Most Powerful Test against High Dimensional Free Alternatives

Yi He[1,2], Sombut Jaidee[2], and Jiti Gao[2]

[1]University of Amsterdam
[2]Monash University

August 19, 2020

Abstract

We develop a powerful quadratic test for the overall significance of many covariates in a dense regression model in the presence of nuisance parameters. By equally weighting the sample moments, the test is asymptotically correct in high dimensions even when the number of coefficients is larger than the sample size. Our theory allows a non-parametric error distribution and weakly exogenous nuisance variables, in particular autoregressors in many applications. Using random matrix theory, we show that the test has the optimal asymptotic testing power among a large class of competitors against local dense alternatives whose direction is free in the eigenbasis of the sample covariance matrix among regressors. The asymptotic results are adaptive to the covariates' cross-sectional and temporal dependence structure and do not require a limiting spectral law of their sample covariance matrix. In the most general case, the nuisance estimation may play a role in the asymptotic limit and we give a robust modification for these irregular scenarios. Monte Carlo studies suggest a good power performance of our proposed test against high dimensional dense alternative for various data generating processes. We apply the test to detect the significance of over one hundred exogenous variables in the FRED-MD database for predicting the monthly growth in the US industrial production index.

Keywords: High–dimensional linear model; null hypothesis; uniformly power test

JEL classification: C12, C21, C55

1 Introduction

When the data dimension exceeds the sample size, the classical variance ratio statistics (e.g. F statistic) are degenerate and therefore infeasible for testing many regression coefficients simultaneously. Even with a smaller data dimension but comparable to the sample size, the traditional quadratic tests still suffer in weak power; see, e.g., Zhong and Chen (2011). One existing solution is to consider the sparse models where the true model only deviates from the null hypothesis in only a few components. For example, by using higher criticism methods (Donoho and Jin, 2004; Hall and Jin, 2010; Zhong et al., 2013), by adding a non-trivial power enhancement component sensitive to the sparsity (Fan et al., 2015; Kock and Preinerstorfer, 2019), or by detecting the extreme behavior of marginal p values or t statistics (**??**Chernozhukov et al., 2019), one may improve the testing power in many applications. For more general sparse inference theory for high dimensional linear mean regression we also refer to **??**, **?**, **?**, **?**, **?** and many references therein.

While the sparsity assumption is convenient, it may not be always available in economic applications. In particular, Giannone et al. (2017) observe non-sparsity among the original covariates for five out of six important economic data sets. When the true model is dense, Goeman et al. (2006) propose a powerful score test against local departures from the null. The rationale behind is a version of Neyman–Pearson lemma under an empirical Bayesian model, by taking into account the likelihood of the random regression coefficients. Strictly, their approach requires knowledge of the error distribution. This is an ambitious task in high dimensions, as the sample residuals may be degenerate for useful statistical inference. A list of follow-up works, such as Goeman et al. (2011) and Guo and Chen (2016), extend this approach to generalized linear models with some prior knowledge (such as the variance) of the error distribution. U-statistic based tests are also available in some independent models; see, e.g., Zhong and Chen (2011) and Cui et al. (2018). For a simple linear hypothesis in dense models, Zhu and Bradic (2018) develop a test using an implicit sparse condition on the projected covariance matrix among predictors. **?** extend the inference theory with heteroscedasticity for low-dimensional parameters in the presence of many nuisance parameters but less than the sample size. Our review is not exhaustive, and we also refer to the references of the aforementioned papers.

We follow the dense modeling strategy and are mostly interested in the high dimensional data with the number of unrestricted coefficients larger than the sample size. This means that we are typically testing a large (sub)set of original covariates rather than their unknown sparse representation if there is any. Therefore, our test procedure is more interpretable from an economic perspective and is free from tuning parameters. Using random matrix theory, we relax the parametric conditions on regression errors and allow nuisance variables, particularly autoregressors, to enter the estimation procedure. To our knowledge, our results are novel and the scope of the applications is much wider in dealing with nuisance parameters and non-Gaussian models.

We start from the autoregressive model to be used in our empirical study, where we test the overall significance of a large set of exogenous predictors. We then provide direct relaxations for general nuisance variables as linear combinations of lagged and contemporary information. A natural approach is to estimate the nuisance coefficients with the restrictions and then test on the residuals. This is similar to the score test in Goeman et al. (2011) using a Gaussian likelihood, but we show that the approach generalizes and does not require specific knowledge about the error distribution. For generality we work with nonrandom coefficients in our asymptotic theory, while our free alternatives (to be defined later on) are originated from the exchangeable Baysian model. To relax the independence assumption between the autoregressors and errors, we standardize the test statistic into a martingale form and establish its asymptotic normality using martingale central limit theorem. Adapting to the non-sparse cross sectional dependence structure and unknown temporal dependence structure among regressors, we study strongly exogenous variables, that is, weak predictors uncorrelated with the shocks to the response variable. A more sophisticated (i.e. weaker) exogeneity assumption is possible using more general martingale theory (see, e.g., Hall and Heyde, 1980, Chapter 3.3) with the cost of greater complications, but we leave the relaxations as future work.

In order for the limiting power to be nontrivial, our study is based on the local alternatives with weak signal length converging to zero at a proper rate depending on the data dimension and the sample size. This rate turns out to be similar to that at the detection boundary in, e.g., **?** and Arias-Castro et al. (2011) between the sparse and dense Gaussian models, although we allow a general error distribution here. Using random matrix theory, we derive the asymptotic power of our proposed test and compare it with that of a large class of other quadratic tests. In the spirit of Ledoit and Wolf (2012), we construct the competing quadratic statistics based on spectral transformations of the large dimensional weighting matrix for testing our moment conditions, including some naive case equivalent to the classical F statistic asymptotically as the data dimension diverges. The nuisance estimation, even for a small number of parameters, can have a substantial effect for irregular time-dependent high-dimensional covariates in the most general case. Examples include the high-dimensional moving-average or autoregressive predictors sharing common lagged coefficients across all dimensions, as shown in our simulations. For an unified inference, in the end of our theory we establish a robust method which is novel to our knowledge.

Equally weighting the sample moments yields the optimal asymptotic power for free alternatives under regularity conditions, when comparing with the competitors in our asymptotic theory. Roughly, we call a regression coefficient vector free if its direction is unrelated to the spectral information of the sample covariance matrix of the associated predictors; we give the mathematical definition in the next section. The free models play a crucial role in random matrix theory for characterizing the eigenvector asymptotics of large sample covariance matrix; see, e.g., Bai et al. (2007), Ledoit and Péché (2011), Xia et al. (2013), Pan (2014), Xi et al. (2020) and many references therein. When the regressors are independent and identically distributed over time, the aforementioned papers show that the class of free alternatives coincides under the eigenbasis of the population covariance matrix and that of the sample covariance matrix; in particular, when regressors are orthogonal and properly standardized, all directions over the unit sphere are free. More straightforward examples include the sample paths from the (exchangeable) stochastic coefficients model in, e.g., Goeman et al. (2011) and Dobriban and Wager (2018). As noted above, for generality, we unify the definition in a frequentist framework and consider only deterministic alternatives in our asymptotic theory.

Last but not least, our asymptotic theory uses the observable empirical spectral distribution of the large sample covariance matrix rather than its limit. Hence, our statistical methods are data adaptive and remain valid even when the empirical spectral distribution diverges. The asymptotic results remain true by substituting in the limiting distribution, if there is any, for such as the data generating processes in Marčenko and Pastur (1967), Yin (1986), Silverstein and Bai (1995), Silverstein (1995), Zhang (2006), El Karoui (2009), Jin et al. (2009), Zheng and Li (2011), Pan et al. (2014), Liu et al. (2015), Xia and Zheng (2018) and many references therein.

The rest of the paper is organized as follows. We develop the asymptotic theory in Section 2. Section 3 presents a simulation study that demonstrates the finite-sample performance of our optimal test. In Section 4 we provide a macroeconomic application using the FRED-MD database, which is designed for high dimensional empirical analysis. We conclude the paper in Section **??**. We sketch the mathematical proofs of the main theorems and the corollaries in the end. To save space, the technical details including the proofs of the lemmas and propositions are provided in the supplement. Besides, in the supplement we report more simulation and empirical analysis results, verify the technical conditions for some examples and give the most general asymptotic limits. Throughout, for any matrix A, we denote is $(t+l,t)$-th element as $A(t+l,t)$, its transpose by A^T, its trace by $\operatorname{tr}(A)$, its spectral norm by $\|A\|_{sp} = \sup_{\|u\|=1}\|Au\|$, and its Frobenius norm by $\|A\| = \sqrt{\operatorname{tr}(A^TA)}$. When A is symmetric, we denote its smallest and largest eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively; if A is also positive semi-definite, we use $\lambda_{\max}(A)$ to denote its spectral norm. We denote by '$\xrightarrow{\mathbb{P}}$' the convergence in probability and by '$\xrightarrow{d}$' the convergence in distribution. With a slight abuse of notation, we denote by $Z_n \xrightarrow{d} \mathcal{N}(0,1)$ if the sequence (or array) of random variables $\{Z_n\}$ converges in distribution to some standard normal variable. Unless specified otherwise, all asymptotic results hold as the sample size $n \to \infty$ in the probability space with the largest sigma algebra.

2 Asymptotic theory

As a motivating example, suppose we observe responses $y_1, \ldots, y_n \in \mathbb{R}$ and the initial values $\{y_0, y_{-1}, \ldots, y_{1-d}\}$ generated by an autoregressive regression model

$$y_t = \theta_0 + \sum_{i=1}^{d} \theta_i y_{t-i} + x_t^T\beta + \varepsilon_t, \tag{2.1}$$

where $x_t = (x_{t,1}, \ldots, x_{t,p})^T \in \mathbb{R}^p$ are observable exogenous variables with unknown coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$; ε_t are unobservable regression errors with zero mean and unknown variance. Let $z_t := (1, y_{t-1}, \ldots, y_{t-d})^T \in \mathbb{R}^{d+1}$ collect the lagged dependent variables whose coefficients $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$ we always estimate. Therefore, we can rewrite the model into a general form given by

$$y_t = z_t^T\theta + x_t^T\beta + \varepsilon_t, \quad t = 1, \ldots, n. \tag{2.2}$$

We postpone the extension beyond the autoregressive model to Subsection 2.3. For cross-sectional or more general data sets, we allow the nuisance variables $z_t := (1, z_{t,1}, \ldots, z_{t,d})$ to contain both the lagged and current information in such a way that

$$z_{t,i} = \alpha_i + \sum_{l=1}^{\infty} \psi_i(l) w_{t-l} + r_{t,i}, \quad w_t = x_t^T \beta + \varepsilon_t, \tag{2.3}$$

with mean $\alpha_i = \mathbb{E} z_{t,i}$, moving average coefficients $\{\psi_i(l) : l = 1, 2, \ldots\}$ that may be different for different nuisance variables, and contemporary random shocks $r_{t,i}$. Note that the coefficient vector β may enter both equations (2.2) and (2.3). We use only the former one for our testing statistic, to avoid estimating the infinite sequence(s) of moving average coefficients. At the moment our focus is on the autoregressive model (2.1) as a starting point.

Our null hypothesis is that all the exogenous variables x_t are irrelevant, that is,

$$H_0 : \ \beta = \mathbf{0}_p, \tag{2.4}$$

where $\mathbf{0}_p$ denotes the p-dimensional vector of all zeros. The zeros are not special. By rewriting the linear regression model, one may replace the null value by an arbitrary non-zero coefficient vector. Similarly, one may map the coefficients to their linear combinations by transforming the variables. We use the zeros as null values for presentation convenience. We consider a large dimension p comparable to the sample size n:

Assumption 1. The dimension $p = p(n) \to \infty$ and $p/n \to c \in (0, \infty)$ as the sample size $n \to \infty$.

This asymptotic regime is standard in random matrix theory (see, e.g., the surveys in Bai and Silverstein, 2010), which is useful in many economic applications with comparable p and n (see, e.g., Stock and Watson, 2002 and Ledoit and Wolf, 2017). The concentration ratio p/n, usually larger than 1, plays an important role in our asymptotic limits. While the assumption rules out the case $p/n \to \infty$ in general, our simulations in the next section suggest that our asymptotic approximation performs well for a wide range of p/n in finite samples. Indeed, our asymptotic theory actually allows the dimension p to diverge at a higher rate (i.e. a very large p/n in practice) under the null hypothesis but we maintain the current setup to simplify the power theory for the non-sparse alternatives. To avoid unnecessary complications, we assume that the nuisance dimension d is fixed although our proofs actually allow for a slow rate of divergence.

Note that our model is indexed by the sample size n as the dimension $p = p(n)$ diverges, but we suppress this in the subscripts whenever no confusion arises. Throughout we assume the following exogeneity and identification conditions.

Assumption 2. The following conditions hold:

(a) The regression errors $\varepsilon_t = \sigma_n \eta_t$, for some unknown variance $\sigma_n^2 = \sigma_n^2(x_1, \cdots, x_n)$ bounded away from zero almost surely, and $\{\eta_t, \mathcal{F}_{n,t}\}$ is a martingale difference array such that $\mathbb{E}[\eta_t \mid \mathcal{F}_{n,t-1}] = 0$

and $\mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_{n,t-1}\right] = 1$ where $\mathcal{F}_{n,t-1}$ is the product of sigma-algebras generated by $\{\eta_s : s \leq t-1\}$, $\{z_s : 1 \leq s \leq t\}$ and $\{x_s : 1 \leq s \leq n\}$. For some $\iota \in (0,1]$, $\mathbb{E}\left[|\eta_t^2 - 1|^{1+\iota} \mid \mathcal{F}_{n,0}\right] \leq \kappa_n$ for all $t = 1, \ldots, n$ with probability 1, and $\kappa_n = \kappa_n(x_1, \ldots, x_n) = O_{\mathbb{P}}(1)$.

(b) The regressors $(z_t^T, x_t^T)^T$ are identically distributed over index $t = 1, \ldots, n$, and their population covariance matrix is finite and positive definite for each n; without loss of generality, x_t is demeaned such that $\mathbb{E}[x_t]$ has all zero entries.

(c) All the roots of the d–th degree polynomial equation $1 - \theta_1\lambda - \theta_2\lambda^2 - \ldots - \theta_d\lambda^d = 0$ are greater than 1 in absolute value when considering the autoregressive model (2.1).

Condition (a) only requires slightly more than the second moment of η_t and allows heavy tails. Condition (b) assumes the existence of the second moments among regressors. We allow a complex dependence structure over the (time) index t, without restricting any particular form of stationarity nor mixingness. The zero mean condition is only for presentation convenience here, as we always demean the predictors when pre-processing the data. Conditions (c) is a standard stability condition for autoregressive model. The special case with no autoregressor (i.e. $d = 0$) is included in our theory, although it is not very interesting here.

In the following subsections, we first develop our test statistic and establish its asymptotic distribution under the null hypothesis. Then, we introduce a sequence of free alternatives and derive an asymptotic power property for our test. Extensions to non-free alternatives are also discussed. To show the optimality of our proposed test, we compare it with a class of other quadratic tests using weighted matrixes based on spectral transformations of the large dimensional sample covariance matrix for our moment conditions. Finally, we generalize the results beyond autoregressive models for possibly irregular predictors.

2.1 Testing the null hypothesis

Observe that our problem is equivalent to testing the high dimensional moment condition given by

$$\mathbb{E}\left[x_t(y_t - z_t^T\theta)\right] = \Sigma\beta \overset{H_0}{=} (0, \ldots, 0)^T,$$

as the population covariance matrix $\Sigma := \mathbb{E}\left[x_t x_t^T\right]$ is positive definite, where '$\overset{H_0}{=}$' denotes equality under the null hypothesis. We estimate the nuisance parameters θ from the restricted regression model given by

$$y_t \overset{H_0}{=} z_t^T\theta + \varepsilon_t, \ t = 1, \ldots, n,$$

that is, in matrix form,

$$Y \overset{H_0}{=} Z\theta + \epsilon,$$

where $Y = (y_1, \dots, y_n)^T$ denotes the response vector, $Z = (z_1, \dots, z_n)^T \in \mathbb{R}^{n\times(d+1)}$ denotes the nuisance design matrix, and $\epsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ denotes the error vector. Minimizing the squared errors

$$\sum_{t=1}^{n} (y_t - z_t^T \theta)^2,$$

yields the least-squares estimator

$$\widehat{\theta} = \left(Z^T Z\right)^{-1} Z^T Y,$$

and the residual vector

$$e = (e_1, \dots, e_n)^T = Y - Z\widehat{\theta}.$$

Let $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ be the sample mean of testing variables. Note that $\sum_{t=1}^{n} e_t = 0$ as the intercept term is included in estimation. Substituting the residual vector e for the error vector ϵ, our estimate of the moment vector $\mathbb{E}\left[x_t \varepsilon_t\right]$ is therefore given by

$$\frac{1}{n}\sum_{t=1}^{n} x_t e_t = \frac{1}{n}\sum_{t=1}^{n} (x_t - \bar{x}) e_t = \frac{1}{n}\widetilde{X} e,$$

where $\widetilde{X}$ is the demeaned design matrix given by

$$\widetilde{X} = (x_1 - \bar{x}, \dots, x_n - \bar{x})^T =: (\widetilde{x}_1, \dots, \widetilde{x}_n)^T.$$

Summing up the squared sample moments yields

$$\frac{1}{n^2} e^T \widetilde{X}\widetilde{X}^T e = \frac{1}{n} e^T \left(\frac{1}{n}\widetilde{X}\widetilde{X}^T\right) e =: \frac{1}{n} e^T \underline{S}_n e,$$

where $\underline{S}_n = \frac{1}{n}\widetilde{X}\widetilde{X}^T$. Let $A_n = \underline{S}_n - \operatorname{diag}(\underline{S}_n)$, where $\operatorname{diag}(\underline{S}_n)$ denotes a diagonal matrix with same main diagonal of $\underline{S}_n$. We center the quadratic form asymptotically by removing the diagonal elements, normalize it, and construct the quadratic test statistic

$$Q_n := \frac{1}{\sqrt{2}\,\|A_n\|} e^T A_n e = \frac{1}{\left\|\widetilde{A}_n\right\|} e^T \widetilde{A}_n e,$$

where $\widetilde{A}_n$ denotes the lower triangular part of A_n.

Rewriting Q_n into an (approximate) martingale form

$$Q_n = \frac{\sqrt{2}}{\|A_n\|} \sum_{1\le s<t\le n} e_t e_s \frac{1}{n} \widetilde{x}_t^T \widetilde{x}_s,$$

and applying the martingale limit theorems, we shall establish the asymptotic distribution of Q_n under the null hypothesis (2.4).

Theorem 1 (Oracle test)**.** *Let Assumptions 1 and 2 hold. In addition, suppose that:*

(i) The cross product matrix $\widehat{\Omega} := \frac{1}{n} Z^T Z$ has eigenvalues bounded away from 0.

(ii) $\lambda_{\max}\left(\widetilde{A}_n^T\widetilde{A}_n\right) = o_{\mathbb{P}}\left(\|A_n\|^2\right)$.

(iii) $\sum_{t=1}^{n-l}|A_n(t+l,t)| = o_{\mathbb{P}}\left(n^{\frac{1}{2}}\|A_n\|\right)$ *for each* $l \geq 1$, *where* $A_n(s,t)$ *denotes the* (s,t) *element of* A_n.

Then $Q_n/\sigma_n^2 \xrightarrow{d} \mathcal{N}(0,1)$ *under the null hypothesis* (2.4).

We offer some general discussions on the conditions. Condition (i) is a trivial identification condition that holds with probability approaching one, for example, when $\widehat{\Omega} \xrightarrow{\mathbb{P}} \Omega$ for some positive definite matrix Ω. In the supplementary document (He et al., 2020) we check condition (ii) for our simulation models. Here, we argue that the condition is weak enough to allow much more general data sequences. Observe that $\|A_n\|^2 = 2\,\mathrm{tr}(\widetilde{A}_n^T\widetilde{A}_n)$, and therefore we can rewrite condition (ii) slightly as

$$\lambda_{\max}\left(\widetilde{A}_n^T\widetilde{A}_n\right) = o_{\mathbb{P}}\left(\mathrm{tr}(\widetilde{A}_n^T\widetilde{A}_n)\right).$$

In other words, we assume that the individual eigenvalues of $\widetilde{A}_n^T\widetilde{A}_n$ are asymptotically negligible to their sum. This is essential for a Lindeberg's condition when applying the martingale central limit theorem to random quadratic forms; see, e.g., de Jong (1987), Wu and Shao (2007) and many references therein. We use the lower triangular part $\widetilde{A}_n$ instead of the full matrix A_n to provide a (much) weaker condition, as it is not hard to show that $\lambda_{\max}\left(\widetilde{A}_n^T\widetilde{A}_n\right) \leq \lambda_{\max}\left(A_n^TA_n\right) = \|A_n\|_{sp}^2$. Indeed, even when $\|A_n\|_{sp}^2$ is non-negligible relative to $\|A_n\|^2$, our condition still holds if A_n satisfies the "large p, small n" paradigm (see, e.g., Cui et al. 2018):

$$\mathrm{tr}\left(A_n^4\right) = o_{\mathbb{P}}\left(\mathrm{tr}^2\left(A_n^2\right)\right),$$

where we may replace A_n^2 by $\widetilde{A}_n^T\widetilde{A}_n$ if necessary.

We call condition (iii) a regular scenario, which allows the low-dimensional estimation error of nuisance parameters to die away in high dimensions. For general purposes, we relax this condition to allow for an irregular data sequence and develop a robust method in Subsection 2.3. By Cauchy–Schwarz inequality, it suffices to assume that each subdiagonal of A_n is negligible in the sense that

$$\sum_{t=1}^{n-l} A_n^2(t+l,t) = o_p(\|A_n\|^2), \text{ for each } l \geq 1, \tag{2.5}$$

as required in Wu and Shao (2007), equation (13). We can verify condition (2.5), for example, for independent sequence $\{x_t\}$ with covariance matrix $\Sigma = \mathbb{E}[x_tx_t^T]$ such that $\mathrm{tr}(\Sigma^2) = o(np)$, when $\|A_n\|^2$ diverges at the rate of p under regularity conditions. In applications, as we suggested above, the condition can be justified from data as the weighting matrix A_n is observable.

A feasible test with size $\alpha \in (0,1)$ is therefore to reject the null if

$$Q_n > \widehat{\sigma}^2\Phi^{-1}(1-\alpha), \tag{2.6}$$

where Φ^{-1} denotes the quantile function of a standard normal variable, and $\widehat{\sigma}^2$ is a consistent estimator of σ_n^2 in the sense that $\widehat{\sigma}^2/\sigma_n^2 \xrightarrow{\mathbb{P}} 1$. When the true model is sparse, a consistent estimator of the error variance in (ultra)high dimension is available using the refitted cross-validation techniques (Fan et al., 2012). For non-sparse models but with regression coefficients in some special directions, an alternative estimator is available by solving special moment conditions (Dicker, 2014). The optimal estimation method of the error variance in high dimensional regression model is beyond the scope of this paper. Here, for a statistical testing purpose, we propose to simply use the restricted least-squares estimator

$$\widehat{\sigma}^2 = \frac{1}{n-(d+1)} e^T e, \tag{2.7}$$

where e is the null residual vector as above. Our estimator is feasible even when $p > n$. To summarize, we provide a final corollary for our feasible test.

Corollary 1 (Feasible test)**.** *Under the conditions of Theorem 1 and using the variance estimator (2.7), our rejection rule (2.6) is asymptotically correct, that is,*

$$\mathbb{P}\left(Q_n > \widehat{\sigma}^2 \Phi^{-1}(1-\alpha)\right) \to \alpha,$$

for any size $\alpha \in (0,1)$.

2.2 Power theory for free alternatives

In order for the limiting power to be nontrivial, we consider the local alternatives with weak signal length $\|\beta\|^2 = \sum_{i=1}^p \beta_i^2$ asymptotically proportional to $\sqrt{p}/n$. The factor n^{-1} comes from the low-dimensional case (when p is small) and the extra factor $p^{1/2}$ turns out to be an appropriate rate in high dimensions from our proofs. See Introduction for more details about the rate. Our local alternative is therefore given by

$$H_1: \ h^2 := \lim_{n\to\infty} \frac{n}{\sqrt{p}} \|\beta\|^2 \in (0,\infty). \tag{2.8}$$

Under the null hypothesis (2.4) we have $h^2 = 0$.

As mentioned in the introduction, we specify more structures on β while allowing it to be non-sparse. Let the spectral decomposition of the sample covariance matrix be

$$S_n := \frac{1}{n} \widetilde{X}^T \widetilde{X} = U_n \Lambda_n U_n^T,$$

where $\Lambda_n = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal matrix with the eigenvalues of S_n on the diagonal, and $U_n = (u_1, \ldots, u_p)$ is an orthogonal matrix whose columns are the corresponding eigenvectors. Define the empirical spectral distribution of S_n by

$$F^{S_n}(x) = \frac{1}{p} \sum_{i=1}^p \mathbb{1}\left(\lambda_i \le x\right).$$

Observe that our weighting matrix $\underline{S}_n$ shares the same set of positive eigenvalues with S_n, and their empirical spectral distributions satisfy the equation

$$F^{\underline{S}_n} = \left(1 - \frac{p}{n}\right) I_{[0,\infty)} + \frac{p}{n} F^{S_n},$$

where $I_{[0,\infty)}$ is a step function with value 1 on $[0,\infty)$ and zero otherwise. We are interested in the free alternatives such that the projections of coefficient vector β on the eigenbasis of S_n are asymptotically independent of the eigenvalues. More precisely, define the weighted empirical spectral distribution

$$F^{S_n}(x;\beta) := \frac{1}{\beta^T\beta}\sum_{i=1}^{p}\left(u_i^T\beta\right)^2 \mathbb{1}\left(\lambda_i \leq x\right),$$

where $u_i^T\beta$ is the i-th eigenbasis coordinate of β and $\beta^T\beta = \sum_{i=1}^{p}\left(u_i^T\beta\right)^2$ by Pythagorean theorem. Note that $F^{S_n}(x;\beta)$ only depends on the direction of β. We assume that $F^{S_n}(x;\beta)$ satisfies the following condition:

Assumption 3. The direction of coefficient vector β is free in the sense that

$$F^{S_n}(x;\beta) - F^{S_n}(x) \to 0$$

almost surely for all $x \in [0,\infty)$ unless $\beta = \mathbf{0}_p$.

In other words, the eigenvectors of S_n do not contain the information of the underlying regression vector β in large samples. This is an interesting case where the eigenmatrix U_n behaves as it is(were) asymptotically uniformly distributed over orthogonal matrices; see, e.g., Bai et al. (2007), Pan (2014), Xia et al. (2013), Xi et al. (2020), and Chapter 10 of Bai and Silverstein (2010) for more discussions of this interesting property.

Theorem 2 (Oracle Power)**.** *Under the conditions of Theorem 1 and Assumption 3, suppose moreover that:*

(i) $\lambda_{\max}(S_n) = O_{\mathbb{P}}(1)$, $\lambda_{\max}(\Sigma) = O(1)$ *and* $\lambda_{\max}\left(\mathbb{E}\left[\bar{x}\bar{x}^T\right]\right) = O(1)$.

(ii) $\operatorname{var}\left[x_t^T x_t\right] = o_{\mathbb{P}}(n^2)$, *or more generally, the diagonal elements of the* $n\times n$ *matrix* $\underline{S}_n$ *concentrate around their average with a vanishing sample variance, that is,*

$$\frac{1}{n}\sum_{t=1}^{n}\left(\underline{S}_n(t,t) - \frac{1}{n}\operatorname{tr}(\underline{S}_n)\right)^2 \xrightarrow{\mathbb{P}} 0, \tag{2.9}$$

where $\underline{S}_n(t,t)$ *denotes the* t*-th diagonal element of* $\underline{S}_n$.

(iii) $\mathbb{E}(x_t^T\xi_n)^4$ *is bounded in* n *for each sequence of unit vectors* $\{\xi_n\}$.

Under the null (2.4) *with* $h = 0$ *or under the local alternatives* (2.8),

$$\frac{Q_n}{\sigma_n^2} - \frac{h^2}{\sqrt{2}\sigma_n^2}\varpi_n \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\varpi_n = \sqrt{\int x^2 dF^{S_n}(x) - \frac{p}{n}\left(\int x dF^{S_n}(x)\right)^2} = \sqrt{\frac{n}{p}\cdot\operatorname{var}\left[\underline{\Lambda}_n \mid F^{\underline{S}_n}\right]}, \tag{2.10}$$

and $\underline{\Lambda}_n$ *is a random variable from the (random) spectral distribution* $F^{\underline{S}_n}$, *provided that* ϖ_n *is bounded away from zero with probability approaching 1.*

Remark 1 (General alternatives). The results remain true for non-free alternatives, that is, the alternative beyond Assumption 3, with a general form

$$\varpi_n = \frac{\int x^2 dF^{S_n}(x;\beta) - \frac{p}{n}\int x dF^{S_n}(x;\beta) \int x dF^{S_n}(x)}{\sqrt{\int x^2 dF^{S_n}(x) - \frac{p}{n}(\int x dF^{S_n}(x))^2}}$$

depending on the unknown direction of the underlying coefficients β. For example, consider the adaptive coefficient vector $\beta = U_n \Lambda_n^{s/2} \mathbf{1}_p$ (or those in the same direction) as in the first set of simulations in Goeman et al. (2006). It is not hard to verify that for all $s > 0$

$$\varpi_n = \sqrt{\frac{n}{p} \cdot \mathrm{var}\left[\underline{\Lambda}_n \mid F^{\underline{S}_n}\right]} \cdot \frac{p}{n} \frac{\mathrm{cov}\left[\underline{\Lambda}_n, \underline{\Lambda}_n^{1+s} \mid F^{\underline{S}_n}\right]}{\mathrm{var}\left[\underline{\Lambda}_n \mid F^{\underline{S}_n}\right] \cdot \mathbb{E}\left[\underline{\Lambda}_n^{s} \mid F^{\underline{S}_n}\right]},$$

and the same holds for $s \leq 0$ if we set $0^s = 0$. In general, when β is arbitrary (but not free), the departure ϖ_n does not have a tractable form to produce an interesting theory here, and thus we leave it for future study.

Remark 2 (Time-variation adjustment). The kurtosis condition (iii) is not necessary if the concentration condition (2.9) holds in the strictest sense: all the diagonal elements of $\underline{S}_n$ are equal. This is possible by construction if we substitute $\widetilde{x}_t$ by the time-variation adjusted data (Zheng and Li, 2011) given by

$$\widetilde{x}_{t,\mathrm{adj}} = \frac{\widetilde{x}_t}{\|\widetilde{x}_t\|} \cdot \sqrt{\mathrm{tr}(S_n)}.$$

The theorem remains true for the adjusted data if $\max_{1\leq t\leq n} \left| \frac{\widetilde{x}_t^T \widetilde{x}_t}{\mathrm{tr}(S_n)} - 1 \right| \xrightarrow{\mathbb{P}} 0$; see the examples in Lemma 7, Lemma 8 or Corollary 3 in El Karoui (2009). We include more discussions in the supplementary document (He et al., 2020).

We offer more general remarks on the conditions. Condition (i) is a convenient way to control the magnitudes of the associated quadratic forms. With more technical arguments we may replace them by weaker moment conditions on the eigenvalues to allow for the case $p/n \to \infty$, see Chen and Pan, 2012, but we do not pursue the details here. In the supplementary document (He et al., 2020) we check the condition for examples including cross-sectional (in)dependent, moving-average and autoregressive predictors. The predictors can be cross-sectionally dependent in a weak sense according to the definition in Chudik et al. (2011); see also Onatski (2012) for some special factor models. We use the boundedness of $\lambda_{\max}(S_n)$ for the convergence of the moments of $F^{S_n}(x;\beta)$ towards that of $F^{S_n}(x)$; see, e.g., Bai and Silverstein (1998) for some sufficient conditions for bounding $\lambda_{\max}(S_n)$ in models associated with independent data. In practice, one may calculate $\lambda_{\max}(S_n)$ and justify the condition from the independent data case. The last part of the condition controls the de-meaning effect in our estimation. For the aforementioned example models we typically have $\lambda_{\max}\left(\mathbb{E}\left[\bar{x}\bar{x}^T\right]\right) = \frac{1}{n} O\left(\lambda_{\max}(\Sigma)\right) = O(n^{-1})$; we require a much weaker rate in general to allow for complicated time–series dependence structures.

Conditions (ii) is a cross–sectional concentration condition. The first part of the condition is for understanding the probabilistic behaviour, while the data may justify the second part. The condition is trivial for a high dimensional Gaussian vector $x_t \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$ with covariance matrix Σ having bounded eigenvalues in n. More generally, we allow the linear model $x_t = \Sigma^{1/2} f_t$, where f_t has independent entries with a bounded fourth moment. A latent factor model is also possible, such as $x_t = \Phi_1 \eta_t + v_t$, where all the entries of $\eta_t \in \mathbb{R}^k$ and $v_t \in \mathbb{R}^p$ are independent and $\Phi_1 \in \mathbb{R}^{p \times k}$ is a factor loading matrix. The factor dimension k may be small in a sparse model or may diverge to infinity in a dense model. We summarize all these examples into a class of affine models below:

Proposition 1. *The following affine model satisfies condition* (2.9) *in Theorem 2 under Assumption 1, provided condition (i) therein:*

(a) The random vectors $x_t = \Phi f_t$ are identically distributed, not necessarily independent, with covariance matrix $\Sigma = \Phi\Phi^T \in \mathbb{R}^{p \times p}$, where the loading matrix $\Phi = [\phi_1, \ldots, \phi_{p+k}] \in \mathbb{R}^{p \times (p+k)}$ may be asymmetric, the latent components $f_t \in \mathbb{R}^{(p+k)}$ has identity covariance matrix by construction, and the number of extra components $k = k(n) \geq 0$ may be bounded or diverge to infinity at an arbitrary rate.

(b) The unobserved entries of $f_t = (f_{t,1}, \ldots, f_{t,p+k})$ are mutually exogenous such that $\mathbb{E}\left[f_{t,i} \mid f_{t,l}, l \neq i\right] = 0$ and $\mathbb{E}|f_{t,i}|^4$ are bounded by some large constant not depending on n for all i.

Otherwise, condition (2.9) *is trivial if we adjust for the time variation; see Remark 2.*

We allow high level dependence among the latent components beyond the mean. The conditions could be replaced by martingale analogies if there exists a natural ordering for the components. Condition (b) may be relaxed using less order of moments or replaced by more general concentration conditions like in Lemma 7, Lemma 8, or Corollary 3 in El Karoui (2009), but we do not pursue further details here. Assumption 1 and the conditions (i) in Theorem 2 simplify some technical arguments but are not really necessary for the above proposition, and we may relax them with the following spectral conditions: $\|\Sigma\|^2 = o(n^2)$, and $\lambda_{\max}(S_n) \cdot \lambda_{\max}\left(\mathbb{E}\left[\bar{x}\bar{x}^T\right]\right) = o_{\mathbb{P}}(n^2)$.

The above proposition has a much wider scope beyond a factor model. For example, it includes the high-dimensional moving average model in Appendix B (He et al., 2020) given by

$$x_t = \psi w_{t-1} + w_t = \left[\psi T^{1/2}, T^{1/2}\right] \begin{bmatrix} v_{t-1} \\ v_t \end{bmatrix} =: \Phi f_t$$

where $\psi \in (-1, 1)$ is a scalar lagged coefficient, and $w_t = T^{1/2} v_t$ for some covariance matrix T with bounded spectral norm and the entries of random vectors $\{v_t\}$ are i.i.d. with zero mean, unit variance and bounded kurtosis. By similar arguments but with a truncation trick, one may also apply the proposition to the high-dimensional autoregressive model

$$x_t = \phi x_{t-1} + w_t,$$

where $\phi \in (-1, 1)$ is a scalar lagged coefficient and w_t is the same innovations as above. Take an arbitrary diverging integer sequence $K = K(n) \to \infty$, and decompose its infinite-order moving average representation as

$$x_t = \sum_{l=0}^{\infty} \phi^l w_{t-l} = \sum_{l=0}^{K} \phi^l w_{t-l} + \sum_{l=K+1}^{\infty} \phi^l w_{t-l}.$$

Then we can rewrite the leading part

$$\sum_{l=0}^{K} \phi^l w_t = \left[\phi^K \Sigma^{1/2}, \ldots, \phi\Sigma^{1/2}, \Sigma^{1/2}\right] \begin{bmatrix} v_{t-K} \\ \vdots \\ v_t \end{bmatrix} =: \Phi f_t,$$

and verify that reminder term is asymptotically negligible. See He et al. (2020) for details.

Our asymptotic approximation in Theorem 2 is data adaptive, and allows a diverging sequence of the empirical spectral distributions F^{S_n}. If a limiting spectral distribution, say, F does exist, we can easily deduce the following corollary:

Corollary 2 (Limiting spectral distribution)**.** *Under the conditions of Theorem 2, suppose moreover that F^{S_n} tends to some non-degenerate law F with probability one. The asymptotic result remains true by substituting the limit F for F^{S_n} in (2.10).*

Searching for the limiting distribution function F is an active research area in random matrix theory, that traces back to at least Marčenko and Pastur (1967): if the entries $\{x_{t,j} : t = 1, \ldots, n, j = 1, \ldots, p\}$ are i.i.d. random variables with variance τ^2, the limit $F(x)$ exists and has the density function

$$f(x) = \frac{1}{\sqrt{2\pi x c \tau^2}} \sqrt{(b-x)(x-a)} \text{ if } x \in (a, b) \text{ and otherwise zero,} \tag{2.11}$$

and has a point mass $1-1/c$ at the origin if $c > 1$, where $b = \tau^2(1+\sqrt{c})^2$, $a = \tau^2(1-\sqrt{c})^2$ and again $p/n \to c \in (0, \infty)$. For this particular limit, we can verify that $\varpi_n \to \tau^2$ with probability one. That is, the limiting power is stable over the concentration ratio p/n. We refer to Theorem 3.10 and Theorem 4.1 in Bai and Silverstein (2010) for generalization to non-i.i.d. models. When $\{x_t\}$ is a high dimensional autoregressive and moving average (ARMA) time series, or satisfies certain temporal dependence condition, Pan et al. (2014) have established the limiting spectral distribution $F(x)$; see also Zhang (2006). Further studies in linear time series we refer to, e.g., Jin et al. (2009), Liu et al. (2015) and many references therein. For elliptical and high-frequency data, we refer to El Karoui (2009), Zheng and Li (2011) and Xia and Zheng (2018). To extend these results to large dimensional sample correlation matrix (i.e. the sample covariance matrix for standardized data), we refer to El Karoui (2009) again and Gao et al. (2017).

Observe that the variance estimator (2.7) is still consistent under the local alternatives (2.8). Our feasible test therefore achieves the oracle testing power, in an adaptive sense:

Corollary 3 (Feasible power)**.** *Under the conditions of Theorem 2, we have*

$$\mathbb{P}\left(Q_n > \widehat{\sigma}^2\Phi^{-1}(1-\alpha) \mid X_n\right) - \Phi\left(\Phi^{-1}(\alpha) + \frac{h^2}{\sqrt{2}\sigma_n^2}\varpi_n\right) \xrightarrow{\mathbb{P}} 0$$

for any size $\alpha \in (0,1)$.

Note that our feasible test has a non-trivial power even when $p > n$. When the true signal length is not negligible in the variance estimation (2.7), the bias of our restricted estimator may (slightly) reduce the finite–sample power relative to the theoretical limit. As noted above, in such case one may improve the power by using any better variance estimator in some special cases (Fan et al., 2012; Dicker, 2014). We show in simulations that our restricted estimator provides a good performance in sufficiently high dimensions (and large samples), and leave the finite-sample improvements for future research.

Finally, we show that the proposed unweighted test is uniformly most powerful for the free alternatives among a large class of quadratic tests, under regularity conditions. To motivate our competing tests, first consider the weighted quadratic statistic in the case $p < n$ given by

$$\widetilde{Q}_n = \frac{1}{n}e^T\widetilde{X}S_n^{-1}\widetilde{X}^Te.$$

Standardizing the residuals by $\widehat{\sigma}$ gives the F-test when $z_t = 1$; see, e.g., Zhong and Chen (2011) and Wang and Cui (2013) for the power analysis of F-test when $p/n \to c \in (0,1)$ and $\{x_t\}$ is an i.i.d. sequence. When $p > n$, however, the F test has no testing power as $\widetilde{Q}_n/\widehat{\sigma}^2 \equiv \frac{n-(d+1)}{n}$ is degenerate. To compare F statistic with our equally weighted test statistic, one may again remove the diagonal elements in the weighting matrix, standardize the quadratic form, and consider the test statistic: $Q_n = \frac{1}{\sqrt{2}\|A_n\|}e^TA_ne$, where $A_n = \frac{1}{n}\widetilde{X}S_n^{-1}\widetilde{X}^T - \text{diag}\left(\frac{1}{n}\widetilde{X}S_n^{-1}\widetilde{X}^T\right)$ with a slight abuse of the notation. We can verify that Theorem 1 remains true when $p/n \to c \in (0,1)$ and $\lambda_{\min}(S_n)$ is bounded away from zero. Moreover, it is elementary to show that this testing procedure is asymptotically equivalent to the F-test when $z_t = 1$.

Now, in the spirit of Ledoit and Wolf (2012), consider an arbitrary non-negative weighting function δ on $[0,\infty)$ and the associated weighing matrix

$$W_n(\delta) = \frac{1}{n}\widetilde{X}\delta(S_n)\widetilde{X}^T,$$

where $\delta(S_n)$ is a matrix that transforms the eigenvalues of S_n by the function δ but keeps the eigenvectors. Removing the diagonal elements gives $A_n(\delta) = W(\delta) - \text{diag}(W(\delta))$ and standardizing the quadratic form again yields the test statistic:

$$Q_n(\delta) = \frac{1}{\sqrt{2}\,\|A_n(\delta)\|}e^TA_n(\delta)e.$$

It is clear that our equally-weighted test and the first weighted test (F test) both belong to this class asymptotically, but corresponding to different weighting functions $\delta(x) = 1$ and $\delta(x) = x^{-1}\mathbb{1}(x > 0)$ respectively. Furthermore, we can verify that the null distribution remains true by substituting the weighting matrix $A_n = A_n(\delta)$ everywhere.

Our aim in this section is to search for an optimal weighting function δ with the highest testing power among this universe. Using random matrix theory, we can derive the limiting distribution of $Q_n(\delta)$ under our local alternatives (2.8).

Theorem 3. *Suppose the conditions of Theorem 1 and Theorem 2 hold after substituting $A_n(\delta)$ for A_n therein, and moreover that the sample variance of the diagonal elements of $W_n(\delta)$ tends to zero. Under the null (2.4) with $h = 0$ or the local alternatives (2.8),*

$$\frac{Q_n(\delta)}{\sigma_n^2} - \frac{h^2}{\sqrt{2}\sigma_n^2}\varpi_n(\delta) \xrightarrow{d} \mathcal{N}\left(0, 1\right),$$

where

$$\begin{aligned}\varpi_n(\delta) =& \frac{\int x^2\delta(x)dF^{S_n}(x) - \frac{p}{n}\int x dF^{S_n}(x)\cdot\int x\delta(x)dF^{S_n}(x)}{\sqrt{\int x^2\delta^2(x)dF^{S_n}(x) - \frac{p}{n}\left(\int x\delta(x)dF^{S_n}(x)\right)^2}} \\ =& \varpi_n \cdot \operatorname{corr}\left[\underline{\Lambda}_n\delta(\underline{\Lambda}_n), \underline{\Lambda}_n \mid F^{\underline{S}_n}\right],\end{aligned}$$

with ϖ_n and $\underline{\Lambda}_n$ from Theorem 2, if provided that $\operatorname{var}\left[\underline{\Lambda}_n\delta(\underline{\Lambda}_n) \mid F^{\underline{S}_n}\right]$ *is bounded away from 0 almost surely.*

Like in Theorem 2, we require a concentration condition on the diagonal elements of $W_n(\delta)$. We can verify the condition directly from the data, as the weighting matrix $W_n(\delta)$ is observable for any given δ. From a population perspective, we argue that this condition is natural at least in the independent model. Let $\lambda_{\min}^{+}(S_n)$ denote the smallest positive eigenvalue of S_n.

Proposition 2. *The empirical spectral distribution F^{S_n} tends to a limit F solving the equation in Silverstein (1995), and the sample variance of the diagonal elements of $W_n(\delta)$ tends to zero for every function $\delta : [0, \infty) \to \mathbb{R}$ continuous on $[\liminf_n \lambda_{\min}^{+}(S_n), \limsup_n \lambda_{\max}(S_n)]$ when:*

(i) $x_t = \Sigma^{1/2} f_t$, where $\{f_{t,i} : t = 1, \dots, n, i = 1, \dots, p\}$ is a double array of i.i.d. random variables with zero mean, unit variance and $4 + \iota$ moment bounded in n for some $\iota > 0$;

(ii) the covariance matrix Σ is non-negative definite with spectral norm bounded in n, and with empirical spectral distribution $H_n \xrightarrow{w} H$ a proper distribution function.

Hence, Theorem 3 remains true with either F^{S_n} or its limit F, under the conditions of Theorem 1 and Theorem 2 after substituting $A_n(\delta)$ for A_n therein.

Recall that the variance estimator (2.7) is still consistent under the local alternatives (2.8). The asymptotic power of our feasible weighted test follows:

Corollary 4 (Power of weighted tests)**.** *Under the conditions of Theorem 3,*

$$\mathbb{P}\left(Q_n(\delta) > \widehat{\sigma}^2\Phi^{-1}(1-\alpha) \mid X_n\right) - \Phi\left(\Phi^{-1}(\alpha) + \frac{h^2}{\sqrt{2}\sigma_n^2}\varpi_n(\delta)\right) \xrightarrow{\mathbb{P}} 0,$$

for any size $\alpha \in (0, 1)$.

Now, maximizing the asymptotic power of our test (2.6) is equivalent to maximizing the asymptotic departure $\varpi_n(\delta)$ with respect to δ. Note that the (random) correlation coefficient is smaller than 1 almost surely unless $\delta(x)$ is a constant for $x > 0$ (or at least on the spectrum of S_n). In other words, our equally weighted test maximizes the asymptotic test power, for the local free alternatives.

2.3 Towards a more general model and the robust approach

In the previous subsections, we have studied autoregressive model (2.1) involving the nuisance variables $z_t = (1, y_{t-1}, \ldots, y_{t-d})$ with

$$y_{t-i} = B^i y_t = B^i \left(1 - \theta(B)\right)^{-1} (\theta_0 + w_t), \quad w_t = x_t^T \beta + \varepsilon_t,$$

where B denotes the lag operator and $\theta(B) = \theta_1 B + \theta_2 B^2 + \ldots + \theta_d B^d$. By condition (iii) in Assumption 2, we can expand that $(1-\theta(B))^{-1} = \sum_{l=0}^{\infty} \psi(l) B^l$ for some lagged coefficients $\{\psi(l)\}$ with $\psi(0) = 1$, and represent y_{t-i} in an infinite-order moving average form given by

$$y_{t-i} = \alpha + \sum_{l=1}^{\infty} \psi(l+i-2) w_{t-l}, \quad i = 1, \ldots, d,$$

with a common mean $\alpha = \mathbb{E} y_t = \sum_{l=0}^{\infty} \psi(l)\theta_0$. Clearly, the autoregressor y_{t-i} is a special case of the nuisance variables (2.3) with $\alpha_i = \alpha$, $\psi_i(l) = \psi(l+i-2)$, and no current shocks $r_{i,t}$.

From now on, we consider the universal model (2.2) with the general nuisance variables given in (2.3). Throughout we assume that the past effect dies away with absolutely summable moving coefficients, that is, $\sum_{l=1}^{\infty} |\psi_i(l)| < \infty$ for all $i = 1, \ldots, d$. Note that we have added some contemporary information

$$r_t = (r_{t,1}, \ldots, r_{t,d})^T,$$

and for identification purposes only we assume that $\mathbb{E}\left[r_t \mid x_{t-l}^T \beta, \varepsilon_{t-l}, l = 1, 2, \ldots\right] = \mathbf{0}_d$. We extend the sigma algebra $\mathcal{F}_{n,t}$ in Assumption 2 with that generated by $\{r_t\}$ if necessary. As noted in the introduction, it is possible to relax the strong exogeneity condition using more general martingale theory with the cost of more complications but we leave it for future works.

Theorem 4. *Theorem 1 and Corollary 1 remain true. Furthermore, we may relax condition (iii) in Theorem 1 as follows:*

$$\sum_{t=1}^{n-l} A_n(t+l, t) = o_{\mathbb{P}}\left(n^{\frac{1}{2}} \|A_n\|\right). \tag{2.12}$$

Condition (2.12) is only slightly weaker than the condition (iii) in Theorem 1, if the off-diagonal elements cancel each other. It is not necessary for nuisance vectors $\{z_t\}$ containing only contemporary information, that is, $\psi_i(l) = 0$ for all l and i. More generally, we can relax the condition as follows:

$$\sum_{t=1}^{n-l} A_n(t+l, t) \mathbb{1}(\psi_i(l) \neq 0) = o_p(n^{1/2} \|A_n\|), \text{ for all } i = 1, \ldots, d.$$

Nevertheless, according to our simulations, the condition may exclude some special time-series covariates $\{x_t\}$ such as a high-dimensional autoregressive or moving average sequence with common coefficients across all dimensions. For these particular examples and more irregular scenarios, we relax the condition and provide a robust null distribution as follows:

Theorem 5. *Assume the conditions of Theorem 1 except condition (iii) therein. Let*

$$\Psi_i = \sum_{l=1}^{\infty} \psi_i(l) L_n^l = \sum_{l=1}^{n} \psi_i L_n^l, \tag{2.13}$$

and L_n be the $n \times n$ lower shift matrix with ones on the subdiagonal and zeros elsewhere. Under the null hypothesis (2.4),

$$\frac{Q_n}{\sigma_n^2 \sqrt{1-\rho_n^2}} \xrightarrow{d} \mathcal{N}(0,1),$$

with the irregularity coefficient $\rho_n^2 = \|\mu_n\|^2$ depending on

$$\begin{aligned}\mu_n =& \frac{1}{n^{1/2} \left\|\widetilde{A}_n\right\| \sigma_n} \mathbb{E}\left[\Omega^{-\frac{1}{2}} Z^T \widetilde{A}_n \epsilon \mid \widetilde{A}_n\right] \\ =& \frac{\sigma_n}{n^{1/2} \left\|\widetilde{A}_n\right\|} \Omega^{-1/2} \left[0, \operatorname{tr}\left(\Psi_1^T \widetilde{A}_n\right), \ldots, \operatorname{tr}\left(\Psi_d^T \widetilde{A}_n\right)\right]^T,\end{aligned}$$

if provided that ρ_n^2 is bounded away from 1 almost surely and $\widehat{\Omega} \xrightarrow{\mathbb{P}} \Omega = \mathbb{E}\left[z_t z_t^T\right]$.

The null distribution has a smaller asymptotic variance if the coefficient ρ_n^2 is non-negligible. The variance loss is due to the estimation of nuisance parameters. The condition (iii) in Theorem 1 implies that $\rho_n^2 \xrightarrow{\mathbb{P}} 0$ and therefore simplifies the asymptotic limit. We note that $\rho_n^2 < 1$ with probability 1 in each sample, and excluding the boundary is a rather technical requirement. For the extreme event with $\rho_n^2 \to 1$, we may need a different testing procedure as our test statistic $Q_n \xrightarrow{\mathbb{P}} 0$; we leave this for future study.

Now, for both regular and irregular scenarios, one may reject the null (2.4) if

$$Q_n > \widehat{\sigma}^2 \sqrt{1-\widehat{\rho}^2} \Phi^{-1}(1-\alpha), \tag{2.14}$$

where α is the size of the test, and $\widehat{\sigma}^2$ and $\widehat{\rho}^2$ are consistent estimators of in the sense that $\widehat{\sigma}^2/\sigma_n^2 \xrightarrow{\mathbb{P}} 1$ and $(1-\widehat{\rho}_n^2)/(1-\rho_n^2) \xrightarrow{\mathbb{P}} 1$. In what follows, we again use variance estimator $\widehat{\sigma}^2$ given in (2.7). For ρ_n^2 we propose the restricted estimator

$$\widehat{\rho}^2 = \frac{e^T \widetilde{A}_n^T P_Z \widetilde{A}_n e}{e^T \widetilde{A}_n^T \widetilde{A}_n e}, \tag{2.15}$$

using the restricted residual e as for the variance estimator (2.7). Observe that $\widehat{\rho}^2$ is between 0 and 1 as well by construction. To our knowledge, estimating the coefficient ρ_n^2 is a novel statistical topic and it is beyond the scope of this paper to investigate the optimal estimation method. We prove that this restricted estimator is consistent under the null and immediately deduce the following corollary:

Corollary 5. *Under the conditions of Theorem 5 and using the estimators (2.7) and (2.15), our general rejection rule (2.14) is asymptotically correct.*

For completeness, we also generalize the power theory for the local free alternatives. We start again from relaxing the structure of nuisance variables for our usual test (2.6):

Theorem 6. *Theorem 2 and Theorem 3 remain true under the conditions of Theorem 4, if provided that*

$$\lambda_{\max}\left(\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_i^T \mid X\right]\right) = o_{\mathbb{P}}\left(np^{-1/2}\right), \text{ or } \mathbf{v}_i^T\mathbf{v}_i = o_{\mathbb{P}}\left(np^{-1/2}\right), \tag{2.16}$$

where $\mathbf{v}_i = (v_{1,i}, \ldots, v_{n,i})^T$ and $v_{t,i} = \sum_{s\le 0}\psi_i(t-s)\varepsilon_s + \sum_{s\le 0}\psi_i(t-s)x_s^T\beta + r_{t,i}$, for all $i = 1\ldots, d$. Hence, our proposed test is again the most powerful one against the local free alternatives asymptotically.

We may relax the small-o order in (2.16) to be big-o order by carefully checking the proofs, but we keep the stronger one for our next theorem. Note that $\mathbf{v}_i$ depends only on the initial values $\{\varepsilon_s, x_s^T\beta : s \le 0\}$ and the contemporary shocks $\{r_{t,i} : t = 1, \ldots, n\}$. For the autoregressive process (2.1), we can show that, for every i, the entries of $\mathbf{v}_i$ satisfy the homogeneous linear difference equation:

$$v_{t,i} = \theta_1 v_{t-1,i} + \ldots + \theta_d v_{t-d,i}, \; t \ge d+1. \tag{2.17}$$

and therefore $\{v_{t,i}\}$ is square summable in probability, that is, $\mathbf{v}_i^T\mathbf{v}_i = \sum_{t=1}^n v_{t,i}^2 = O_{\mathbb{P}}(1)$. Then the spectral condition (2.16) is trivial when $p = o(n^2)$.

Finally, we develop the power theory for our robust test (2.14). Let $\xi_n = \beta/\|\beta\|$ denote the direction of β under the alternatives or an arbitrary unit vector when $\beta = \mathbf{0}_p$.

Theorem 7. *Suppose all the conditions of Theorem 6 hold under both the null and alternatives, and we relax the condition (iii) in Theorem 1. In addition, suppose the direction ξ_n is also free in the sense that*

$$\xi_n^T\left(\frac{1}{n}\widetilde{X}^T\Psi_i^T\widetilde{X}\right)\xi_n - \frac{1}{p}\operatorname{tr}\left(\frac{1}{n}\widetilde{X}^T\Psi_i^T\widetilde{X}\right) \xrightarrow{\mathbb{P}} 0, \; i = 1, \ldots, d. \tag{2.18}$$

Under the null (2.4) with $h = 0$ or under the local alternatives (2.8),

$$\frac{Q_n}{\sigma_n^2\sqrt{1-\rho_n^2}} - \frac{h^2}{\sqrt{2}\sigma_n^2}\sqrt{1-\rho_n^2}\varpi_n \xrightarrow{d} \mathcal{N}(0,1).$$

Relaxing the additional freeness condition (2.18) leads to a more technical but not very useful limit that depends on the unknown direction of β. Hence, for similar reasons in Remark 1, we postpone the relaxations to the supplementary document (He et al., 2020).

By showing the consistency of $\widehat{\rho}^2$ under the local alternatives, we can deduce the following corollary:

Corollary 6. *Under the conditions of Theorem 7 and using the estimators (2.7) of σ^2 and the estimator (2.15) of ρ^2,*

$$\mathbb{P}\left(Q_n > \widehat{\sigma}^2\sqrt{1-\widehat{\rho}^2}\Phi^{-1}(1-\alpha)|X_n\right) - \Phi\left(\Phi^{-1}(\alpha) + \frac{h^2}{\sqrt{2}\sigma_n^2}\sqrt{1-\rho_n^2}\varpi_n\right) \xrightarrow{\mathbb{P}} 0$$

for any size $\alpha \in (0,1)$.

Remark 3. We may generalize the results for a general weight function δ. Substituting A_n by $A_n(\delta)$ in Theorem 5, we can extend the definition of $\rho_n^2 = \rho_n^2(\delta)$ and $\mu_n = \mu_n(\delta)$. To avoid confusions, from now on we denote ρ_n^2 and μ_n as the values associated with $\delta(x) \equiv 1$. Under the conditions of Theorem 6 where the conditions also hold for $A_n(\delta)$ replacing A_n,

$$\frac{Q_n(\delta)}{\sigma_n^2\sqrt{1-\rho_n^2(\delta)}} - \frac{h^2}{\sqrt{2}\sigma_n^2}\frac{\varpi_n(\delta) - \rho_n(\delta, 1)\cdot\varpi_n}{\sqrt{1-\rho_n^2(\delta)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where ϖ_n and $\varpi_n(\delta)$ are given in Theorem 2 and 3 respectively, and $\rho_n(\delta, 1) := \mu_n^T(\delta)\mu_n$ lies between -1 and 1 almost surely. Recall that $\varpi_n(\delta) \le \varpi_n$, and therefore the departure

$$\frac{\varpi_n(\delta) - \rho_n(\delta, 1)\cdot\varpi_n}{\sqrt{1-\rho_n^2(\delta)}} \le \frac{1-\rho_n(\delta,1)}{\sqrt{1-\rho_n^2(\delta)}}\varpi_n,$$

where the upper bound approaches $\sqrt{1-\rho_n^2}\varpi_n$ when $\mu_n(\delta) - \mu_n \to \mathbf{0}_d$. In other words, our equally-weighted test (with $\delta(x) \equiv 1$) maximizes the asymptotic power locally against the competitors with a similar nuisance effect $\mu_n(\delta) = \mu_n + o_{\mathbb{P}}(1)$. Optimizing the power beyond such neighborhood may be numerically feasible with the estimator (2.15) of $\rho_n^2(\delta)$ and the estimator of $\rho_n(\delta, 1)$ given by

$$\widehat{\rho}_n(\delta, 1) = \frac{e^T \widetilde{A}_n^T(\delta) P_Z \widetilde{A}_n e}{\sqrt{e^T \widetilde{A}_n^T(\delta)\widetilde{A}_n(\delta) e}\sqrt{e^T \widetilde{A}_n^T \widetilde{A}_n e}},$$

where $\widetilde{A}_n(\delta)$ is the lower triangular part of $A_n(\delta)$. It requires much more works to establish the uniform convergence of the asymptotic approximations over a large class of functions δ, that are certainly interesting for future studies.

3 Simulation

In this section, we study the empirical size and the power performance of the proposed tests using a Monte Carlo experiment. Throughout we choose an asymptotic size of $\alpha = 5\%$, and calculate the finite-sample size and power over 5000 replications using the default seed in MATLAB 2019b. Without loss of generality, we first generate the exogenous predictors and errors with zero means, and then generate the target variable using autoregressive model (2.1) with intercept $\theta_0 = 0$. However, this is unknown to the statistician who always demeans the predictors in each sample and estimate the intercept. We fix the order of autoregressive $d = 3$, and use the autoregressive coefficients $(\theta_1, \theta_2, \theta_3) = (0.30, 0.08, 0.11)$ calibrated from our empirical application. We simulate independent innovations η_t from the standardized student t-distribution with five degrees of freedom, that are independent of the regressors. Then we generate the regression errors $\varepsilon_t = \sigma_n \eta_t$ with an adaptive variance $\sigma_n^2 = \varpi_n/\sqrt{2}$ in each sample, and therefore the asymptotic power only depends on the length of the coefficient β and, in general, the irregularity coefficient ρ_n^2; see Corollary 3 and 6 in Section 2.

In each replication, we consider exogenous variables from four data generating processes:

Table 1: Size and power (%) of the tests against uniform stochastic coefficient (i) at level $\alpha = 5\%$ with $p/n = \frac{1}{4}, \frac{1}{2}, 1, 2, 4$ and $\sqrt{p}/n = 0.05$. The columns are for: (i) the feasible test using $\widehat{\sigma}_n^2$ and assuming $\rho_n^2 = 0$, (i^o) the oracle test using the true variance σ_n^2 and assuming $\rho_n^2 = 0$, (i*) the robust test using $\widehat{\sigma}_n^2$ and $\widehat{\rho}_n^2$.

	IID			CSD			MA1			AR1		
p/n	(i)	(i^o)	(i*)	(i)	(i^o)	(i*)	(i)	(i^o)	(i*)	(i)	(i^o)	(i*)
$H_0 : \|\beta\|^2 = 0$												
25/100	5.4	5.2	5.7	5.6	5.6	6.0	5.7	6.1	6.1	6.2	6.1	6.6
100/200	5.5	5.2	5.7	5.8	6.2	5.9	5.7	5.7	6.2	5.8	5.5	6.4
400/400	5.4	5.4	5.5	5.9	6.1	6.0	4.8	5.1	5.7	4.4	4.5	5.4
1600/800	5.6	5.6	5.6	6.1	5.9	6.1	4.4	4.5	5.8	4.1	4.3	6.3
6400/1600	5.4	5.2	5.4	5.8	5.8	5.8	3.0	3.1	5.2	2.7	2.8	5.7
$H_a^1 : \|\beta\|^2 = 1 \times \frac{\sqrt{p}}{n}$												
25/100	23.2	25.5	23.7	23.4	25.0	24.1	23.9	25.3	25.2	24.3	25.2	25.8
100/200	24.9	27.1	25.2	24.9	26.5	25.2	24.7	26.2	26.1	25.8	26.8	27.5
400/400	25.6	28.5	25.8	26.8	28.0	26.9	23.2	24.9	25.4	21.3	22.5	23.8
1600/800	26.1	28.7	26.1	26.3	28.8	26.4	19.2	21.5	23.1	19.1	21.0	24.3
6400/1600	24.2	27.7	24.3	25.5	28.0	25.6	13.7	15.5	20.7	13.2	14.7	21.6
$H_a^2 : \|\beta\|^2 = 2 \times \frac{\sqrt{p}}{n}$												
25/100	45.1	50.0	45.8	44.1	46.5	44.8	42.5	45.6	43.8	43.3	45.6	45.4
100/200	50.5	56.0	51.3	49.7	53.4	50.1	47.7	50.8	49.3	48.7	53.0	50.8
400/400	53.8	60.5	54.0	54.1	58.3	54.4	49.2	54.0	51.9	47.0	51.8	50.8
1600/800	54.9	61.9	55.0	56.8	62.1	56.9	44.1	49.9	49.7	42.5	47.9	49.6
6400/1600	55.2	62.9	55.3	57.7	63.5	57.8	35.8	42.0	46.7	34.0	39.5	46.4
$H_a^3 : \|\beta\|^2 = 5 \times \frac{\sqrt{p}}{n}$												
25/100	85.1	90.1	85.8	78.6	82.7	79.3	77.0	81.0	78.1	76.4	80.4	77.7
100/200	92.3	95.8	92.4	89.5	92.4	89.6	87.3	90.9	88.0	87.8	91.1	88.7
400/400	95.2	98.0	95.2	94.4	96.8	94.5	92.9	95.9	93.6	91.4	94.7	92.6
1600/800	96.9	98.7	96.9	96.9	98.5	96.9	93.0	96.8	95.1	93.4	96.5	95.3
6400/1600	97.4	98.9	97.4	98.2	99.2	98.2	91.0	95.6	94.8	89.0	94.3	94.0

(1) $x_t = v_t$, where $\{v_t\}$ have i.i.d. standardized t_5 entries independent over time t;

(2) $x_t = T_p^{1/2} v_t$ with the error covariance matrix $T_p \in \mathbb{R}^{p \times p}$ equaling to the Toeplitz matrix with i, j-th elements $\rho^{|i-j|}$ and $\rho = 0.5$;

(3) $x_t = 0.3 T_p^{1/2} v_{t-1} + T_p^{1/2} v_t$, a high dimensional moving average model;

(4) $x_t = 0.3 x_{t-1} + T_p^{1/2} v_t$, a high dimensional autoregressive model.

We generate the direction of the regression coefficients in the following ways respectively:

(i) uniformly over the $\mathbb{R}^p$ unit sphere;

(ii) the average direction of the eigenvectors of the population correlation matrix T_p.

The first case is a stochastic coefficient model that is always free. The second one is a deterministic coefficient model, that is free at least for the independent models (1) and (2), see Bai et al. (2007) and Pan (2014), and shows similar performance as that of a free model in the time-series models (3) and (4). In the supplementary document (He et al., 2020), we have also implemented the adaptive regression directions as in the first set of simulations in Goeman et al. (2011) which shows that our general asymptotic approximation applies for non-free alternatives as well; see Remark 1.

We compare the size and power for the local departure level

$$h^2 = \frac{n}{\sqrt{p}} \|\beta\|^2 = 0, 1, 2, 5,$$

corresponding to the coefficient vector

$$\beta = \|\beta\| \, \xi = h \left(\frac{\sqrt{p}}{n} \right)^{\frac{1}{2}} \xi,$$

where ξ denotes the direction we generated above in case (i) or (ii). We vary the concentration ratios $p/n = \frac{1}{4}, \frac{1}{2}, 1, 2, 4$ to obtain a wide range of (p, n), while fixing the order of local alternatives $\sqrt{p}/n = 0.05$ or $\sqrt{p}/n = 0.1$.

Table 1 and Table 2 report the results for different directions, with $\sqrt{p}/n = 0.05$. We report the size and power for three different tests: the feasible test using the estimated variance $\widehat{\sigma}_n^2$ for regular scenarios (i.e. assuming $\rho_n^2 = 0$), the oracle test using the true variance σ_n^2 for regular scenarios (i.e. assuming $\rho_n^2 = 0$), and the robust test using the estimated variance $\widehat{\sigma}_n^2$ and the estimated irregularity coefficient $\widehat{\rho}_n^2$ for both regular and irregular scenarios. Overall, we observe that the feasible test performs similarly to the oracle test. The size and power of the feasible tests and oracle ones are close to the asymptotic level for independent models, but become smaller for time-series predictors as suggested by our theory especially for larger concentration ratio p/n. On the other hand, the robust test maintains a stable size and power across all scenarios. While by construction the robust test always has a larger size and power than the feasible test, the differences are small for independent predictors according to our theory.

Table 2: Size and power (%) of the tests against deterministic coefficient (ii) at level $\alpha = 5\%$ with $p/n = \frac{1}{4}, \frac{1}{2}, 1, 2, 4$ and $\sqrt{p}/n = 0.05$. The columns are for: (ii) the feasible test using $\widehat{\sigma}_n^2$ and assuming $\rho_n^2 = 0$, (ii^o) the oracle test using the true variance σ_n^2 and assuming $\rho_n^2 = 0$, (ii*) the robust test using $\widehat{\sigma}_n^2$ and $\widehat{\rho}_n^2$.

	IID			CSD			MA1			AR1		
p/n	(ii)	(ii^o)	(ii*)	(ii)	(ii^o)	(ii*)	(ii)	(ii^o)	(ii*)	(ii)	(ii^o)	(ii*)
$H_0 : \|\beta\|^2 = 0$												
25/100	5.6	5.1	5.8	5.9	5.7	6.2	5.8	6.1	6.3	5.9	5.9	6.3
100/200	5.6	5.7	5.7	6.2	6.3	6.3	6.0	6.0	6.5	5.3	5.1	5.8
400/400	5.2	5.4	5.3	6.0	5.9	6.1	4.7	4.7	5.3	5.0	4.9	6.1
1600/800	5.7	5.5	5.7	6.0	6.0	6.0	4.5	4.7	6.0	3.8	4.0	5.7
6400/1600	4.9	4.9	4.9	5.4	5.4	5.4	3.0	2.9	5.3	2.5	2.4	5.3
$H_a^1 : \|\beta\|^2 = 1 \times \frac{\sqrt{p}}{n}$												
25/100	23.5	24.9	24.2	24.9	25.8	25.5	24.4	25.6	25.6	24.1	24.9	25.4
100/200	24.9	27.3	25.2	26.8	28.2	27.3	25.0	26.5	26.5	23.7	24.8	25.5
400/400	25.3	28.6	25.5	25.8	27.5	26.0	22.4	24.4	24.7	22.3	24.4	24.8
1600/800	25.6	28.5	25.7	24.7	27.0	24.8	19.2	21.3	23.6	18.1	20.2	22.6
6400/1600	24.3	27.8	24.5	26.3	28.6	26.4	15.0	16.8	21.9	13.4	15.2	20.8
$H_a^2 : \|\beta\|^2 = 2 \times \frac{\sqrt{p}}{n}$												
25/100	45.1	48.8	45.8	45.9	47.9	46.5	44.2	47.3	45.6	43.9	46.6	45.4
100/200	50.0	55.6	50.4	50.6	53.8	51.1	48.4	52.6	50.2	46.5	50.5	48.7
400/400	52.7	59.1	52.9	54.0	58.8	54.3	47.4	53.0	50.7	47.5	52.2	50.8
1600/800	54.0	60.8	54.1	54.5	59.8	54.6	44.5	49.8	49.8	42.0	48.3	49.7
6400/1600	54.4	61.0	54.6	57.0	62.8	57.1	37.5	43.4	47.8	33.0	39.1	45.4
$H_a^3 : \|\beta\|^2 = 5 \times \frac{\sqrt{p}}{n}$												
25/100	83.0	88.5	83.6	82.7	86.3	83.2	80.5	84.9	81.3	80.7	85.2	81.9
100/200	92.0	95.6	92.2	90.5	93.8	90.7	88.9	92.5	89.8	87.6	91.7	88.6
400/400	95.0	97.9	95.1	95.0	97.1	95.0	92.6	95.7	93.6	92.6	95.6	93.7
1600/800	96.7	98.7	96.7	97.1	98.5	97.1	93.2	95.8	94.5	92.6	95.8	94.5
6400/1600	97.9	99.3	97.9	98.1	99.1	98.1	91.6	95.5	94.9	88.7	94.2	94.1

In the supplementary document (He et al., 2020) we repeat the analysis for a larger order of $\sqrt{p}/n = 0.1$. The conclusions are qualitatively the same. The power difference between the feasible and oracle tests becomes slightly larger, as the error variance estimator contains a larger finite-sample upward bias under the alternatives.

4 Application

Our empirical application is to test whether the exogenous macroeconomic variables at the 'FRED-MD' database:

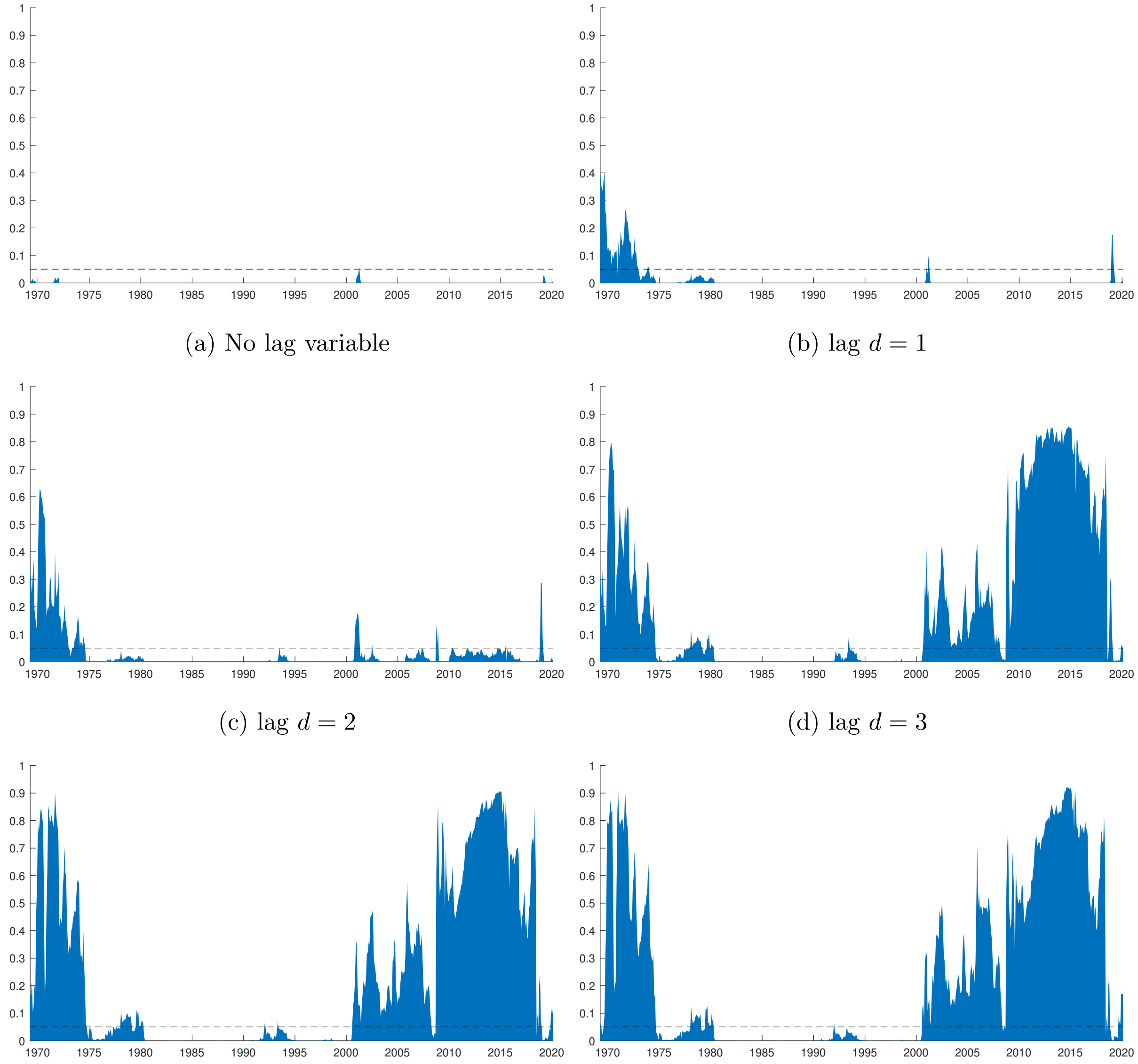
https://research.stlouisfed.org/econ/mccracken/fred-databases/

are overall significant for forecasting the monthly growth rate, on a percentage scale, of US industrial production index, an important indicator of macroeconomic activity. Our response variable is $y_t = \log(\mathrm{IP}_t/\mathrm{IP}_{t-1}) \times 100$, where IP_t denotes the US industrial production index for the month t. The database has similar predictive content as that in Stock and Watson (2002), and it is regularly updated through the Federal Reserve Economic Data (FRED).

Our data set includes monthly observations of the industrial production index (INDPRO) and 127 other predictors from January, 1959 to February, 2020. We transform the raw datasets into stationary forms and remove the data outliers using the MATLAB codes provided on the above website; see also McCracken and Ng (2016) for more details of the method. Our tests use rolling windows of sample size $n = 120$ months equaling to a time span of ten years. In each window we drop the variables with missing values, leaving approximately $p \approx 120$ one-month lagged standardized predictors besides lagged response variables. Note that the standard F test is (almost) degenerate here, because the number of exogenous variables is larger than (or very close to) the sample size. In the supplementary document (He et al., 2020) we show that the results are very similar after adjusting for time variations (see Remark 2). Hence, we only report the results without time variation adjustment in this section.

We compare the rolling window p-values for different autoregressive order d between 0 and 5 in Figure 1. The dashed lines indicate our benchmark significance level $\alpha = 5\%$. We only report the p-values for the robust tests here, and the findings are similar for the standard tests as shown in the supplementary document (He et al., 2020). Clearly, ignoring the impact of autoregressors shows almost all rejections and likely overoptimistic outcomes for predictability. By including more autoregressors, the p-values stabilize for $d \geq 3$ and time-varying patterns emerge. Overall, the coefficients are jointly significant most of the time between year 1975 and year 2000, but more recently during the first half year of 2019.

Figure 1: Ten years ($n = 120$) rolling windows monthly robust p values between March, 1969 and February, 2020 for different number of lags $d = 0, 1, 2, 3, 4, 5$.



(a) No lag variable

(b) lag $d = 1$

(c) lag $d = 2$

(d) lag $d = 3$

(e) lag $d = 4$

(f) lag $d = 5$

5 Proofs

We first sketch the mathematical proofs of the main theorems, and then prove all the corollaries. The technical and detailed proofs of the lemmas and Propositions 1 and 2 are available in the supplement (He et al., 2020). It is not hard to check throughout that our proofs do not depend on the error variance σ_n^2, if one re-parameterize β as β/σ_n, θ as θ/σ_n, and divide the variables y_t and ε_t by σ_n. To simplify presentations, we assume throughout that $\sigma_n^2 = 1$ and use ε_t everywhere instead of η_t without loss of generality. Throughout this section we denote $P_Z = Z(Z^TZ)^{-1}Z^T$ as the projection matrix on the column space of Z. Enlarging the probability space, if necessary, we define all random variables in a common probability space with probability measure $\mathbb{P}$ for presentation convenience.

5.1 Proofs of Theorem 1

Throughout this subsection we assume the conditions of Theorem 1. Under the null hypothesis, we can decompose that

$$\begin{aligned}
Q_n =& \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T(I-P_Z)\widetilde{A}_n(I-P_Z)\epsilon\\
=& \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T A_n\epsilon - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z\widetilde{A}_n\epsilon - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z\widetilde{A}_n^T\epsilon + \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z A_n P_Z\epsilon\\
=:& \widetilde{Q}_n + T_1 + T_2 + T_3,
\end{aligned}$$

where $\widetilde{Q}_n$ has a martingale form:

$$\frac{\sqrt{2}}{\|A_n\|}\sum_{t=1}^n \varepsilon_t\left(\sum_{s=1}^{t-1}\varepsilon_s\frac{1}{n}\widetilde{x}_s^T\widetilde{x}_t\right) =: \sum_{t=1}^n \Delta_t$$

and Δ_t is a martingale difference array such that $\mathbb{E}(\Delta_t|\mathcal{F}_{n,t-1}) = 0$. We shall show that $\widetilde{Q}_n \xrightarrow{d} \mathcal{N}(0,1)$ using martingale central limit theorem, and show that $T_1, T_2, T_3 \xrightarrow{\mathbb{P}} 0$.

We need some lemmas for our proof in the end. We begin with some fundamental inequalities and their useful implications here. The first lemma is an elementary result combining Markov inequality and the law of iterated expectations. We use it frequently to control the asymptotic bounds of perturbation terms.

Lemma 1 (Markov inequality). *For an arbitrary sequence of measurable statistics θ_n and a sequence of sub-sigma-algebra $\mathcal{F}_n$, the Markov inequality implies that $|\theta_n| = O_{\mathbb{P}}\left(\mathbb{E}\left[|\theta_n| \mid \mathcal{F}_n\right]\right)$. Note that $\mathbb{E}\left[|\theta_n| \mid \mathcal{F}_n\right]$ is a $\mathcal{F}_n$-measurable random variable in general.*

The second is a concentration inequality for quadratic forms. The results are well-known for quadratic forms in i.i.d. variables (see, e.g., Lemma B.26 in Bai and Silverstein, 2010), and the proof for martingale difference array is very similar. We provide the proof in the supplementary document for completeness.

Lemma 2 (Concentration inequality for martingale quadratic forms)**.** *Let $\{\varepsilon_t, \mathcal{F}_{n,t} : t = 1, \ldots, n\}$ be a martingale difference array with common conditional variance $\mathbb{E}\left[\varepsilon_t^2 \mid \mathcal{F}_{n,t-1}\right] = 1$, and A be any $n \times n$ real matrix measurable by $\mathcal{F}_{n,0}$. If further given that $\mathbb{E}\left[|\varepsilon_t^2 - 1|^{1+\iota} \mid \mathcal{F}_{n,0}\right] \leq \kappa_n$ a.s. for some $\iota \in [0, 1]$, $\forall t = 1, \ldots, n$,*

$$\mathbb{E}\left[\left|\epsilon^T A\epsilon - \operatorname{tr}(A)\right|^{1+\iota} \mid \mathcal{F}_{n,0}\right] \leq M\left(\kappa_n \sum_{t=1}^{n} |A(t,t)|^{1+\iota} + \|A\|^{1+\iota}\right)$$

almost surely for some absolute constant M depending only on ι, where $A(t,t)$ is the t-th diagonal element of A. Hence,

$$\epsilon^T A\epsilon - \operatorname{tr}(A) = O_{\mathbb{P}}\left(\kappa_n^{\frac{1}{1+\iota}}\left(\sum_{t=1}^{n} |A(t,t)|\right)^{\frac{1}{1+\iota}} \max_{1\leq t\leq n} |A(t,t)|^{\frac{\iota}{1+\iota}} + \|A\|\right).$$

The next inequality is essentially Lemma S.3 in Lam (2016). We add a necessary condition (which is missing therein), and provide the detailed proof in the supplementary document.

Lemma 3 (A trace inequality)**.** *For an arbitrary symmetric $p \times p$ matrix A and a non-negative definite $p \times p$ matrix B,*

$$|\operatorname{tr}(AB)| \leq \|A\|_{sp} \operatorname{tr}(B).$$

The next two lemmas verify the conditions for our martingale central limit theorem, Corollary 3.1 in Hall and Heyde (1980).

Lemma 4. *The asymptotic negligibility condition is satisfied is such a way that*

$$\max_{1\leq t\leq n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}\right] \xrightarrow{\mathbb{P}} 0.$$

Lemma 5. *The conditional variance for the martingale converges to 1, that is,*

$$\sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}\right] = \frac{1}{\left\|\widetilde{A}_n\right\|^2} \sum_{t=1}^{n}\left(\sum_{s=1}^{t-1} \varepsilon_s \frac{1}{n} \widetilde{x}_s^T \widetilde{x}_t\right)^2 \xrightarrow{\mathbb{P}} 1.$$

Our last two lemmas control the effect of nuisance variables on the residuals.

Lemma 6. $\epsilon^T P_Z \epsilon = O_{\mathbb{P}}(d/\lambda_{\min}(\widehat{\Omega}))$.

Lemma 7. $Z^T \widetilde{A}_n \epsilon = o_{\mathbb{P}}\left(\sqrt{n}\left\|\widetilde{A}_n\right\|\right)$

Now we can prove the null distribution using the above lemmas.

Proof of Theorem 1. We first apply Corollary 3.1 in Hall and Heyde (1980) to show that $\widetilde{Q}_n \xrightarrow{d} \mathcal{N}(0, 1)$. As the conditional variance converges in Lemma 5, it remains to verify the conditional Lindeberg condition, that is, $\sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mathbb{1}(|\Delta_t| > \delta) \mid \mathcal{F}_{n,t-1}\right] \xrightarrow{\mathbb{P}} 0$ for every small constant $\delta > 0$ (with a slight abuse of notation). Let $\delta > 0$, and we have

$$\begin{aligned}\sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mathbb{1}(|\Delta_t| > \delta) \mid \mathcal{F}_{n,t-1}\right] =& \sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}, \Delta_t^2 > \delta^2\right] \mathbb{P}\left(\Delta_t^2 > \delta^2 \mid \mathcal{F}_{n,t-1}\right) \\ \leq& \sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}, \Delta_t^2 > \delta^2\right] \cdot \max_{1 \leq t \leq n} \mathbb{P}\left(\Delta_t^2 > \delta^2 \mid \mathcal{F}_{n,t-1}\right).\end{aligned}$$

Using the law of iterated expectations,

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}, \Delta_t^2 > \delta^2\right] | \mathcal{F}_{n,0}\right] = \sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,0}\right] = 1,$$

and therefore $\sum_{t=1}^{n} \mathbb{E}\left[\Delta_t^2 | \mathcal{F}_{n,t-1}, \Delta_t^2 > \delta^2\right] = O_{\mathbb{P}}(1)$ by Lemma 1. On the other hand, using Markov inequality and Lemma 4,

$$\max_{1 \leq t \leq n} P\left(\Delta_t^2 > \delta^2 \mid \mathcal{F}_{n,t-1}\right) \leq \frac{1}{\delta^2} \max_{1 \leq t \leq n} \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}\right] \xrightarrow{\mathbb{P}} 0.$$

This completes the proof of $\widetilde{Q}_n \xrightarrow{d} \mathcal{N}(0,1)$.

In the following, we shall show that $T_1, T_2, T_3 \xrightarrow{\mathbb{P}} 0$. We start from the difficult term T_2. By Cauchy–Schwarz inequality,

$$\begin{aligned}T_2^2 = \frac{1}{\left\|\widetilde{A}_n\right\|^2}\left(\epsilon^T P_Z \widetilde{A}_n^T \epsilon\right)^2 \leq& \epsilon^T P_Z \epsilon \cdot \frac{1}{\|A_n\|^2} \epsilon^T \widetilde{A}_n P_Z \widetilde{A}_n^T \epsilon \\ =& \left(\lambda_{\min}(\widehat{\Omega})\right)^{-1} O_{\mathbb{P}}(d) \cdot \frac{1}{\|A_n\|^2} \epsilon^T \widetilde{A}_n P_Z \widetilde{A}_n^T \epsilon, \qquad (5.1)\end{aligned}$$

where we apply Lemma 6 in the last step. Furthermore, the last quadratic form

$$\epsilon^T \widetilde{A}_n P_Z \widetilde{A}_n^T \epsilon = \frac{1}{n} \epsilon^T \widetilde{A}_n Z \widehat{\Omega}^{-1} Z^T \widetilde{A}_n^T \epsilon \leq \left(\lambda_{\min}(\widehat{\Omega})\right)^{-1} \frac{1}{n} \epsilon^T \widetilde{A}_n Z Z^T \widetilde{A}_n^T \epsilon.$$

Using the martingale property, a direct calculation yields that

$$\begin{aligned}&\mathbb{E}\left[\epsilon^T \widetilde{A}_n Z Z^T \widetilde{A}_n^T \epsilon \mid X\right] = \sum_{i=1}^{d} \mathbb{E}\left[\left(\sum_{1 \leq s < t \leq n} z_{s,i} \varepsilon_t \frac{1}{n} \widetilde{x}_t^T \widetilde{x}_s\right)^2 \mid X\right] \\ =& \sum_{i=1}^{d} \mathbb{E}\left[\left(\sum_{t=1}^{n}\left(\sum_{s=1}^{t-1} z_{s,i} \frac{1}{n} \widetilde{x}_t^T \widetilde{x}_s\right)^2\right)^2 \mid X\right] \\ =& \mathbb{E}\left[\operatorname{tr}\left(\widetilde{A}_n Z Z^T \widetilde{A}_n^T\right) \mid X\right] = \operatorname{tr}\left(\widetilde{A}_n^T \widetilde{A}_n \mathbb{E}\left[Z Z^T \mid X\right]\right) \leq \lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right) \operatorname{tr}\left(\mathbb{E}\left[Z Z^T \mid X\right]\right),\end{aligned}$$

where $X = [x_1, x_2, \cdots, x_n]^\top$, and the last step follows from the trace inequality in Lemma 3. Exchanging between trace and expectation operations,

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbb{E}\left[Z Z^T \mid X\right]\right)\right] = \operatorname{tr}(\mathbb{E}\left[Z Z^T\right]) = nd,$$

and therefore $\operatorname{tr}\left(\mathbb{E}\left[ZZ^T \mid X\right]\right) = O_p(nd)$ by Lemma 1 as ZZ^T is non-negative definite. Hence, by Lemma 1 again,

$$\epsilon^T \widetilde{A}_n ZZ^T \widetilde{A}_n^T \epsilon = O_{\mathbb{P}}\left(\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right) nd\right) \tag{5.2}$$

Collecting all bounds and substituting in (5.1) yields that

$$T_2^2 = \left(\lambda_{\min}(\widehat{\Omega})\right)^{-2} O_{\mathbb{P}}(d^2) \frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\|A_n\|^2} \xrightarrow{\mathbb{P}} 0.$$

Similarly, we can show that

$$T_1^2 = \left(\lambda_{\min}(\widehat{\Omega})\right)^{-2} O_{\mathbb{P}}(d) \cdot \frac{1}{n\|A_n\|^2} \epsilon^T \widetilde{A}_n^T ZZ^T \widetilde{A}_n \epsilon \xrightarrow{\mathbb{P}} 0,$$

where the last step follows from Lemma 7. Finally, by Cauchy–Schwarz inequality

$$\begin{aligned} |T_3| = \frac{\sqrt{2}}{\|A_n\|} \left|\epsilon^T P_Z \widetilde{A}_n P_Z \epsilon\right| \leq & \frac{\sqrt{2}}{\|A_n\|} \sqrt{\epsilon^T P_Z \epsilon \cdot \epsilon^T P_Z \widetilde{A}_n^T \widetilde{A}_n P_Z \epsilon} \\ =& O_p\left(\sqrt{\frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\|A_n\|^2}} \epsilon^T P_Z \epsilon\right) \xrightarrow{\mathbb{P}} 0, \end{aligned}$$

where we have applied Lemma 6 in the last step. □

5.2 Proofs of Theorems 2 and 3

We only prove Theorem 3, as the proof of Theorem 2 is easier by substituting function δ everywhere by a constant function on $[0, \infty)$. Let $\mathbb{S}_n = \frac{1}{n}X^T X$ and $\underline{\mathbb{S}}_n = \frac{1}{n}XX^T$, using the raw design matrix $X = [x_1, \ldots, x_n]^T$. For any matrix A, we denote its (i,j)-th element by $A(i,j)$. Recall that $\mathbf{1}_n$ denotes n-dimensional all-ones vector. Throughout we assume all the conditions of Theorem 3, which imply that $\|A_n\|_{sp} \leq \|S_n\|_{sp} = O_{\mathbb{P}}(1)$ and $\|A_n\| / \sqrt{p} = \varpi_n + o_{\mathbb{P}}(1)$ is bounded away from 0 with arbitrarily high probability.

We only need to prove for the case where $\beta \neq \mathbf{0}$; the case for $\beta = \mathbf{0}$ is shown in Theorem 1. For presentation convenience, we write A_n in short of $A_n(\delta)$, ϖ_n in short of $\varpi_n(\delta)$, and W_n in short of $W_n(\delta)$. Expand that

$$\begin{aligned} Q_n =& \frac{1}{\sqrt{2}\|A_n\|} \left\{\epsilon^T(I - P_Z) + \beta^T X^T (I - P_Z)\right\} A_n \left\{(I - P_Z)\epsilon + (I - P_Z)X\beta\right\} \\ =& \frac{1}{\sqrt{2}\|A_n\|} \epsilon^T (I - P_Z) A_n (I - P_Z)\epsilon + \frac{\sqrt{2}}{\|A_n\|} \epsilon^T (I - P_Z) A_n (I - P_Z) X\beta \\ &+ \frac{1}{\sqrt{2}\|A_n\|} \beta^T X^T (I - P_Z) A_n (I - P_Z) X\beta =: T_1 + T_2 + T_3. \end{aligned} \tag{5.3}$$

Following the proof of Theorem 1, we can show $T_1 \xrightarrow{\mathbb{P}} \mathcal{N}(0,1)$ by noting that Lemma 7 remains true under the local alternatives:

Lemma 8. $\frac{1}{\sqrt{n}\|\widetilde{A}_n\|} Z^T \widetilde{A}_n \epsilon = o_{\mathbb{P}}\left(1 + \|\beta\|^2\right)$. *Hence, Lemma 7 remains true under the alternatives.*

For the theorem, it remains to show that:

(*) $T_3 - \frac{h^2}{\sqrt{2}} \varpi_n \xrightarrow{\mathbb{P}} 0$.

(**) $T_2 \xrightarrow{\mathbb{P}} 0$.

We shall prove these statements one-by-one. We need the following lemma.

Lemma 9. $\beta^T \widetilde{X}^T P_Z \widetilde{X} \beta = \lambda_{\min}^{-1}(\widehat{\Omega}) \cdot O_p\left(\|\beta\|^2 + n\|\beta\|^4\right)$.

Proof of Statement (*). Substituting $W_n = \frac{1}{n}\widetilde{X}\delta(S_n)\widetilde{X}^T$ and noting that the diagonal elements W_n concentrate around their average in terms of sample variance,

$$\begin{aligned}\frac{1}{p}\|A_n\|^2 =& \frac{1}{p}\operatorname{tr}(W_n)^2 - \frac{1}{p}\operatorname{tr}(\operatorname{diag}(W_n))^2 \\ =& \frac{1}{p}\operatorname{tr}(W_n)^2 - \frac{n}{p}\left(\frac{1}{n}\operatorname{tr}(W_n)\right)^2 + o_{\mathbb{P}}(1) \\ =& \int x^2\delta^2(x)dF^{S_n}(x) - \frac{p}{n}\left(\int x\delta(x)dF^{S_n}(x)\right)^2 + o_{\mathbb{P}}(1),\end{aligned}$$

where the leading term is the square of the numerator in ϖ_n and it is bounded away from zero with arbitrarily high probability. Further expand that

$$\begin{aligned}\sqrt{\frac{2}{p}}\|A_n\|T_3 =& \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T A_n \widetilde{X}\beta + \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T P_Z A_n P_Z \widetilde{X}\beta - 2\frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T P_Z A_n \widetilde{X}\beta \\ =:& J_1 + J_2 + J_3.\end{aligned} \tag{5.4}$$

It suffices to show that

$$J_1 - h^2\left(\int x^2\delta(x)dF^{S_n}(x) - \frac{p}{n}\int x dF^{S_n}(x)\cdot\int x\delta(x)dF^{S_n}(x)\right) \xrightarrow{\mathbb{P}} 0,$$

and $J_2, J_3 \xrightarrow{\mathbb{P}} 0$. Decompose

$$\begin{aligned}J_1 =& \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T W_n \widetilde{X}\beta - \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T\left(\frac{1}{n}\operatorname{tr}(W_n)I_n\right)\widetilde{X}\beta \\ &- \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T\left(\operatorname{diag}(W_n) - \frac{1}{n}\operatorname{tr}(W_n)I_n\right)\widetilde{X}\beta =: J_{1,1} + J_{1,2} + J_{1,3}.\end{aligned} \tag{5.5}$$

Let $h_n^2 = n/\sqrt{p}\|\beta\|^2$. We can rewrite

$$J_{1,1} = \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T\left(\frac{1}{n}\widetilde{X}\delta(S_n)\widetilde{X}^T\right)\widetilde{X}\beta = h_n^2\int x^2\delta(x)dF^{S_n}(x;\beta),$$

and

$$\begin{aligned}J_{1,2} =& -\sqrt{p}\|\beta\|\frac{1}{\|\beta\|^2}\beta^T S_n\beta\cdot\frac{1}{p}\operatorname{tr}(S_n\delta(S_n))\cdot \\ =& -h_n^2\cdot\frac{p}{n}\cdot\int x dF^{S_n}(x;\beta)\int x\delta(x)dF^{S_n}(x).\end{aligned}$$

Denote d_t as the t-th diagonal element of W_n. Expanding the quadratic form and applying Cauchy–Schwarz inequality,

$$\begin{aligned}|J_{1,3}| =& \frac{1}{\sqrt{p}} \left| \sum_{t=1}^{n} \left(d_t - \frac{1}{n}\sum_{t=1}^{n} d_t \right) \left(\widetilde{x}_t^T \beta \right)^2 \right| \\ \leq & \frac{1}{\sqrt{p}} \left(\sum_{t=1}^{n} \left(d_t - \frac{1}{n}\sum_{t=1}^{n} d_t \right)^2 \sum_{t=1}^{n} \left(\widetilde{x}_t^T \beta \right)^4 \right)^{\frac{1}{2}} \\ =& h_n^2 \left(\frac{1}{n}\sum_{t=1}^{n} \left(d_t - \frac{1}{n}\sum_{t=1}^{n} d_t \right)^2 \frac{1}{n}\sum_{t=1}^{n} \left(\widetilde{x}_t^T \xi_n \right)^4 \right)^{\frac{1}{2}}\end{aligned}$$

where $\xi_n = \beta / \left\| \beta \right\|$ is the direction of the coefficient β. Note that by assumption we have

$$\frac{1}{n}\sum_{t=1}^{n} \left(d_t - \frac{1}{n}\sum_{t=1}^{n} d_t \right)^2 = o_{\mathbb{P}}(1),$$

and thus to show $J_{1,3} \xrightarrow{\mathbb{P}} 0$ it remains to verify that

$$\frac{1}{n}\sum_{t=1}^{n} \left(\widetilde{x}_t^T \xi_n \right)^4 = O_{\mathbb{P}}(1).$$

For some absolute constant M,

$$\left(\widetilde{x}_t^T \xi_n \right)^4 = \left(x_t^T \xi_n - \bar{x}^T \xi_n \right)^4 \leq M \left\{ \left(x_t^T \xi_n \right)^4 + \left(\bar{x}^T \xi_n \right)^4 \right\}.$$

It suffices to show that $\frac{1}{n}\sum_{t=1}^{n} \left(x_t^T \xi_n \right)^4 = O_{\mathbb{P}}(1)$ and $\left(\bar{x}^T \xi_n \right)^4 = O_{\mathbb{P}}(1)$. The first part immediately follows from Lemma 1 and condition (iv). Moreover,

$$\left(\bar{x}^T \xi_n \right)^4 = \left(\xi_n^T \bar{x} \bar{x}^T \xi_n \right)^2 = \left(O_{\mathbb{P}}(1) \right)^2 = O_{\mathbb{P}}(1),$$

as $\mathbb{E}\left[\xi_n^T \bar{x} \bar{x}^T \xi_n \right] = \xi_n^T \mathbb{E}\left[\bar{x} \bar{x}^T \right] \xi_n = O(1)$. This completes proof for $J_{1,3} \xrightarrow{\mathbb{P}} 0$. Summing up the limits of $J_{1,1}$, $J_{1,2}$ and $J_{1,3}$ yields (5.2). Next, using Lemma 9,

$$|J_2| \leq \frac{1}{\sqrt{p}} \left\| A_n \right\|_{sp} \beta^T \widetilde{X}^T P_Z \widetilde{X} \beta = \left\| A_n \right\|_{sp} O_{\mathbb{P}} \left(\frac{1}{\sqrt{p}} \left\| \beta \right\|^2 + h_n^2 \left\| \beta \right\|^2 \right) \xrightarrow{\mathbb{P}} 0.$$

Combining with Cauchy–Schwarz inequality it follows that

$$|J_3| \leq 2 J_1 J_2 \xrightarrow{\mathbb{P}} 0.$$

Substituting h_n^2 by its limit h^2 completes the proof. □

Proof of Statement (**). First we decompose that

$$\begin{aligned}T_2 =& \frac{\sqrt{2}}{\left\| A_n \right\|} \epsilon^T A_n \widetilde{X} \beta - \frac{\sqrt{2}}{\left\| A_n \right\|} \epsilon^T P_Z A_n \widetilde{X} \beta - \frac{\sqrt{2}}{\left\| A_n \right\|} \epsilon^T A_n P_Z \widetilde{X} \beta + \frac{\sqrt{2}}{\left\| A_n \right\|} \epsilon^T P_Z A_n P_Z \widetilde{X} \beta \\ =& T_{2,1} + T_{2,2} + T_{2,3} + T_{2,4}.\end{aligned}$$

Note that

$$\beta^T \widetilde{X}^T A_n^2 \widetilde{X} \beta \leq \|A_n\|_{sp}^2 \beta^T \widetilde{X}^T \widetilde{X} \beta = n \|A_n\|_{sp}^2 \beta^T S_n \beta.$$

Hence,

$$\mathbb{E}\left[T_{2,1}^2 \mid X\right] = \frac{2}{\|A_n\|^2} \beta^T \widetilde{X}^T A_n^2 \widetilde{X} \beta = \frac{2\|A_n\|_{sp}^2}{\|A_n\|^2/n} \cdot O_p(\|\beta\|^2) \xrightarrow{\mathbb{P}} 0,$$

and therefore $T_{2,1} \xrightarrow{\mathbb{P}} 0$ by Lemma 1. Combining with Cauchy–Schwarz inequality and Lemma 6,

$$T_{2,2}^2 \leq \epsilon^T P_Z \epsilon \cdot \frac{2}{\|A_n\|^2} \beta^T \widetilde{X}^T A_n^2 \widetilde{X} \beta = O_p(d) \cdot o_p(1) \xrightarrow{\mathbb{P}} 0.$$

Next, decompose $A_n = \widetilde{A}_n + \widetilde{A}_n^T$, and

$$T_{2,3} = -\frac{1}{\left\|\widetilde{A}_n\right\|} \epsilon^T \widetilde{A}_n P_Z \widetilde{X} \beta - \frac{1}{\left\|\widetilde{A}_n\right\|} \epsilon^T \widetilde{A}_n^T P_Z \widetilde{X} \beta =: T_{2,3,1} + T_{2,3,2}. \tag{5.6}$$

By Cauchy–Schwarz inequality,

$$\begin{aligned}
(\epsilon^T \widetilde{A}_n P_Z \widetilde{X} \beta)^2 &= \left(\frac{1}{n} \epsilon^T \widetilde{A}_n Z \widehat{\Omega}^{-1} Z^T \widetilde{X} \beta\right)^2 \\
&\leq \frac{1}{n^2} \epsilon^T \widetilde{A}_n Z \widehat{\Omega}^{-1} Z^T \widetilde{A}_n^T \epsilon \cdot \beta^T \widetilde{X}^T Z \widehat{\Omega}^{-1} Z^T \widetilde{X} \beta \\
&\leq \left(\lambda_{\min}(\widehat{\Omega})\right)^{-1} \frac{1}{n} \epsilon^T \widetilde{A}_n Z Z^T \widetilde{A}_n^T \epsilon \cdot \beta^T \widetilde{X}^T P_Z \widetilde{X} \beta.
\end{aligned}$$

We invoke from (5.2) that

$$\epsilon^T \widetilde{A}_n Z Z \widetilde{A}_n^T Z \epsilon = O_{\mathbb{P}}\left(\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right) \cdot nd\right) = o_{\mathbb{P}}(n \|A_n\|^2).$$

Combining with Lemma 9, we have that

$$T_{2,3,1}^2 = \frac{1}{n\|A_n\|^2} \cdot o_{\mathbb{P}}(n\|A_n\|^2) \cdot O_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.$$

Similarly,

$$(\epsilon^T \widetilde{A}_n^T P_Z \widetilde{X} \beta)^2 \leq \left(\lambda_{\min}(\widehat{\Omega})\right)^{-1} \frac{1}{n} \epsilon^T \widetilde{A}_n^T Z Z^T \widetilde{A}_n \epsilon \cdot \beta \widetilde{X}^T P_Z \widetilde{X} \beta$$

and therefore, combining Lemma 8 and Lemma 9,

$$T_{2,3,2}^2 = o_{\mathbb{P}}(1 + \|\beta\|^2) \cdot O_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.$$

Finally, using Cauchy–Schwarz inequality and Lemma 6 and 9,

$$\begin{aligned}
T_{2,4}^2 \leq& \frac{2}{\|A_n\|^2} \epsilon^T P_Z \epsilon \cdot \beta^T \widetilde{X}^T P_Z A_n^2 P_Z \widetilde{X} \beta \\
\leq& \frac{2\|A_n\|_{sp}^2}{\|A_n\|^2} \cdot \epsilon^T P_Z \epsilon \cdot \beta^T \widetilde{X}^T P_Z \widetilde{X} \beta = o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

This completes the proof. □

5.3 Proof of Theorems 4 and 5

The proof of Theorem 4 is straightforward and is available in the supplement (He et al., 2020) for completeness: the first part is completely analogous to that in the subsection 5.1; the second part follows from Theorem 5 (to be proved below) and we verify that $\rho_n^2 = \mu_n^T \mu_n \xrightarrow{\mathbb{P}} 0$ in this case.

To prove Theorem 5, we need the following lemma:

Lemma 10. *Under the conditions of Theorem 5,*

$$\left\| \frac{1}{\sqrt{n}\left\|\widetilde{A}_n\right\|} Z^T \widetilde{A}_n \epsilon - \Omega^{1/2}\mu_n \right\|^2 \xrightarrow{\mathbb{P}} 0.$$

Proof of Theorem 5. We first expand our test statistic as

$$\begin{aligned}
Q_n =& \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T (I - P_Z)\widetilde{A}_n (I-P_Z)\epsilon \\
=& \frac{1}{\left\|\widetilde{A}_n\right\|}\left(\epsilon^T \widetilde{A}_n \epsilon - \frac{1}{n}\epsilon^T Z\Omega^{-1}Z^T\widetilde{A}_n\epsilon\right) \\
& - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T \frac{1}{n} Z\left(\widehat{\Omega}^{-1} - \Omega^{-1}\right)Z^T \widetilde{A}_n\epsilon - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z \widetilde{A}_n^T \epsilon + \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z \widetilde{A}_n P_Z \epsilon \\
=& \left(\frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T \widetilde{A}_n \epsilon - \frac{1}{\sqrt{n}}\epsilon^T Z \Omega^{-1/2}\mu_n\right) - \frac{1}{\sqrt{n}}\epsilon^T Z\Omega^{-1/2}\left(\frac{1}{\sqrt{n}\left\|\widetilde{A}_n\right\|}\Omega^{-1/2}Z^T\widetilde{A}_n\epsilon - \mu_n\right) \\
& - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T \frac{1}{n} Z\left(\widehat{\Omega}^{-1} - \Omega^{-1}\right)Z^T \widetilde{A}_n\epsilon - \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z \widetilde{A}_n^T \epsilon + \frac{1}{\left\|\widetilde{A}_n\right\|}\epsilon^T P_Z \widetilde{A}_n P_Z \epsilon \\
=:& \widetilde{Q}_n - T_{1,1} - T_{1,2} + T_2 + T_3,
\end{aligned} \tag{5.7}$$

where $T_2 = o_{\mathbb{P}}(1)$ and $T_3 = o_{\mathbb{P}}(1)$ are the same in the proof of Theorem 1, and

$$\widetilde{Q}_n = \sum_{t=1}^n \varepsilon_t \left(\sum_{s=1}^{t-1} \frac{1}{\left\|\widetilde{A}_n\right\|}\varepsilon_s \frac{1}{n}\widetilde{x}_s^T \widetilde{x}_t - \frac{1}{\sqrt{n}} z_t^T \Omega^{-1}\mu_n\right) =: \sum_{t=1}^n \Delta_t$$

and Δ_t is a martingale difference array such that $\mathbb{E}(\Delta_t | \mathcal{F}_{n,t-1}) = 0$ and $\mathbb{E}\left(\Delta_t^2 | \mathcal{F}_{n,0}\right) = 1 - \rho_n^2 > 0$.

It suffices to show that $T_{1,1}, T_{1,2} \xrightarrow{\mathbb{P}} 0$ and $\widetilde{Q}_n / \sqrt{1-\rho_n^2} \xrightarrow{d} \mathcal{N}(0,1)$. By Cauchy–Schwarz inequality,

$$T_{1,1}^2 \le \frac{1}{n}\epsilon^T Z\Omega^{-2}Z^T\epsilon \cdot \left\| \frac{1}{\sqrt{n}\left\|\widetilde{A}_n\right\|} Z^T \widetilde{A}_n \epsilon - \Omega^{1/2}\mu_n \right\|^2.$$

By Lemma 10, $\left\| \frac{1}{\sqrt{n}\left\|\widetilde{A}_n\right\|} Z^T \widetilde{A}_n \epsilon - \Omega^{1/2}\mu_n \right\|^2 = o_{\mathbb{P}}(1)$. In addition,

$$\frac{1}{n}\epsilon^T Z\Omega^{-2}Z^T\epsilon \le (\lambda_{\min}(\Omega))^{-2}\frac{1}{n}\epsilon^T ZZ^T\epsilon = (\lambda_{\min}(\Omega))^{-2}\frac{1}{n}\sum_{t=1}^n \varepsilon_t^2 z_t^T z_t$$

and, by using the martingale condition

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 z_t^T z_t\right] = \mathbb{E}(z_t^T z_t) = \operatorname{tr}(\Omega) = O(1).$$

Hence, $T_{1,1}^2 = O_{\mathbb{P}}(1)\cdot o_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0$. Furthermore, using the spectral decomposition of $\widehat{\Omega}^{-1}-\Omega = \sum_{i=1}^{d+1}\widetilde{\lambda}_i\widetilde{u}_i\widetilde{u}^T$ and Cauchy–Schwarz inequality, it is not hard to show that

$$|T_{1,2}| \leq \left(\sum_{i=1}^{d+1}|\widetilde{\lambda}_i|\right)\cdot\sqrt{\frac{1}{n}\epsilon^T ZZ^T\epsilon}\cdot\sqrt{\frac{1}{\left\|\widetilde{A}_n\right\|^2}\epsilon^T\widetilde{A}_n^T\frac{1}{n}ZZ^T\widetilde{A}_n\epsilon}.$$

Taking squares of both sides, and recall from above that $\frac{1}{n}\epsilon^T ZZ^T\epsilon = O_{\mathbb{P}}(1)$ and noting that $|\widetilde{\lambda}_i| = o_{\mathbb{P}}(1)$,

$$T_{1,2}^2 = o_{\mathbb{P}}\left(\frac{1}{\left\|\widetilde{A}_n\right\|^2}\epsilon^T\widetilde{A}_n^T\frac{1}{n}ZZ^T\widetilde{A}_n\epsilon\right) = \lambda_{\max}(\widehat{\Omega})\cdot o_{\mathbb{P}}\left(\frac{1}{\left\|\widetilde{A}_n\right\|^2}\epsilon^T\widetilde{A}_n^T\widetilde{A}_n\epsilon\right)\xrightarrow{\mathbb{P}} 0,$$

where in the last equality we used the identity $\left\|\frac{1}{n}ZZ^T\right\|_{sp} = \lambda_{\max}(\widehat{\Omega})$, and for the convergence we applied Lemma 1 with the fact that $\mathbb{E}\left[\frac{1}{\left\|\widetilde{A}_n\right\|^2}\epsilon^T\widetilde{A}_n^T\widetilde{A}_n\epsilon \mid \mathcal{F}_{n,0}\right] = \frac{\operatorname{tr}\left(\widetilde{A}_n^T\widetilde{A}_n\right)}{\left\|\widetilde{A}_n\right\|^2} = 1$.

It remains to show that $\widetilde{Q}_n/\sqrt{1-\rho_n^2}\xrightarrow{d}\mathcal{N}(0,1)$. Following the proof of Theorem 1, it suffices to verify that:

(*) $\max_{1\leq t\leq n}\mathbb{E}\left[\Delta_t^2\mid\mathcal{F}_{n,t-1}\right]\xrightarrow{\mathbb{P}} 0$.

(**) $\frac{1}{1-\rho_n^2}\sum_{t=1}^{n}\mathbb{E}\left[\Delta_t^2\mid\mathcal{F}_{n,t-1}\right]\xrightarrow{\mathbb{P}} 1$, or $\sum_{t=1}^{n}\mathbb{E}\left[\Delta_t^2\mid\mathcal{F}_{n,t-1}\right]-\left(1-\rho_n^2\right)\xrightarrow{\mathbb{P}} 0$.

By Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\Delta_t^2\mid\mathcal{F}_{n,t-1}\right] &= \left(\sum_{s=1}^{t-1}\frac{1}{\left\|\widetilde{A}_n\right\|}\varepsilon_s\frac{1}{n}\widetilde{x}_s^T\widetilde{x}_t-\frac{1}{\sqrt{n}}z_t^T\Omega^{-1/2}\mu_n\right)^2\\ &\leq \frac{2}{\left\|\widetilde{A}_n\right\|^2}\left(\sum_{s=1}^{t-1}\varepsilon_s\frac{1}{n}\widetilde{x}_s^T\widetilde{x}_t\right)^2+\frac{2}{n}\left(z_t^T\Omega^{-1/2}\mu_n\right)^2 =: J_{t,1}+J_{t,2}.\end{aligned}$$

From Lemma 4 we already know that $\max_{1\leq t\leq n}J_{t,1}\xrightarrow{\mathbb{P}} 0$. It remains to show that $\max_{1\leq t\leq n}J_{t,2}\xrightarrow{\mathbb{P}} 0$. By Cauchy–Schwarz inequality,

$$J_{t,2}\leq\mu_n^T\mu_n\cdot\frac{2}{n}z_t^T\Omega^{-1}z_t = \rho_n^2\cdot\frac{2}{n}z_t^T\Omega^{-1}z_t.$$

By a direct calculation,

$$\mathbb{E}\left[z_t^T\Omega^{-1}z_t\right] = \operatorname{tr}\left(\mathbb{E}\left[\Omega^{-1}z_t z_t^T\right]\right) = \operatorname{tr}\left(I_d\right) = d<\infty.$$

Then using, for example, Lemma 11.2 in Owen (2001) and noting the independent condition is not necessary for the Borel–Cantelli arguments therein, we have $\max_{1\leq t\leq n}z_t^T\Omega^{-1}z_t = o(n)$. It then follows that $\max_{1\leq t\leq n}J_{t,2} = o(1)$.

It remains to verify the condition (**). Let $b_t = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^d$ denote the unit vector with t-th entry equaling to 1 and all other entries equaling to 0. By a direct calculation,

$$\mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}\right] = \left(\frac{\widetilde{A}_n}{\left\|\widetilde{A}_n\right\|}\epsilon - \frac{1}{\sqrt{n}}Z\Omega^{-1/2}\mu_n\right)^T b_t b_t^T \left(\frac{\widetilde{A}_n}{\left\|\widetilde{A}_n\right\|}\epsilon - \frac{1}{\sqrt{n}}Z\Omega^{-1/2}\mu_n\right).$$

Hence,

$$\begin{aligned}\sum_{t=1}^n \mathbb{E}\left[\Delta_t^2 \mid \mathcal{F}_{n,t-1}\right] =& \left(\frac{\widetilde{A}_n}{\left\|\widetilde{A}_n\right\|}\epsilon - \frac{1}{\sqrt{n}}Z\Omega^{-1/2}\mu_n\right)^T \left(\frac{\widetilde{A}_n}{\left\|\widetilde{A}_n\right\|}\epsilon - \frac{1}{\sqrt{n}}Z\Omega^{-1/2}\mu_n\right)\\ =&\epsilon^T \frac{\widetilde{A}_n^T\widetilde{A}_n\epsilon}{\left\|\widetilde{A}_n\right\|^2} - \mu_n^T\mu_n - 2\mu_n^T\left(\frac{1}{\sqrt{n}\left\|A_n\right\|}\Omega^{-1/2}Z^T\widetilde{A}_n\epsilon - \mu_n\right)\\ &+ \mu_n^T\left(\Omega^{-1/2}\widehat{\Omega}\Omega^{-1/2} - I_d\right)\mu_n =: R_1 - R_2 - 2R_3 + R_4.\end{aligned}$$

From Lemma 5 we already know that $R_1 \xrightarrow{\mathbb{P}} 1$. Moreover, recall that $R_2 = \rho_n^2$ and

$$|R_4| \le R_2 \left\|\Omega^{-1/2}\widehat{\Omega}^{-1}\Omega^{-1/2} - I_d\right\|_{sp} \xrightarrow{\mathbb{P}} 0.$$

It remains to prove that $R_3 \xrightarrow{\mathbb{P}} 0$. Using Cauchy–Schwarz inequality and Lemma 10,

$$\begin{aligned}R_3^2 \le& \mu_n^T\mu_n \cdot \left\|\frac{1}{\sqrt{n}\left\|A_n\right\|}\Omega^{-1/2}Z^T\widetilde{A}_n\epsilon - \mu_n\right\|^2\\ \le& 1 \cdot (\lambda_{\min}(\Omega))^{-1}\left\|\frac{1}{\sqrt{n}\left\|A_n\right\|}Z^T\widetilde{A}_n\epsilon - \Omega^{1/2}\mu_n\right\|^2 \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Our proof is now complete. □

5.4 Proof of Theorems 6 and 7

The proof of Theorem 6 is completely analogous to that of Theorem 3 and the details are available in the supplement (He et al., 2020) for completeness.

To prove Theorem 7, we need to generalize Lemma 10:

Lemma 11. *Lemma 10 remains true under the alternatives.*

Proof of Theorem 7. We prove the general results in Remark 3, and the theorem is the special case with $\delta(x) = 1$. We invoke the decomposition $Q_n(\delta) = T_1 + T_2 + T_3$ in (5.3). Recall from (5.7) and the proof of Theorem 5 that we can rewrite $T_1 = \widetilde{Q}_n(\delta) - T_{1,1} - T_{1,2}$, where $\widetilde{Q}_n(\delta)/\sqrt{1-\rho_n^2(\delta)} \xrightarrow{d} \mathcal{N}(0,1)$ and $T_{1,1}, T_{1,2} \xrightarrow{\mathbb{P}} 0$ using Lemma 11. Next, invoking the proof of statements (*) and (**) in the proof of Theorem 3 respectively, we can show that $T_3 = \frac{h^2}{\sqrt{2}}\varpi_n(\delta) + o_{\mathbb{P}}(1)$ and $T_2 = T_{2,3,2} + o_{\mathbb{P}}(1)$ by generalizing Lemma 9 to be

$$\beta^T\widetilde{X}^T P_Z\widetilde{X}\beta = O_p\left(\|\beta\|^2 + n\|\beta\|^4 + 1\right); \tag{5.8}$$

see the proof of Theorem 6 in the supplement for the proof of the last equation.

It remains to show that $T_{2,3,2} + h^2 \frac{\varpi_n}{\sqrt{2}} \mu_n^T(\delta)\mu_n \xrightarrow{\mathbb{P}} 0$. Note that from (5.6) we have

$$\begin{aligned} T_{2,3,2} :=& - \frac{1}{\left\| \widetilde{A}_n(\delta) \right\|} \epsilon^T \widetilde{A}_n^T(\delta) P_Z \widetilde{X}\beta \\ =& - \left(\frac{1}{\sqrt{n} \left\| \widetilde{A}_n \right\|} Z^T \widetilde{A}_n \epsilon \right)^T \widehat{\Omega}^{-1} \left(\frac{1}{\sqrt{n}} Z^T \widetilde{X}\beta \right) \\ =& - \left(\Omega^{1/2}\mu_n(\delta) + o_{\mathbb{P}}(1) \right)^T \left(\Omega^{-1} + o_{\mathbb{P}}(1) \right) \left(\frac{1}{\sqrt{n}} Z^T \widetilde{X}\beta \right). \end{aligned}$$

It suffices to show that

$$\frac{1}{\sqrt{n}} Z^T \widetilde{X}\beta - h^2 \frac{\varpi_n}{\sqrt{2}} \Omega^{1/2}\mu_n \xrightarrow{\mathbb{P}} 0.$$

Recall from the proof of Theorem 3 that, for $\delta(x) \equiv 1$, $\left\| \widetilde{A}_n \right\| / \sqrt{p} = \| A_n \| / \sqrt{2p} = \varpi_n / \sqrt{2} + o_{\mathbb{P}}(1)$. Then substituting $\varpi_n / \sqrt{2}$ by $\left\| \widetilde{A}_n \right\| / \sqrt{p}$, substituting h^2 by $\frac{n}{\sqrt{p}} \|\beta\|^2$, and using the additional freeness assumption in the theorem, we only need to show that

$$\frac{1}{\sqrt{n}} Z^T \widetilde{X}\beta - \frac{1}{\sqrt{n}} \left[0, \beta^T \widetilde{X}^T \Psi_1^T \widetilde{X}\beta, \ldots, \beta^T \widetilde{X}^T \Psi_d^T \widetilde{X}\beta \right] \xrightarrow{\mathbb{P}} 0,$$

that is,

$$\frac{1}{\sqrt{n}} \mathbf{z}_i^T \widetilde{X}\beta - \frac{1}{\sqrt{n}} \beta^T \widetilde{X}^T \Psi_i^T \widetilde{X}\beta \xrightarrow{\mathbb{P}} 0, \ \forall i = 1, \ldots, d.$$

Now using the expansion

$$\mathbf{z}_i := (z_{1,i}, \ldots, z_{n,i})^T = \alpha_i \mathbf{1}_n + \Psi_i \epsilon + \Psi_i X\beta + \mathbf{v}_i, \ i = 1, \ldots, d,$$

we have that

$$\frac{1}{\sqrt{n}} \mathbf{z}_i^T \widetilde{X}\beta - \frac{1}{\sqrt{n}} \beta^T \widetilde{X}^T \Psi_i^T \widetilde{X}\beta = \frac{1}{\sqrt{n}} \epsilon^T \Psi_i^T \widetilde{X}\beta + \frac{1}{\sqrt{n}} \mathbf{v}_i^T \widetilde{X}\beta =: R_1 + R_2.$$

Noting that $\left\| \Psi_i \Psi_i^T \right\|_{sp} \le \left(\sum_{l=1}^\infty |\psi_i(l)| \right)^2 < \infty$,

$$\mathbb{E}\left[R_1^2 \mid X \right] = \frac{1}{n} \beta^T \widetilde{X}^T \Psi_i \Psi_i^T \widetilde{X}\beta \le \left\| \Psi_i \Psi_i^T \right\|_{sp} \beta^T S_n \beta = O(1) \cdot O_{\mathbb{P}}\left(\|\beta\|^2 \right) \xrightarrow{\mathbb{P}} 0.$$

Moreover, we have either that, by Cauchy–Schwarz inequality,

$$R_2^2 \le \mathbf{v}_i^T \mathbf{v}_i \cdot \beta^T S_n \beta = o_{\mathbb{P}}(np^{-1/2}) \cdot O_{\mathbb{P}}\left(\|\beta\|^2 \right) \xrightarrow{\mathbb{P}} 0,$$

or, by the definition of spectral norm,

$$\mathbb{E}\left[R_2^2 \mid X \right] \le \lambda_{\max}\left(\left\| \mathbb{E}\left[\mathbf{v}_i \mathbf{v}_i^T | X \right] \right\|_{sp} \right) \beta^T S_n \beta \xrightarrow{\mathbb{P}} 0.$$

While R_2^2 and $\mathbb{E}\left[R_2^2 \mid X \right]$ above are actually defined with different sub-sigma-algebras, we extend the definition of the random variables in the largest probability space with probability measure $\mathbb{P}$ for presentation convenience. In either case we have $R_2 \xrightarrow{\mathbb{P}} 0$. Our proof is now complete. □

5.5 Proof of Corollaries 1–6

Proof of Corollary 1. It suffices to prove the consistency of the variance estimator (2.7). A direct calculation yields the matrix expression given by

$$\widehat{\sigma}^2 = \frac{1}{n-(d+1)}\epsilon^T\left(I-P_Z\right)\epsilon = \frac{n}{n-(d+1)}\frac{1}{n}\epsilon^T\epsilon - \frac{1}{n-(d+1)}\epsilon^T P_Z \epsilon =: T_1 + T_2.$$

Using the martingale law of large number and noting that $d/n \to 0$, we can show that $T_1 \xrightarrow{\mathbb{P}} 1$. It remains to show that $T_2 \xrightarrow{\mathbb{P}} 0$, which follows from Lemma 6. □

Proof of Corollary 2 . Note that the support of F^{S_n} is bounded with probability tending to 1. It follows from Portmanteau Theorem (e.g. Theorem 2.1 in Billingsley) that $\varpi_n \xrightarrow{\mathbb{P}} \varpi$, where ϖ has the same expression but uses F instead of F^{S_n}. □

Proof of Corollary 3 and 4. We only need to prove under the local alternatives (2.8) that $\widehat{\sigma}_n^2/\sigma_n^2 \xrightarrow{\mathbb{P}} 1$. Expanding

$$\begin{aligned}
&\frac{1}{n-(d+1)}y^T\left(I-P_Z\right)y\\
=&\frac{1}{n-(d+1)}\left(X\beta+\epsilon\right)^T\left(I-P_Z\right)\left(X\beta+\epsilon\right)\\
=&\frac{1}{n-(d+1)}\beta^T\widetilde{X}^T\left(I-P_Z\right)\widetilde{X}\beta + \frac{2}{n-(d+1)}\beta^T\widetilde{X}^T\left(I-P_Z\right)e + \frac{1}{n-(d+1)}e^Te\\
=:&T_1+T_2+T_3.
\end{aligned}$$

Note that Lemma 6 holds under the alternatives as well, and then by carefully checking the proof of Corollary 1, we already have $T_3 \xrightarrow{\mathbb{P}} 1$. Using the spectral norm inequality,

$$\begin{aligned}
T_1 \le& \|I-P_Z\|_{sp}\cdot\frac{1}{n-(d+1)}\beta^T\widetilde{X}^T\widetilde{X}\beta\\
\le& 1\cdot\frac{n}{n-(d+1)}\beta^T S_n\beta = O_{\mathbb{P}}\left(\|\beta\|^2\right) \xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

Finally, by Cauchy–Schwarz inequality $T_2^2 \le T_1\cdot T_3 \xrightarrow{\mathbb{P}} 0$. □

Proof of Corollary 5. It suffices to show that $\widehat{\rho}_n^2 - \rho_n^2 = \widehat{\rho}_n^2 - \|\mu_n\|^2 \xrightarrow{\mathbb{P}} 0$. Let

$$\widehat{\rho}_n^2 = \frac{e^T\widetilde{A}_n^T P_Z\widetilde{A}_n e/\left\|\widetilde{A}_n\right\|^2}{e^T\left(\widetilde{A}_n^T\widetilde{A}_n\right)e/\left\|\widetilde{A}_n\right\|^2} =: \frac{\Delta_1}{\Delta_2}.$$

It suffices to show that: (*) $\Delta_1 - \|\mu_n\|^2 \xrightarrow{\mathbb{P}} 0$; and (**) $\Delta_2 - 1 \xrightarrow{\mathbb{P}} 0$.

We expand that

$$\Delta_1 = \frac{\epsilon^T\widetilde{A}_n^T P_Z\widetilde{A}_n\epsilon}{\left\|\widetilde{A}_n\right\|^2} - \frac{2\epsilon^T P_Z\widetilde{A}_n^T P_Z\widetilde{A}_n\epsilon}{\left\|\widetilde{A}_n\right\|^2} + \frac{\epsilon^T P_Z\widetilde{A}_n^T P_Z\widetilde{A}_n P_Z\epsilon}{\left\|\widetilde{A}_n\right\|^2} =: \Delta_{1,1} - 2\Delta_{1,2} + \Delta_{1,3}.$$

By Lemma 10 and the assumption that $\widehat{\Omega} \xrightarrow{\mathbb{P}} \Omega$,

$$\begin{aligned}\Delta_{1,1} &= \left(\frac{1}{\sqrt{n}\,\|A_n\|} Z^T \widetilde{A}_n \epsilon\right)^T \widehat{\Omega}^{-1} \left(\frac{1}{\sqrt{n}\,\|A_n\|} Z^T \widetilde{A}_n \epsilon\right) \\ &= \left(\Omega^{1/2}\mu_n + o_{\mathbb{P}}(1)\right)^T \left(\Omega^{-1} + o_{\mathbb{P}}(1)\right) \left(\Omega^{1/2}\mu_n + o_{\mathbb{P}}(1)\right) = \mu_n^T \mu_n + o_{\mathbb{P}}(1).\end{aligned}$$

On the other hand, using the definition of spectral norms and Lemma 6,

$$0 \le \Delta_{1,3} \le \|P_Z\|_{sp} \frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\left\|\widetilde{A}_n\right\|^2} \cdot \epsilon^T P_Z \epsilon = 1 \cdot o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0.$$

Now applying Cauchy–Schwarz inequality we also have that

$$|\Delta_{1,2}|^2 \le \Delta_{1,1}\Delta_{1,3} \xrightarrow{\mathbb{P}} 0.$$

This completes the proof of statement (*). The proof of statement (**) is similar. We expand that

$$\Delta_2 = \frac{\epsilon^T \widetilde{A}_n^T \widetilde{A}_n \epsilon}{\left\|\widetilde{A}_n\right\|^2} - 2\frac{\epsilon^T P_Z \widetilde{A}_n^T \widetilde{A}_n \epsilon}{\left\|\widetilde{A}_n\right\|^2} + \frac{\epsilon^T P_Z \widetilde{A}_n^T \widetilde{A}_n P_Z \epsilon}{\left\|\widetilde{A}_n\right\|^2} =: \Delta_{2,1} - 2\Delta_{2,2} + \Delta_{2,3}.$$

Note that the diagonal elements of $\widetilde{A}_n^T \widetilde{A}_n$ are nonnegative and bounded by $\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)$, and their sum $\operatorname{tr}\left(\widetilde{A}_n^T \widetilde{A}_n\right) = \left\|\widetilde{A}_n\right\|^2$. Using Lemma 2,

$$\begin{aligned}\Delta_{2,1} - 1 =& O_{\mathbb{P}}\left(\left(\frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\left\|\widetilde{A}_n\right\|^2}\right)^{\frac{\iota}{1+\iota}} + \frac{\left\|\widetilde{A}_n^T \widetilde{A}_n\right\|}{\left\|\widetilde{A}_n\right\|^2}\right) \\ =& o_{\mathbb{P}}(1) + O_{\mathbb{P}}\left(\frac{\lambda_{\max}^{1/2}\left(\widetilde{A}_n^T \widetilde{A}_n\right) \cdot \sqrt{\operatorname{tr}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}}{\left\|\widetilde{A}_n\right\|^2}\right) \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Finally, by the definition of spectral norm and Lemma 6,

$$0 \le \Delta_{2,3} \le \frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\left\|\widetilde{A}_n\right\|^2} \cdot \epsilon^T P_Z \epsilon = o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) \xrightarrow{\mathbb{P}} 0,$$

and by Cauchy–Schwarz inequality $\Delta_{2,2}^2 \le \Delta_{2,1}\Delta_{2,3} \xrightarrow{\mathbb{P}} 0$. This completes the proof. □

Proof of Corollary 6. We need to prove that $\widehat{\rho}_n^2 - \rho_n^2 \xrightarrow{\mathbb{P}} 0$ under the alternatives. Let

$$\widehat{\rho}_n^2 = \frac{e^T \widetilde{A}_n^T P_Z \widetilde{A}_n e / \left\|\widetilde{A}_n\right\|^2}{e^T \left(\widetilde{A}_n^T \widetilde{A}_n\right) e / \left\|\widetilde{A}_n\right\|^2} =: \frac{\widetilde{\Delta}_1}{\widetilde{\Delta}_2}.$$

It suffices to show that: (*) $\widetilde{\Delta}_1 - \rho_n^2 \xrightarrow{\mathbb{P}} 0$; and (**) $\widetilde{\Delta}_2 - 1 \xrightarrow{\mathbb{P}} 0$.

By a direct calculation,

$$\begin{aligned}\widetilde{\Delta}_1 =&\frac{\epsilon^T(I-P_Z)\widetilde{A}_n^T P_Z \widetilde{A}_n (I-P_Z)\epsilon}{\left\|\widetilde{A}_n\right\|^2} + \frac{\beta^T \widetilde{X}^T (I-P_Z)\widetilde{A}_n^T P_Z \widetilde{A}_n (I-P_Z)\widetilde{X}\beta}{\left\|\widetilde{A}_n\right\|^2}\\ &+ 2\frac{\beta^T \widetilde{X}^T (I-P_Z)\widetilde{A}_n^T P_Z \widetilde{A}_n (I-P_Z)\epsilon}{\left\|\widetilde{A}_n\right\|^2} =: \Delta_1 + R_1 + R_2.\end{aligned}$$

Recall from the proof of Corollary 5 that $\Delta_1 - \rho_n^2 \xrightarrow{\mathbb{P}} 0$. For statement (*) it remains to show that $R_1 \xrightarrow{\mathbb{P}} 0$, as then by Cauchy–Schwarz inequality we have $R_2^2 \leq 4\Delta_1 R_1 \xrightarrow{\mathbb{P}} 0$. Observe that $\lambda_{\max}(P_Z) = \lambda_{\max}(I - P_Z) = 1$. Then, using the definition of spectral norm,

$$R_1 \leq \frac{\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right)}{\left\|\widetilde{A}_n\right\|^2/n} \beta^T S_n \beta = O_{\mathbb{P}}\left(\|\beta\|^2\right) \xrightarrow{\mathbb{P}} 0,$$

where we used the facts that $\lambda_{\max}\left(\widetilde{A}_n^T \widetilde{A}_n\right) \leq \lambda_{\max}^2(S_n) = O_{\mathbb{P}}(1)$ and $\left\|\widetilde{A}_n\right\|^2/n = \|A_n\|^2/(2n) = \frac{p}{2n}\left(\varpi_n + o_{\mathbb{P}}(1)\right)$ which is bounded away from 0 with probability tending to 1.

The proof of statement (**) is completely analogous, after replacing $\widetilde{A}_n^T P_Z \widetilde{A}_n$ by $\widetilde{A}_n^T \widetilde{A}_n$ everywhere above. We omit the details. □

References

Arias-Castro, E., E. J. Candès, and Y. Plan (2011, 10). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Ann. Statist. 39*(5), 2533–2556.

Bai, Z. D., B. Q. Miao, and G. M. Pan (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab. 35*, 1532–1572.

Bai, Z. D. and J. W. Silverstein (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab. 26*, 316–345.

Bai, Z. D. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices* (Second ed.). Springer, New York.

Chen, B. B. and G. M. Pan (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli 18*, 1405–1420.

Chernozhukov, V., D. Chetverikov, and K. Kato (2019). Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud. 86*, 1867–1900.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and strong cross-section dependence and estimation of large panels. *Econom. J. 14*, C45–C90.

Cui, H., W. Guo, and W. Zhong (2018). Test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *Ann. Statist. 46*, 958–988.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory Related Fields 75*, 261–277.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika 101*, 269–284.

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: ridge regression and classification. *Ann. Statist. 46*, 247–279.

Donoho, D. and J. Jin (2004, 06). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist. 32*(3), 962–994.

El Karoui, N. (2009). Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab. 19*, 2362–2405.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 74*, 37–65.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*, 1497–1541.

Gao, J., X. Han, G. M. Pan, and Y. Yang (2017). High dimensional correlation matrices: the central limit theorem and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 79*(3), 677–693.

Giannone, D., M. Lenza, and G. E. Primiceri (2017). Economic predictions with big data: The illusion of sparsity. CEPR Discussion Paper No. DP12256.

Goeman, J. J., S. A. van De Geer, and H. C. van Houwelingen (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 68*, 477–493.

Goeman, J. J., H. C. van Houwelingen, and L. Finos (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika 98*, 381–390.

Guo, B. and S. X. Chen (2016). Tests for high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 78*, 1079–1102.

Hall, P. and C. C. Heyde (1980). *Martingale limit theory and its application.* Academic Press, New York.

Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist. 38*, 1686–1732.

Hayashi, F. (2000). *Econometrics.* Princeton Univ. Press.

He, Y., S. Jaidee, and J. Gao (2020). Supplement to "most powerful test against high dimensional free alternatives".

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis.* Cambridge Univ. Press.

Jin, B., C. Wang, B. Miao, and M.-N. L. Huang (2009). Limiting spectral distribution of large-dimensional sample covariance matrices generated by VARMA. *J. Multivariate Anal. 100*, 2112–2125.

Kock, A. B. and D. Preinerstorfer (2019). Power in high-dimensional testing problems. *Econometrica 87*, 1055–1069.

Lam, C. (2016). Nonparametric eigenvalue-regularized precision or covariance matrix estimator. *Ann. Statist. 44*, 928–953.

Ledoit, O. and S. Péché (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields 151*, 233–264.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist. 40*, 1024–1060.

Ledoit, O. and M. Wolf (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *Rev. Financ. Stud. 30*, 4349–4388.

Liu, H., A. Aue, and D. Paul (2015). On the Marčenko-Pastur law for linear time series. *Ann. Statist. 43*, 675–712.

Marčenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik 1*, 457–483.

McCracken, M. W. and S. Ng (2016). FRED-MD: a monthly database for macroeconomic research. *J. Bus. Econom. Statist. 34*, 574–589.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics 168*, 244–258.

Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC, New York.

Pan, G. M. (2014). Comparison between two types of large sample covariance matrices. *Ann. Inst. Henri Poincaré Probab. Stat. 50*, 655–677.

Pan, G. M., J. Gao, and Y. Yang (2014). Testing independence among a large number of high-dimensional random vectors. *J. Amer. Statist. Assoc. 109*, 600–612.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal. 55*, 331–339.

Silverstein, J. W. and Z. D. Bai (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal. 54*, 175–192.

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

Wang, S. and H. Cui (2013). Generalized F test for high dimensional linear regression coefficients. *J. Multivariate Anal. 117*, 134–149.

Wu, W. B. and X. Shao (2007). A limit theorem for quadratic forms and its applications. *Econometric Theory 23*, 930–951.

Xi, H., F. Yang, and J. Yin (2020). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *Ann. Statist. 48*, 953–982.

Xia, N., Y. Qin, and Z. D. Bai (2013). Convergence rates of eigenvector empirical spectral distribution of large dimensional sample covariance matrix. *Ann. Statist. 41*, 2572–2607.

Xia, N. and X. Zheng (2018). On the inference about the spectral distribution of high-dimensional covariance matrix based on high-frequency noisy observations. *Ann. Statist. 46*, 500–525.

Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal. 20*, 50–68.

Zhang, L. (2006). *Spectral analysis of large dimensional random matrices.* Ph. D. thesis, National Univ. Singapore.

Zheng, X. and Y. Li (2011). On the estimation of integrated covariance matrices of high dimensional diffusion processes. *Ann. Statist. 39*, 3121–3151.

Zhong, P.-S. and S. X. Chen (2011). Tests for high-dimensional regression coefficients with factorial designs. *J. Amer. Statist. Assoc. 106*, 260–274.

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

Zhu, Y. and J. Bradic (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc. 113*, 1583–1600.