{n,F} - (\tilde{H}_L')^{-1}\right) + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n) \\
&= O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).
\end{aligned}$$

We finish the proof by combining this and (B.2.17). □

Lemma B.2.16. *Recall the definitions in (B.2.10). Let* $D_n = \text{Blockdiag}(D_{n,F}, D_{n,X})$*, where* $D_{n,F}$ *and* $D_{n,X}$ *are defined in Lemmas B.2.5*

and B.2.12, respectively. Let Assumption 3 and H_0 hold. Then

$$\max_{j=(j_1,j_2)\in J} n^{-1}\sum_{i=1}^{n}(\hat{\Psi}_{i,j}D_{n,j_2}-\Psi_{i,j})^2 = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n).$$

Proof. Recall, from the definitions in (B.2.10),

$$\hat{\Psi}_{(F),i} = \begin{cases} \hat{v}_i\left[(n_Y^{-1}n)\hat{L}_i' - \hat{\Lambda}_i'\hat{Q}_F\right] & \text{for } 1\le i\le n_Y \\ -\hat{v}_i\hat{\Lambda}_i'\hat{Q}_F & \text{for } n_Y+1\le i\le n \end{cases} \tag{B.2.18}$$

and

$$\Psi_{(F),i} = \begin{cases} v_i\left[(n_Y^{-1}n)L_i'\Sigma_L^{-1}(\tilde{H}_L')^{-1} - \Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\right] & \text{for } 1\le i\le n_Y \\ -v_i\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1} & \text{for } n_Y+1\le i\le n. \end{cases} \tag{B.2.19}$$

Thus, by the triangular inequality, we have

$$\begin{aligned} &\max_{1\le i\le n_Y}\|\hat{\Psi}_{(F),i}D_{n,F}-\Psi_{(F),i}\|_\infty \\ &\le \|\hat{v}-v\|_\infty \underbrace{\max_{1\le i\le n_Y}\|[(n_Y^{-1}n)\hat{L}_i'-\hat{\Lambda}_i'\hat{Q}_F]D_{n,F}\|_\infty}_{=:G_1} \\ &\quad + \|v\|_\infty \underbrace{\max_{1\le i\le n_Y}\left\|(n_Y^{-1}n)\left[\hat{L}_i'D_{n,F}-L_i'\Sigma_L^{-1}(\tilde{H}_L')^{-1}\right]\right\|_\infty}_{=:G_2} \\ &\quad + \|v\|_\infty \underbrace{\max_{1\le i\le n_Y}\left\|\hat{\Lambda}_i'\hat{Q}_FD_{n,F}-\Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\right\|_\infty}_{=:G_3}. \end{aligned} \tag{B.2.20}$$

Notice that

$$\begin{aligned} G_2 &= (n_Y^{-1}n)\|\hat{L}D_{n,F}-L\Sigma_L^{-1}(\tilde{H}_L')^{-1}\|_\infty \\ &\overset{(i)}{=} (n_Y^{-1}n)\|(LH_L+\Delta_L)D_{n,F}-L\Sigma_L^{-1}(\tilde{H}_L')^{-1}\|_\infty \\ &\le (n_Y^{-1}n)\left(\|L[H_LD_{n,F}-\Sigma_L^{-1}(\tilde{H}_L')^{-1}]\|_\infty + \|\Delta_LD_{n,F}\|_\infty\right) \\ &\overset{(ii)}{\le} (n_Y^{-1}n)\left(\|L\|_\infty\|H_LD_{n,F}-\Sigma_L^{-1}(\tilde{H}_L')^{-1}\|\sqrt{p_Y} + \|\Delta_LD_{n,F}\|_\infty\right) \end{aligned}$$

$$\overset{(iii)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n),$$

where (i) holds by Lemma B.2.4(1), (ii) holds by Lemma B.2.1(1) and (iii) holds by the bounds for $\|H_L D_{n,F} - \Sigma_L^{-1}(\tilde{H}_L')^{-1}\|$ (Lemma B.2.5), $\|L\|_\infty$ (Lemma B.2.2) and $\|\Delta_L D_{n,F}\|_\infty = \|\Delta_L\|_\infty$ (Lemma B.2.6). By Lemma B.2.15, $G_3 = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. By the triangular inequality,

$$G_1 \le G_2 + G_3 + \left\|(n_Y^{-1}n)L_i'\Sigma_L^{-1}(\tilde{H}_L')^{-1} - \Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\right\|_\infty \overset{(i)}{=} O_{\mathbb{P}}(\log^{O(1)} n),$$

where (i) holds by Lemma B.2.2. By (B.2.20), together with the bounds for G_1, G_2 and G_3, we have

$$\begin{aligned}\max_{1\le i\le n_Y}\|\hat{\Psi}_{(F),i}D_{n,F} - \Psi_{(F),i}\|_\infty &\le \|\hat{v}-v\|_\infty O_{\mathbb{P}}(\log^{O(1)} n) + \|v\|_\infty O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\\ &\overset{(i)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n), \qquad \text{(B.2.21)}\end{aligned}$$

where (i) holds by Lemmas B.2.10 and B.2.2. From (B.2.18) and (B.2.19), we have

$$\begin{aligned}&\max_{n_Y+1\le i\le n}\|\hat{\Psi}_{(F),i}D_{n,F} - \Psi_{(F),i}\|_\infty\\
&\le \|\hat{v}-v\|_\infty \max_{n_Y+1\le i\le n}\|\hat{\Lambda}_i'\hat{Q}_F D_{n,F}\|_\infty\\
&\quad + \|v\|_\infty \max_{n_Y+1\le i\le n}\|\hat{\Lambda}_i'\hat{Q}_F D_{n,F} - \Lambda_i'\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty\\
&\le \|\hat{v}-v\|_\infty\left(\|\Lambda\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty + \|\hat{\Lambda}\hat{Q}_F D_{n,F} - \Lambda\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty\right)\\
&\quad + \|v\|_\infty\|\hat{\Lambda}\hat{Q}_F D_{n,F} - \Lambda\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty\\
&\overset{(i)}{\le} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\left(\|\Lambda\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|_\infty + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\right)\\
&\quad + O_{\mathbb{P}}(\log^{O(1)} n)O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\\
&\overset{(ii)}{\le} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\|\Lambda\|_\infty\cdot\|\Sigma_\Lambda^{-1}Q_F(\tilde{H}_L')^{-1}\|\sqrt{p_Y} + O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)\\
&\overset{(iii)}{=} O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n), \qquad \text{(B.2.22)}\end{aligned}$$

where (i) holds by Lemmas B.2.10, B.2.2 and B.2.15 and (ii) holds by Lemma B.2.1 and (iii) holds by Lemma B.2.2.

By (B.2.21) and (B.2.22), $\max_{1\le i\le n}\|\hat{\Psi}_{(F),i}D_{n,F} - \Psi_{(F),i}\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$. By an analogous argument, we can show that $\max_{1\le i\le n}\|\hat{\Psi}_{(X),i}D_{n,X} - \Psi_{(X),i}\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n)$ with $D_{n,X}$ defined in Lemma B.2.12. Therefore,

$$\begin{aligned}
&\max_{1\le i\le n}\|\hat{\Psi}_i D_n - \Psi_i\|_\infty \\
&= \left(\max_{1\le i\le n}\|\hat{\Psi}_{(F),i}D_{n,F} - \Psi_{(F),i}\|_\infty\right) \vee \left(\max_{1\le i\le n}\|\hat{\Psi}_{(X),i}D_{n,X} - \Psi_{(X),i}\|_\infty\right) \\
&= O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).
\end{aligned}$$

Thus, the desired result follows by

$$\begin{aligned}
\max_{j=(j_1,j_2)\in J} n^{-1}\sum_{i=1}^{n}(\hat{\Psi}_{i,j}D_{n,j_2} - \Psi_{i,j})^2 &\le \max_{1\le i\le n}\max_{j=(j_1,j_2)\in J}\left|\hat{\Psi}_{i,j}D_{n,j_2} - \Psi_{i,j}\right|^2 \\
&= \max_{1\le i\le n}\|\hat{\Psi}_i D_n - \Psi_i\|_\infty^2 = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n).
\end{aligned}$$

□

Proof of Theorem 2.3.1. Recall the definitions in (B.2.10). We define

$$\begin{cases}
\hat{S}_n = n^{1/2} M_{\hat{Z}}[\hat{F}D_{n,F}, \hat{X}D_{n,X}] \\
\Upsilon_i = \Psi_i \\
\hat{\Upsilon}_i = \hat{\Psi}_i \text{Blockdiag}(D_{n,F}, D_{n,X}) \\
\tilde{S}_n^{\hat{\Upsilon}} = n^{-1/2}\sum_{i=1}^{n}(\hat{\Upsilon}_i - \bar{\hat{\Upsilon}})\xi_i \qquad \text{with } \bar{\hat{\Upsilon}} = n^{-1}\sum_{i=1}^{n}\hat{\Upsilon}_i,
\end{cases}$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d $N(0,1)$ random variables independent of the data (defined in (B.2.10)), $D_{n,F}$ and $D_{n,X}$ are defined in Lemmas B.2.5 and B.2.12, respectively.

Since $D_{n,F}$ and $D_{n,X}$ are diagonal matrices with diagonal entries taking values in $\{-1,1\}$, we have $\|\hat{S}_n\|_\infty = \|S_n\|_\infty$ and $\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty = \|S_n^{BS}\|_\infty$, where S_n and S_n^{BS} are defined in Algorithm 3. Therefore, we only need to show the following

claim:

$$\limsup_{n\to\infty} \sup_{\eta\in(0,1)} \left| \mathbb{P}\left[\|\hat{S}_n\|_\infty > \mathcal{Q}\left(1-\eta, \|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right)\right] - \eta \right| = 0, \tag{B.2.23}$$

where $\mathcal{G}_n$ is the σ-algebra defined in (B.2.10) and $\mathcal{Q}\left(1-\eta, \|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty\right) = \inf\left\{x \in \mathbb{R} \mid \mathbb{P}\left(\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty > x \mid \mathcal{G}_n\right) \leq \alpha\right\}$. By Proposition B.1.1, it suffices to verify the following conditions:

(i) There exist constants $q_1, q_2 > 0$ such that $\|\hat{S}_n - S_n^{\Upsilon}\|_\infty = O_{\mathbb{P}}(n^{-q_1})$ and $\max_{1\leq j\leq p} n^{-1}\sum_{i=1}^n (\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2 = O_{\mathbb{P}}(n^{-q_2})$, where $S_n^{\Upsilon} = n^{-1/2}\sum_{i=1}^n \Upsilon_i$ and $\hat{\Upsilon}_{i,j}$ and Υ_i denote the jth component of $\hat{\Upsilon}_i$ and Υ_i, respectively.

(ii) There exist a constant $r > 2$ and an $\mathcal{F}_n$-measurable positive random variable B_n such that, almost surely, $n^{-1}\sum_{i=1}^n \mathbb{E}(|\Upsilon_{i,j}|^3 \mid \mathcal{F}_n) \leq B_n$, $n^{-1}\sum_{i=1}^n \mathbb{E}(|\Upsilon_{i,j}|^4 \mid \mathcal{F}_n) \leq B_n^2$ and $\mathbb{E}(\max_{1\leq j\leq p} |\Upsilon_{i,j}|^r \mid \mathcal{F}_n) \leq 2B_n^r$.

(iii) There exists a constant $b > 0$ such that $\mathbb{P}\left(\min_{1\leq j\leq p} n^{-1}\sum_{i=1}^n \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n) > b\right) \to 1$, $\max_{1\leq j\leq p} |n^{-1}\sum_{i=1}^n [\Upsilon_{i,j}^2 - \mathbb{E}(\Upsilon_{i,j}^2 \mid \mathcal{F}_n)]| = o_{\mathbb{P}}(1)$ and $\max_{1\leq j\leq p} |n^{-1}\sum_{i=1}^n \Upsilon_{i,j}| = o_{\mathbb{P}}(1)$.

(iv) $n^{2/r-1}B_n^2 \log^3(p\vee n) = o_{\mathbb{P}}(1)$, $n^{-1}B_n^2 \log^7(p\vee n) = o_{\mathbb{P}}(1)$, $n^{q_2}/\log^4 p \to \infty$ and $n^{2q_1}/\log p \to \infty$.

To see Condition (i), we notice that, by Lemmas B.2.11 and B.2.12, we have

$$\|\hat{S}_n - S_n^{\Upsilon}\|_\infty = \left\| \left[n^{1/2} M_{\hat{Z}} \hat{F} D_{n,F} - S_n^{\Psi(F)}, n^{1/2} M_{\hat{Z}} \hat{X} D_{n,X} - S_n^{\Psi(X)} \right] \right\|_\infty = O_{\mathbb{P}}(n^{-1/2}\log^{O(1)} n).$$

Hence, $\|\hat{S}_n - S_n^{\Upsilon}\|_\infty = O_{\mathbb{P}}(n^{-1/3})$. By Lemma B.2.16, $\max_{1\leq j\leq p} n^{-1}\sum_{i=1}^n (\hat{\Upsilon}_{i,j} - \Upsilon_{i,j})^2 = O_{\mathbb{P}}(n^{-1}\log^{O(1)} n) = O_{\mathbb{P}}(n^{-1/3})$. Thus, Condition (i) holds with $q_1 = q_2 = 1/3$.

Applying Lemma B.2.13 with $q = 8$, we have that Condition (ii) holds with $r = 8$ and $B_n = O_{\mathbb{P}}(n^{1/4}\log^{O(1)} n)$. Condition (iii) holds by Lemma B.2.14. Since

$p = T(p_Y + p_W) = O(n)$, Condition (iv) follows by simple computations. The proof is complete. □

Proof of Theorem 2.3.2. The proof consists of two steps. First, we show that, under H_1, the test statistic diverges at the rate $\sqrt{n}$. In the second step, we show that the critical value diverges at the rate $\log^{O(1)} n$.

Step 1: show $\sqrt{n}\|M_{\hat{Z}}[\hat{F}, \hat{X}]\|_\infty$ diverges at the rate $\sqrt{n}$ under H_1.

Recall $\zeta = [F, X]$ and $\Sigma_\zeta = \zeta'\zeta/T$ and let $r_0 := p_Y + p_W - k_0$ and $a_n := s_r(\Lambda Z')$. By Lemma 2.2.1, under H_1, $r = p_Y + p_W - k_1 > r_0$. Hence,

$$\begin{aligned} a_n = s_r(Z\Lambda') = s_r\left([FL', XR']\right) &= s_r\left\{\zeta \text{Blockdiag}(L', R')\right\} \\ &\overset{(i)}{\geq} s_r\left(\zeta\right) s_{p_Y+p_W-r+1}\left\{\text{Blockdiag}(L', R')\right\} \\ &= s_r\left(\zeta\right) s_{p_C+1}\left\{\text{Blockdiag}(L', R')\right\}, \end{aligned} \tag{B.2.24}$$

where (i) holds by Lemma B.3.1(2). By Assumption 3, with probability approaching one, $s_r(\zeta) = \sqrt{T \cdot s_r(\Sigma_\zeta)} \geq \sqrt{T/\kappa}$. Notice that

$$\begin{aligned} s_{p_C+1}\left\{\text{Blockdiag}(L', R')\right\} &= \sqrt{s_{p_C+1}\left[\text{Blockdiag}(n_Y\Sigma_L, n_W\Sigma_R)\right]} \\ &\overset{(i)}{\geq} \sqrt{\kappa^{-1}(n_Y \wedge n_W)} \quad \text{with probability approaching one,} \end{aligned}$$

where (i) holds by the fact that the eigenvalues of Σ_L and Σ_R are bounded below by κ^{-1} (Assumption 3). These observations, together with (B.2.24), imply that there exists a constant $C_1 > 0$ such that

$$\mathbb{P}\left(a_n \geq C_1 n\right) \to 1. \tag{B.2.25}$$

By the definition of SVD, $\|M_{\hat{Z}}\chi'\| = s_{r_0+1}(\chi')$. It follows, by Lemma B.3.1(1) and $r > r_0$ (thus $r \geq r_0 + 1$)), that

$$\|M_{\hat{Z}}\chi'\| + \|v'\| = s_{r_0+1}(\chi') + s_1(-v') \geq s_{r_0+1}(\chi' - v') = s_{r_0+1}(\Lambda Z') \geq s_r(\Lambda Z') = a_n. \tag{B.2.26}$$

By $\chi' = \zeta \text{Blockdiag}(L', R') + v'$ and the sub-multiplicative property of the

spectral norm,

$$\|M_{\hat{Z}}\chi'\| \leq \|M_{\hat{Z}}[F,X]\| \cdot \|\text{Blockdiag}(L',R')\| + \|v\|. \tag{B.2.27}$$

It follows, by (B.2.26) and (B.2.27), that

$$\begin{aligned}\|M_{\hat{Z}}[F,X]\| \geq \frac{a_n - 2\|v\|}{\|\text{Blockdiag}(L',R')\|} &= \frac{a_n - 2\|v\|}{\sqrt{\|\text{Blockdiag}(n_Y\Sigma_L, n_W\Sigma_R)\|}} \\ &\overset{(i)}{\geq} \frac{a_n - 2\|v\|}{\sqrt{\max\left\{n_Y\|\Sigma_L\|, n_W\|\Sigma_R\|\right\}}},\end{aligned} \tag{B.2.28}$$

where (i) holds by Lemma B.3.1(5). Since $\|v\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$ (Lemma B.2.2) and the eigenvalues of Σ_L and Σ_R are bounded, it follows by (B.2.25) and (B.2.28), that there exists a constant $C_2 > 0$ such that

$$\mathbb{P}\left(\|M_{\hat{Z}}[F,X]\| > \sqrt{n}C_2\right) \to 1. \tag{B.2.29}$$

By Lemma B.2.4(2) and (4), $\hat{F} = FH_F + \Delta_F$, where $H_F \in \mathbb{R}^{p_Y \times p_Y}$ is a (random) matrix with singular values bounded below by a positive constant with probability approaching one and $\|\Delta_F\| = O_{\mathbb{P}}(\log^{O(1)} n)$. By a similar argument (omitted), $\hat{X} = XH_X + \Delta_X$, where $H_X \in \mathbb{R}^{p_W \times p_W}$ and $\|\Delta_X\| = O_{\mathbb{P}}(\log^{O(1)} n)$ have the same property as H_F and Δ_X, respectively. By Lemma B.3.1(5), there exists a constant $C_3 > 0$ such that

$$\mathbb{P}\left(s_{p_Y+p_W}\left[\text{Blockdiag}(H_F, H_X)\right] > C_3\right) \to 1. \tag{B.2.30}$$

Moreover,

$$\begin{aligned}\|M_{\hat{Z}}[\hat{F},\hat{X}]\| &= \|M_{\hat{Z}}\left([F,X]\text{Blockdiag}(H_F,H_X) + [\Delta_F,\Delta_X]\right)\| \\ &\geq \|M_{\hat{Z}}[F,X]\text{Blockdiag}(H_F,H_X)\| - \|M_{\hat{Z}}[\Delta_F,\Delta_X]\| \\ &\overset{(i)}{\geq} \|M_{\hat{Z}}[F,X]\| \cdot s_{p_Y+p_W}\left[\text{Blockdiag}(H_F,H_X)\right] - \|M_{\hat{Z}}[\Delta_F,\Delta_X]\| \\ &\overset{(ii)}{\geq} \|M_{\hat{Z}}[F,X]\| \cdot s_{p_Y+p_W}\left[\text{Blockdiag}(H_F,H_X)\right] - O_{\mathbb{P}}(\log^{O(1)} n),\end{aligned} \tag{B.2.31}$$

where (i) holds by Lemma B.3.1 and the fact that $\|M_{\hat{Z}}[F,X]\| = s_1\{M_{\hat{Z}}[F,X]\}$ and (ii) follows by $\|M_{\hat{Z}}[\Delta_F,\Delta_X]\| \leq \|[\Delta_F,\Delta_X]\|$ ($M_{\hat{Z}}$ being a projection matrix). By the sub-additivity of probability measures, we have

$$\begin{aligned}
&\mathbb{P}\left(\|M_{\hat{Z}}[\hat{F},\hat{X}]\| \geq 2\sqrt{n}C_2/C_3\right) \\
&\geq \mathbb{P}\left(\|M_{\hat{Z}}[F,X]\|/C_3 - O_{\mathbb{P}}(\log^{O(1)} n) \geq 2\sqrt{n}C_2/C_3\right) \\
&\qquad - \mathbb{P}\left(s_{p_Y+p_W}\left[\text{Blockdiag}(H_F,H_X)\right] \leq C_3\right) \\
&\overset{(i)}{\to} 1,
\end{aligned} \tag{B.2.32}$$

where (i) holds by (B.2.29), (B.2.30) and (B.2.31).

Recall the elementary inequalty $\|A\| \leq \|\text{vec}A\|_2 \leq \sqrt{\dim(\text{vec}A)}\|A\|_\infty$ for any matrix A. Hence, $\|M_{\hat{Z}}[F,X]\| \leq \sqrt{T(p_Y+p_W)}\|M_{\hat{Z}}[F,X]\|_\infty$. This and (B.2.32) imply that there exists a constant $C_4 > 0$ such that

$$\mathbb{P}\left(\|M_{\hat{Z}}[\hat{F},\hat{X}]\|_\infty > C_4\right) \to 1. \tag{B.2.33}$$

Step 2: show that the critical value diverges at the rate $\log^{O(1)} n$.

Recall $\hat{\Psi}_i$, $\tilde{S}_n^{\hat{\Psi}}$ and $\mathcal{G}_n$ defined in (B.2.10). Notice that $\tilde{S}_n^{\hat{\Psi}} = S_n^{BS}$, where S_n^{BS} is defined in Algorithm 3. Notice that $\tilde{S}_n^{\hat{\Psi}}$, conditional on $\mathcal{G}_n$, is a zero mean Gaussian vector with its entries having a maximal variance bounded above by $\max_{1\leq i\leq n}\|\hat{\Psi}_i\|_\infty^2$. In other words, $\tilde{S}_n^{\hat{\Upsilon}}/\max_{1\leq i\leq n}\|\hat{\Psi}_i\|_\infty \in \Xi(1,2,\mathcal{G}_n)$. By Lemma B.3.6, $\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty/\max_{1\leq i\leq n}\|\hat{\Psi}_i\|_\infty = O_{\mathbb{P}}(\sqrt{\log n})$. From the proofs of Lemmas B.2.4 and B.2.7, it is not hard to show that, under H_1, $\max_{1\leq i\leq n}\|\hat{\Psi}_i\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$. Therefore, $\|\tilde{S}_n^{\hat{\Upsilon}}\|_\infty = O_{\mathbb{P}}(\log^{O(1)} n)$.

By (B.2.33), the test statistic $\sqrt{n}\|M_{\hat{Z}}[\hat{F},\hat{X}]\|_\infty$ diverges at the rate $\sqrt{n}$. Since the critical value only diverges at the rate $\log^{O(1)} n$, the probability that the test rejects the null hypothesis converges to one. The proof is complete. □

B.3 Technical tools

Lemma B.3.1. *The following hold.*
(1) Let $A, B \in \mathbb{R}^{n_1 \times n_2}$ *be two matrices. If* $i + j - 1 \leq \min\{n_1, n_2\}$, *then* $s_{i+j-1}(A + B) \leq s_i(A) + s_j(B)$.
(2) Let $A \in \mathbb{R}^{n_1 \times n_0}$ *and* $B \in \mathbb{R}^{n_0 \times n_2}$. *If* $1 \leq i \leq n_0$, *then* $s_i(AB) \geq s_i(A)s_{n_0-i+1}(B)$.
(3) Let $A, B \in \mathbb{R}^{n_1 \times n_2}$ *be two matrices. If* $\text{rank}B \leq k$ *and* $1 \leq j \leq \min\{n_1, n_2\} - k$, *then* $s_j(A) \geq s_{j+k}(A + B) \geq s_{2k+j}(A)$.
(4) Let $A \in \mathbb{R}^{n_1 \times n_2}$. *Suppose that* $B \in \mathbb{R}^{n_1 \times m}$ *consists of the first* m *columns of* A *with* $m \leq n_2$. *Then for* $1 \leq j \leq m \wedge n_1$, $s_j(B) \leq s_j(A)$.
(5) Let $A \in \mathbb{R}^{n_1 \times n_1}$ *and* $B \in \mathbb{R}^{n_2 \times n_2}$. *Then* $s_{n_1+n_2}[\text{Blockdiag}(A, B)] \geq \min\{s_{n_1}(A), s_{n_2}(B)\}$ *and* $s_1[\text{Blockdiag}(A, B)] \leq \max\{s_1(A), s_1(B)\}$.

Proof. Part (1) and (4) are Fact 6(b) and Fact 3, respectively, in Chapter 17.4 of Hogben (2006). Part (2) follows by Lemma 3 of Wang and Xi (1997). Part (3) follows by applying part (1): $s_j(A) = s_j(A) + s_{k+1}(B) \geq s_{j+k}(A + B)$ and $s_{j+k}(A + B) = s_{j+k}(A + B) + s_{k+1}(-B) \geq s_{2k+j}(A)$.

To see part (5), let $\lambda_{\max}$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues, respectively. Notice that

$$\begin{aligned}
\left(s_{n_1+n_2}[\text{Blockdiag}(A, B)]\right)^2 &= \lambda_{\min}\left(\text{Blockdiag}(A'A, B'B)\right) \\
&= \min_{x_1'x_1 + x_2'x_2 = 1} x_1'A'Ax_1 + x_2'B'Bx_2 \\
&\geq \min_{x_1'x_1 + x_2'x_2 = 1} \lambda_{\min}(A'A)\|x_1\|_2^2 + \lambda_{\min}(B'B)\|x_2\|_2^2 \\
&\geq \min\left\{\lambda_{\min}(A'A), \lambda_{\min}(B'B)\right\} \\
&= \min\left\{\left(s_{n_1}(A)\right)^2, \left(s_{n_2}(B)\right)^2\right\}.
\end{aligned}$$

This proves the first claim in part (5). Notice that

$$\begin{aligned}
\left(s_1[\text{Blockdiag}(A, B)]\right)^2 &= \lambda_{\max}\left(\text{Blockdiag}(A'A, B'B)\right) \\
&= \max_{x_1'x_1 + x_2'x_2 = 1} x_1'A'Ax_1 + x_2'B'Bx_2 \\
&\geq \max_{x_1'x_1 + x_2'x_2 = 1} \lambda_{\max}(A'A)\|x_1\|_2^2 + \lambda_{\max}(B'B)\|x_2\|_2^2
\end{aligned}$$

$$\geq \max\left\{\lambda_{\max}(A'A), \lambda_{\max}(B'B)\right\}$$
$$= \max\left\{\left(s_1(A)\right)^2, \left(s_1(B)\right)^2\right\}.$$

This proves the second claim in part (5). The proof is complete. □

Lemma B.3.2. *Let X and Y be two random vectors. Then $\forall t, \varepsilon > 0$, $|\mathbb{P}(\|X\|_\infty > t) - \mathbb{P}(\|Y\|_\infty > t)| \leq \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty \in (t-\varepsilon, t+\varepsilon])$.*

Proof. The result holds by the following observations using the triangular inequality: (1) $\mathbb{P}(\|X\|_\infty > t) \leq \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty > t - \varepsilon) = \mathbb{P}(\|X - Y\|_\infty > \varepsilon) + \mathbb{P}(\|Y\|_\infty > t) + \mathbb{P}(\|Y\|_\infty \in (t-\varepsilon, t])$ and (2) $\mathbb{P}(\|X\|_\infty > t) \geq \mathbb{P}(\|Y\|_\infty > t + \varepsilon) - \mathbb{P}(\|X-Y\|_\infty > \varepsilon) = \mathbb{P}(\|Y\|_\infty > t) - \mathbb{P}(\|Y\|_\infty \in (t, t+\varepsilon]) - \mathbb{P}(\|X-Y\|_\infty > \varepsilon)$. □

Lemma B.3.3. *Let X and Y be two random vectors and $\mathcal{F}$ and $\mathcal{G}$ two σ-algebras. Define $F_X(x) = \mathbb{P}(\|X\|_\infty \leq x \mid \mathcal{F})$ and $F_Y(x) = \mathbb{P}(\|Y\|_\infty \leq x \mid \mathcal{G})$. Then $\forall \varepsilon > 0$, $\sup_{\alpha \in (0,1)} \left|\mathbb{P}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \mid \mathcal{F}\right) - \alpha\right| \leq \varepsilon + \mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > \varepsilon \mid \mathcal{F}\right)$.*

Proof. For simplicity, we use $\mathbb{P}_{|\mathcal{F}}(\cdot)$ to denote $\mathbb{P}(\cdot \mid \mathcal{F})$. Fix $\alpha \in (0,1)$ and notice that

$$\begin{aligned}
&\mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha)\right) \\
&\leq \mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| \leq \varepsilon\right) \\
&\quad + \mathbb{P}_{|\mathcal{F}}\left(\sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > \varepsilon\right) \\
&\overset{(i)}{\leq} \mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_X^{-1}(1-\alpha-\varepsilon)\right) + \mathbb{P}_{|\mathcal{F}}\left(\sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > \varepsilon\right) \\
&= \alpha + \varepsilon + \mathbb{P}_{|\mathcal{F}}\left(\sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| > \varepsilon\right), \qquad \text{(B.3.1)}
\end{aligned}$$

where (i) follows from Lemma A.1(ii) in Romano and Shaikh (2012) (if $\sup_{x \in \mathbb{R}}[F_Y(x) - F_X(x)] \leq \varepsilon$ then $F_X^{-1}(1-\alpha-\varepsilon) \leq F_Y^{-1}(1-\alpha)$). Also notice that

$$\mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha)\right)$$

$$
\begin{aligned}
&\geq \mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_Y^{-1}(1-\alpha) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \leq \varepsilon\right) \\
&\overset{(i)}{\geq} \mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_X^{-1}(1-\alpha+\varepsilon) \text{ and } \sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| \leq \varepsilon\right) \\
&\geq \mathbb{P}_{|\mathcal{F}}\left(\|X\|_\infty > F_X^{-1}(1-\alpha+\varepsilon)\right) - \mathbb{P}_{|\mathcal{F}}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right) \\
&= \alpha - \varepsilon - \mathbb{P}_{|\mathcal{F}}\left(\sup_{x\in\mathbb{R}}|F_X(x)-F_Y(x)| > \varepsilon\right) \qquad \text{(B.3.2)}
\end{aligned}
$$

where (i) follows from Lemma A.1(ii) in Romano and Shaikh (2012) (if $\sup_{x\in\mathbb{R}}[F_X(x)-F_Y(x)] \leq \varepsilon$ then $F_Y^{-1}(1-\alpha) \leq F_X^{-1}(1-\alpha+\varepsilon)$). The desired result follows by (B.3.1) and (B.3.2). □

Lemma B.3.4. *Let $Y = (Y_1, \cdots, Y_p)'$ be a random vector and $\mathcal{F}$ a σ-algebra. If $\mathbb{E}(Y \mid \mathcal{F}) = 0$, $Y \mid \mathcal{F}$ is Gaussian and $\min \mathbb{E}(Y_j^2 \mid \mathcal{F}) \geq b$ a.s. for some constant $b > 0$, then there exists a constant $C_b > 0$ depending only on b such that $\forall \varepsilon > 0$.*

$$
\sup_{x\in\mathbb{R}} \mathbb{P}\left(\|Y\|_\infty \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}\right) \leq C_b \varepsilon \sqrt{\log p} \quad a.s.
$$

Proof. By Nazarov's anti-concentration inequality (Lemma A.1 in Chernozhukov, Chetverikov, and Kato (2014)), there exists a constant C_b' depending only on b such that almost surely, $\sup_{x\in\mathbb{R}} \mathbb{P}(\max_{1\leq j\leq p} Y_j \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}) \leq 2C_b'\varepsilon\sqrt{\log p}$ and $\sup_{x\in\mathbb{R}} \mathbb{P}(\max_{1\leq j\leq p}(-Y_j) \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}) \leq 2C_b'\varepsilon\sqrt{\log p}$.

Since $\|Y\|_\infty = \max\{\max_{1\leq j\leq p} Y_j, \max_{1\leq j\leq p}(-Y_j)\}$, the desired result follows by $\sup_{x\in\mathbb{R}} \mathbb{P}(\|Y\|_\infty \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}) \leq \sup_{x\in\mathbb{R}} \mathbb{P}(\max_{1\leq j\leq p} Y_j \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}) + \sup_{x\in\mathbb{R}} \mathbb{P}(\max_{1\leq j\leq p}(-Y_j) \in (x-\varepsilon, x+\varepsilon] \mid \mathcal{F}) \leq 4C_b'\varepsilon\sqrt{\log p}$. □

Lemma B.3.5. *The following hold.*

(1) Let $X \in \mathbb{R}^{m_X}$ be a random vector whose jth entry is denoted by X_j. Suppose that there exist constants $b, \gamma > 0$ such that $\forall j \in \{1, \cdots, m_X\}$, X_j has an exponential-type tail with parameter (b, γ). Then for any nonrandom vector $a \in \mathbb{R}^{m_X}$, $a'X$ has an exponential-type tail with parameter $\left(b\|a\|_1 \log^{1/\gamma}(\|a\|_0 + 2), \gamma\right)$.

(2) Let $\{X_j\}_{j=1}^{m_X}$ be a sequence of random variables. Suppose that constants $b, \gamma > 0$ satisfy that $\forall j \in \{1, \cdots, m_X\}$, X_j has an exponential-type tail with parameter (b, γ). Let $q > 0$ be any nonrandom number. Then there exists a constant $C_{\gamma,q} > 0$ depend-

ing only on γ and q such that $\mathbb{E}\max_{1\le j\le m_X}|X_j|^q \le C_{\gamma,q} m_X b^q$ and $\mathbb{E}|X_j|^q \le C_{\gamma,q} b^q$ $\forall j \in \{1,\cdots,m_X\}$.

(3) Let X_1 and X_2 be two random variables having exponential-type tails with parameters (b_1,γ_1) and (b_2,γ_2), respectively. Then X_1X_2 has an exponential-type tail with parameter $\left(2^{1/\gamma_0}b_1b_2,\gamma_0\right)$, where $\gamma_0 = \gamma_1\gamma_2(\gamma_1+\gamma_2)^{-1}$

(4) Let X have an exponential-type tail with parameter (b_X,γ_X). Then $\forall a\in\mathbb{R}$, $X-a$ has an exponential-type tail with parameter $(b_X+|a|,\gamma_X)$.

(5) Let $K_1,K_2,K_3,K>0$ be constants such that $\forall d\ge K_1$, $\mathbb{P}(|X|\ge d)\le K_2\exp(-(d/K_3)^K)$. Then X has an exponential-type tail with parameter (c,K), where $c>0$ is a constant depending only on K_1, K_2, K_3 and K.

Proof. **Proof of part (1).** Let $A_0 := \{i \mid a_i \neq 0\}$. Then by Holder's inequality and the union bound, $\mathbb{P}(|a'X|>x) \le \mathbb{P}(\|a\|_1\max_{i\in A_0}|X_i|>x) \le \sum_{i\in A_0}\mathbb{P}(\|a\|_1|X_i|>x) \le \|a\|_0\exp\left[1-(xb^{-1}\|a\|_1^{-1})^\gamma\right]$. If $\|a\|_0=1$, then the result follows by $b\|a\|_1 < b\|a\|_1\log^{1/\gamma}(3)$. For $\|a\|_0>1$, we let $c = b\|a\|_1\log^{1/\gamma}\|a\|_0 < b\|a\|_1\log^{1/\gamma}(\|a\|_0+2)$. For $x\le c$, $\mathbb{P}(|a'X|>x)\le 1\le\exp(1-(x/c)^\gamma)$. Since $\mathbb{P}(|a'X|>x)\le\|a\|_0\exp\left[1-(xb^{-1}\|a\|_1^{-1})^\gamma\right]$, it suffices to show that $\forall x>c$, $\log\|a\|_0-(xb^{-1}\|a\|_1^{-1})^\gamma \le 1-(xc^{-1})^\gamma$. This is to say that $x^\gamma \ge (\log\|a\|_0-1)/((b\|a\|_1)^{-\gamma}-c^{-\gamma})$ $\forall x>c$. By simple computations, one can show that $c^\gamma = (\log\|a\|_0-1)/((b\|a\|_1)^{-\gamma}-c^{-\gamma})$. Part (1) follows.

Proof of part (2). Notice that, by the union bound, $\mathbb{P}(\max_{1\le j\le m_X}|X_j|>x)\le\sum_{j=1}^{m_X}\mathbb{P}(|X_j|>x)\le m_X\exp\left[1-(x/b)^\gamma\right]$. Then

$$\mathbb{E}\max_{1\le j\le m_X}|X_j|^q \overset{(i)}{=} \int_0^\infty \mathbb{P}\left(\max_{1\le j\le m_X}|X_j|^q>x\right)dx = \int_0^\infty \mathbb{P}\left(\max_{1\le j\le m_X}|X_j|>x^{1/q}\right)dx$$
$$\overset{(ii)}{\le} m_X\int_0^\infty \exp\left[1-\left(x^{1/q}/b\right)^\gamma\right]dx \overset{(iii)}{=} m_Xb^q\left(q\gamma^{-1}\int_0^\infty e^{1-d}d^{q/\gamma-1}dd\right),$$

where (i) follows by the identity $\mathbb{E}X=\int_0^\infty\mathbb{P}(X>x)dx$ for any non-negative random variable X, (ii) follows by $\mathbb{P}(\max_{1\le j\le m_X}|X_j|>x)\le m_X\exp\left[1-(x/b)^\gamma\right]$ and (iii) follows by a change of variable $d=\left(x^{1/q}/b\right)^\gamma$. The bound for $\mathbb{E}\max_{1\le j\le m_X}|X_j|^q$ follows with $C_{\gamma,q}=q\gamma^{-1}\int_0^\infty e^{1-d}d^{q/\gamma-1}dd$. The bound for $\mathbb{E}|X_j|^q$ follows by the same reasoning with $\max_{1\le j\le m_X}|X_j|$ replaced by $|X_j|$. This completes the proof

for part (2).

Proof of part (3). The proof of Lemma A.2 of Fan, Liao, and Mincheva (2011) implies that $\forall \gamma \in (0, \gamma_0)$, $X_1 X_2$ has an exponential-type tail with parameter (b_3, γ) $\forall b_3 > b_0 \max\{(\gamma/\gamma_0)^{1/\gamma_0}, (1+\log 2)^{1/\gamma_0}\}$, where $\gamma_0 = \gamma_1\gamma_2(\gamma_1+\gamma_2)^{-1}$ and $b_0 = b_1 b_2$. It is easy to check that $2^{(\gamma_1+\gamma_2)\gamma_1^{-1}\gamma_2^{-1}} b_1 b_2 > b_0 \max\{(\gamma/\gamma_0)^{1/\gamma_0}, (1+\log 2)^{1/\gamma_0}\}$. Part (3) follows.

Proof of part (4). Let $c = b_X + |a|$. Notice that $\mathbb{P}(|X-a| > t) \leq \mathbb{P}(|X| + |a| > t) = \mathbb{P}(|X| > t - |a|)$. For $t \in (0, c]$, $\mathbb{P}(|X| > t - |a|) \leq 1 \leq \exp[1 - (t/c)^{\gamma_X}]$. For $t > c$, $t - |a| > 0$ and $\mathbb{P}(|X| > t - |a|) \leq \exp[1 - ((t-|a|)/b_x)^{\gamma_X}]$. It is easy to check that $(t-|a|)/b_x \geq t/c$ $\forall t > c$. Part (4) follows.

Proof of part (5). It is easy to see that there exist constants $c_1, c_2 > 0$ such that $c_1 \geq K_1$, $c_2 \geq K_3 \vee c_1$ and $\log K_2 - 1 \leq (K_3^{-K} - c_2^{-K})c_1^K$. Now we verify that we can take $c = c_2$. For $d \in (0, c_2]$, $\mathbb{P}(|X| > d) \leq 1 \leq \exp(1 - (x/c_2)^K)$; for $d > c_2$, $\mathbb{P}(|X| > d) \leq K_2 \exp(-(d/K_3)^K)$ and it is straight-forward to check that $\forall d > c_2$, $K_2 \exp(-(d/K_3)^K) \leq \exp(1 - (d/c_2)^K)$. Part (5) follows. □

Lemma B.3.6. *Let $\{W_j\}_{j\in J}$ be random variables. If there exist constant $b, \gamma > 0$ such that $\forall j \in J$, W_j has an exponential-type tail with parameter (b, γ), then $\max_{j\in J} |W_j| = O_{\mathbb{P}}(\log^{1/\gamma} |J|)$, where $|J|$ is the cardinality of J.*

Proof. By the union bound, we have

$$\begin{aligned}\mathbb{P}\left(\max_{j\in J} |W_j| > (\log|J|)^{1/\gamma} x\right) &\leq \sum_{j\in J} \mathbb{P}\left(|W_j| > (\log|J|)^{1/\gamma} x\right) \\ &\leq |J| \exp\left[1 - \left((\log|J|)^{1/\gamma} x/b\right)^{\gamma}\right] = \exp\left[1 + (1 - (x/b)^{\gamma}) \log|J|\right].\end{aligned}$$

Hence, for any $\varepsilon > 0$, one can choose large enough x such that the right-hand side of the above display is smaller than ε. The result follows. □

Lemma B.3.7. *Let $\mathcal{F}_n$ be a σ-algebra and $\{W_t\}_{t=1}^T$ be random variables with $\mathbb{E}(W_t \mid \mathcal{F}_n) = 0$. Suppose that the following hold:*
(i) There exist constants $\gamma_1, b_1 > 0$ such that $\forall t \in [T]$, $W_t \in \Xi(b_1, \gamma_1, \mathcal{F}_n)$
(ii) There exist constants $\gamma_2, b_2 > 0$ such that $\alpha_n(t \mid \mathcal{F}_n) \leq \exp(-b_2 t^{\gamma_2})$ and

$\gamma := \left[\gamma_1^{-1} + \gamma_2^{-1}\right]^{-1} < 1$ *a.s, where*

$$\alpha_n(t \mid \mathcal{F}_n) := \sup\Big\{\Big|\mathbb{P}(A \mid \mathcal{F}_n)\mathbb{P}(B \mid \mathcal{F}_n) - \mathbb{P}(A \bigcap B \mid \mathcal{F}_n)\Big| : A \in \sigma\left(\{(W_s, \cdots, W_s) \mid s \leq \iota\}\right), B \in \sigma\left(\{(W_s, \cdots, W_s) \mid s \geq \iota + t\}\right) \text{ and } \iota \in \mathbb{N}\Big\}.$$

Then $T^{-1/2}\sum_{t=1}^{T} W_t \in \Xi(b_*, \gamma, \mathcal{F}_n)$*, where* $b_* > 0$ *is a constant depending only on* γ_1*,* γ_2*,* b_1 *and* b_2*.*

Proof. Let $K > 0$ be a constant to be chosen later. By Theorem 1 in Merlevède, Peligrad, and Rio (2011) (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), there exist constants $C_1, C_2, C_3, C_4, C_5 > 0$ depending only on γ_1, γ_2, b_1 and b_2, such that $\forall d \geq K$,

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} W_t\right| > dT^{1/2} \mid \mathcal{F}_n\right) \leq \underbrace{T\exp\left(-C_1 T^{\gamma/2} d^{\gamma}\right)}_{J_{1,T}(d)} + \underbrace{\exp\left(-\frac{C_2 d^2 T}{1 + C_3 T}\right)}_{J_{2,T}(d)} + \underbrace{\exp\left[-C_4 d^2 \exp\left(C_5 \frac{(T^{1/2}d)^{\gamma/(1-\gamma)}}{[\log(T^{1/2}d)]^{\gamma}}\right)\right]}_{J_{3,T}(d)} \text{ a.s.}$$

It is not hard to see that one can choose a large enough K such that $\forall d \geq K$, $J_{1,T}(d) \leq \exp(-C_1 d^{\gamma})$, $J_{3,T}(d) \leq J_{1,T}(d)$ and $J_{2,T}(d) \leq \exp(-C_6 d^2)$, where $C_6 = C_2/(1 + C_3)$. Hence, $\forall d \geq K$, $J_{1,T}(d) + J_{2,T}(d) + J_{3,T}(d) \leq 2\exp(-C_1 d^{\gamma}) + \exp(-C_6 d^2)$. Since $\gamma < 1$, we can enlarge K, if necessary, such that $\forall d \geq K$, $\exp(-C_6 d^2) \leq \exp(-C_1 d^{\gamma})$.

Hence, $\forall d \geq K$, $\mathbb{P}\left(T^{-1/2}\left|\sum_{t=1}^{T} W_t\right| > d \mid \mathcal{F}_n\right) \leq 3\exp\left(-C_1 d^{\gamma}\right)$ a.s. Thus, the desired result follows by Lemma B.3.5. □

Lemma B.3.8. *Let* $\{x_{i,j}\}_{(i,j)\in[n]\times J}$ *be an array of random variables and* $\mathcal{F}_n$ *be a* σ*-algebra. Suppose the following hold:*

(i) Condition on $\mathcal{F}_n$*,* x_i *is independent across* i*, where* $x_i = \{x_{i,j} \mid j \in J\}$*.*

(ii) $\mathbb{E}(x_{i,j} \mid \mathcal{F}_n) = 0$ $\forall (i,j) \in [n] \times J$*.*

(iii) There exist constants $b, \gamma > 0$ *such that* $\forall (i,j) \in [n] \times J$ *and* $\forall x > 0$*,* $\mathbb{P}(|x_{i,j}| >$

$x \mid \mathcal{F}_n) \leq \exp\left(1 - (x/b)^\gamma\right)$ *a.s.*

(iv) $\forall 0 < c < \infty$, $n^{-c} \log |J| \to 0$, *where* $|J|$ *denotes the cardinality of* J.

Then $\max_{j \in J} \left|\sum_{i=1}^n \left[x_{i,j} - \mathbb{E}(x_{i,j} \mid \mathcal{F}_n)\right]\right| = O_{\mathbb{P}}(\sqrt{n \log |J|})$.

Proof. Fix an arbitrary $\varepsilon > 0$. Let $\tilde{x}_{i,j} = x_{i,j} - \mathbb{E}(x_{i,j} \mid \mathcal{F}_n)$. By Lemma B.3.5(2) and (4) applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$, we have that there exists a constant $b_1 > 0$ depending only on b and γ such that $\forall d > 0\, \forall (i,j) \in [n] \times J$, $\mathbb{P}(|\tilde{x}_{i,j}| > x \mid \mathcal{F}_n) \leq \exp\left(1 - (x/b_1)^\gamma\right)$ a.s.

Then by Theorem 1 in Merlevède, Peligrad, and Rio (2011) (applied to the conditional probability measure $\mathbb{P}(\cdot \mid \mathcal{F}_n)$), there exist positive constants C_1, C_2, C_3, C_4, C_5 and r depending only on bM_ε and γ such that $r < 1$ and $\forall d > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^n \tilde{x}_{i,j}\right| > d\sqrt{n \log |J|} \mid \mathcal{F}_n\right) \\
&\leq n \exp\left[-C_1 \left(d\sqrt{n \log |J|}\right)^r\right] + \exp\left[-\frac{C_2 n d^2 \log |J|}{1 + nC_3}\right] \\
&+ \exp\left\{-C_4 d^2 \log |J| \exp\left[C_5 \log^{-r}\left(d\sqrt{n \log |J|}\right)\left(d\sqrt{n \log |J|}\right)^{r/(1-r)}\right]\right\} \ a.s.
\end{aligned}
$$

Then, by the union bound, we have that

$$
\begin{aligned}
&\mathbb{P}\left(\max_{j \in J}\left|\sum_{i=1}^n \tilde{x}_{i,j}\right| > d\sqrt{n \log |J|} \mid \mathcal{F}_n\right) \\
&\leq \sum_{j \in J} \mathbb{P}\left(\max_{j \in J}\left|\sum_{i=1}^n \tilde{x}_{i,j}\right| > d\sqrt{n \log |J|} \mid \mathcal{F}_n\right) \\
&\leq |J| n \exp\left[-C_1 \left(d\sqrt{n \log |J|}\right)^r\right] + |J| \exp\left[-\frac{C_2 n d^2 \log |J|}{1 + nC_3}\right] \\
&+ |J| \exp\left\{-C_4 d^2 \log |J| \exp\left[C_5 \log^{-r}\left(d\sqrt{n \log |J|}\right)\left(d\sqrt{n \log |J|}\right)^{r/(1-r)}\right]\right\} \ a.s.
\end{aligned}
$$

By assumption (iv), the first and third terms in the above display go to zero for any $d > 0$. Hence, we can choose a large constant $d_* > 0$ and $n_* \in \ltimes$ such that $\forall n \geq n_*$, $\mathbb{P}\left(\max_{j \in J} |\sum_{i=1}^n \tilde{x}_{i,j}| > d_* \sqrt{n \log |J|} \mid \mathcal{F}_n\right) \leq \varepsilon$ a.s. The result follows by the law of iterated expectations. □

B.4 An example of difficult low-dimensional asymptotics

By Lemma 2.2.1, H_0 in (2.2.3) holds if and only if $\text{rank}[F, X] = p_Y + p_W - k_0$, which is equivalent to the condition $\text{rank}\Sigma_\zeta = p_Y + p_W - k_0$. Hence, one way of testing H_0 is to test the rank of a low-dimensional matrix $\Sigma_\zeta \in \mathbb{R}^{p\times p}$, where $p = p_Y + p_W$. Let $\lambda_{k_0}(\cdot)$ denote the sum of the smallest k_0 eigenvalues. We can test whether $\lambda_{k_0}(\Sigma_\zeta) = 0$.

For simplicity, assume that (F, X, L, R) is nonrandom with $\Sigma_L = I_{p_Y}$ and $\Sigma_R = I_{p_W}$ so the normalization imposed in PCA is correct. A natural estimator for Σ_ζ is $\hat{\Sigma}_\zeta = T^{-1}\hat{\zeta}'\hat{\zeta}$, where $\hat{\zeta} = [\hat{F}, \hat{X}]$. It is not hard to show that $T\text{vec}(\hat{\Sigma}_\zeta - \Sigma_\zeta) \to^d N(a, \Omega)$ for some vector $a \in \mathbb{R}^{p^2}$ and matrix $\Omega \in \mathbb{R}^{p\times p}$. It is well known that, under H_0, $\lambda_{k_0}(\cdot)$ is a smooth function in a neighborhood of Σ_ζ.

The first difficulty is that a is nonzero and is not straight-forward to estimate. It is not hard to show that $a = \text{vec}(\text{Blockdiag}(n_Y^{-1}L'\mathbb{E}(ee')L, n_W^{-1}R'\mathbb{E}(uu')R))$. Let $\hat{e} = Y - \hat{L}\hat{F}'$ and $\hat{u} = W - \hat{R}\hat{X}'$. Notice that by construction, $\hat{L}'\hat{e} = 0$ and $\hat{R}'\hat{u} = 0$. This means that $n_Y^{-1}\hat{L}'\hat{e}\hat{e}'\hat{L} = 0$ and $n_W^{-1}\hat{R}'\hat{u}\hat{u}'\hat{R} = 0$. Therefore, it is not clear how to consistently estimate a.

The second difficulty is more problematic and arises due to the singularity of Ω, which requires us to take into account the asymptotic distribution of the error in approximating the distribution of $T\text{vec}(\hat{\Sigma}_\zeta - \Sigma_\zeta)$ with $N(a, \Omega)$. To see this, let $\lambda'_{k_0}(\cdot)$ denote the derivative of $\lambda_{k_0}(\cdot)$. By the first order Taylor's expansion,

$$T[\lambda_{k_0}(\hat{\Sigma}_\zeta) - \lambda_{k_0}(\Sigma_\zeta)] = T\lambda'_{k_0}(\Sigma_\zeta)\text{vec}(\hat{\Sigma}_\zeta - \Sigma_\zeta) + O_{\mathbb{P}}(T^{-1}).$$

Hence, the limiting distribution of $T[\lambda_{k_0}(\hat{\Sigma}_\zeta) - \lambda_{k_0}(\Sigma_\zeta)]$ is Gaussian with variance $[\lambda'_{k_0}(\Sigma_\zeta)]'\Omega\lambda'_{k_0}(\Sigma_\zeta)$. Since under H_0, $\lambda_{k_0}(\hat{\Sigma}_\zeta) \geq 0 = \lambda_{k_0}(\Sigma_\zeta)$, the limiting distribution of $T[\lambda_{k_0}(\hat{\Sigma}_\zeta) - \lambda_{k_0}(\Sigma_\zeta)]$ cannot be Gaussian with nonzero variance. Therefore, $\lambda'_{k_0}(\Sigma_\zeta)\Omega = 0$. Since $\lambda'_{k_0}(\Sigma_\zeta)$ can be shown to be nonzero, Ω is singular.

Since $\lambda'_{k_0}(\Sigma_\zeta)a \neq 0$, we need to remove the bias in $\hat{\Sigma}_\zeta$ and consider

$$T^{3/2}[\lambda_{k_0}(\hat{\Sigma}_\zeta - T^{-1}\hat{a}) - \lambda_{k_0}(\Sigma_\zeta)] = T^{3/2}\lambda'_{k_0}(\Sigma_\zeta)\text{vec}(\hat{\Sigma}_\zeta - \Sigma_\zeta - T^{-1}\hat{a}) + O_{\mathbb{P}}(T^{-1/2}),$$

where $\hat{a}$ is an estimate for a. Suppose that we have strong approximation $T\text{vec}(\hat{\Sigma}_\zeta - \Sigma_\zeta) = a + J_1 + J_{2,T}$, where $J_1 \sim N(0, \Omega)$ and $J_{2,T} = o_{\mathbb{P}}(1)$. Then, by the above display and $\lambda'_{k_0}(\Sigma_\zeta)\Omega = 0$, we have

$$T^{3/2}[\lambda_{k_0}(\hat{\Sigma}_\zeta - T^{-1}\hat{a}) - \lambda_{k_0}(\Sigma_\zeta)] = \lambda'_{k_0}(\Sigma_\zeta)\left[\sqrt{T}(a - \hat{a}) + \sqrt{T}J_{2,T}\right] + o_{\mathbb{P}}(1).$$

Although we might be able to find an estimator $\hat{a}$ such that it is possible to derive the asymptotics of $\sqrt{T}(\hat{a} - a)$, dealing with $\sqrt{T}J_{2,T}$, which does not vanish, is much harder. This is because $J_{2,T}$ contains the error of approximating terms such as $n_Y^{-1}\sum_{i=1}^{n_Y}\sum_{t=1}^{T} e_{i,t}\zeta_t L'_i$ with Gaussian distributions. Consequently, we need to resort to higher order Edgeworth expansions. Due to the intertemporal dependence in $e_{i,t}$, these expansions could be very complicated; see Gotze and Hipp (1994) and Lahiri (2010). In addition, we have to take into account the dependence between $\sqrt{T}(a - \hat{a})$ and $\sqrt{T}J_{2,T}$.

The third difficulty is that since Ω is singular with $\lambda'_{k_0}(\Sigma_\zeta)\Omega = 0$, we have reasons to suspect that the product of $\lambda'_{k_0}(\Sigma_\zeta)$ and some higher order terms in the Edgeworth expansion is also degenerate. For some estimators $\hat{a}$, $\lambda'_{k_0}(\Sigma_\zeta)\sqrt{T}(a - \hat{a})$ might also be degenerate in the limit. In some cases, we also need to include higher order terms in the Taylor expansion of the function $\lambda_{k_0}(\cdot)$. Hence, quite complicated arguments and possibly additional assumptions are required to determine which orders in expansions are needed.

These difficulties still arise even if one replaces $\lambda_{k_0}(\cdot)$ with other functions, such as canonical correlations.

Appendix C

Proofs for Chapter 3

In the rest of the article, we use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the minimal and maximal eigenvalues of a matrix, respectively. For a random variable, let $\|\cdot\|_{L^r(P)}$ denote the $L^r(P)$-norm, i.e., $\|z_i\|_{L^r(P)} = [Ez_i^r]^{1/r}$. For a vector $x = (x_1, \cdots, x_p)^\top \in \mathbb{R}^p$, let $\mathcal{M}(x)$ denotes its support $\{i \mid x_i \neq 0\}$.

C.1 Proof Theorems 3.2.1 and 3.2.2

Proof of Theorem 3.2.1. Under H_0 in (3.1.2), $l_i(g_0) = z_i(\varepsilon_i + w_i^\top \beta_*)$. Notice that

$$\sigma_l^2 = El_i(g_0)^2 = \sigma_z^2\sigma_\varepsilon^2 + Ez_i^2(w_i^\top\beta_*)^2 \geq \sigma_z^2\sigma_\varepsilon^2.$$

Hence, $s_n^2 := \sum_{i=1}^n E(l_i(g_0))^2 \geq n\sigma_z^2\sigma_\varepsilon^2$. It follows that

$$\begin{aligned}\frac{\sum_{i=1}^n E|l_i(g_0)|^3}{s_n^3} &\leq \frac{E|z_i(\varepsilon_i + w_i^\top\beta_*)|^3}{n^{1/2}\sigma_\varepsilon^3\sigma_z^3} \\ &\overset{(i)}{\leq} \frac{\sqrt{\|z_i\sigma_z^{-1}\|_{L^6(P)}^6\|\varepsilon_i + w_i^\top\beta_*\|_{L^6(P)}^6}}{n^{1/2}c^3} \overset{(ii)}{=} o(1),\end{aligned}$$

where (i) follows by Holder's inequality and (ii) follows by Assumption 4 and Minkowski's inequality $\|\varepsilon_i + w_i^\top\beta_*\|_{L^6(P)} \leq \|\varepsilon_i\|_{L^6(P)} + \|w_i^\top\beta_*\|_{L^6(P)} = O(1)$. By Lyapunov's CLT (Theorem 11.1.4 of Athreya and Lahiri (2006)), $\sum_{i=1}^n l_i(g_0)/s_n \to^d \mathcal{N}(0,1)$.

By Slutsky's lemma, it suffices to show that $s_n/\sqrt{n^{-1}\sum_{i=1}^n l_i(g_0)^2} \to^p 1$. Notice that this is equivalent to the condition

$$n^{-1}\sum_{i=1}^n \left(\frac{l_i(g_0)^2}{El_i(g_0)^2} - 1\right) = o_P(1). \tag{C.1.1}$$

By Markov's inequality, we have that, for any $M > 0$,

$$P\left[\left(n^{-1}\sum_{i=1}^n \left(\frac{l_i(g_0)^2}{El_i(g_0)^2} - 1\right)\right)^2 > M\right] \leq M^{-1}n^{-1}E\left(\frac{l_i(g_0)^2}{El_i(g_0)^2} - 1\right)^2 \tag{C.1.2}$$

$$\overset{(i)}{\leq} 2M^{-1}n^{-1}\left[\frac{El_i(g_0)^4}{\left[El_i(g_0)^2\right]^2} + 1\right], \tag{C.1.3}$$

where (i) holds by the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

By Holder's inequality and Assumption 4,

$$El_i(g_0)^4\sigma_z^{-4} \leq \sqrt{\|z_i\sigma_z^{-1}\|_{L^8(P)}^8 \|\varepsilon_i + w_i^\top\beta_*\|_{L^8(P)}^8} < C_0$$

for some constant $C_0 > 0$, depending only on C. Since $El_i(g_0)^2 \geq \sigma_z^2\sigma_\varepsilon^2 \geq \sigma_z^2 c^2$, we have

$$El_i(g_0)^4 / \left[El_i(g_0)^2\right]^2 \leq C_0 c^{-4} < \infty.$$

This, together with (C.1.3), implies (C.1.1). The proof is complete. □

Proof of Theorem 3.2.2. Since the eigenvalues of Σ_X are bounded away from zero and infinity, we have $\sigma_z^2 = Ez_i^2 = b^\top\Sigma_X b = (a^\top\Omega_X a)^{-1} \asymp \|a\|_2^{-2}$. It follows, by $\sqrt{n}|h_n|/\|a\|_2 \to \infty$, that

$$\sqrt{n}|h_n|\sigma_z \to \infty. \tag{C.1.4}$$

It should be noted that when $a^\top\beta_* = g_0 + h_n$, we have $l_i(g_0) = z_i(\varepsilon_i + w_i^\top\beta_*) + z_i^2 h_n$. Also note that (C.1.3) in the proof of Theorem 3.2.1 still holds, in that for all $M > 0$,

$$P\left[\left(n^{-1}\sum_{i=1}^n \left(\frac{l_i(g_0)^2}{El_i(g_0)^2} - 1\right)\right)^2 > M\right] \leq 2M^{-1}n^{-1}\left[\frac{El_i(g_0)^4}{\left[El_i(g_0)^2\right]^2} + 1\right]. \tag{C.1.5}$$

Observe that, by Assumption 4,

$$\left\|\frac{l_i(g_0)}{\sigma_z(\sigma_z|h_n|\vee 1)}\right\|_{L^4(P)} \le \left\|\frac{z_i(\varepsilon_i + w_i^\top\beta_*)}{\sigma_z(\sigma_z|h_n|\vee 1)}\right\|_{L^4(P)} + \left\|\frac{z_i^2 h_n}{\sigma_z(\sigma_z|h_n|\vee 1)}\right\|_{L^4(P)} \quad \text{(C.1.6)}$$

$$\le \left\|\frac{z_i(\varepsilon_i + w_i^\top\beta_*)}{\sigma_z}\right\|_{L^4(P)} + \left\|z_i\sigma_z^{-1}\right\|_{L^8(P)}^2 \frac{\sigma_z|h_n|}{\sigma_z|h_n|\vee 1} \quad \text{(C.1.7)}$$

$$\le \left\|z_i\sigma_z^{-1}\right\|_{L^8(P)}^2 \left\|\varepsilon_i + w_i^\top\beta_*\right\|_{L^8(P)}^2 + O(1) = O(1). \quad \text{(C.1.8)}$$

Observe that

$$El_i(g_0)^2 = E\left(z_i\varepsilon_i + z_i(w_i^\top\beta_* + z_i h_n)\right)^2 = E(z_i^2\varepsilon_i^2) + E(z_i^2(w_i^\top\beta_* + z_i h_n)^2) \ge \sigma_z^2\sigma_\varepsilon^2.$$

Also, we have $El_i(g_0)^2 \ge [El_i(g_0)]^2 = \sigma_z^4 h_n^2$. Hence,

$$El_i(g_0)^2 \ge (\sigma_z^4 h_n^2 \vee \sigma_z^2\sigma_\varepsilon^2) = \sigma_z^2(\sigma_z^2 h_n^2 \vee \sigma_\varepsilon^2).$$

This, together with (C.1.8) and Assumption 4, implies that

$$\frac{El_i(g_0)^4}{\left[El_i(g_0)^2\right]^2} \le \frac{O(1)\left[\sigma_z(\sigma_z|h_n|\vee 1)\right]^4}{\left[\sigma_z^2(\sigma_z^2 h_n^2 \vee \sigma_\varepsilon^2)\right]^2} \le O(1)\frac{\sigma_z^4 h_n^4 \vee 1}{\sigma_z^4 h_n^4 \vee c^4} \le O(1)(1\vee c^{-4}). \quad \text{(C.1.9)}$$

It follows, by (C.1.5) and (C.1.9), that $n^{-1}\sum_{i=1}^n (l_i(g_0)^2/El_i(g_0)^2 - 1) = o_P(1)$, which means that

$$\frac{n^{-1}\sum_{i=1}^n l_i(g_0)^2}{El_i(g_0)^2} = 1 + o_P(1). \quad \text{(C.1.10)}$$

By Markov's inequality, we have that, $\forall M > 0$,

$$P\left[\left(\frac{n^{-1/2}\sum_{i=1}^n (l_i(g_0) - El_i(g_0))}{\sqrt{El_i(g_0)^2}}\right)^2 > M\right] \le M^{-1}\frac{E\left[l_i(g_0) - El_i(g_0)\right]^2}{El_i(g_0)^2}$$

$$= M^{-1}\frac{El_i(g_0)^2 - [El_i(g_0)]^2}{El_i(g_0)^2} \le M^{-1}.$$

Hence,

$$\frac{n^{-1/2}\sum_{i=1}^n (l_i(g_0) - El_i(g_0))}{\sqrt{El_i(g_0)^2}} = O_P(1). \quad \text{(C.1.11)}$$

Lastly, we observe that

$$\begin{aligned} El_i(g_0)^2 &= E\left[z_i(\varepsilon_i + w_i^\top \beta_*) + z_i^2 h_n\right]^2 \\ &\overset{(i)}{\leq} 2Ez_i^2(\varepsilon_i + w_i^\top \beta_*)^2 + 2Ez_i^4 h_n^2 \\ &\overset{(ii)}{\leq} 2\sigma_z^2\sqrt{E(z_i\sigma_z^{-1})^4 E(\varepsilon_i + w_i^\top \beta_*)^4} + 2E(z_i\sigma_z^{-1})^4\sigma_z^4 h_n^2 \\ &\overset{(iii)}{\leq} O(1) + O(1)\sigma_z^4 h_n^2, \end{aligned}$$

where (i) follows according to the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, (ii) follows by Holder's inequality and (iii) is determined by Assumption 4. Since $El_i(g_0) = \sigma_z^2 h_n$, we have

$$\left|\frac{n^{-1/2}\sum_{i=1}^n El_i(g_0)}{\sqrt{El_i(g_0)^2}}\right| \geq \frac{\sqrt{n}\sigma_z^2|h_n|}{\sqrt{O(1) + O(1)\sigma_z^4 h_n^2}} = \left(\frac{O(1)}{n\sigma_z^2 h_n^2} + O(n^{-1})\right)^{-1/2} \to \infty, \tag{C.1.12}$$

where the last step follows by (C.1.4). The desired result follows by (C.1.12), (C.1.11) and (C.1.10), together with Slutsky's lemma. □

C.2 Proof of Theorem 3.3.1

In the rest of the article, we recall the definitions from Section 3.3: $z_i = a^\top x_i/(a^\top a)$, $w_i = (I_p - aa^\top/(a^\top a))x_i$, $\pi_* = U_a^\top \beta_*$ and $\tilde{w}_i = U_a^\top w_i$.

We need to derive some auxiliary results before we can prove Theorem 3.3.1. The proof of the following lemma is similar to that of Theorem 7.1 of Bickel, Ritov, and Tsybakov (2009) and thus is omitted.

Lemma C.2.1. *Let $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$. Let $\hat{\xi}$ be any vector satisfying $\|n^{-1}X^\top(Y - X\hat{\xi})\|_\infty \leq \eta$. Suppose that there exists ξ_* such that $\|n^{-1}X^\top(Y - X\xi_*)\|_\infty \leq \eta$ and $\|\hat{\xi}\|_1 \leq \|\xi_*\|_1$. If $s_* = \|\xi_*\|_0$ and*

$$\min_{J_0\subseteq\{1,\cdots,p\},|J_0|\leq s_*}\ \min_{\delta\neq 0,\|\delta_{J_0^c}\|_1\leq\|\delta_{J_0}\|_1} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_0}\|_2} \geq \kappa, \tag{C.2.1}$$

then $\|\delta\|_1 \leq 8\eta s_*\kappa^{-2}$ and $\delta^\top X^\top X\delta/n \leq 16\eta^2 s_*\kappa^{-2}$, where $\delta = \hat{\xi} - \xi_*$.

Lemma C.2.2. *Suppose that Assumption 5 and H_0 in (3.1.2) hold. Consider the optimization problem* (3.3.6). *Let $v_i = y_i - z_i g_0$, $\sigma_v^2 = Ev_i^2$ and $\rho_* = \sigma_\varepsilon/\sigma_v$. There exists a constant $C > 0$, such that for any $\eta, \lambda > C\sqrt{n^{-1}\log p}$, $\rho_0 \leq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{-1/2}$, we have*

$$P\left((\pi_*, \rho_*) \text{ and } \gamma_* \text{ are in the feasible region in } (3.3.6)\right) \to 1.$$

Proof. Let $V = Y - Zg_0$ and notice that under Assumption 5, $Z - \tilde{W}\gamma_* = u$ and $E\tilde{w}_i u_i = 0$. Since $u_i\sigma_u^{-1} \sim \mathcal{N}(0,1)$ is independent of $\tilde{w}_i \sim \mathcal{N}(0, \Sigma_{\tilde{W}})$ with the eigenvalues of $\Sigma_{\tilde{W}} = E\tilde{W}^\top\tilde{W}/n$ bounded away from zero and infinity, it follows that there exists a constant that upper bounds the sub-exponential norm of each entry of $\tilde{w}_i u_i \sigma_u^{-1}$. To see this, note that, by the moment generating function of $\mathcal{N}(0,1)$, for $t > 0$,

$$E\exp(t\tilde{w}_{i,j}u_i\sigma_u^{-1}) = E[E(\exp(t\tilde{w}_{i,j}u_i\sigma_u^{-1}) \mid \tilde{w}_{i,j})] = E\exp(\tilde{w}_{i,j}^2 t^2/2).$$

Since $\tilde{w}_{i,j}^2$ has bounded sub-exponential norm (by Lemma 5.14 of Vershynin (2010)), Lemma 5.15 of Vershynin (2010) implies that for small enough t, $E\exp(t\tilde{w}_{i,j}u_i\sigma_u^{-1}) = E\exp(\tilde{w}_{i,j}^2 t^2/2)$ is bounded by some constant. Hence, Equation (5.16) in Vershynin (2010) implies that $\tilde{w}_{i,j}u_i\sigma_u^{-1}$ has bounded the sub-exponential norm.

By Proposition 5.16 of Vershynin (2010) and the union bound, we have that $\forall t_0 > 0$,

$$P\left(\|n^{-1}\tilde{W}^\top u\sigma_u^{-1}\|_\infty > t_0\sqrt{n^{-1}\log p}\right) \leq 2p\exp\left[-\min\left(\frac{t_0^2\log p}{K^2}, \frac{t_0\sqrt{n\log p}}{K}\right)\right],$$

where $K > 0$ is a constant depending only on the constants in Assumption 5. Hence, there exists a constant $M_1 > 0$, such that $P\left(\|n^{-1}\tilde{W}^\top u\|_\infty > M_1\sigma_u\sqrt{n^{-1}\log p}\right) \to 0$. It follows that

$$P\left(\|n^{-1}\tilde{W}^\top(Z - \tilde{W}\gamma_*)\|_\infty > 2c_3^{-1/2}M_1\sqrt{n^{-1}\log p}n^{-1/2}\|Z\|_2\right)$$

$$\leq P\left(\|n^{-1}\tilde{W}^\top u\|_\infty > M_1\sigma_u\sqrt{n^{-1}\log p}\right) + P\left(2\frac{\sqrt{c_3}\sigma_u}{\sigma_z} \geq \frac{n^{-1/2}\|Z\|_2}{\sigma_z}\right) \overset{(i)}{=} o(1), \tag{C.2.2}$$

where (i) follows by $2\sqrt{c_3}\sigma_u/\sigma_z \geq 2$ (Assumption 5) and $n^{-1/2}\|Z\sigma_z^{-1}\|_2 = 1 + o_P(1)$. By the Law of Large Numbers : $n^{-1}\|Z\sigma_z^{-1}\|_2^2$ is the average of n independent $\chi^2(1)$ random variables.

Notice that under H_0 in (3.1.2), $V - \tilde{W}\pi_* = \varepsilon$. By an analogous argument, there exists a constant $M_3 > 0$ such that

$$P\left(\|n^{-1}\tilde{W}^\top(V - \tilde{W}\pi_*)\|_\infty > M_3\rho_*\sqrt{n^{-1}\log p}n^{-1/2}\|V\|_2\right) \to 0. \tag{C.2.3}$$

Since $V = \tilde{W}\pi_* + \varepsilon = W\beta_* + \varepsilon$, we have that $\sigma_v^2 = \beta_*^\top\Sigma_W\beta_* + \sigma_\varepsilon^2$. Assumption 5 implies that

$$\beta_*^\top\Sigma_X\beta_* + \sigma_\varepsilon^2 = \sigma_y^2 \leq \sigma_\varepsilon^2/c_3$$

and thus $\beta_*^\top\Sigma_X\beta_* \leq (c_3^{-1} - 1)\sigma_\varepsilon^2$. Therefore,

$$\|\beta_*\|_2^2 \leq (\beta_*^\top\Sigma_X\beta_*)/\lambda_{\min}(\Sigma_X) \leq c_1^{-1}\beta_*^\top\Sigma_X\beta_* \leq c_1^{-1}(1 - c_3^{-1})\sigma_\varepsilon^2.$$

Observe that $\Sigma_W = M_a\Sigma_X M_a$, where $M_a = I_p - aa^\top/(a^\top a)$ is a projection matrix. Hence, $\lambda_{\max}(\Sigma_W) \leq \lambda_{\max}(\Sigma_X)$ and thus

$$\sigma_v^2 = \beta_*^\top\Sigma_W\beta_* + \sigma_\varepsilon^2 \leq \lambda_{\max}(\Sigma_X)\|\beta_*\|_2^2 + \sigma_\varepsilon^2 \leq [1 + c_2c_1^{-1}(c_3^{-1} - 1)]\sigma_\varepsilon^2.$$

Since $\rho_0 \leq [1 + c_2c_1^{-1}(c_3^{-1} - 1)]^{-1/2} \leq \sigma_\varepsilon/\sigma_v$, we have that $\sigma_\varepsilon^2 = \rho_*\sigma_v\sigma_\varepsilon \geq \sigma_v^2\rho_*\rho_0$. By the Law of Large Numbers ,

$$n^{-1}V^\top(V - \tilde{W}\pi_*) = n^{-1}V^\top\varepsilon = (1 + o_P(1))\sigma_\varepsilon^2,$$

$n^{-1}\|V\|_2^2 = (1 + o_P(1))\sigma_v^2$ and thus

$$P\left(\frac{n^{-1}V^\top(V - \tilde{W}\pi_*)}{\rho_0\rho_* n^{-1}\|V\|_2^2} \geq \frac{1}{2}\right) = P\left(\frac{\sigma_\varepsilon^2(1 + o_P(1))}{\sigma_v^2(1 + o_P(1))\rho_0\rho_*} \geq \frac{1}{2}\right) \to 1. \tag{C.2.4}$$

The desired result follows by (C.2.2), (C.2.3), (C.2.4) and the fact that

$$\rho_* = \sigma_\varepsilon \sigma_v^{-1} \geq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{-1/2} \geq \rho_0.$$

□

Lemma C.2.3. *If* $n^{-1}V^\top(V - \tilde{W}\hat{\pi}) \geq \bar{\eta}$, *then* $n^{-1}(V - \tilde{W}\hat{\pi})^\top(V - \tilde{W}\hat{\pi}) \geq \bar{\eta}^2/(n^{-1}V^\top V)$.

Proof. Since $n^{-1}V^\top(V - \tilde{W}\hat{\pi}) \geq \bar{\eta}$, we have that, for any $t \geq 0$,

$$\begin{aligned}
&n^{-1}(V - \tilde{W}\hat{\pi})^\top(V - \tilde{W}\hat{\pi}) \\
&\geq n^{-1}(V - \tilde{W}\hat{\pi})^\top(V - \tilde{W}\hat{\pi}) + t\left(\bar{\eta} - n^{-1}V^\top(V - \tilde{W}\hat{\pi})\right) \\
&\overset{(i)}{\geq} \min_{\gamma}\left\{n^{-1}(V - \tilde{W}\gamma)^\top(V - \tilde{W}\gamma) + t\left(\bar{\eta} - n^{-1}V^\top(V - \tilde{W}\gamma)\right)\right\} \\
&= t\bar{\eta} - \frac{1}{4}t^2 n^{-1}V^\top V,
\end{aligned}$$

where (i) follows by the first-order condition of quadratic optimizations. The desired result follows by maximizing the last line with respect to t with $t = 2\bar{\eta}/\left(n^{-1}V^\top V\right)$. □

We now proceed to prove Theorem 3.3.1.

Proof of Theorem 3.3.1. Let $V = Y - Zg_0$, $s_* = \|\gamma_*\|_0$, $\eta_\pi = \eta n^{-1/2}\|V\|_2$ and $\lambda_\gamma = \lambda n^{-1/2}\|Z\|_2$. Notice that

$$\begin{aligned}
&n^{-1/2}(V - \tilde{W}\hat{\pi})^\top(Z - \tilde{W}\hat{\gamma}) \\
&\qquad = \underbrace{n^{-1/2}(V - \tilde{W}\hat{\pi})^\top u}_{I_1} + \underbrace{n^{-1/2}(V - \tilde{W}\hat{\pi})^\top\tilde{W}(\gamma_* - \hat{\gamma})}_{I_2}. \quad \text{(C.2.5)}
\end{aligned}$$

Since the eigenvalues of $E\tilde{W}^\top\tilde{W}/n$ is bounded away from zero and infinity, it follows, as a simple consequence of Theorem 6 in Rudelson and Zhou (2013), that there exists a constant $\kappa > 0$, such that $P(\mathcal{D}_n(s_*, \kappa)) \to 1$, where

$$\mathcal{D}_n(s_*, \kappa) = \left\{\min_{J_0 \subseteq \{1,\cdots,p\}, |J_0| \leq s_*} \min_{\delta \neq 0, \|\delta_{J_0^c}\|_1 \leq \|\delta_{J_0}\|_1} \frac{\|\tilde{W}\delta\|_2}{\sqrt{n}\|\delta_{J_0}\|_2} > \kappa\right\}. \quad \text{(C.2.6)}$$

Define the event $\mathcal{M} = \Big\{(\pi_*, \rho_*)$ and γ_* are in the feasible region in (3.3.6)$\Big\}$. By Lemma C.2.2, with appropriate choice of tuning parameters as specified in the theorem, we have $P(\mathcal{M}) \to 1$ and thus

$$P\left(\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)\right) \to 1. \tag{C.2.7}$$

We apply Lemma C.2.1 with (Y, X, ξ_*) replaced by $(Z, \tilde{W}, \gamma_*)$ and obtain that, on the event $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$,

$$\|\hat{\gamma} - \gamma_*\|_1 \leq 8\lambda_\gamma s_* \kappa^{-2} \text{ and } n^{-1/2}\|\tilde{W}(\hat{\gamma} - \gamma_*)\|_2 \leq 4\lambda_\gamma \sqrt{s_*}\kappa^{-1}. \tag{C.2.8}$$

Thus, on $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$, we have the bound

$$|I_2| \leq n^{1/2}\|n^{-1}\tilde{W}^\top(V - \tilde{W}\hat{\pi})\|_\infty \|\hat{\gamma} - \gamma_*\|_1 \leq 8n^{1/2}\lambda_\gamma \eta_\pi s_* \kappa^{-2},$$

where in the last step we utilized

$$\|n^{-1}\tilde{W}^\top(V - \tilde{W}\hat{\pi})\|_\infty \leq \eta\hat{\rho}n^{-1/2}\|V\|_2 \leq \eta_\pi$$

with $\hat{\rho} \leq 1$, from the constraints in optimization problem (3.3.6). Moreover, by constraints in (3.3.6) and Lemma C.2.3, on $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$, we have that

$$\hat{\sigma}_\varepsilon \geq \rho_0 \hat{\rho} n^{-1/2}\|V\|_2/2 \geq \rho_0^2 n^{-1/2}\|V\|_2/2$$

and thus, by $\sigma_u \geq c_3\sigma_z$,

$$\left|\frac{I_2}{\hat{\sigma}_\varepsilon \sigma_u}\right| \leq \frac{16n\lambda_\gamma \eta_\pi s_* \kappa^{-2}}{c_3 \sigma_z \rho_0^2 \|V\|_2} = \frac{16\sqrt{n}\lambda\eta s_* \kappa^{-2}}{c_3 \rho_0^2} \times \frac{n^{-1/2}\|Z\|_2}{\sigma_z} \overset{(i)}{=} o_P(1), \tag{C.2.9}$$

where (i) follows by $\rho_0^{-1} = O(1)$ and $\lambda, \eta \asymp \sqrt{n^{-1}\log p}$ with $s_* = o(\sqrt{n}/\log p)$ and $n^{-1/2}\|Z\|_2/\sigma_z = 1 + o_P(1)$. Observe that by the Law of Large Numbers : $n^{-1}\|Z\sigma_z^{-1}\|_2^2$ is the average of n independent $\chi^2(1)$ random variables.

For I_1, notice that under H_0 in (3.1.2), u is independent of $\{V, \tilde{W}\}$. Since $\hat{\pi}$ and $\hat{\sigma}_\varepsilon$ are computed using $\{V, \tilde{W}\}$, it follows that u is independent of $V - \tilde{W}\hat{\pi}$

and $\hat{\sigma}_\varepsilon$. Thus, under H_0,

$$I_1 \hat{\sigma}_\varepsilon^{-1} \sigma_u^{-1} \mid (V, \tilde{W}) \sim \mathcal{N}(0,1)$$

and thus $I_1 \hat{\sigma}_\varepsilon^{-1} \sigma_u^{-1} \sim \mathcal{N}(0,1)$. This, together with (C.2.9), implies that, under H_0,

$$S_n \frac{\hat{\sigma}_u}{\sigma_u} = \frac{n^{-1/2}(V - \tilde{W}\hat{\pi})^\top (Z - \tilde{W}\hat{\gamma})}{\hat{\sigma}_\varepsilon \sigma_u} \to^d \mathcal{N}(0,1). \tag{C.2.10}$$

By (C.2.8) and $s_* = o(\sqrt{n}/\log p)$,

$$\begin{aligned} \left|\hat{\sigma}_u - n^{-1/2}\|u\|_2\right| &= \left|n^{-1/2}\|Z - \tilde{W}\hat{\gamma}\|_2 - n^{-1/2}\|u\|_2\right| \\ &\leq n^{-1/2}\|Z - \tilde{W}\hat{\gamma} - u\|_2 \\ &= n^{-1/2}\|\tilde{W}(\hat{\gamma} - \gamma_*)\|_2 \\ &= O_P(\lambda n^{-1/2}\|Z\|_2 \sqrt{s_*}) = o_P(n^{-3/4}\|Z\|_2). \end{aligned} \tag{C.2.11}$$

Therefore,

$$\begin{aligned} \frac{|\hat{\sigma}_u - \sigma_u|}{\sigma_u} &\leq \frac{\left|\hat{\sigma}_u - n^{-1/2}\|u\|_2\right|}{\sigma_u} + \left|n^{-1/2}\|u\sigma_u^{-1}\|_2 - 1\right| \\ &\overset{(i)}{=} \frac{o_P(n^{-3/4}\|Z\|_2)}{\sigma_u} + o_P(1) \\ &\overset{(ii)}{=} o_P(1), \end{aligned} \tag{C.2.12}$$

where (i) follows by the Law of Large Numbers ($n^{-1}\|u\sigma_u^{-1}\|_2^2$ is the average of n independent $\chi^2(1)$ random variables) and (ii) follows by $\sigma_z/\sigma_u \leq c_3^{-1}$ (Assumption 5) and $n^{-1/2}\|Z\|_2/\sigma_z = 1 + o_P(1)$ (as argued in (C.2.9)).

By (C.2.12), $\hat{\sigma}_u/\sigma_u = 1 + o_P(1)$ and the desired result follows by (C.2.10) and Slutsky's lemma. □

C.3 Proof of Theorem 3.3.2

We need the some auxiliary results before we prove Theorem 3.3.2.

Lemma C.3.1. *Let Assumption 5 hold. In (3.3.3),* $\sigma_u = (a^\top \Omega_X a)^{-1/2}$.

Proof. We define $\tilde{\gamma}_* = U_a \gamma_* \in \mathbb{R}^p$ and observe that $z_i = \tilde{w}_i^\top \gamma_* + u_i = w_i^\top \tilde{\gamma}_* + u_i$ and $Ew_i u_i = U_a^\top E\tilde{w}_i u_i = 0$. Thus, $\sigma_u^2 = Ez_i^2 - \tilde{\gamma}_*^\top Ew_i w_i^\top \tilde{\gamma}_*$.

Let $M_a = I_p - aa^\top/(a^\top a)$. Recall that $z_i = a^\top x_i/(a^\top a)$ and $w_i = M_a x_i$. Since $Ew_i z_i = Ew_i w_i^\top \tilde{\gamma}_*$,

$$Ew_i w_i^\top = M_a \Sigma_X M_a$$

and $Ew_i z_i = M_a \Sigma_X a \|a\|_2^{-2}$, we have

$$M_a \Sigma_X M_a \tilde{\gamma}_* = M_a \Sigma_X a \|a\|_2^{-2}$$

and thus, $M_a \Sigma_X (M_a \tilde{\gamma}_* - a\|a\|_2^{-2}) = 0$. Since M_a is the projection matrix onto the $(p-1)$-dimensional linear space orthogonal to a, there exists $k_1 \in \mathbb{R}$ with

$$\Sigma_X (M_a \tilde{\gamma}_* - a\|a\|_2^{-2}) = k_1 a,$$

implying that $M_a \tilde{\gamma}_* = k_1 \Omega_X a + \|a\|_2^{-2} a$. Next we aim to identify k_1. Observe that

$$\begin{aligned}
\tilde{\gamma}_*^\top M_a \Sigma_X M_a \tilde{\gamma}_* &\overset{(i)}{=} (M_a \tilde{\gamma}_*)^\top (M_a \Sigma_X M_a \tilde{\gamma}_*) \\
&\overset{(ii)}{=} (k_1 \Omega_X a + \|a\|_2^{-2} a)^\top \Sigma_X a \|a\|_2^{-2} \\
&= k_1 + \|a\|_2^{-4} a^\top \Sigma_X a.
\end{aligned}$$

where (i) and (ii) follow by $M_a^2 = M_a$ and $M_a \Sigma_X M_a \tilde{\gamma}_* = M_a \Sigma_X a \|a\|_2^{-2}$, respectively. Together with

$$\begin{aligned}
\tilde{\gamma}_*^\top M_a \Sigma_X M_a \tilde{\gamma}_* &= (M_a \tilde{\gamma}_*)^\top \Sigma_X (M_a \tilde{\gamma}_*) \\
&= (k_1 \Omega_X a + \|a\|_2^{-2} a)^\top \Sigma_X (k_1 \Omega_X a + \|a\|_2^{-2} a),
\end{aligned}$$

we can solve for the unknown k_1. The above display allows us to obtain $k_1 = -(a^\top \Omega_X a)^{-1}$ and thus

$$\tilde{\gamma}_*^\top M_a \Sigma_X M_a \tilde{\gamma}_* = \|a\|_2^{-4} a^\top \Sigma_X a - (a^\top \Omega_X a)^{-1}.$$

Since

$$\sigma_u^2 = Ez_i^2 - \tilde{\gamma}_*^\top E w_i w_i^\top \tilde{\gamma}_* = Ez_i^2 - \tilde{\gamma}_*^\top M_a \Sigma_X M_a \tilde{\gamma}_*$$

and $Ez_i^2 = \|a\|_2^{-4} a^\top \Sigma_X a$, we have $\sigma_u^2 = (a^\top \Sigma_X a)^{-1}$. The proof is complete. □

Lemma C.3.2. *Let Assumption 5 hold. Suppose that at least one of the following conditions holds:*
(1) $\|a\|_0 \vee \|\beta_*\|_0 = o(\sqrt{n}/\log p)$ *or*
(2) $\mathcal{M}(a) \bigcap \mathcal{M}(\beta_*) = \emptyset$ *and* $\|\beta_*\|_0 = o(\sqrt{n}/\log p)$.
Then, $\|\pi_*\|_0 = o(\sqrt{n}/\log p)$.

Proof. We denote $s_a = \|a\|_0$ and $s_\beta = \|\beta_*\|_0$. Without loss of generality, we assume that $a = (a_0^\top, 0)^\top \in \mathbb{R}^p$ with $\|a_0\|_0 = s_a$. Let $U_{a_0} \in \mathbb{R}^{s_a \times (s_a - 1)}$ satisfy $U_{a_0}^\top U_{a_0} = I_{s_a - 1}$ and $U_{a_0} U_{a_0}^\top = I_{s_a} - a_0 a_0^\top / (a_0^\top a_0)$. It is easy to verify that

$$I_p - aa^\top/(a^\top a) = \begin{pmatrix} I_{s_a} - a_0 a_0^\top/(a_0^\top a_0) & 0 \\ 0 & I_{p-s_a} \end{pmatrix}$$

and

$$U_a = \begin{pmatrix} U_{a_0} & 0 \\ 0 & I_{p-s_a} \end{pmatrix} \in \mathbb{R}^{p \times (p-1)}.$$

It then follows that

$$\pi_* = U_a^\top \beta_* = \begin{pmatrix} U_{a_0}^\top \beta_{*,\mathcal{M}(a)} \\ \beta_{*,[\mathcal{M}(a)]^c} \end{pmatrix} \tag{C.3.1}$$

Take note that

$$\|\pi_*\|_0 = \|U_{a_0}^\top \beta_{*,\mathcal{M}(a)}\|_0 + \|\beta_{*,[\mathcal{M}(a)]^c}\|_0 \le (s_a - 1) + \|\beta_*\|_0.$$

This proves the result under condition (1). Under condition (2), $\beta_{*,\mathcal{M}(a)} = 0$ and thus $\|\pi_*\|_0 = \|\beta_{*,[\mathcal{M}(a)]^c}\|_0 = \|\beta_*\|_0$. This proves the result under condition (2). □

Lemma C.3.3. *Let Assumption 5 and* $H_{1,n}$ *in (3.3.10) hold. Let* $v_i = y_i - z_i g_0$,

$\sigma_v^2 = Ev_i^2$, $\bar{\rho}_* = [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{-1/2}$ *and* $h_n = n^{-1/2}(a^\top \Omega_X a)^{1/2}\sigma_\varepsilon d$. *Then,*

$$\sigma_v = O(1).$$

Moreover, there exists a constant $C > 0$, *such that* $\forall \eta, \lambda > C\sqrt{n^{-1}\log p}$ *and* $\forall \rho_0 \leq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{-1/2}$, *we have*

$$P\left((\pi_* + \gamma_* h_n, \bar{\rho}_*, \gamma_*) \text{ is feasible in } (3.3.6)\right) \to 1.$$

Proof. Under $H_{1,n}$, $v_i = \tilde{w}_i^\top \pi_* + \varepsilon_i + z_i h_n$. Consequently,

$$\|v_i - \tilde{w}_i^\top \pi_* - \varepsilon_i\|_{L^2(P)} = \|z_i h_n\|_{L^2(P)} = \sqrt{E z_i^2 h_n^2}.$$

Observe that

$$\begin{aligned} E z_i^2 h_n^2 &= \|a\|_2^{-4} a^\top \Sigma_X a h_n^2 \\ &= \|a\|_2^{-4} (a^\top \Sigma_X a)(a^\top \Omega_X a)(a^\top \Omega_X a)^{-1} h_n^2 \\ &\overset{(i)}{\leq} (c_2 c_1^{-1})(a^\top \Omega_X a)^{-1} h_n^2 = (c_2 c_1^{-1}) n^{-1} \sigma_\varepsilon^2 d^2 = o(1), \end{aligned}$$

where (i) holds by Assumption 5. Hence, by the triangular inequality applied to $L^2(P)$-norm, we have $\sigma_v = \|v_i\|_{L^2(P)} = \|\tilde{w}_i^\top \pi_* + \varepsilon_i\|_{L^2(P)} + o(1)$. By the same argument as in the proof of Lemma C.2.2,

$$\|\tilde{w}_i^\top \pi_* + \varepsilon_i\|_{L^2(P)}^2 = \pi_*^\top E\tilde{w}_i \tilde{w}_i^\top \pi_* + \sigma_\varepsilon = \beta_*^\top \Sigma_W \beta_* + \sigma_\varepsilon^2 \leq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]\sigma_\varepsilon^2.$$

The first claim follows by $\sigma_v \leq [1 + c_2 c_1^{-1}(c_3^{-1} - 1)]^{1/2}\sigma_\varepsilon + o(1) = O(1)$.

Notice that under $H_{1,n}$, the analysis for the feasibility of γ_* is the same as under H_0. Thus, by the argument in the proof of Lemma C.2.2, for some constant $M_1 > 0$, we have

$$P\left(\|n^{-1}\tilde{W}^\top (Z - \tilde{W}\gamma_*)\|_\infty > M_1\sqrt{n^{-1}\log p} n^{-1/2}\|Z\|_2\right) \to 0. \tag{C.3.2}$$

(3.3.3) implies that, under $H_{1,n}$, $v_i = \tilde{w}_i^\top \pi_* + \varepsilon_i + z_i h_n = \tilde{w}_i^\top (\pi_* + \gamma_* h_n) +$

$\varepsilon_i + u_i h_n$. Thus,

$$n^{-1}\tilde{W}^\top(V - \tilde{W}(\pi_* + \gamma_* h_n)) = n^{-1}\sum_{i=1}^n \tilde{w}_i\varepsilon_i + n^{-1}\sum_{i=1}^n \tilde{w}_i u_i h_n.$$

By a similar argument as in the proof of Lemma C.2.2, entries of $\tilde{w}_i\varepsilon_i$ and $\tilde{w}_i u_i \sigma_u^{-1}$ have bounded sub-exponential norms. As in the proof of Lemma C.2.2, we can use Proposition 5.16 of Vershynin (2010) and the union bound to conclude that for some constant $M_2 > 0$ we have $P(\|n^{-1}\tilde{W}^\top u\sigma_u^{-1}\|_\infty > M_2\sqrt{n^{-1}\log p}) \to 0$ and $P(\|n^{-1}\tilde{W}^\top\varepsilon\|_\infty > M_2\sqrt{n^{-1}\log p}) \to 0$. It follows that

$$\begin{aligned} & P\left(\|n^{-1}\tilde{W}^\top(V - \tilde{W}(\pi_* + \gamma_* h_n))\|_\infty > 2M_2\sqrt{n^{-1}\log p}\right) \qquad \text{(C.3.3)} \\ = \; & P\left(\|n^{-1}\tilde{W}^\top(\varepsilon + uh_n)\|_\infty > 2M_2\sqrt{n^{-1}\log p}\right) \\ \leq \; & P\left(\|n^{-1}\tilde{W}^\top\varepsilon\|_\infty > M_2\sqrt{n^{-1}\log p}\right) \\ + \; & P\left(\|n^{-1}\tilde{W}^\top u\sigma_u^{-1}\|_\infty|\sigma_u h_n| > M_2\sqrt{n^{-1}\log p}\right) \overset{(i)}{=} o(1), \end{aligned}$$

where (i) holds by $|\sigma_u h_n| = n^{-1/2}\sigma_\varepsilon|d| = o(1)$ (by Lemma C.3.1 and the definition of h_n).

Notice that

$$Ev_i(u_i h_n + \varepsilon_i)\sigma_\varepsilon^{-2} = E(u_i h_n + \varepsilon_i)^2\sigma_\varepsilon^{-2} = 1 + \sigma_\varepsilon^{-2}\sigma_u^2 h_n^2 = 1 + n^{-1}d^2 = 1 + o(1).$$

By the Law of Large Numbers ,

$$n^{-1}V^\top(V - \tilde{W}(\pi_* + \gamma_* h_n))\sigma_\varepsilon^{-2} = E(u_i h_n + \varepsilon_i)^2\sigma_\varepsilon^{-2} + o_P(1) = 1 + o_P(1).$$

In the display above, the first $o_P(1)$ term is equal to $n^{-1}\sigma_\varepsilon^{-2}(\pi_* + \gamma_* h_n)^\top\tilde{W}^\top(\varepsilon + h_n u)$. Since $\tilde{W}$ is uncorrelated with (ε, u), this term is the partial sum of zero-mean independent random variables. Since $\pi_* + h_n\gamma_*$ has bounded L^2-norm by Bernstein's inequality, we have that this term is $o_P(1)$. The Law of Large Numbers also implies

that $n^{-1}\|V\|_2^2\sigma_v^{-2} = 1 + o_P(1)$. Hence,

$$
\begin{aligned}
&P\left(\frac{n^{-1}V^\top(V - \tilde{W}(\pi_* + \gamma_* h_n))}{n^{-1}\|V\|_2^2} > \frac{1}{2}\rho_0\bar{\rho}_*\right) \\
&= P\left(\frac{\sigma_\varepsilon^2(1+o_P(1))}{\sigma_v^2(1+o_P(1))} > \frac{1}{2}\rho_0\bar{\rho}_*\right) \\
&\overset{(i)}{\geq} P\left(\frac{(1+o_P(1))}{[1 + c_2c_1^{-1}(c_3^{-1}-1)](1+o_P(1))} > \frac{1}{2}\rho_0\bar{\rho}_*\right) \overset{(ii)}{\geq} 1 + o(1), \qquad \text{(C.3.4)}
\end{aligned}
$$

where (i) follows by $\sigma_v \leq [1 + c_2c_1^{-1}(c_3^{-1}-1)]^{1/2}\sigma_\varepsilon + o(1)$ (shown at the beginning of the proof) and (ii) follows by $\rho_0 \leq \bar{\rho}_* = [1 + c_2c_1^{-1}(c_3^{-1}-1)]^{-1/2}$. According to (C.3.2), (C.3.3) and (C.3.4), $P\left((\pi_* + \gamma_* h_n, \bar{\rho}_*, \gamma_*) \text{ is feasible in } (3.3.6)\right) \to 1$. The proof is complete. □

Now we are ready to prove Theorem 3.3.2.

Proof of Theorem 3.3.2. Let $V = Y - Zg_0$, $s_* = \|\gamma_*\|_0 + \|\pi_*\|_0$, $h_n = n^{-1/2}(a^\top\Omega_X a)^{1/2}\sigma_\varepsilon d$, $\lambda_\gamma = \lambda n^{-1/2}\|Z\|_2$, $\eta_\pi = \eta n^{-1/2}\|V\|_2$ and $\sigma_v^2 = EV^\top V/n$. Notice that $\|\gamma_*\|_0 \leq s_*$ and $\|\pi_* + \gamma_* h_n\|_0 \leq s_*$. By Lemmas C.3.1 and C.3.2,

$$
s_* = o(\sqrt{n}/\log p) \quad \text{and} \quad h_n = n^{-1/2}\sigma_u^{-1}\sigma_\varepsilon d. \qquad \text{(C.3.5)}
$$

Under $H_{1,n}$, $V = Zh_n + \tilde{W}\pi_* + \varepsilon = uh_n + \tilde{W}(\gamma_* h_n + \pi_*) + \varepsilon$ and thus

$$
\begin{aligned}
&n^{-1/2}(V - \tilde{W}\hat{\pi})^\top(Z - \tilde{W}\hat{\gamma}) \qquad \text{(C.3.6)} \\
&\quad = n^{-1/2}(V - \tilde{W}\hat{\pi})^\top\tilde{W}(\gamma_* - \hat{\gamma}) + n^{-1/2}(V - \tilde{W}\hat{\pi})^\top u \\
&\quad = \underbrace{n^{-1/2}(V - \tilde{W}\hat{\pi})^\top\tilde{W}(\gamma_* - \hat{\gamma})}_{I_1} + \underbrace{n^{-1/2}\varepsilon^\top u}_{I_2} \\
&\qquad\qquad + \underbrace{n^{-1/2}h_n u^\top u}_{I_3} + \underbrace{n^{-1/2}(\pi_* - \hat{\pi} + \gamma_* h_n)^\top\tilde{W}^\top u}_{I_4}.
\end{aligned}
$$

We next treat each of the four terms in the decomposition above separately.

As argued in the proof of Theorem 3.3.1, there exists a constant $\kappa > 0$ such that $P(\mathcal{D}_n(s_*, \kappa)) \to 1$. Define the event $\mathcal{M} = \{(\pi_* +$

$\gamma_* h_n, \bar{\rho}_*, \gamma_*)$ is feasible in (3.3.6)$\}$. By Lemma C.3.3, $P(\mathcal{M}) \to 1$ and thus

$$P\left(\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)\right) \to 1. \tag{C.3.7}$$

Since $\hat{\gamma}$ does not depend on whether $h_n = 0$, we conclude, as argued in the proof of Theorem 3.3.1, that $\hat{\sigma}_u/\sigma_u = 1 + o_P(1)$ and that on the event $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$,

$$\|\hat{\gamma} - \gamma_*\|_1 \leq 8\lambda_\gamma s_* \kappa^{-2} \text{ and } n^{-1/2}\|\tilde{W}(\hat{\gamma} - \gamma_*)\|_2 \leq 4\lambda_\gamma \sqrt{s_*}\kappa^{-1}. \tag{C.3.8}$$

By the definition of $\hat{\pi}$,

$$\|n^{-1}\tilde{W}^\top (V - \tilde{W}\hat{\pi})\|_\infty \leq \eta\hat{\rho} n^{-1/2}\|V\|_2 \leq \eta_\pi;$$

thus, by (C.3.8),

$$\begin{aligned}
\frac{|I_1|}{\hat{\sigma}_u \sigma_\varepsilon} &\leq \frac{\sqrt{n}\|n^{-1}\tilde{W}^\top (V - \tilde{W}\hat{\pi})\|_\infty \|\hat{\gamma} - \gamma_*\|_1}{\hat{\sigma}_u \sigma_\varepsilon} \\
&\leq \frac{8\sqrt{n}\eta_\pi \lambda_\gamma s_* \kappa^{-2}}{\sigma_u(1 + o_P(1))\sigma_\varepsilon} \\
&\overset{(i)}{=} \frac{(s_* n^{-1/2} \log p) O_P(\sigma_v \sigma_z)}{\sigma_u \sigma_\varepsilon} \overset{(ii)}{=} o_P(1),
\end{aligned} \tag{C.3.9}$$

where (i) follows by $n^{-1}\|Z\|_2^2 \sigma_z^{-2} = 1 + o_P(1)$ and $n^{-1}\|V\|_2^2 \sigma_v^{-2} = 1 + o_P(1)$ (by the Law of Large Numbers since both $n^{-1}\|Z\|_2^2 \sigma_z^{-2}$ and $n^{-1}\|V\|_2^2 \sigma_v^{-2}$ are averages of n independent $\chi^2(1)$ random variables) and (ii) holds by (C.3.5),$\sigma_u/\sigma_z \geq c_3$, $\sigma_\varepsilon \geq c_1$ and $\sigma_v = O(1)$ (Lemma C.3.3).

By CLT, $I_2/\sigma_u \to^d \mathcal{N}(0, \sigma_\varepsilon^2)$. Since $\hat{\sigma}_u/\sigma_u = 1 + o_P(1)$, the Slutsky's lemma implies that

$$\frac{I_2}{\hat{\sigma}_u \sigma_\varepsilon} \to^d \mathcal{N}(0, 1). \tag{C.3.10}$$

By (C.3.5),

$$n^{-1/2} h_n u^\top u/(\hat{\sigma}_u \sigma_\varepsilon) = d(\sigma_u/\hat{\sigma}_u)(n^{-1} u^\top u \sigma_u^{-2}).$$

Notice that $n^{-1}u^\top u\sigma_u^{-2}$ is the average of n independent $\chi^2(1)$ random variables. It

follows, by the Law of Large Numbers and $\sigma_u/\hat{\sigma}_u = 1 + o_P(1)$, that

$$\frac{I_3}{\hat{\sigma}_u \sigma_\varepsilon} = d + o_P(1). \tag{C.3.11}$$

On the event $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$, we have that

$$\|\hat{\pi}\|_1 \leq \|\pi_* + \gamma_* h_n\|_1,$$

$$\|n^{-1}\tilde{W}^\top (V - \tilde{W}\hat{\pi})\|_\infty \leq \eta \hat{\rho} n^{-1/2} \|V\|_2 \leq \eta_\pi$$

and

$$\|n^{-1}\tilde{W}^\top (V - \tilde{W}(\pi_* + \gamma_* h_n))\|_\infty \leq \eta \rho_* n^{-1/2} \|V\|_2 \leq \eta_\pi.$$

We apply Lemma C.2.1 with (Y, X, ξ_*), replaced by $(V, \tilde{W}, \pi_* + \gamma_* h_n)$, and obtain

$$\|\hat{\pi} - (\pi_* + \gamma_* h_n)\|_1 \leq 8\eta_\pi s_* \kappa^{-2} \text{ and } n^{-1/2} \left\|\tilde{W}[\hat{\pi} - (\pi_* + \gamma_* h_n)]\right\|_2 \leq 4\eta_\pi \sqrt{s_*}\kappa^{-1}. \tag{C.3.12}$$

Observe that, on the event $\mathcal{M} \bigcap \mathcal{D}_n(s_*, \kappa)$, $\|n^{-1}\tilde{W}^\top u\|_\infty = \|n^{-1}\tilde{W}^\top (Z - \tilde{W}\gamma_*)\|_\infty \leq \lambda_\gamma$ and thus

$$\begin{aligned}
\frac{|I_4|}{\hat{\sigma}_u \sigma_\varepsilon} &\leq \frac{\sqrt{n}\|n^{-1}\tilde{W}^\top u\|_\infty \|\hat{\pi} - (\pi_* + \gamma_* h_n)\|_1}{\sigma_u \sigma_\varepsilon} \cdot \frac{\sigma_u}{\hat{\sigma}_u} \\
&\overset{(i)}{\leq} \frac{8\sqrt{n}\lambda_\gamma \eta_\pi s_* \kappa^{-2}}{\sigma_u \sigma_\varepsilon} \cdot (1 + o_P(1)) \\
&= O_P(s_* n^{-1} \log p) \frac{n^{-1/2}\|V\|_2}{\sigma_\varepsilon} \cdot \frac{n^{-1/2}\|Z\|_2}{\sigma_u} \overset{(ii)}{=} o_P(1),
\end{aligned} \tag{C.3.13}$$

where (i) follows by (C.3.12) and $\hat{\sigma}_u/\sigma_u = 1 + o_P(1)$ and (ii) follows by the same argument as (C.3.9). By Slutszky's lemma, together with (C.3.6), (C.3.9), (C.3.10), (C.3.11), (C.3.13), we have

$$S_n \frac{\hat{\sigma}_\varepsilon}{\sigma_\varepsilon} = \frac{n^{-1/2}(V - \tilde{W}\hat{\pi})^\top (Z - \tilde{W}\hat{\gamma})}{\hat{\sigma}_u \sigma_\varepsilon} \to^d \mathcal{N}(d, 1). \tag{C.3.14}$$

Since σ_ε is bounded away from zero, it remains to show that $\hat{\sigma}_\varepsilon = \sigma_\varepsilon + o_P(1)$.

Note that

$$\begin{aligned}\left|\hat{\sigma}_\varepsilon - n^{-1/2}\|\varepsilon\|_2\right| &= \left|n^{-1/2}\|V - \tilde{W}\hat{\pi}\|_2 - n^{-1/2}\|\varepsilon\|_2\right| \\ &\leq n^{-1/2}\|V - \tilde{W}\hat{\pi} - \varepsilon\|_2 \\ &\overset{(i)}{\leq} n^{-1/2}\|\tilde{W}(\pi_* + \gamma_* h_n - \hat{\pi})\|_2 + n^{-1/2}\|u\sigma_u^{-1}\|_2|\sigma_u h_n| \\ &\overset{(ii)}{=} O_P(\eta_\pi\sqrt{s_*}) + O_P(1)n^{-1/2}\sigma_\varepsilon|d| \overset{(iii)}{=} o_P(1),\end{aligned}$$

where (i) holds by $V = \tilde{W}(\pi_* + \gamma_* h_n) + uh_n + \varepsilon$, (ii) holds by (C.3.12) and $n^{-1/2}\|u\sigma_u^{-1}\|_2 = O_P(1)$ (by the Law of Large Numbers) and (iii) holds by $\eta_\pi\sqrt{s_*} = n^{-1/2}\|V\|_2\sqrt{s_* n^{-1}\log p}$, (C.3.5), $n^{-1/2}\|V\|_2 = O_P(1)$ (argued in (C.3.9)).

Law of Large Numbers also implies that $n^{-1/2}\|\varepsilon\|_2^2 = \sigma_\varepsilon^2 + o_P(1)$. This, together with the above display, implies that $\hat{\sigma}_\varepsilon = \sigma_\varepsilon + o_P(1)$. Hence, by Slutszky's lemma and (C.3.14), $S_n \to^d \mathcal{N}(d, 1)$. It follows that

$$\begin{aligned}P\left(|S_n| > \Phi\left(1 - \frac{\alpha}{2}\right)\right) &= P\left(S_n < -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) + P\left(S_n > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \\ &= P\left(S_n - d < -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - d\right) \\ &\qquad + P\left(S_n - d > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - d\right) \\ &\to \Phi\left(-\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - d\right) \\ &\qquad + 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - d\right).\end{aligned}$$

The desired result follows by noticing that

$$1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - d\right) = \Phi\left(-\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + d\right).$$

□

Bibliography

Ahn, Seung C, and Alex R Horenstein. 2013. "Eigenvalue ratio test for the number of factors". *Econometrica* 81 (3): 1203–1227.

Ahn, Seung C, Young H Lee, and Peter Schmidt. 2013. "Panel data models with multiple time-varying individual effects". *Journal of Econometrics* 174 (1): 1–14.

Alvarez, Javier, and Manuel Arellano. 2003. "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators". *Econometrica* 71 (4): 1121–1159.

Amengual, Dante, and Mark W Watson. 2007. "Consistent estimation of the number of dynamic factors in a large N and T panel". *Journal of Business & Economic Statistics* 25 (1): 91–96.

Andrews, Donald. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point". *Econometrica* 61 (4): 821–56.

Andrews, Donald WK. 2005. "Cross-section Regression with Common Shocks". *Econometrica* 73 (5): 1551–1585.

Ang, Andrew. 2014. *Asset Management: A Systematic Approach to Factor Investing*. Oxford University Press (UK).

Ang, Andrew, and Geert Bekaert. 2007. "Stock return predictability: Is it there?" *Review of Financial studies* 20 (3): 651–707.

Ang, Andrew, Robert J Hodrick, Yuhang Xing, and Xiaoyan Zhang. 2006. "The cross-section of volatility and expected returns". *The Journal of Finance* 61 (1): 259–299.

Arellano, Manuel, and Jinyong Hahn. 2007. "Understanding bias in nonlinear panel models: Some recent developments". *Econometric Society Monographs* 43:381.

Athreya, Krishna B, and Soumendra N Lahiri. 2006. *Measure Theory and Probability Theory*. Springer Science & Business Media.

Bai, Jushan. 2003. "Inferential theory for factor models of large dimensions". *Econometrica*: 135–171.

— . 2009. "Panel data models with interactive fixed effects". *Econometrica* 77 (4): 1229–1279.

Bai, Jushan, and Yuan Liao. 2012. "Efficient estimation of approximate factor models via regularized maximum likelihood". *Available at SSRN 2152416*.

Bai, Jushan, and Serena Ng. 2006a. "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions". *Econometrica* 74 (4): 1133–1150.

— . 2002. "Determining the number of factors in approximate factor models". *Econometrica* 70 (1): 191–221.

— . 2012. "Determining the number of primitive shocks in factor models". *Journal of Business & Economic Statistics*.

— . 2006b. "Evaluating latent and observed factors in macroeconomics and finance". *Journal of Econometrics* 131 (1): 507–537.

— . 2008a. "Extremum estimation when the predictors are estimated from large panels". *Annals of Economics and Finance* 9 (2): 201–222.

— . 2008b. "Large Dimensional Factor Analysis". *Foundations and Trends® in Econometrics* 3 (2): 89–163.

Bai, Jushan, and Pierre Perron. 2003. "Computation and analysis of multiple structural change models". *Journal of applied econometrics* 18 (1): 1–22.

— . 1998. "Estimating and testing linear models with multiple structural changes". *Econometrica*: 47–78.

Bai, Jushan, and Peng Wang. 2016. "Econometric Analysis of Large Factor Models". *Annual Review of Economics* 8 (1).

Bakshi, Gurdip, and Nikunj Kapadia. 2003. "Delta-hedged gains and the negative market volatility risk premium". *Review of Financial Studies* 16 (2): 527–566.

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the long run: A potential resolution of asset pricing puzzles". *The Journal of Finance* 59 (4): 1481–1509.

Belloni, A., V. Chernozhukov, and C. Hansen. 2016. "LASSO Methods for Gaussian Instrumental Variables Models". *to appear in the Journal of Royal Statistical Society: Series B* (). arXiv: 1012.1297 [stat.ME].

Belloni, A., V. Chernozhukov, and K. Kato. 2015. "Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems". *Biometrika* 102 (1): 77–94. doi:10.1093/biomet/asu056. eprint: http://biomet.oxfordjournals.org/content/102/1/77.full.pdf+html. http://biomet.oxfordjournals.org/content/102/1/77.abstract.

Belloni, A., V. Chernozhukov, and L. Wang. 2011. "Square-root lasso: pivotal recovery of sparse signals via conic programming". *Biometrika* 98 (4): 791–806. doi:10.1093/biomet/asr043.

Belloni, Alexandre, and Victor Chernozhukov. 2011. "L1-penalized quantile regression in high-dimensional sparse models". *The Annals of Statistics* 39 (1): 82–130.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on treatment effects after selection among high-dimensional controls". *The Review of Economic Studies* 81 (2): 608–650.

Bernanke, Ben S, and Jean Boivin. 2003. "Monetary policy in a data-rich environment". *Journal of Monetary Economics* 50 (3): 525–546.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach". *The Quarterly Journal of Economics* 120 (1): 387–422.

Bickel, Peter J, Yaácov Ritov, and Alexandre B Tsybakov. 2009. "Simultaneous analysis of Lasso and Dantzig selector". *The Annals of Statistics*: 1705–1732.

Blomström, Magnus, Robert E Lipsey, and Mario Zejan. 1996. "Is Fixed Investment the Key to Economic Growth?" *The Quarterly Journal of Economics* 111 (1): 269–276.

Boivin, Jean, and Serena Ng. 2006. "Are more data always better for factor analysis?" *Journal of Econometrics* 132 (1): 169–194.

Bollerslev, Tim, Michael Gibson, and Hao Zhou. 2011. "Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities". *Journal of econometrics* 160 (1): 235–245.

Bond, Steve, Asli Leblebicioglu, and Fabio Schiantarelli. 2010. "Capital accumulation and growth: a new look at the empirical evidence". *Journal of Applied Econometrics* 25 (7): 1073–1099.

Boneva, Lena, Oliver Linton, and Michael Vogt. 2015. "A semiparametric model for heterogeneous panel data with fixed effects". *Journal of Econometrics*.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data". *Econometrica* 83 (3): 1147–1184. ISSN: 1468-0262. doi:10.3982/ECTA11319.

Bühlmann, Peter, and Sara van de Geer. 2015. "High-dimensional inference in misspecified linear models". *Electron. J. Statist.* 9 (1): 1449–1473. doi:10.1214/15-EJS1041.

Bühlmann, Peter, and Sara Van De Geer. 2011. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bühlmann, Peter, et al. 2013. "Statistical significance in high-dimensional linear models". *Bernoulli* 19 (4): 1212–1242.

Cai, T. T., and Z. Guo. 2016. "Accuracy Assessment for High-dimensional Linear Regression". *arXiv preprint arXiv:1603.03474* ().

Cai, T Tony, and Zijian Guo. 2015. "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity". *arXiv preprint arXiv:1506.05539*.

Cai, T Tony, and Tiefeng Jiang. 2011. "Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices". *The Annals of Statistics* 39 (3): 1496–1525.

Campbell, John Y, and John H Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior". *Journal of Political Economy* 107 (2).

Campbell, John Y, Andrew Wen-Chuan Lo, Archie Craig MacKinlay, et al. 1997. *The econometrics of financial markets*. Vol. 2. princeton University press Princeton, NJ.

Campbell, John Y, and Robert J Shiller. 1988a. "Stock prices, earnings, and expected dividends". *The Journal of Finance* 43 (3): 661–676.

— . 1988b. "The dividend-price ratio and expectations of future dividends and discount factors". *Review of financial studies* 1 (3): 195–228.

Campbell, John Y, and Samuel B Thompson. 2008. "Predicting excess stock returns out of sample: Can anything beat the historical average?" *Review of Financial Studies* 21 (4): 1509–1531.

Candes, Emmanuel, and Terence Tao. 2007. "The Dantzig selector: statistical estimation when p is much larger than n". *The Annals of Statistics*: 2313–2351.

Carrasco, Marine, and Xiaohong Chen. 2002. "Mixing and moment properties of various GARCH and stochastic volatility models". *Econometric Theory* 18 (1): 17–39.

Chen, Jia, Jiti Gao, and Degui Li. 2013. "Estimation in partially linear single-index panel data models with fixed effects". *Journal of Business & Economic Statistics* 31 (3): 315–330.

Chen, Liang. 2012. "Identifying observed factors in high dimensional factor models". *Universidad Carlos III de Madrid, Working paper*.

Chen, Xiaohong, Qi-Man Shao, and Wei Biao Wu. 2016. "Self-normalized Cramér-Type Moderate Deviations under Dependence". *The Annals of Statistics* 44 (4): 1593–1617.

Cheng, Xu, Zhipeng Liao, and Frank Schorfheide. 2016. "Shrinkage estimation of high-dimensional factor models with structural instabilities". *The Review of Economic Studies*: rdw005.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2014. "Central limit theorems and bootstrap in high dimensions". *arXiv preprint arXiv:1412.3661*.

— . 2013. "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors". *The Annals of Statistics* 41 (6): 2786–2819.

Chernozhukov, Victor, Christian Hansen, and Martin Spindler. 2015. "Valid post-selection and post-regularization inference: An elementary, general approach". *Annual Review of Economics*. doi:10.1146/annurev-economics-012315-015826.

Dangl, Thomas, and Michael Halling. 2012. "Predictive regressions with time-varying coefficients". *Journal of Financial Economics* 106 (1): 157–181.

DeLong, Bradford, and Lawrence H Summers. 1991. "Equipment Investment and Economic Growth". *Quarterly Journal of Economics* 106 (2): 445–502.

Donoho, David L, and Iain M Johnstone. 1994. "Minimax risk over lp-balls for lp-error". *Probability Theory and Related Fields* 99 (2): 277–303.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin. 2012. "A quasi–maximum likelihood approach for large, approximate dynamic factor models". *Review of Economics and Statistics* 94 (4): 1014–1024.

Fama, Eugene F, and Kenneth R French. 2015. "A five-factor asset pricing model". *Journal of Financial Economics* 116 (1): 1–22.

— . 1989. "Business conditions and expected returns on stocks and bonds". *Journal of financial economics* 25 (1): 23–49.

— . 2016. "Dissecting anomalies with a five-factor model". *Review of Financial Studies* 29 (1): 69–103.

— . 1988. "Dividend yields and expected stock returns". *Journal of financial economics* 22 (1): 3–25.

— . 1992. "The cross-section of expected stock returns". *the Journal of Finance* 47 (2): 427–465.

Fan, Jianqing, and Runze Li. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties". *Journal of the American statistical Association* 96 (456): 1348–1360.

Fan, Jianqing, Yuan Liao, and Martina Mincheva. 2011. "High dimensional covariance matrix estimation in approximate factor models". *Annals of statistics* 39 (6): 3320.

Fan, Jianqing, and Rui Song. 2010. "Sure independence screening in generalized linear models with NP-dimensionality". *The Annals of Statistics* 38 (6): 3567–3604.

Feller, William. 1968. *An introduction to probability theory and its applications: volume I.* Vol. 3. John Wiley & Sons London-New York-Sydney-Toronto.

Foerster, Andrew T, Pierre-Daniel G Sarte, and Mark W Watson. 2011. "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production". *Journal of Political Economy* 119 (1): 1–38.

Forgy, E. W. 1965. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". *Biometrics* 21:768–769.

Forni, Mario, and Marco Lippi. 1997. *Aggregation and the microfoundations of dynamic macroeconomics*. Oxford University Press.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. 2012. "The generalized dynamic factor model". *Journal of the American Statistical Association.*

— . 2004. "The generalized dynamic factor model consistency and rates". *Journal of Econometrics* 119 (2): 231–255.

— . 2000. "The generalized dynamic-factor model: Identification and estimation". *Review of Economics and statistics* 82 (4): 540–554.

Frank, Murray Z, and Vidhan K Goyal. 2009. "Capital structure decisions: which factors are reliably important?" *Financial management* 38 (1): 1–37.

Freund, Caroline L, and Diana Weinhold. 2004. "The effect of the Internet on international trade". *Journal of international economics* 62 (1): 171–189.

Freyberger, Joachim. 2012. "Nonparametric panel data models with interactive fixed effects".

Gautier, Eric, and Alexandre B Tsybakov. 2013. "Pivotal estimation in high-dimensional regression via linear programming". In *Empirical Inference*, 195–204. Springer.

Geer, Sara Van de, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. 2014. "On asymptotically optimal confidence regions and tests for high-dimensional models". *The Annals of Statistics* 42 (3): 1166–1202.

Giacomini, Raffaella, and Barbara Rossi. 2009. "Detecting and predicting forecast breakdowns". *The Review of Economic Studies* 76 (2): 669–705.

— . 2010. "Forecast comparisons in unstable environments". *Journal of Applied Econometrics* 25 (4): 595–620.

— . 2006. "How stable is the forecasting performance of the yield curve for output growth?" *Oxford Bulletin of Economics and Statistics* 68 (s1): 783–795.

Giraud, Christophe, Sylvie Huet, and Nicolas Verzelen. 2012. "High-Dimensional Regression with Unknown Variance". *Statist. Sci.* 27, no. 4 (): 500–518. doi:10.1214/12-STS398.

Glosten, Lawrence R, Ravi Jaganathan, and David E Runkle. 1993. "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks". *Journal of Finance* 48 (5): 1779–1801.

Gorman, William M. 1981. "Some engel curves". *Essays in the theory and measurement of consumer behaviour in honor of sir Richard Stone.*

Gorodetskii, VV. 1978. "On the strong mixing property for linear sequences". *Theory of Probability & Its Applications* 22 (2): 411–413.

Gotze, F, and C Hipp. 1994. "Asymptotic distribution of statistics in time series". *The Annals of Statistics*: 2062–2088.

Goyal, Amit, Christophe Pérignon, and Christophe Villa. 2008. "How common are common return factors across the NYSE and Nasdaq?" *Journal of Financial Economics* 90 (3): 252–271.

Goyal, Amit, and Pedro Santa-Clara. 2003. "Idiosyncratic risk matters!" *The Journal of Finance* 58 (3): 975–1007.

Goyal, Amit, and Ivo Welch. 2008. "A comprehensive look at the empirical performance of equity premium prediction". *Review of Financial Studies* 21 (4): 1455–1508.

— . 2003. "Predicting the equity premium with dividend ratios". *Management Science* 49 (5): 639–654.

Graham, John R, and Campbell R Harvey. 2001. "The theory and practice of corporate" nance: evidence from the" eld". *Journal of Financial Economics* 60 (187): 243.

Hahn, Jinyong, and Guido Kuersteiner. 2002. "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large". *Econometrica* 70 (4): 1639–1657.

Hahn, Jinyong, and Hyungsik Roger Moon. 2006. "Reducing bias of MLE in a dynamic panel model". *Econometric Theory* 22 (03): 499–512.

Hall, Peter, Jiashun Jin, and Hugh Miller. 2014. "Feature selection when there are many influential features". *Bernoulli* 20, no. 3 (): 1647–1671. doi:10.3150/13-BEJ536.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle". *Econometrica* 57 (2): 357–84.

Hansen, Karsten T, James J Heckman, and Kathleen J Mullen. 2004. "The effect of schooling and ability on achievement test scores". *Journal of Econometrics* 121 (1): 39–98.

Harris, Milton, and Artur Raviv. 1991. "The theory of capital structure". *the Journal of Finance* 46 (1): 297–355.

Haugen, Robert A, and Josef Lakonishok. 1988. *The incredible January effect: The stock market's unsolved mystery*. Business One Irwin.

Henkel, Sam James, J Spencer Martin, and Federico Nardari. 2011. "Time-varying short-horizon predictability". *Journal of Financial Economics* 99 (3): 560–580.

Hodrick, Robert J. 1992. "Dividend yields and expected stock returns: Alternative procedures for inference and measurement". *Review of Financial studies* 5 (3): 357–386.

Hogben, Leslie. 2006. *Handbook of linear algebra*. CRC Press.

Hsiao, Cheng. 2014. *Analysis of panel data*. Vol. 54. Cambridge university press.

Hsiao, Cheng, Trent W Appelbe, and Christopher R Dineen. 1993. "A general framework for panel data models with an application to Canadian customer-dialed long distance telephone service". *Journal of Econometrics* 59 (1-2): 63–86.

Islam, Nazrul. 1995. "Growth empirics: a panel data approach". *The Quarterly Journal of Economics*: 1127–1170.

Javanmard, Adel, and Andrea Montanari. 2014a. "Confidence intervals and hypothesis testing for high-dimensional regression". *The Journal of Machine Learning Research* 15 (1): 2869–2909.

Javanmard, Adel, and Andrea Montanari. 2015. "De-biasing the Lasso: Optimal Sample Size for Gaussian Designs". *ArXiv e-prints* (). arXiv: 1508.02757 [math.ST].

Javanmard, Adel, and Andrea Montanari. 2014b. "Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory". *Information Theory, IEEE Transactions on* 60 (10): 6522–6554.

Jia, Jinzhu, and Karl Rohe. 2012. "Preconditioning to comply with the irrepresentable condition". *arXiv preprint arXiv:1208.5584*.

Jin, Jiashun, and Zheng T.. Ke. 2014. "Rare and Weak effects in Large-Scale Inference: methods and phase diagrams". *ArXiv e-prints* (). arXiv: 1410.4578 [math.ST].

Jones, Charles I. 1995. "Time series tests of endogenous growth models". *The Quarterly Journal of Economics*: 495–525.

Jones, Charles P, Douglas K Pearce, and Jack W Wilson. 1987. "Can tax-loss selling explain the January effect? A note". *The Journal of Finance* 42 (2): 453–461.

Jurado, Kyle, Sydney C Ludvigson, and Serena Ng. 2015. "Measuring Uncertainty". *The American Economic Review* 105 (3): 1177–1216.

Kapetanios, George. 2010. "A testing procedure for determining the number of factors in approximate factor models with large datasets". *Journal of Business & Economic Statistics* 28 (3): 397–409.

Keim, Donald B, and Robert F Stambaugh. 1986. "Predicting returns in the stock and bond markets". *Journal of Financial Economics* 17 (2): 357–390.

Koijen, Ralph, and Stijn Van Nieuwerburgh. 2011. "Predictability of Returns and Cash Flows". *Annual Review of Financial Economics* 3 (1): 467–491.

Kramer, Charles. 1994. "Macroeconomic Seasonality and the January Effect". *Journal of Finance*: 1883–1891.

Lahiri, Soumendra Nath, et al. 2010. "Edgeworth expansions for studentized statistics under weak dependence". *The Annals of Statistics* 38 (1): 388–434.

Lettau, Martin, and Stijn Van Nieuwerburgh. 2008. "Reconciling the return predictability evidence". *Review of Financial Studies* 21 (4): 1607–1652.

Lewbel, Arthur. 1991. "The rank of demand systems: theory and nonparametric estimation". *Econometrica: Journal of the Econometric Society*: 711–730.

Lin, Chang-Ching, and Serena Ng. 2012. "Estimation of panel data models with parameter heterogeneity when group membership is unknown". *Journal of Econometric Methods* 1 (1): 42–55.

Litan, Robert E, and Alice M Rivlin. 2001. "Projecting the economic impact of the Internet". *The American Economic Review* 91 (2): 313–317.

Lloyd, Stuart. 1982. "Least squares quantization in PCM". *IEEE transactions on information theory* 28 (2): 129–137.

Lord, FM, MR Novick, and Allan Birnbaum. 1968. "Statistical theories of mental test scores."

Ludvigson, Sydney C, and Serena Ng. 2009. "Macro factors in bond risk premia". *Review of Financial Studies* 22 (12): 5027–5067.

— . 2007. "The empirical risk–return relation: A factor analysis approach". *Journal of Financial Economics* 83 (1): 171–222.

Mankiw, N Gregory, David Romer, and David N Weil. 1992. "A Contribution to the Empirics of Economic Growth". *The Quarterly Journal of Economics* 107 (2): 407–437.

McCracken, Michael W, and Serena Ng. 2015. "FRED-MD: a monthly database for macroeconomic Research". *Journal of Business & Economic Statistics*, no. just-accepted.

Meinshausen, Nicolai, Lukas Meier, and Peter Bühlmann. 2012. "P-values for high-dimensional regression". *Journal of the American Statistical Association*.

Merlevède, Florence, Magda Peligrad, and Emmanuel Rio. 2011. "A Bernstein type inequality and moderate deviations for weakly dependent sequences". *Probability Theory and Related Fields* 151 (3-4): 435–474.

Meyn, Sean P, and Richard L Tweedie. 2012. *Markov chains and stochastic stability*. Springer Science & Business Media.

Mitra, Ritwik, and Cun-Hui Zhang. 2014. "The Benefit of Group Sparsity in Group Inference with De-biased Scaled Group Lasso". *arXiv preprint arXiv:1412.4170*.

Moon, Hyungsik Roger, and Peter CB Phillips. 2004. "GMM estimation of autoregressive roots near unity with panel data". *Econometrica* 72 (2): 467–522.

Moon, Hyungsik Roger, and Martin Weidner. 2015. "Linear regression for panel with unknown number of factors as interactive fixed effects". *Econometrica*.

Nickl, Richard, Sara van de Geer, et al. 2013. "Confidence sets in sparse regression". *The Annals of Statistics* 41 (6): 2852–2876.

Ning, Yang, and Han Liu. 2014. "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models". *arXiv preprint arXiv:1412.8765*.

Onatski, Alexei. 2009. "Testing hypotheses about the number of factors in large factor models". *Econometrica* 77 (5): 1447–1479.

Paye, Bradley S, and Allan Timmermann. 2006. "Instability of return prediction models". *Journal of Empirical Finance* 13 (3): 274–315.

Peng, Hanxiang, and Anton Schick. 2012. "Asymptotic normality of quadratic forms with random vectors of increasing dimension". *Preprint*.

Pesaran, M Hashem. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure". *Econometrica* 74 (4): 967–1012.

Pesaran, M Hashem, and Elisa Tosetti. 2011. "Large panels with common factors and spatial correlation". *Journal of Econometrics* 161 (2): 182–202.

Pesaran, M Hashem, and Takashi Yamagata. 2008. "Testing slope homogeneity in large panels". *Journal of Econometrics* 142 (1): 50–93.

Pettenuzzo, Davide, and Allan Timmermann. 2011. "Predictability of stock returns and asset allocation under structural breaks". *Journal of Econometrics* 164 (1): 60–78.

Phillips, Peter CB, and Hyungsik R Moon. 1999. "Linear regression limit theory for nonstationary panel data". *Econometrica* 67 (5): 1057–1111.

Phillips, Peter CB, and Donggyu Sul. 2003. "Dynamic panel estimation and homogeneity testing under cross section dependence". *The Econometrics Journal* 6 (1): 217–259.

Piterbarg, Vladimir I. 2012. *Asymptotic methods in the theory of Gaussian processes and fields*. Vol. 148. American Mathematical Society.

Pouzo, Demian. 2015. "Bootstrap consistency for quadratic forms of sample averages with increasing dimension". *Electronic Journal of Statistics* 9 (2): 3046–3097.

Pritchard, Jonathan K. 2001. "Are rare variants responsible for susceptibility to complex diseases?" *The American Journal of Human Genetics* 69 (1): 124–137.

Qian, Junhui, and Le Wang. 2012. "Estimating semiparametric panel data models by marginal integration". *Journal of Econometrics* 167 (2): 483–493.

Qu, Zhongjun, and Pierre Perron. 2007. "Estimating and testing structural changes in multivariate regressions". *Econometrica* 75 (2): 459–502.

Rajan, Raghuram G, and Luigi Zingales. 1995. "What Do We Know about Capital Structure? Some Evidence from International Data". *Journal of Finance* 50 (5): 1421–60.

Rapach, David E, Jack K Strauss, and Guofu Zhou. 2010. "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy". *Review of Financial Studies* 23 (2): 821–862.

Rapach, David E, and Mark E Wohar. 2006. "Structural breaks and predictive regression models of aggregate US stock returns". *Journal of Financial Econometrics* 4 (2): 238–274.

Raskutti, Garvesh, Martin J Wainwright, and Bin Yu. 2011. "Minimax rates of estimation for high-dimensional linear regression over-balls". *Information Theory, IEEE Transactions on* 57 (10): 6976–6994.

Robinson, Peter M, et al. 1988. "Root-N-Consistent Semiparametric Regression". *Econometrica* 56 (4): 931–54.

Romano, Joseph P, and Azeem M Shaikh. 2012. "On the uniform asymptotic validity of subsampling and the bootstrap". *The Annals of Statistics* 40 (6): 2798–2822.

Rosenberg, Barr. 1972. "The estimation of stationary stochastic regression parameters reexamined". *Journal of the American Statistical Association* 67 (339): 650–654.

Ross, Stephen A. 1976. "The arbitrage theory of capital asset pricing". *Journal of economic theory* 13 (3): 341–360.

Rossi, Barbara. 2013. "Advances in Forecasting under Instability". *Handbook of Economic Forecasting* 2:1203–1324.

Rudelson, Mark, and Shuheng Zhou. 2013. "Reconstruction from anisotropic random measurements". *Information Theory, IEEE Transactions on* 59 (6): 3434–3447.

Shao, Xiaofeng. 2010. "A self-normalized approach to confidence interval construction in time series". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (3): 343–366. ISSN: 1467-9868. doi:10.1111/j.1467-9868.2009.00737.x. http://dx.doi.org/10.1111/j.1467-9868.2009.00737.x.

Stock, James H, and Mark W Watson. 1998. *Diffusion indexes*. Tech. rep. National bureau of economic research.

— . 1996. "Evidence on structural instability in macroeconomic time series relations". *Journal of Business & Economic Statistics* 14 (1): 11–30.

— . 2015. "Factor models and structural vector autoregressions in macroeconomics". *Handbook of macroeconomics* 2.

— . 2002a. "Forecasting using principal components from a large number of predictors". *Journal of the American statistical association* 97 (460): 1167–1179.

— . 2006. "Forecasting with many predictors". *Handbook of economic forecasting* 1:515–554.

— . 2002b. "Macroeconomic forecasting using diffusion indexes". *Journal of Business & Economic Statistics* 20 (2): 147–162.

— . 2007. "Why has US inflation become harder to forecast?" *Journal of Money, Credit and banking* 39 (s1): 3–33.

Su, Liangjun, and Qihui Chen. 2013. "Testing homogeneity in panel data models with interactive fixed effects". *Econometric Theory* 29 (06): 1079–1135.

Su, Liangjun, Sainan Jin, and Yonghui Zhang. 2015. "Specification test for panel data models with interactive fixed effects". *Journal of Econometrics* 186 (1): 222–244.

Sun, Tingni, and Cun-Hui Zhang. 2012. "Scaled sparse linear regression". *Biometrika*: ass043.

Swamy, Paravastu AVB. 1970. "Efficient Inference in a Random Coefficient Regression Model". *Econometrica* 38 (2): 311–23.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society (Series B)* 58:267–288.

Titman, Sheridan, and Roberto Wessels. 1988. "The determinants of capital structure choice". *The Journal of finance* 43 (1): 1–19.

van de Geer, S., and J. Jankova. 2016. *Semi-parametric efficiency bounds and efficient estimation for high-dimensional models*. arXiv: `1601 . 00815 [math.ST]`.

Vershynin, Roman. 2010. "Introduction to the non-asymptotic analysis of random matrices". *arXiv preprint arXiv:1011.3027*.

Viceira, Luis M. 1997. "Testing for structural change in the predictability of asset returns". *Manuscript, Harvard University* 4 (3.5): 3–0.

Wang, Bo-Ying, and Bo-Yan Xi. 1997. "Some inequalities for singular values of matrix products". *Linear algebra and its applications* 264:109–115.

Wang, Xiangyu, David Dunson, and Chenlei Leng. 2016. "DECOrrelated feature space partitioning for distributed sparse regression". *arXiv preprint arXiv:1602.02575*.

Ward, Rachel A. 2009. "Compressed Sensing With Cross Validation". *Information Theory, IEEE Transactions on* 55 (12): 5773–5782. ISSN: 0018-9448.

Welch, Ivo. 2004. "Capital Structure and Stock Returns". *Journal of Political Economy* 112 (1): 106–131.

Yuan, Ming, and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1): 49–67.

Zaffaroni, Paolo. 2009. "Generalized least squares estimation of panel with common shocks". In *XXIX-th European Meeting of Statisticians, 2013, Budapest Contents*, 41.

Zhang, Cun-Hui, and Stephanie S Zhang. 2014. "Confidence intervals for low dimensional parameters in high dimensional linear models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 217–242.

Zhao, Peng, and Bin Yu. 2006. "On Model Selection Consistency of Lasso". *J. Mach. Learn. Res.* 7 (): 2541–2563. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=1248547.1248637.

Zou, Hui. 2006. "The adaptive lasso and its oracle properties". *Journal of the American statistical association* 101 (476): 1418–1429.