MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

Most Powerful Test against High Dimensional Free Alternatives

Yi He, Sombut Jaidee and Jiti Gao

April 2020

Working Paper 13/20

AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

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

March 19, 2020

Abstract

We propose a powerful quadratic test for the overall significance of many weak exogenous variables in a dense autoregressive model. By shrinking the classical weighting matrix on the sample moments to be identity, 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 estimation of the autoregressive coefficients. 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 predictors' cross-sectional and temporal dependence structure, and do not require a limiting spectral law of their sample covariance matrix. The method extends beyond autoregressive models, and allows more general nuisance parameters. Monte Carlo studies suggest a good power performance of our proposed test against high dimensional dense alternative for various data generating processes. We apply our tests to detect the overall significance of over one hundred exogenous variables in the latest 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 (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 t statistics (Chernozhukov et al., 2018), one may improve the testing power in many applications. While the sparsity assumption is interesting, it may be debatable in many economic applications and may not be always available. In particular, Giannone et al. (2017) have observed non-sparsity for five out of six important economic data sets.

When the true model is dense, Goeman et al. (2006) have proposed 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) have developed a test using an implicit sparse condition on the projected covariance matrix among predictors. 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 case where the number of unrestricted coefficients is larger than the sample size. Our theory allows for lower but diverging dimensions, and we do not restrict the order of data dimension unless necessary. Using random matrix theory, we relax the distribution and dependence 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.

A natural approach is to estimate the autoregressive 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 non-random 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 works.

In order for the limiting power to be nontrivial, throughout we study the local alternatives with weak signal length converging to zero at a proper rate depending on the data dimension and the sample size. 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 in high dimensions.

We show that our proposed test, a unweighted quadratic test, has the optimal asymptotic power against all considered competitors, when the direction of the regression coefficients is free in the eigenbasis of the sample covariance matrix. Roughly, a coefficient vector is free if its direction is unrelated to the spectral information of sample covariance matrix; 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. (2019) 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 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.

Our asymptotic theory uses the observable empirical spectral distribution of the large sample covariance matrix rather than its limit. Hence, our results are data adaptive and remain true even when the empirical spectral distribution diverges. The asymptotic results remain true by substituting in the limiting distribution (if there is any), for the data generating processes in, e.g., Marčenko and Pastur (1967), Yin (1986), Silverstein and Bai (1995), Silverstein (1995), Zhang (2006), 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 defer all the mathematical proofs to the end. 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 $\mathrm{tr}(A)$, its spectral norm by $\|A\|_{sp} = \sup_{u\neq \mathbf{0}} \frac{\|u^T A\|_2}{\|u\|_2}$, its Frobenius norm by $\|A\|_F = \sqrt{\mathrm{tr}(A^T A)}$. When A is symmetric, we denote its smallest and largest absolute eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively. We denote by '$\xrightarrow{P}$' the convergence in probability and by '$\xrightarrow{d}$' the convergence in distribution.

2 Asymptotic theory

Suppose we observe responses $y_1, \dots, y_n \in \mathbb{R}$ and the initial values $\{y_0, y_{-1}, \dots, y_{1-d}\}$ 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 $x_t = (x_{t,1}, \dots, x_{t,p})^T \in \mathbb{R}^p$ are observable exogenous variables with unknown coefficients $\beta = (\beta_1, \dots, \beta_p)^T$; ε_t are unobservable regression errors with zero mean and unknown variance. Let $z_t := (1, y_{t-1}, \dots, y_{t-d})^T \in \mathbb{R}^{d+1}$ collect the nuisance variables whose coefficients $\theta = (\theta_0, \theta_1, \dots, \theta_d)^T$ we always estimate. Therefore, we can rewrite the model into a general form given by

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

We shall discuss later on how to extend our theory beyond the autoregressive model in the subsection 2.3, which may be useful for cross sectional data sets as well.

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

$$H_0: \ \beta = \mathbf{0}_p, \tag{2.3}$$

where $\mathbf{0}_p$ denotes the p-dimensional all zeros vector. The zeros are not special. By rewriting the linear regression model appropriately, 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 properly. We use the zeros as null values for presentation convenience. We consider a large dimensional asymptotic regime with a diverging dimension p:

Assumption 1. The dimension $p = p(n) \to \infty$ as the sample size $n \to \infty$.

The concentration ratio p/n, usually larger than 1, plays an important role in our asymptotic theory. Unless specified otherwise, we allow an arbitrary rate of dimension p while we typically require that $p/n \to c \in (0, \infty)$ for technical reasons. Our simulations in the next section support our asymptotic approximations over a wide range of p/n in finite samples. To avoid unnecessary complications, we assume that the nuisance dimension d is fixed although our proofs actually allow 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 adapted sequence $\{\varepsilon_t, \mathcal{F}_t\}$ introduces a martingale difference sequence (array) such that $\mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$, where $\mathcal{F}_{t-1}$ is the product of sigma-algebras generated by $\{\varepsilon_s : s \leq t-1\}$, $\{z_s : 1 \leq s \leq t\}$ and $\{x_s : 1 \leq s \leq n\}$. Furthermore, the errors have a common conditional variance $\mathbb{E}\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] = \sigma_n^2 \in (0, \infty)$, and $\mathbb{E}\left[|\varepsilon_t^2/\sigma_n^2 - 1|^2 \mid \mathcal{F}_0\right] \leq \kappa_n$ almost surely with κ_n measurable by $\mathcal{F}_0$.

(b) The regressors $(z_t^T, x_t^T)^T$ are identically distributed over index $t = 1, \dots, n$, 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 $1 - \theta_1\lambda - \theta_2\lambda^2 - \dots - \theta_d\lambda^d = 0$ are greater than 1 in absolute value.

Similar to Lam (2016), in condition (a) we introduce a (stochastic) bound of the fourth moment for each n but allow it to diverge if necessary; see also Phillips and Jin (2014) for similar fourth moment conditions. With the cost of greater complications, it is reasonable to reduce the order of moment conditions to be $\mathbb{E}\left[|\varepsilon_t^2/\sigma_n^2 - 1|^{1+\iota}|\mathcal{F}_0\right] \leq \kappa_n^{(1+\iota)/2}$ for some $\iota > 0$ under (slightly) stronger regularity conditions on the predictors; see Remark 1 below. Condition (b) assumes the existence of the second moments among regressions. 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 null distribution. Then, we introduce the free alternatives and derive the asymptotic power of 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.

2.1 Testing the null hypothesis

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

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

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

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

that is, in matrix form,

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

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

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

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

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

Summing up the squared sample moments yields

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

Let $A_n = \underline{S}_n - \text{diag}\,(\underline{S}_n)$, where $\text{diag}(\underline{S}_n)$ 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

$$Q_n := \frac{1}{\sqrt{2}\, \|A_n\|} e^T A_n e.$$

Rewriting Q_n into an (approximate) martingale form

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

and applying the martingale limit theorems, we shall establish the asymptotic distribution of Q_n under the null hypothesis (2.3). Denote the cross product of the scaled design matrix $\frac{1}{\sqrt{n}} Z$ by

$$\widehat{\Omega} = \frac{1}{n} Z^T Z = \frac{1}{n} \sum_{t=1}^{n} z_t z_t^T .$$

Let $\widetilde{A}_n$ be a strictly lower triangular matrix with same lower diagonal elements of A_n.

Theorem 1 (Oracle test)**.** *Under Assumptions 1 and 2, suppose moreover that:*

(i) The cross product matrix $\widehat{\Omega}$ has eigenvalues bounded away from 0.

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

(iii) $\kappa_n = O_{\mathbb{P}}(1)$*, or more generally* $\kappa_n \left\| \text{diag}(\widetilde{A}_n^T \widetilde{A}_n) \right\|^2 = o_{\mathbb{P}}\left(\|A_n\|^4 \right)$.

(iv) $\sum_{t=1}^{n-l} |A_n(t+l, t)| = o_p(n^{1/2} \|A_n\|)$ *for each* $l \geq 1$.

Then $Q_n / \sigma_n^2 \xrightarrow{d} \mathcal{N}(0, 1)$ *under the null hypothesis* (2.3).

Remark 1. When $\mathbb{E}\left[|\varepsilon_t^2/\sigma_n^2 - 1|^{1+\iota} \mid \mathcal{F}_0\right] \le \kappa_n^{(1+\iota)/2}$ for some $\iota > 0$ and κ_n measurable by $\mathcal{F}_0$, the results remain true without the fourth moment condition if condition (iii) is replaced by $\kappa_n \left\|\operatorname{diag}(\widetilde{A}_n^T \widetilde{A}_n)\right\|_{1+\iota} = o_{\mathbb{P}}\left(\|A_n\|^2\right)$, where $\|A\|_{1+\iota} = (\sum_{i,j} |A(i,j)|^{1+\iota})^{1/(1+\iota)}$ for any matrix A.

Condition (i) is a trivial identification condition that holds, for example, when $\widehat{\Omega} \xrightarrow{P} \Omega$ for some positive definite matrix Ω. Note that our test statistic is affine invariant with respect to the regressors z_t. The theorem remains true if the condition holds for z_t after some non-singular affine transformation. Condition (ii) is weaker than the typical Noether's condition $\|A_n\|_{sp} = o_{\mathbb{P}}(\|A_n\|)$ for the asymptotic normality of quadratic forms with zero diagonals; see, e.g., de Jong (1987), Wu and Shao (2007) and many references therein. The second part of condition (iii) is redundant if $\kappa_n = O_{\mathbb{P}}(1)$, as

$$\left\|\operatorname{diag}(\widetilde{A}_n^T \widetilde{A}_n)\right\| \le \left(\operatorname{tr}\left(\widetilde{A}_n^T \widetilde{A}_n\right)^2\right)^{1/2} \le \left\|\widetilde{A}_n^T \widetilde{A}_n\right\|_{sp}^{1/2} \left\|\widetilde{A}_n\right\| = o_{\mathbb{P}}\left(\|A_n\|^2\right),$$

where the first step follows the majorization property of eigenvalues against diagonal elements, the second step follows a simple max inequality and the last step follows from the second part of condition (ii). Condition (iv) is a weak mixing condition removing some spurious correlation between $\{x_t\}$ and the autoregressors in large samples; by Cauchy–Schwarz inequality it is weaker than requiring $\sum_{t=1}^{n-l} A_n^2(t+l,t) = o_p(\|A_n\|^2)$ like in Wu and Shao (2007), equation (13). In the regular cases (such that in Theorem 2 below), $\|A_n\|$ usually diverges at the rate of $\sqrt{p}$, and therefore we can show that an i.i.d. sequence $\{x_t\}$ naturally satisfies the condition if $\lambda_{\max}(\Sigma) = o_p(n^{1/2})$ and $\lambda_{\max}(S_n) = o_{\mathbb{P}}(n^{1/2})$. In practice, as the matrix A_n is observable, the condition may be justified by the data.

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

$$Q_n > \widehat{\sigma}^2 \Phi^{-1}(1-\alpha), \tag{2.4}$$

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 $\widehat{\sigma}^2/\sigma_n^2 \xrightarrow{P} 1$. 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 error variance in high dimensional regression model is beyond the scope of this paper. Here, given our null hypothesis, we propose to simply use the restricted least-squares estimator

$$\widehat{\sigma}^2 = \frac{1}{n-(d+1)} e^T e, \tag{2.5}$$

where e is the null residual vector as above. Our estimator is convenient to implement and remains feasible even when $p > n$. To summarize, we provide the following corollary for our feasible test.

Corollary 1 (Feasible test)**.** *Under the conditions of Theorem 1 and using the variance estimator (2.5), our feasible test is asymptotically correct, that is,* $\mathbb{P}\left(Q_n > \widehat{\sigma}^2 \Phi^{-1}(1-\alpha)\right)$ *converges to* α *for all* $\alpha \in (0,1)$*.*

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. Our local alternative is therefore given by

$$H_1: \ h^2 := \lim_{n\to\infty} \frac{n}{\sqrt{p}} \|\beta\|^2 \in (0,\infty), \tag{2.6}$$

while under the null hypothesis (2.3) we have $h^2 = 0$.

As announced 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 $\Lambda_n = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ 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 random coordinates of the standardized 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. 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 $F^{S_n}(x; \beta) - F^{S_n}(x) \xrightarrow{a.s.} 0$ 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 asymptotically uniform distributed over orthogonal matrices; see, e.g., Bai et al. (2007), Pan (2014), Xia et al. (2013), Xi et al. (2019), and Bai and Silverstein (2010), Chapter 10 for more discussions of this interesting property.

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

(i) The concentration ratio $p/n \to c \in (0, \infty)$.

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

(iii) $\operatorname{var}\left(x_t^T x_t\right) = o_{\mathbb{P}}(n^2)$, 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,

$$\frac{1}{n} \sum_{t=1}^{n} \left(\underline{S}_n(t,t) - \frac{1}{n} \operatorname{tr}(\underline{S}_n)\right)^2 \xrightarrow{P} 0. \tag{2.7}$$

(iv) $\mathbb{E}(x_t^T \xi_n)^4$ is bounded for every sequence of unit vectors ξ_n.

Under the null (2.3) with $h = 0$ or under the local alternatives (2.6),

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

where

$$\varpi_n = \sqrt{\int x^2 dF^{S_n}(x) - \frac{p}{n}\left(\int x dF^{S_n}(x)\right)^2} = \sqrt{\frac{n}{p}\cdot \mathrm{var}\left(\Lambda_n | F^{\underline{S}_n}\right)} > 0, \qquad (2.8)$$

and Λ_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 tending to 1.

Remark 2 (General alternatives). The results remain true for non-free alternatives, that is, the alternative beyond Assumption 3, with a general form

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

which depends on the direction of the underlying coefficients β. As this general form is not tractable to produce interesting theory here, we leave it for future study.

We offer some remarks on the conditions. Condition (i) is a standard asymptotic regime 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). While condition (ii) rules out the case $p/n \to \infty$ in general (Chen and Pan, 2012), our simulations in the next section suggest that our asymptotic approximation perform well for a wide range of p/n in finite samples.

Condition (ii) implies that the support of the empirical spectral distributions of Σ, S_n and $\mathbb{E}(\bar{x}-\mu)(\bar{x}-\mu)^T$ are all bounded in n (with arbitrarily high probability). This means that the predictors may 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 independent models. In practice, one may calculate $\lambda_{\max}(S_n)$ and justify the condition from the data. The last part of the condition controls the de-meaning effect in our estimation. It is trivial for independent

x_t's with $\lambda_{\max}\left(\mathbb{E}\left[\bar{x}\bar{x}^T\right]\right) = \frac{1}{n}\lambda_{\max}(\Sigma) = O(n^{-1})$; we require a much weaker rate in general to allow for time dependence.

Conditions (iii) is a cross-sectional concentration conditions. The first part of the condition is for understanding the probabilistic behavior, 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, but note that they are unnecessary for our theory.

Proposition 1. *The following affine model satisfies condition* (2.7) *in Theorem 2, provided conditions (i) and (ii) therein:*

(a) The random vectors $x_t = \Phi f_t$ are identically distributed 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 $f_t \in \mathbb{R}^{(p+k)}$ are standardized 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 $\mathbb{E}(f_{t,i} | f_{t,l}, l \neq i, l = 1, \ldots, k) = 0$ and $\mathbb{E}\left(|f_{t,i}|^4\right)$ are bounded by some large constant not depending on n for all i.

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. The conditions (i) and (ii) in Theorem 2 simplify some technical arguments but not necessary for the above proposition, and we may relax them with the following spectral conditions: $\|\Sigma\|^2 = o(n^2)$, and $\lambda_{\max}(S_n) \cdot \lambda_{\max}\left(\mathbb{E}\left[\bar{x}\bar{x}^T\right]\right) = o_{\mathbb{P}}(n^2)$.

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 (2.8).*

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

$$f(x) = \frac{1}{\sqrt{2\pi}xc\tau^2}\sqrt{(b-x)(x-a)} \text{ if } x \in (a,b) \text{ and otherwise zero,} \tag{2.9}$$

where $b = \tau^2(1+\sqrt{c})^2$, $a = \tau^2(1-\sqrt{c})^2$ and again $p/n \to c \in (0,\infty)$. In this case, it is easy to 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 high-frequency data, we refer to Zheng and Li (2011) and Xia and Zheng (2018).

Observe that the variance estimator (2.5) is still consistent under the local alternatives (2.6), our feasible test achieves the oracle testing power, in an adaptive sense:

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

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

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 (2.5), our restricted estimator may tend to be (slightly) biased upward and therefore may reduce the finite-sample power relative to the theoretical limit. As noted above, in such case one may improve the power by using 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 as future works.

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

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

Standardizing the residuals by $\widehat{\sigma}^2$ gives the F-test when $z_t = 1$; see, e.g., 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 unweighted test statistic, one may again remove the diagonal elements in the weighting matrix, standardize the quadratic form, and consider the test statistic

$$Q_n = \frac{1}{\sqrt{2}\,\|A_n\|} e^T A_n e$$

where

$$A_n = \frac{1}{n}\widetilde{X} S_n^{-1} \widetilde{X}^T - \operatorname{diag}\left(\frac{1}{n}\widetilde{X} S_n^{-1} \widetilde{X}^T\right)$$

with a slight abuse of the notation. It is easy to 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 $A_n(\delta) = W(\delta) - \operatorname{diag}(W(\delta))$ and standardizing the quadratic form again yields the test statistic

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

It is clear that our unweighted test and the aforementioned 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, it is easy to verify that the null distribution remains true by substituting the weighting matrix $A_n = A_n(\delta)$ everywhere, provided being well-defined in large samples.

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 (2.6).

Theorem 3. *Let $\delta : [0, \infty) \to \mathbb{R}$ be an arbitrary function continuous on $(0, \infty)$ and with an arbitrary constant value $\delta(0)$. 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 (2.3) with $h = 0$ or the local alternatives (2.6),*

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

where

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

with ϖ_n and Λ_n from Theorem 2, if provided that $\operatorname{var}(\Lambda_n\delta(\Lambda_n)|F^{\underline{S}_n})$ is bounded away from 0 almost surely.

Like in Theorem 2, we require a concentration condition on the diagonal elements of $W_n(\delta)$. It is easiest to 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:

Proposition 2. *The sample variance of the diagonal elements of $W_n(\delta)$ tends to zero for all function δ continuous on $(0, \infty)$ with an arbitrary constant value $\delta(0)$, as well as F^{S_n} tends to a limit F solving the equation in Silverstein (1995) when:*

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

(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 (2.5) is still consistent under the local alternatives (2.6). The asymptotic power of our feasible weighted test follows:

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

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

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

Now, maximizing the asymptotic power of our test is equivalent to maximizing the asymptotic departure $\varpi_n(\delta)$ with respect to δ. Noting the (random) correlation coefficient is smaller than 1 almost surely unless $\delta(x)$ is a constant for $x > 0$. In other words, our proposed unweighted test maximizes the the asymptotic test power, for the local free alternatives.

2.3 Towards a more general model

In this section, we discuss how to generalize the results beyond autoregressive models. Specifically, we consider the general model (2.2) with some general, possibly nonstationary, nuisance variables $z_t = (1, z_{t,1}, \ldots, z_{t,d})$ such that

$$z_{t,i} = \alpha_i + \sum_{l=1}^{t-1} \psi_i(l)\varepsilon_{t-l} + v_{t,i}$$

with mean $\alpha_i = \mathbb{E}z_{t,i}$, nonrandom moving average coefficients $\psi_i(l)$, and exogenous shocks $v_{t,i}$ (possibly dependent on $\{x_{t,i}\}$) such that the regression error $\{\varepsilon_t\}$ is a martingale difference sequence adapted to the product of the natural filtration $\{\mathcal{F}_t\}_{t\geq 0}$ in Assumption 2 and the sigma algebra generated by $\{v_{t,i} : t = 1, \ldots, n, i = 1, \ldots, d\}$. Our autoregressive regression model (2.1) is a special case with $v_{t,i}$ as linear combinations of the initial values $\{\varepsilon_t : t \leq 0\}$, that is, measurable by $\mathcal{F}_0$. Another interesting case is that $\psi_i(l) = 0$ for all $1 \leq l < n$, when $z_{t,i}$ is one of our exogenous predictors. In general, we allow the nuisance variables contains a mixture of lagged information and exogenous shocks. 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 if the lagged coefficients $\psi_i(l)$ are absolutely summable, that is, $\sum_{l=1}^{\infty} |\psi_i(l)| < \infty$ for all $i = 1, \ldots, d$.*

Our conditions are sufficient, but not always necessary. When the nuisance input vector z_t contains only exogenous predictors, that is, $\psi_i(l) = 0$ for all l and i, our results remain true without condition (iv) in Theorem 1. More specifically, our theory holds under a weaker form of condition (iv) given by

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

Theorem 5. *Under the conditions of Theorem 4, Theorem 2 and Theorem 3 remain true if the spectrum of the autocorrelation matrix $\{v_{t,i} : t = 1, \ldots, n\}$ is bounded in probability over the sample paths of X, that is,*

$$\lambda_{\max} \left(\mathbb{E} \left(\mathbf{v}_i \mathbf{v}_i^T \mid X \right) \right) = o_{\mathbb{P}}(n) \tag{2.10}$$

with $\mathbf{v}_i = (v_{1,i}, \ldots, v_{n,i})^T$ for all $i = 1 \ldots, d$.

Inverting the autoregressive process (2.1) into a moving average form and by direct calculations, we can show that $\lambda_{\max} \left(\mathbb{E} \left(\mathbf{v}_i \mathbf{v}_i^T \mid X \right) \right) = O_{\mathbb{P}} \left(1 + n \|\beta\|^2 \right) = O_{\mathbb{P}}(\sqrt{p})$ under our local alternatives. Therefore, the spectral condition (2.10) is natural for the autoregressive models, when p is comparable to n (or up to the order of n^2 exclusive). Note that $\left\| \left(\mathbb{E}(\mathbf{v}_i \mathbf{v}_i^T | X) \right)^{1/2} \right\|^2 = \mathrm{tr}(\mathbb{E}(\mathbf{v}_i \mathbf{v}_i^T | X)) = \sum_{t=1}^{n} \mathbb{E}(v_{t,i}^2 | X) = O_{\mathbb{P}} \left(\sum_{t=1}^{n} \mathbb{E} v_{t,i}^2 \right) = O_{\mathbb{P}}(n)$. A sufficient condition for (2.10) is a Noether type condition that $\lambda_{\max}(\Omega_X) = O_{\mathbb{P}}(\|\Omega_X\|)$ where $\Omega_X = \left(\mathbb{E}(\mathbf{v}_i \mathbf{v}_i^T | X) \right)^{1/2}$.

3 Simulation

In this section we study the empirical size and the power performance of the proposed test using a Monte Carlo experiment. 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 demean the predictors in each sample and estimate the intercept. We fix the order of autoregressive $d = 3$, and use the autoregressive coefficients $(\theta_1, \theta_2, \theta_3) = (0.3, 0.08, 0.1)$ calibrated from our empirical application. We simulate independent errors from the standardized student t-distribution with five degrees of freedom that are independent of the regressors. We consider exogenous variables from three data generating processes:

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

(2) $x_t = T^{1/2} v_t$ with the error covariance matrix T equaling to the Toeplitz matrix with i, j-th elements $\rho^{|i-j|}$ with $\rho = 0.5$;

(3) $x_t = T^{1/2} v_t - 0.1 T^{1/2} v_{t-1}$ follows a high dimensional first-order moving average process, where the lagged coefficient -0.1 is calibrated from the economic data in our empirical analysis.

The first data generating process (DGP) is a simple i.i.d. model. We introduce the cross-sectional dependence in the second DGP, and the time dependence as well in the third DGP.

For each pair of (p, n), we generate the direction of regression coefficients, say, ξ uniformly over the $\mathbb{R}^p$ unit sphere. We compare the results for a common asymptotic departure

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

corresponding to the coefficient vector

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

The ratio between p and n, rather than their magnitudes, plays a key role in our limit theory. Hence, rather than selecting data sizes explicitly, we control the concentration ratio $p/n = \frac{1}{5}, \frac{1}{2}, 1, 2, 5$. When p and n are comparable, we report the results for the local alternatives with a weak order of $\sqrt{p}/n = 0.05$ and 0.1 in Table 1 and 2 respectively. Throughout we choose a significance level of 5%, and generate 2000 replications.

Table 1: Size and power of the tests at level $\alpha = 5\%$ with $\sqrt{p}/n = 0.05$ and $p/n = \frac{1}{5}, \frac{1}{2}, 1, 2, 5$ for (IID) i.i.d. model, (CSD) cross-sectionally dependent model, (MA1) first-order moving average model.

			Estimated variance			True variance		
p	n	p/n	IID	CSD	MA1	IID	CSD	MA1
			$H_0 : \|\beta\|^2 = 0$					
16	80	1/5	4.9%	7.3%	6.2%	4.9%	6.6%	6.1%
100	200	1/2	5.7%	5.5%	5.8%	5.4%	5.8%	5.5%
400	400	1	4.9%	5.7%	5.2%	5.0%	5.5%	5.6%
1600	800	2	5.0%	5.8%	4.7%	5.1%	5.8%	5.0%
10000	2000	5	4.9%	5.0%	4.7%	5.2%	4.7%	4.6%
			$H_a^1 : \|\beta\|^2 = 1 \times \frac{\sqrt{2p}}{n\varpi_n}$					
16	80	1/5	23.9%	30.8%	29.9%	25.1%	32.0%	30.1%
100	200	1/2	24.5%	33.0%	29.1%	26.8%	34.1%	30.6%
400	400	1	26.6%	26.5%	25.7%	28.4%	28.5%	27.6%
1600	800	2	26.0%	24.8%	25.8%	30.1%	26.1%	28.1%
10000	2000	5	25.6%	23.2%	22.6%	28.7%	25.1%	25.8%
			$H_a^2 : \|\beta\|^2 = 2 \times \frac{\sqrt{2p}}{n\varpi_n}$					
16	80	1/5	44.6%	53.1%	52.1%	49.1%	54.8%	54.4%
100	200	1/2	51.2%	59.1%	57.4%	56.1%	62.4%	61.6%
400	400	1	53.6%	55.7%	54.0%	60.1%	60.5%	59.3%
1600	800	2	54.9%	55.2%	54.6%	62.2%	60.4%	59.6%
10000	2000	5	54.4%	53.3%	50.0%	61.5%	58.4%	57.2%
			$H_a^3 : \|\beta\|^2 = 5 \times \frac{\sqrt{2p}}{n\varpi_n}$					
16	80	1/5	82.9%	88.1%	88.1%	87.8%	91.1%	90.7%
100	200	1/2	92.2%	94.9%	94.9%	95.9%	96.9%	96.9%
400	400	1	96.2%	95.7%	96.1%	98.3%	97.5%	98.3%
1600	800	2	97.0%	97.0%	96.8%	98.9%	98.8%	98.9%
10000	2000	5	97.3%	98.0%	97.5%	99%	99.1%	98.9%

Table 2: Size and power of the tests at level $\alpha = 5\%$ with $\sqrt{p}/n = 0.1$ and $p/n = \frac{1}{5}, \frac{1}{2}, 1, 2, 5$ for (IID) i.i.d. model, (CSD) cross-sectionally dependent model, (MA1) first-order moving average model.

			Estimated variance			True variance		
p	n	p/n	IID	CSD	MA1	IID	CSD	MA1
			$H_0 : \|\beta\|^2 = 0$					
4	20	1/5	4.6%	4.8%	4.5%	4.6%	4.7%	4.7%
25	50	1/2	5.8%	5.7%	4.9%	6.9%	6.0%	4.7%
100	100	1	4.3%	5.6%	5.9%	4.6%	6.2%	5.9%
400	200	2	5.5%	5.5%	4.9%	5.6%	6.0%	5.2%
2500	500	5	5.4%	5.4%	4.8%	6.0%	5.5%	5.1%
			$H_a^1 : \|\beta\|^2 = 1 \times \frac{\sqrt{2p}}{n\varpi_n}$					
4	20	1/5	17.0%	34.6%	32.2%	19.6%	37.0%	34.3%
25	50	1/2	21.6%	29.5%	29.0%	26.0%	32.9%	31.7%
100	100	1	22.4%	31.6%	29.8%	26.9%	34.8%	34.1%
400	200	2	22.4%	20.2%	17.1%	27.7%	23.7%	21.1%
2500	500	5	22.0%	27.0%	20.0%	27.1%	32.2%	24.8%
			$H_a^2 : \|\beta\|^2 = 2 \times \frac{\sqrt{2p}}{n\varpi_n}$					
4	20	1/5	27.7%	55.9%	54.2%	34.0%	61.6%	60.3%
25	50	1/2	38.7%	49.9%	51.4%	47.3%	56.8%	58.2%
100	100	1	41.6%	57.4%	56.2%	52.8%	65.7%	63.0%
400	200	2	43.0%	37.3%	36.6%	54.0%	45.1%	45.3%
2500	500	5	44.6%	54.1%	46.6%	59.8%	63.8%	57.5%
			$H_a^3 : \|\beta\|^2 = 5 \times \frac{\sqrt{2p}}{n\varpi_n}$					
4	20	1/5	51.6%	83.7%	83.9%	64.7%	90.0%	90.3%
25	50	1/2	71.6%	82.2%	83.7%	82.9%	89.9%	90.1%
100	100	1	80.3%	91.5%	90.9%	92.0%	96.3%	95.9%
400	200	2	83.3%	77.5%	76.9%	93.8%	89.0%	87.9%
2500	500	5	87.7%	93.3%	91.1%	96.9%	98.2%	96.9%

The empirical sizes of the feasible test range between 4.3% and 6.2%, all close to the nominal level 5% except one extreme small-sample case of 7.3%. The power improves as the departure level τ_n^2 grows, and stabilizes as the concentration ratio p/n grows. The power are similar across the data generating processes, because of our adaptive choice of signal length. Comparing the tables we observe that the powers are more sensitive to the departure order $\sqrt{p}/n$ in small samples, while they are generally more robust against concentration ratio p/n for smaller departure order $\sqrt{p}/n$.

To detect the power loss due to the restricted variance estimator under the alternatives, we also compare the results with the oracle test using the true error variance $\sigma_n^2 = 1$. While the variance estimation (slightly) reduces the empirical power, the disadvantage becomes more negligible as the data dimensions (and sample size) grow.

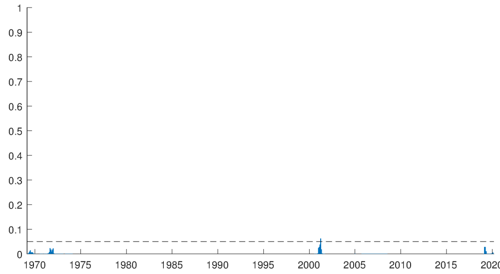
4 Application

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

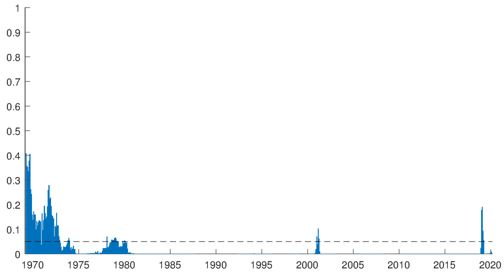
https://research.stlouisfed.org/econ/mccracken/fred-databases/

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

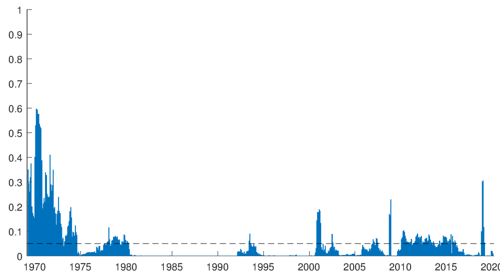
Our data set includes monthly observations of the industrial production index (INDPRO) and 127 other predictors from January, 1959 to February, 2020. We transform the raw datasets into stationary forms and remove the data outliers using the MATLAB codes provided on the above website; see also McCracken 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 $d = 0, 1, 2, 3, 4, 5$ lagged response variables. Note that the standard F test is (almost)



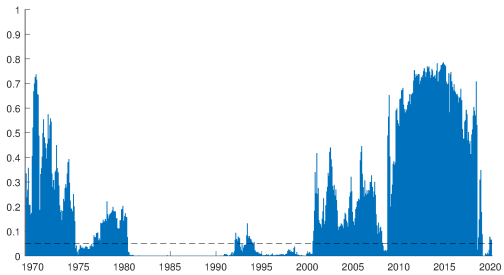
(a) No lag variable



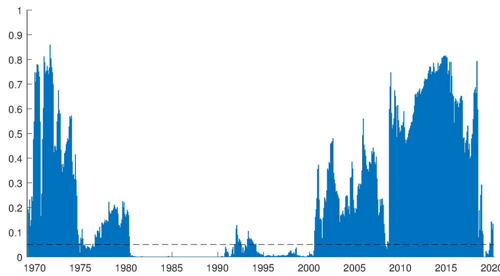
(b) lag $d = 1$



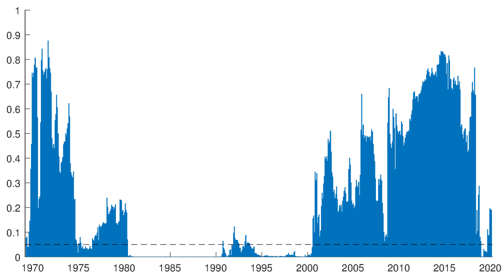
(c) lag $d = 2$



(d) lag $d = 3$



(e) lag $d = 4$



(f) lag $d = 5$

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

degenerate here are therefore not reported, because the number of exogenous variables is larger than (or very close to) the sample size.

We always estimate the autoregressive coefficients as nuisance parameters, and test the overall significance of the exogenous predictors. Figure 1 compares the bar plots of our rolling window p values for different autoregressive order d between 0 and 5. The dashed lines indicate our benchmark significance level $\alpha = 5\%$. Clearly, ignoring the impact of autoregressors shows almost all rejections and therefore overoptimistic outcomes for predictability. By including more autoregressors, the p-values stabilize for $d \geq 3$ and show more interesting time-varying patterns. Overall, the coefficients are jointly significant during a long period between year 1980 and year 2000, and more recently during the first half year of 2019.

5 Proofs

Note that the test statistic is invariant to a non-singular affine transformation of z_t, and does not depend on the variance σ_n^2 . Hence, without loss of generality, we only need to prove the results for the case $\sigma_n^2 = 1$, and $Ez_{t,i}^2 = 1$.

5.1 Proofs of Theorem 1 and Corollary 1

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

$$\begin{aligned} Q_n =& \frac{1}{\sqrt{2}\,\|A_n\|}\epsilon^T(I-P_Z)A_n(I-P_Z)\epsilon \\ =& \frac{1}{\sqrt{2}\,\|A_n\|}\epsilon^T A\epsilon - \frac{\sqrt{2}}{\|A_n\|}\epsilon^T P_Z A\epsilon + \frac{1}{\sqrt{2}\,\|A_n\|}\epsilon^T P_Z A P_Z \epsilon \\ =:& \widetilde{Q}_n - T_1 + T_2, \end{aligned}$$

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

We need some lemmas for our proof in the end. First we list some fundamental inequalities and their useful implications here. The first is a special case of the well-known Burkholder (1973) inequality:

Lemma 1 (Square function inequality)**.** *Let $\{\Delta_t, \mathcal{F}_t\}$ be a martingale difference sequence (array) and $\mathbb{E}\left(\Delta_t^2|\mathcal{F}_0\right)$ exists, then there exists some absolute constant M such that almost surely*

$$\mathbb{E}\left(\left|\sum_{t=1}^{n}\Delta_t\right|^2|\mathcal{F}_0\right) \leq M\sum_{t=1}^{n}\mathbb{E}\left(\Delta_t^2|\mathcal{F}_0\right).$$

The second is an elementary result combining Markov inequality and the law of iterated expectations, and we omit the proof. We will use it to control the asymptotic bounds of perturbation terms.

Lemma 2 (Markov inequality)**.** *For an arbitrary sequence of statistics θ_n and a sub-sigma-algebra $\mathcal{F}$, the Markov inequalty implies that $\theta_n^2 = O_{\mathbb{P}}(E(\theta_n^2|\mathcal{F}))$.*

The third is an asymptotic bound and a concentration inequality for quadratic forms.

Lemma 3 (Concentration inequality for quadratic forms)**.** *Let $\{\varepsilon_t, \mathcal{F}_t : t = 1, \dots, n\}$ be a martingale difference array with common conditional variance $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = 1$, and A be any $n \times n$ real matrix measurable by $\mathcal{F}_0$.*

$$\mathbb{E}\left(|\epsilon^T A\epsilon||\mathcal{F}_0\right) \leq \|\text{diag}(A)\|_1 + \sqrt{2}\|A\|, \ a.s..$$

If further given that $E\left((\varepsilon_t^2-1)^2|\mathcal{F}_0\right) \leq \kappa_n$,

$$\mathbb{E}\left(\left|\epsilon^T A\epsilon - \text{tr}(A)\right|^2|\mathcal{F}_0\right) \leq M\left(\kappa_n\|\text{diag}(A)\|^2 + \|A\|^2\right)$$

for some absolute constant M.

Proof. Let $A = \{A(s,t) : s,t = 1,\dots,n\}$, where $A(s,t)$ is the entry of A in its s-th row and t-th column. Expanding the quadratic form,

$$\epsilon^T A\epsilon = \sum_{t=1}^n \varepsilon_t^2 A(t,t) + \sum_{1\leq s<t\leq n} \varepsilon_s\varepsilon_t\left(A(s,t) + A(t,s)\right) =: T_1 + T_2.$$

By triangle inequality,

$$\mathbb{E}(|T_1||\mathcal{F}_0) \leq \sum_{t=1}^n \mathbb{E}(\varepsilon_t^2|\mathcal{F}_0)|A(t,t)| = \sum_{t=1}^n |A(t,t)| = \|\text{diag}(A)\|_1.$$

Moreover, by a direct calculation and applying Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}(T_2^2|\mathcal{F}_0) = \sum_{1\leq s<t\leq n}(A(s,t)+A(t,s))^2 \leq & 2\sum_{1\leq s<t\leq n}\left(A^2(s,t)+A^2(t,s)\right) \\ =& 2\|A - \text{diag}(A)\|^2. \end{aligned} \tag{5.1}$$

Hence, combining with Jensen's inequality

$$\mathbb{E}(|\epsilon^T A\epsilon|) \leq \mathbb{E}(|T_1|) + \mathbb{E}(|T_2|) \leq \mathbb{E}(|T_1|) + \sqrt{\mathbb{E}(T_2^2)} \leq \|\text{diag}(A)\|_1 + \sqrt{2}\|A - \text{diag}(A)\|$$

The proof for the second part is similar. We expand the quadratic form again

$$\epsilon^T A\epsilon - \text{tr}(A) = \sum_{t=1}^n (\varepsilon_t^2 - 1)A(t,t) + \sum_{1\leq s<t\leq n} \varepsilon_s\varepsilon_t\left(A(s,t)+A(t,s)\right) =: T_3 + T_2.$$

By Lemma 1, for some constant M

$$\begin{aligned}\mathbb{E}(T_3^2|\mathcal{F}_0) \leq & M\sum_{t=1}^n \mathbb{E}(|\varepsilon_t^2 - 1|^2|\mathcal{F}_0)|A(t,t)|^2 \\ \leq & M\kappa_n \cdot \sum_{t=1}^n |A(t,t)|^2 = M\kappa_n \cdot \|\text{diag}(A)\|^2.\end{aligned}$$

Hence, recalling (5.1) and using Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left(|\epsilon^T A\epsilon - \mathrm{tr}(A)|^2|\mathcal{F}_0\right) \leq & M\mathbb{E}\left(|T_3|^2 + |T_2|^2|\mathcal{F}_0\right) \\ \leq & M\mathbb{E}\left(|T_3|^2|\mathcal{F}_0\right) + M\left(\mathbb{E}\left(|T_3|^2|\mathcal{F}_0\right)\right)^{\frac{1+\iota}{2}} \\ \leq & M\kappa_n \cdot \|\mathrm{diag}(A)\|^2 + M\|A - \mathrm{diag}(A)\|^2\end{aligned}$$

This completes the proof. □

The next inequality is similar to Lemma S.3 in Lam (2016). We add a necessary condition (which is missing therein), and provide the detailed proof for completeness.

Lemma 4 (A trace inequality)**.** *For an arbitrary symmetric $p \times p$ matrix A and a non-negative definite $p \times p$ matrix B,*

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

Proof. Slightly abusing the notation, let $U\Lambda U^T$ be a spectral decomposition of A where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is a diagonal matrix consists of the eigenvalues and $U = [u_1, \ldots, u_p]$ is an orthogonal matrix with columns being the corresponding eigenvectors. Hence, $A = \sum_{j=1}^p \lambda_j u_j u_j^T$, $\sum_{j=1}^p u_j u_j^T = UU^T = I$ and $\|A\|_{sp} = \max_j |\lambda_j|$. Then, noting that B is nonnegative definite,

$$\begin{aligned}|\mathrm{tr}(AB)| = & \left|\sum_{j=1}^p \lambda_j u_j^T B u_j\right| \\ \leq & \sum_{j=1}^p |\lambda_j| u_j^T B u_j \leq \|A\|_{sp} \sum_{j=1}^p u_j^T B u_j = \|A\|_{sp}\,\mathrm{tr}\left(B\sum_{j=1}^p u_j u_j^T\right) = \|A\|_{sp}\,\mathrm{tr}\,(B)\,.\end{aligned}$$

□

The next lemma controls the effect of unobserved past errors in our autoregressive models.

Lemma 5. *All solutions of the homogeneous linear difference equation*

$$v_t = \theta_1 v_{t-1} + \ldots + \theta_d v_{t-d},\ t \geq d+1. \tag{5.2}$$

is squared summable, that is, $\sum_{t=1}^d v_t^2 \leq C$ for some large common constant C regardless of the initial values.

Proof. Let $\lambda_1, \ldots, \lambda_d$ be the (complex) roots of the polynomial equation $1 - \theta_1\lambda - \theta_2\lambda^2 - \ldots - \theta_d\lambda^d = 0$. For some finite constant $C_1, \ldots, C_d$, it is well known the general solution

$$v_t = C_1\lambda_1^{-t} + \ldots + C_d\lambda_d^{-t},$$

and therefore by triangular inequality

$$|v_t| \leq |C_1||\lambda_1^{-1}|^t + \ldots + C_d|\lambda_d^{-1}|^t.$$

Now, by Cauchy–Schwarz inequality

$$\begin{aligned}\sum_{t=1}^n v_t^2 \leq & \sum_{t=1}^n \left(|C_1||\lambda_1^{-1}|^t + \ldots + C_d|\lambda_d^{-1}|^t\right)^2 \\ \leq & \sum_{t=1}^n \left(|\lambda_1^{-1}|^{2t} + \ldots + |\lambda_d^{-1}|^{2t}\right)\left(C_1^2 + \ldots + C_d^2\right) \leq C,\end{aligned}$$

for a sufficiently large constant C. □

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

Lemma 6. *The asymptotic negligibility condition is satisfied is such a way that*

$$\max_{1\leq t\leq n} E(\Delta_t^2|\mathcal{F}_{t-1}) \xrightarrow{P} 0.$$

Proof. 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. Rewrite the conditional variance into a quadratic form given by

$$E(\Delta_t^2|\mathcal{F}_{t-1}) = \frac{2}{\|A_n\|^2}\left(\sum_{s=1}^{t-1}\varepsilon_s A(s,t)\right)^2 = \frac{1}{\left\|\widetilde{A}_n\right\|^2}\epsilon^T \widetilde{A}_n^T b_t b_t^T \widetilde{A}_n \epsilon.$$

It suffices to show that $\max_{1\leq t\leq n} \epsilon^T \widetilde{A}_n^T b_t b_t^T \widetilde{A}_n \epsilon = O_{\mathbb{P}}\left(\left\|\widetilde{A}_n\right\|^2\right)$. From Lemma 3,

$$\begin{aligned}&\mathbb{E}\left(\left|\epsilon^T \widetilde{A}_n^T b_t b_t^T \widetilde{A}_n \epsilon - b_t^T \widetilde{A}_n \widetilde{A}_n^T b_t\right|^2 |\mathcal{F}_0\right) \\ \leq & M\left(\kappa_n \left\|\mathrm{diag}(\widetilde{A}_n^T b_t b_t^T \widetilde{A}_n)\right\|^2 + \left((b_t^T \widetilde{A}_n \widetilde{A}_n^T b_t)^2 - \left\|\mathrm{diag}(\widetilde{A}_n^T b_t b_t^T \widetilde{A}_n)\right\|^2\right)\right) \\ \leq & M\left(\kappa_n \sum_{s=1}^{t-1}|A_n(s,t)|^4 + (b_t^T \widetilde{A}_n \widetilde{A}_n^T b_t)^2\right).\end{aligned}$$

Summing up over t,

$$\begin{aligned}&\mathbb{E}\left(\max_{1\leq t\leq n}\left|\epsilon^T \widetilde{A}_n^T b_t b_t^T \widetilde{A}_n \epsilon - b_t^T \widetilde{A}_n \widetilde{A}_n^T b_t\right|^2 |\mathcal{F}_0\right) \\ \leq & \sum_{t=1}^n \mathbb{E}\left(\left|\epsilon^T \widetilde{A}_n^T b_t b_t^T \widetilde{A}_n \epsilon - b_t^T \widetilde{A}_n \widetilde{A}_n^T b_t\right|^2 |\mathcal{F}_0\right) \\ \leq & M\left(\kappa_n \sum_{t=1}^n \sum_{s=1}^{t-1}|A_n(s,t)|^4 + \left\|\mathrm{diag}(\widetilde{A}_n \widetilde{A}_n^T)\right\|^2\right).\end{aligned}$$

Furthermore,

$$\sum_{t=1}^{n}\sum_{s=1}^{t-1}|A_n(s,t)|^4 \le \sum_{t=1}^{n}\left(\sum_{s=1}^{t-1}A_n^2(s,t)\right)^2 = \left\|\mathrm{diag}\left(\widetilde{A}_n^T\widetilde{A}_n\right)\right\|^2$$

and therefore

$$\kappa_n\sum_{t=1}^{n}\sum_{s=1}^{t-1}|A_n(s,t)|^4 \le \kappa_n\left\|\mathrm{diag}\left(\widetilde{A}_n^T\widetilde{A}_n\right)\right\|^2 = o_{\mathbb{P}}\left(\|A_n\|^4\right).$$

Note that the diagonal elements are majorized by the eigenvalues (see, e.g., Theorem 4.3.45 in Horn and Johnson, 2012). Combining the trace inequality (Lemma 4),

$$\left\|\mathrm{diag}(\widetilde{A}_n\widetilde{A}_n^T)\right\|^2 \le \mathrm{tr}\left(\widetilde{A}_n\widetilde{A}_n^T\right)^2 = \mathrm{tr}\left(\widetilde{A}_n^T\widetilde{A}_n\right)^2 \le \left\|\widetilde{A}_n^T\widetilde{A}_n\right\|_{sp}\left\|\widetilde{A}_n\right\|^2 = O(\|A_n\|^4).$$

It follows from Lemma 2 that

$$\max_{1\le t\le n}\left|\epsilon^T\widetilde{A}_n^T b_t b_t^T\widetilde{A}_n\epsilon - \mathrm{tr}\left(\widetilde{A}_n^T b_t b_t^T\widetilde{A}_n\right)\right|^2 = o_{\mathbb{P}}\left(\|A_n\|^4\right),$$

or equivalently

$$\max_{1\le t\le n}\left|\epsilon^T\widetilde{A}_n^T b_t b_t^T\widetilde{A}_n\epsilon - b_t^T\widetilde{A}_n\widetilde{A}_n^T b_t\right| = o_{\mathbb{P}}\left(\|A_n\|^2\right).$$

The rest follows by the triangular inequality and

$$\max_{1\le t\le n}\left|b_t^T\widetilde{A}_n\widetilde{A}_n^T b_t\right| \le \left\|\widetilde{A}_n\widetilde{A}_n^T\right\|_{sp} = o_{\mathbb{P}}(\|A_n\|^2),$$

by the definition of spectral norm. □

Lemma 7. *The conditional variance for the martingale converges to 1, that is,*

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

Proof. Noting that $\mathrm{tr}\left(\widetilde{A}_n^T\widetilde{A}_n\right) = \left\|\widetilde{A}_n\right\|^2 = \|A\|^2/2$ and applying Lemma 3,

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{2}{\|A_n\|^2}\sum_{t=1}^{n}\left(\sum_{s=1}^{t-1}\varepsilon_s\frac{1}{n}\widetilde{x}_s^T\widetilde{x}_t\right)^2 - 1\right)^2|\mathcal{F}_0\right)\\
&= \mathbb{E}\left(\left(\frac{2}{\|A_n\|^2}\epsilon\widetilde{A}_n^T\widetilde{A}_n\epsilon - 1\right)^2|\mathcal{F}_0\right)\\
&\le M\left(\kappa_n\frac{\left\|\mathrm{diag}(\widetilde{A}_n^T\widetilde{A}_n)\right\|^2}{\|A_n\|^4} + \frac{\left\|\widetilde{A}_n^T\widetilde{A}_n - \mathrm{diag}(\widetilde{A}_n^T\widetilde{A}_n)\right\|^2}{\|A_n\|^4}\right) \xrightarrow{P} 0.
\end{aligned}$$

The rest follows from Lemma 2. □

Our last lemma bounds the controls the effect of nuisance variables on the residuals.

Lemma 8. $\epsilon^T P_Z \epsilon = O_{\mathbb{P}}(d/\lambda_{\min}(\widehat{\Omega}))$.

Proof. Expanding the quadratic form,

$$\frac{1}{n}\epsilon^T Z Z^T \epsilon = \frac{1}{n}\sum_{t=1}^n \varepsilon_t^2 z_t^T z_t + \frac{1}{n}\sum_{t\neq s} \varepsilon_t \varepsilon_s z_t^T z_s.$$

Taking the expectation on both sides and using the law of iterated expectations,

$$\mathbb{E}\left(\frac{1}{n}\epsilon^T Z Z^T \epsilon\right) = \frac{1}{n}\sum_{t=1}^n \mathbb{E}(z_t^T z_t) = \sum_{i=1}^d \mathbb{E}(z_{t,i}^2) = O(d).$$

It is follows by Lemma 2 that $\frac{1}{n}\epsilon^T Z Z^T \epsilon = O_{\mathbb{P}}(d)$. Finally,

$$\epsilon^T P_Z \epsilon = \frac{1}{n}\epsilon^T Z \widehat{\Omega}^{-1} Z \epsilon \leq \lambda_{\min}^{-1}(\widehat{\Omega}) \cdot \frac{1}{n}\epsilon^T Z Z^T \epsilon = O_{\mathbb{P}}(1/\lambda_{\min}(\widehat{\Omega})) \cdot O_{\mathbb{P}}(d),$$

using the definition of spectral norm. □

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 $\widetilde{Q}_n \xrightarrow{d} N(0,1)$. As the conditional variance converges in Lemma 7, it remains to verify the conditional Lindeberg condition, that is, $\sum_{t=1}^n \mathbb{E}(\Delta_t^2 \mathbb{1}(|\Delta_t| > \eta)|\mathcal{F}_{t-1}) \xrightarrow{P} 0$ for all $\eta > 0$. Let $\eta > 0$, and we have

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

Using the law of iterated expectations,

$$\mathbb{E}\left(\sum_{t=1}^n \mathbb{E}(\Delta_t^2|\mathcal{F}_{t-1}, \Delta_t^2 > \eta^2)|\mathcal{F}_0\right) = \sum_{t=1}^n \mathbb{E}(\Delta_t^2|\mathcal{F}_0) = 1,$$

and therefore $\sum_{t=1}^n \mathbb{E}(\Delta_t^2|\mathcal{F}_{t-1}, \Delta_t^2 > \eta^2) = O_{\mathbb{P}}(1)$ by Lemma 2. On the other hand, using Markov inequality and Lemma 6,

$$\max_{1\leq t\leq n} P(\Delta_t^2 > \eta^2|\mathcal{F}_{t-1}) \leq \frac{1}{\eta^2}\max_{1\leq t\leq n} E(\Delta_t^2|\mathcal{F}_{t-1}) \xrightarrow{P} 0.$$

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

In the following, we shall show that $T_1 \xrightarrow{P} 0$ and $T_2 \xrightarrow{P} 0$. Decompose the weighting matrix $A_n = \widetilde{A}_n + \widetilde{A}_n^T$, where $\widetilde{A}_n$ contains the (non-zero) entries below the diagonal and $\widetilde{A}_n$ contains the (non-zero) entries above the diagonal. We can further expand

$$T_1 = \frac{2}{\|A_n\|}\epsilon^T P_Z \widetilde{A}_n^T \epsilon + \frac{2}{\|A_n\|}\epsilon^T P_Z \widetilde{A}_n \epsilon =: T_{1,1} + T_{1,2}.$$

By Cauchy–Schwarz inequality,

$$\begin{aligned} T_{1,1}^2 = \frac{4}{\|A_n\|^2}\left(\epsilon^T P_Z \widetilde{A}_n^T \epsilon\right)^2 \leq & 4\epsilon^T P_Z \epsilon \cdot \frac{1}{\|A_n\|^2}\epsilon^T \widetilde{A}_n P_Z \widetilde{A}_n^T \epsilon \\ = & \left(\lambda_{\min}(\widehat{\Omega})\right)^{-1} O_{\mathbb{P}}(d) \cdot \frac{1}{\|A_n\|^2}\epsilon^T \widetilde{A}_n P_Z \widetilde{A}_n^T \epsilon, \end{aligned} \tag{5.3}$$

where we apply Lemma 8 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

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

where the last step follows from the trace inequality in Lemma 4. Using the exchangability between trace and expectation operations,

$$\mathbb{E}\left(\operatorname{tr}\left(\mathbb{E}(ZZ^T|X)\right)\right) = \operatorname{tr}(\mathbb{E}(ZZ^T)) = nd,$$

and therefore $\operatorname{tr}\left(\mathbb{E}(ZZ^T|X)\right) = O_p(nd)$ by Lemma 2 as ZZ^T is non-negative definite. Collecting all bounds and substituting in (5.3) yields that

$$T_{1,1}^2 = \left(\lambda_{\min}(\widehat{\Omega})\right)^{-2} O_{\mathbb{P}}(d^2) \frac{\left\|\widetilde{A}_n^T \widetilde{A}_n\right\|_{sp}}{\|A_n\|^2} \xrightarrow{P} 0.$$

Similarly, we can show that

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

but we need more techniques for a further upper bound. We invert the autoregressive process (under the null) into a moving average form given by

$$y_t = \alpha + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\alpha = \theta_0 \cdot \sum_{j=0}^{\infty} \psi_j$ and the sequence $\{\psi_j\}$ in absolutely summable, that is, $\sum_{j=0}^{\infty} |\psi_j| < \infty$ by Proposition 6.3 in Hayashi (2000). Now, for $i = 1, \ldots, d$, let the vector of lagged observations be

$$\mathbf{y}_{-i} = [y_{1-i}, \ldots, y_{t-i}, \ldots, y_{n-i}]^T = \alpha \mathbf{1}_n + \Psi_i \epsilon + \mathbf{v}_i$$

where

$$\Psi_i = \sum_{j=0}^{\infty} \psi_j L_n^{i+j} = \sum_{j=0}^{n-i} \psi_j L_n^{i+j} \tag{5.4}$$

and L_n is the $n \times n$ lower shift matrix with ones on the subdiagonal and zeros elsewhere, and $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,n})$ only depends the initial values $\{\varepsilon_t : t \leq 0\}$ and satisfies the linear difference equation (5.2). Now, by a direct expansion and Cauchy–Schwarz inequality,

$$\begin{aligned}
&\epsilon^T \widetilde{A}_n^T Z Z^T \widetilde{A}_n \epsilon \\
=&\theta_0^2 \epsilon^T \widetilde{A}_n^T \mathbf{1}_n \mathbf{1}_n^T \widetilde{A}_n \epsilon + \sum_{i=1}^{d} \epsilon^T \widetilde{A}_n^T \mathbf{y}_{-i} \mathbf{y}_{-i}^T \widetilde{A}_n \epsilon \\
\leq& \left(\theta_0^2 + 3\alpha^2\right) \epsilon^T \widetilde{A}_n^T \mathbf{1}_n \mathbf{1}_n^T \widetilde{A}_n \epsilon + 3 \sum_{i=1}^{d} \left(\epsilon^T \Psi_i^T \widetilde{A}_n \epsilon\right)^2 + 3 \sum_{i=1}^{d} \left(\epsilon^T \widetilde{A}_n^T \mathbf{v}_i\right)^2 =: J_1 + J_2 + J_3.
\end{aligned}$$

A direct calculation yields that

$$\mathbb{E} J_1 = \left(\theta_0^2 + 3\alpha^2\right) \mathbf{1}_n^T \widetilde{A}_n \widetilde{A}_n^T \mathbf{1}_n \leq \left(\theta_0^2 + 3\alpha^2\right) \mathbf{1}_n^T \mathbf{1}_n \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} = \left(\theta_0^2 + 3\alpha^2\right) n \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp},$$

and therefore $J_1 = O_{\mathbb{P}}\left(n \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp}\right)$. To derive a asymptotic bound of J_2, we first use the first part of Lemma 3 to get that

$$\mathbb{E} \left| \epsilon^T \Psi_i^T \widetilde{A}_n \epsilon \right| \leq \left\| \operatorname{diag}(\Psi_i^T A_n) \right\|_1 + \sqrt{2} \left\| \Psi_i^T \widetilde{A}_n - \operatorname{diag}(\Psi_i^T \widetilde{A}_n) \right\| =: J_{2,1} + J_{2,2}.$$

Recall the the expansion (5.4) and apply the triangular inequality,

$$\begin{aligned}
\left\| \operatorname{diag}(\Psi_i^T A_n) \right\|_1 =& \left\| \sum_{j=0}^{\infty} \psi_j \operatorname{diag}\left(\left(L_n^{i+j}\right)^T \widetilde{A}_n \right) \right\|_1 \\
=& \sum_{t=1}^{n} \left| \sum_{j=0}^{n-t-i} \psi_j A_n(t+i+j, t) \right| \\
\leq& \sum_{t=1}^{n} \sum_{j=0}^{n-t-i} |\psi_j| \left| A_n(t+i+j, t) \right| = \sum_{j=0}^{n-i} |\psi_j| \sum_{t=1}^{n-i-j} \left| A_n(t+i+j, t) \right|,
\end{aligned}$$

where the last line we exchanged the double summations. Let $\eta > 0$. By choosing a sufficiently large K and for a large constant M not depending on K nor n, we have

$$
\begin{aligned}
J_{2,1} \leq& \sum_{j=0}^{K} |\psi_j| \cdot \sum_{t=1}^{n-i-j} |A_n(t+i+j,t)| + \sum_{j=K+1}^{n-i} |\psi_j| \cdot \sum_{t=1}^{n-i-j} |A_n(t+i+j,t)| \\
\leq& M \max_{1\leq l\leq K+i} \sum_{t=1}^{n-l} |A_n(t+l,t)| + \eta \cdot \max_{l>K} \sum_{t=1}^{n-l} |A_n(t+l,t)|
\end{aligned}
$$

Furthermore, by Cauchy–Schwarz inequality,

$$
\max_{l>J} \sum_{t=1}^{n-l} |A_n(t+l,t)| \leq \sqrt{n-l} \sqrt{\max_{l>J} \sum_{t=1}^{n-l} A_n^2(t+l,t)} \leq \sqrt{n} \left\| A_n \right\|.
$$

Using condition (iv) and noting that η can be arbitrarily small, we can show that

$$
J_{2,1} = o_{\mathbb{P}}(n^{1/2} \left\| A_n \right\|).
$$

On the other hand, by triangle equality and trace inequality (Lemma 4),

$$
J_{2,2}^2 \leq \left\| \Psi_i^T \widetilde{A}_n \right\|^2 = \operatorname{tr}(\Psi_i^T \widetilde{A}_n \widetilde{A}_n^T \Psi_i) \leq \left\| \Psi_i \right\|^2 \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} \leq Mn \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp}
$$

for some large constant M, where in the last step we used that $\left\| \Psi_i \right\|^2 \leq n \sum_{j=0}^{\infty} \psi_j^2$. Combining the bounds of $J_{2,1}$ and $J_{2,2}$ and applying the Markov inequality (Lemma 2), it follows that

$$
J_2 = O_{\mathbb{P}}(J_{2,1}^2 + J_{2,2}^2) = o_{\mathbb{P}}(n \left\| A_n \right\|^2) + O_{\mathbb{P}}\left(n \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} \right)
$$

Next, by a direct calculation and applying Lemma 5,

$$
\mathbb{E}\left(J_3 | \mathcal{F}_0\right) = \sum_{i=1}^{d} \mathbf{v}_i^T \widetilde{A}_n \widetilde{A}_n^T \mathbf{v}_i \leq \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} \sum_{i=1}^{d} \mathbf{v}_i^T \mathbf{v}_i = O_{\mathbb{P}}\left(n \left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} \right).
$$

It follows from Lemma 2 that $J_3 = O_{\mathbb{P}}\left(\left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp} \cdot n \right)$. Using Lemma 2 and combining all bounds of the terms J_1, J_2 and J_3,

$$
T_{1,2}^2 = \left(\lambda_{\min}(\widehat{\Omega}) \right)^{-2} O_{\mathbb{P}}(d) \left(\frac{\left\| \widetilde{A}_n \widetilde{A}_n^T \right\|_{sp}}{\left\| A_n \right\|^2} \cdot d + o_{\mathbb{P}}(1) \right) \xrightarrow{P} 0.
$$

This completes the proof of $T_1 \xrightarrow{P} 0$.

Finally, by Cauchy–Schwarz inequality

$$
\begin{aligned}
T_2 = \frac{\sqrt{2}}{\left\| A_n \right\|} \epsilon^T P_Z \widetilde{A}_n P_Z \epsilon \leq& \frac{\sqrt{2}}{\left\| A_n \right\|} \sqrt{\epsilon^T P_Z \epsilon \cdot \epsilon^T P_Z \widetilde{A}_n^T \widetilde{A}_n P_Z \epsilon} \\
=& O_p\left(\sqrt{\frac{\left\| \widetilde{A}_n^T \widetilde{A}_n \right\|_{sp}}{\left\| A_n \right\|^2}} \epsilon^T P_Z \epsilon \right) \xrightarrow{P} 0,
\end{aligned}
$$

where we have applied Lemma 8 in the last step. $\square$

For Corollary 1, it suffices to prove the consistency of the variance estimator (2.5).

Proof fo Corollary 1. Recall that it suffices to prove for the case $\sigma_n^2 = 1$. A direct calculation yields the matrix expression given by

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

where $P_Z = Z(Z^TZ)^{-1}Z^T$ is the projection matrix on the column space of Z. Using the martingale law of large number and noting that $d/n \to 0$, it is easy to show that $T_1 \xrightarrow{P} 1$. It remains to show that $T_2 \xrightarrow{P} 0$, which follows easily from Lemma 8. □

5.2 Proof of Theorem 2 and Theorem 3

We shall only prove Theorem 3, as the proof of Theorem 2 is completely analogous (even easier) by substituting function δ by a constant function on $(0,\infty)$. Throughout we assume the conditions of Theorem 3. Let $\mathbb{S}_n = \frac{1}{n}X^TX$ 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.

Without loss of generality, we only need to prove for the case $\sigma_n^2 = 1$ and $\beta \neq \mathbf{0}$; 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

$$\begin{aligned}
Q_n =& \frac{1}{\sqrt{2}\,\|A_n\|}\left\{\epsilon^T(I-P_Z) + \beta^T X^T(I-P_Z)\right\} A_n \left\{(I-P_Z)\epsilon + (I-P_Z)X\beta\right\}\\
=& \frac{1}{\sqrt{2}\,\|A_n\|}\epsilon^T(I-P_Z)A_n(I-P_Z)\epsilon + \frac{\sqrt{2}}{\|A_n\|}\epsilon^T(I-P_Z)A_n(I-P_Z)X\beta\\
&+ \frac{1}{\sqrt{2}\,\|A_n\|}\beta^T X^T (I-P_Z)A_n(I-P_Z)X\beta =: T_1 + T_2 + T_3.
\end{aligned} \tag{5.5}$$

From the proof of Theorem 1, we already have $T_1 \xrightarrow{P} \mathcal{N}(0,1)$. It remains to show that:

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

(ii) $T_2 \xrightarrow{P} 0$.

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

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

Proof. Using the definition of spectral norm,

$$\beta^T\widetilde{X}P_Z\widetilde{X}\beta = \frac{1}{n}\beta^T\widetilde{X}Z\widehat{\Omega}^{-1}Z^T\widetilde{X}\beta \leq \lambda_{\min}^{-1}(\widehat{\Omega})\frac{1}{n}\beta^T\widetilde{X}ZZ^T\widetilde{X}\beta. \tag{5.6}$$

It suffices to prove that

$$\beta^T \widetilde{X}^T ZZ^T \widetilde{X}\beta = O_p\left(n\left\|\beta\right\|^2 + n^2\left\|\beta\right\|^4\right). \tag{5.7}$$

Inverting the autoregressive process into a moving average form

$$y_t = \alpha + \sum_{j=1}^{\infty} \psi_j \epsilon_{t-j} + \sum_{j=1}^{\infty} \psi_j x_{t-j}^T \beta,$$

and, a direct calculation yields that

$$\begin{aligned}
\mathbf{y}_{-i} :=& [y_{1-i}, \ldots, y_{n-i}]^T \\
=& \alpha \mathbf{1}_n + \Psi_i \epsilon + \Psi_i X\beta + \widetilde{\mathbf{v}}_i \\
=& \alpha \mathbf{1}_n + \Psi_i \epsilon + \Psi_i \widetilde{X}\beta + \Psi_i \mathbf{1}_n \bar{x}^T \beta + \widetilde{\mathbf{v}}_i
\end{aligned}$$

with Ψ_i given in (5.4), and the reminder $\widetilde{\mathbf{v}}_i$ only depends on the initial values $\{\varepsilon_t, x_t : t \leq 0\}$ and satisfies the linear difference equation (5.2). Substituting these moving average forms and noting that $\mathbf{1}_n^T \widetilde{X} = 0$,

$$\begin{aligned}
\beta^T \widetilde{X}^T ZZ^T \widetilde{X}\beta =& \sum_{i=1}^{d} \beta^T \widetilde{X}^T \mathbf{y}_{-i} \mathbf{y}_{-i}^T \widetilde{X}\beta \\
=& \sum_{i=1}^{d} \left(\left(\Psi_i \epsilon + \Psi_i \widetilde{X}\beta + \Psi_i \mathbf{1}_n \bar{x}^T \beta + \widetilde{\mathbf{v}}_i \right)^T \widetilde{X}\beta \right)^2.
\end{aligned}$$

Applying Cauchy–Schwarz inequality,

$$\begin{aligned}
\beta^T \widetilde{X}^T ZZ^T \widetilde{X}\beta =& 4\sum_{i=1}^{d} \left(\epsilon^T \Psi^T \widetilde{X}\beta\right)^2 + 4\sum_{i=1}^{d} \left(\beta^T \widetilde{X}^T \Psi_i^T \widetilde{X}\beta\right)^2 \\
&+ 4\sum_{i=1}^{d} \left(\mathbf{1}_n^T \Psi_i^T \widetilde{X}\beta\right)^2 (\bar{x}^T\beta)^2 + 4\sum_{i=1}^{d} \left(\widetilde{v}_i^T \widetilde{X}\beta\right)^2 =: 4(R_1 + R_2 + R_3 + R_4).
\end{aligned}$$

Using the martingale property of $\{\varepsilon_t\}$,

$$\mathbb{E}\left(R_1 | X\right) = \beta^T \widetilde{X}^T \left(\sum_{i=1}^{d} \Psi_i \Psi_i^T\right) \widetilde{X}\beta \leq \left\|\sum_{i=1}^{d} \Psi_i \Psi_i^T\right\|_{sp} \beta^T \widetilde{X}^T \widetilde{X}\beta.$$

Applying Cauchy–Schwarz inequality on every quadratic term and summing up over i,

$$R_2 \leq \beta^T \widetilde{X}^T \left(\sum_{i=1}^{d} \Psi_i \Psi_i^T\right) \widetilde{X}\beta \cdot \beta^T \widetilde{X}^T \widetilde{X}\beta \leq \left\|\sum_{i=1}^{d} \Psi_i \Psi_i^T\right\|_{sp} \left(\beta^T \widetilde{X}^T \widetilde{X}\beta\right)^2.$$

Similarly,

$$R_3 \leq (\bar{x}^T\beta)^2 \mathbf{1}_n^T \left(\sum_{i=1}^{d} \Psi_i^T \Psi_i\right) \mathbf{1}_n \beta^T \widetilde{X}^T \widetilde{X}\beta \leq n(\bar{x}^T\beta)^2 \left\|\sum_{i=1}^{d} \Psi_i \Psi_i^T\right\|_{sp} \beta^T \widetilde{X}^T \widetilde{X}\beta.$$

Note that

$$\mathbb{E}(\bar{x}^T\beta)^2 = \beta^T\mathbb{E}(\bar{x}\bar{x}^T)\beta \le \|\beta\|^2 \left\|\mathbb{E}(\bar{x}\bar{x}^T)\right\|_{sp}.$$

It follows from Lemma 2 that

$$R_3 \le O_p(1)\cdot\|\beta\|^2\cdot\left\|\sum_{i=1}^d \Psi_i\Psi_i^T\right\|_{sp}\beta^T\widetilde{X}^T\widetilde{X}\beta.$$

Finally, applying Cauchy–Schwarz inequality and applying Lemma 5,

$$R_4 \le \left(\sum_{i=1}^d \widetilde{v}_i^T\widetilde{v}_i\right)\beta^T\widetilde{X}\widetilde{X}\beta = O_p(1)\cdot\beta^T\widetilde{X}\widetilde{X}\beta.$$

Summing up the bounds for the terms R_1–R_4, and noting that

$$\beta^T\widetilde{X}\widetilde{X}^T\beta = n\beta^T S_n\beta = n\|\beta\|^2\int x dF^{S_n}(x;\beta) = O_p(n\|\beta\|^2),$$

and $\left\|\sum_{i=1}^d \Psi_i\Psi_i^T\right\|_{sp} \le d\left\|\Psi_1\Psi_1^T\right\|_{sp}$ completes the proof of (5.7). □

Proof of Statement (i). 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 numerator in ϖ_n and it is bounded away from zero almost surely. Further expand that

$$\begin{aligned}
\sqrt{\frac{2}{p}}\|A_n\|T_3 =& \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T A_n\widetilde{X}\beta + \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T P_Z A_n P_Z\widetilde{X}\beta - 2\frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T P_Z A_n\widetilde{X}\beta\\
=&: J_1 + J_2 + J_3.
\end{aligned} \tag{5.8}$$

It suffices to show that

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

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

$$\begin{aligned}
J_1 =& \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T W_n\widetilde{X}\beta - \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T\left(\frac{1}{n}\operatorname{tr}(W_n)I_n\right)\widetilde{X}\beta\\
&- \frac{1}{\sqrt{p}}\beta^T\widetilde{X}^T\left(\operatorname{diag}(W_n) - \frac{1}{n}\operatorname{tr}(W_n)I_n\right)\widetilde{X}\beta =: J_{1,1} + J_{1,2} + J_{1,3}.
\end{aligned} \tag{5.9}$$

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

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

Next, denoting the 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 $\xi_n = \beta/\|\beta\|$ is the direction of the coefficient β. Note that by assumption we have

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

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

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

For some absolute constant M,

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

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 2 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 $J_{1,3} \xrightarrow{P} 0$. Summing up the limits of $J_{1,1}$, $J_{1,2}$ and $J_{1,3}$ yields (5.2). Next, using Lemma 9,

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

Combining with Cauchy–Schwarz inequality it follows that

$$|J_3| \leq 2J_1J_2 \xrightarrow{P} 0.$$

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

Proof of Statement (ii). First we decompose that

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

Note that

$$\beta^T \widetilde{X} A_n^2 \widetilde{X}\beta \leq \|A_n\|_{sp}^2 \beta^T \widetilde{X}\widetilde{X}\beta = n\|A_n\|_{sp}^2 \beta^T S_n \beta = n\|A_n\|_{sp}^2 \|\beta\|^2 \int x dF^{S_n}(x;\beta).$$

It follows that

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

and therefore $T_{2,1} \xrightarrow{P} 0$ by Lemma 2. Combining the Cauchy–Schwarz inequality and Lemma 8,

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

Similarly, by decomposing $A_n = \widetilde{A}_n + \widetilde{A}_n^T$ and applying the Cauchy–Schwarz

$$\begin{aligned} T_{2,3}^2 =& \frac{2}{\|A_n\|^2}\left(\epsilon^T \widetilde{A}_n P_Z \widetilde{X}\beta + \epsilon^T \widetilde{A}_n^T P_Z \widetilde{X}\beta\right)^2 \\ \leq& \frac{4}{\|A_n\|^2}\left\{\left(\epsilon^T \widetilde{A}_n P_Z \widetilde{X}\beta\right)^2 + \left(\epsilon^T \widetilde{A}_n^T P_Z \widetilde{X}\beta\right)^2\right\}. \end{aligned}$$

Further applying Cauchy–Schwarz,

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

Similarly,

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

Recall from the proof of Theorem 1 that

$$\epsilon^T \widetilde{A}_n ZZ\widetilde{A}_n^T Z\epsilon = \left\|\widetilde{A}_n^T \widetilde{A}_n\right\|_{sp}\cdot nd = o_{\mathbb{P}}(n\|A_n\|^2),$$

and

$$\epsilon^T \widetilde{A}_n^T ZZ\widetilde{A}_n Z\epsilon = o_{\mathbb{P}}(n\|A_n\|^2).$$

Hence, combining with Lemma 9,

$$T_{2,4}^2 \leq \frac{4}{\|A_n\|^2} o_{\mathbb{P}}\left\{\left(\lambda_{\min}(\widehat{\Omega})\right)^{-1}\frac{1}{n}\cdot n\|A_n\|^2\cdot\left(\lambda_{\min}(\widehat{\Omega})\right)^{-1}\left(\|\beta\|^2+n\|\beta\|^4\right)\right\} \xrightarrow{P} 0.$$

□

5.3 Proofs of Corollaries 2, 3 and 4

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) and continuous mapping theorem that $\varpi_n \to \varpi$, where ϖ has the same expression as ϖ_n except using F instead of F^{S_n}. □

Proof of Corollary 3 and 4. We only need to prove $\widehat{\sigma}_n^2/\sigma_n^2 \xrightarrow{P} 1$ under the local alternatives (2.6). Without loss of generality, we assume $\sigma_n^2 = 1$. Expanding

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

From the proof of Corollary 1, we already have $\frac{1}{n-(d+1)}e^Te \xrightarrow{P} 1$. By Cauchy–Schwarz inequality it is elementary to show that $T_2 \leq 2T_1 \cdot T_3 = O(T_1)$. It remains to show that $T_1 \xrightarrow{P} 0$. Using the definition of spectral norm twice,

$$\begin{aligned}
T_1 \leq& \|I-P_Z\|_{sp}\cdot\frac{1}{n-(d+1)}\beta^TX^TX\beta\\
\leq& 1\cdot\frac{n}{n-(d+1)}\|\beta\|^2\lambda_{\max}(S_n) \xrightarrow{P} 0,
\end{aligned}$$

as $\|\beta\|^2 = O(\sqrt{p}/n) = o(1)$ and $\lambda_{\max}(S_n) = O_{\mathbb{P}}(1)$. □

5.4 Proof of Proposition 1

We shall first show that the proposition holds for the oracle matrix $\underline{\mathbb{S}}_n$, and then we substitute it by the observed matrix $\underline{S}_n$. Note that

$$\underline{\mathbb{S}}_n(t,t) = \frac{1}{n}x_t^Tx_t = \frac{1}{n}f_t^T\Phi^T\Phi f_t$$

Noting that $\underline{\mathbb{S}}_n(t,t)$ are identically distributed (not necessarily independent) and, following the proof of Lemma 3,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\operatorname{tr}\left(\Phi^T\Phi\right)\right|^2\right) =&\mathbb{E}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\operatorname{tr}\left(\Phi^T\Phi\right)\right|^2\\ \le& M\widetilde{\kappa}_n\left(\frac{1}{n^2}\sum_{i=1}^{k}\|\phi_i\|^4+\frac{1}{n^2}\left\|\Phi^T\Phi\right\|^2\right)\\ =&M\widetilde{\kappa}_n\left(\frac{1}{n^2}\sum_{i=1}^{k}\|\phi_i\|^4+\frac{1}{n^2}\|\Sigma\|^2\right).\end{aligned}$$

where $\widetilde{\kappa}_n := \max_{i,k}\mathbb{E}|f_{t,i}^2-1|^2 = O(1)$.

Now, using the relations between L_1 and L_2 norms, for some large constant M

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{\mathbb{S}}_n(t,t)\right|^2\\ \le& M\left(\frac{1}{n}\sum_{t=1}^{n}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\operatorname{tr}\left(\Phi^T\Phi\right)\right|^2+\left|\frac{1}{n}\sum_{t=1}^{n}\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\operatorname{tr}\left(\Phi^T\Phi\right)\right|^2\right)\\ \le& M\frac{2}{n}\sum_{t=1}^{n}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\operatorname{tr}\left(\Phi^T\Phi\right)\right|^2 = O_{\mathbb{P}}\left(\frac{1}{n^2}\sum_{i=1}^{k}\|\phi_i\|^4+\frac{1}{n^2}\|\Sigma\|^2\right),\end{aligned}$$

where the last step follows from Lemma 2. Observe that the last term is $o_{\mathbb{P}}$ as $\|\Sigma\|^2 \le n\|\Sigma\|_{sp}^2 = O(n)$ and $\sum_{i=1}^{k}\|\phi_i\|^4 = \left\|\operatorname{diag}(\Phi^T\Phi)\right\|^2 \le \left\|\Phi^T\Phi\right\|^2 = \|\Sigma\|^2 = O(n)$.

Using the identity that $\widetilde{x}_t = x_t - \bar{x}$, we can calculate that

$$\underline{S}_n(t,t) - \underline{\mathbb{S}}_n(t,t) =: -\frac{2}{n}\bar{x}^T x_t + \frac{1}{n}\bar{x}^T\bar{x},$$

and remove the last perturbation term in the demeaned diagonals to get that

$$\underline{S}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{S}_n(t,t) = \left\{\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{\mathbb{S}}_n(t,t)\right\}-\frac{2}{n}\bar{x}^T x_t.$$

It is elementary to show that, for some constant M only depending on ι,

$$\left|\underline{S}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{S}_n(t,t)\right|^2 \le M\left(\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{\mathbb{S}}_n(t,t)\right|^2+\left|\frac{2}{n}\bar{x}^T x_t\right|^2\right)$$

Taking average over t yields that

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}\left|\underline{S}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{S}_n(t,t)\right|^2\\ \le& M\left\{\frac{1}{n}\sum_{t=1}^{n}\left|\underline{\mathbb{S}}_n(t,t)-\frac{1}{n}\sum_{t=1}^{n}\underline{\mathbb{S}}_n(t,t)\right|^2+\frac{1}{n}\sum_{t=1}^{n}\left|\frac{2}{n}\bar{x}^T x_t\right|^2\right\}.\end{aligned}$$

The proposition then follows as

$$\frac{1}{n}\sum_{t=1}^{n}\left|\frac{2}{n}\bar{x}^T x_t\right|^2 = \frac{4}{n^2}\bar{x}^T S_n \bar{x} = \frac{1}{n^2}\|S_n\|_{sp}\|\bar{x}\|^2 = O_{\mathbb{P}}(n^{-1}).$$

5.5 Proof of Proposition 2

Our first lemma follows from the same arguments for equation (3.2) in Bai and Silverstein (1998) by combining Lemma 2.7 and Lemma 2.9 therein. Note that we have also used Jensen's inequality $(E|f_1|^4)^{\alpha/2} \leq E|f_1|^{2\alpha}$ for any $\alpha \geq 2$. We omit the details of the proof.

Lemma 10 (Concentration inequality for quadratic forms)**.** *For A being a $n \times n$ matrix (complex), we have, for any $\alpha \geq 2$*

$$\mathbb{E}\left|f^T A f - \operatorname{tr}(A)\right|^\alpha \leq M E|f_{1,1}|^{2\alpha}\|A\|^\alpha$$

where M is some absolute constant depending only on α.

Lemma 11. $x_t^T x_t/p \xrightarrow{a.s.} \int x dH(x)$.

Proof. Applying Lemma 10 with $\alpha = 2 + \iota/2$ and noting that $E|f_{1,1}|^{4+\iota}$ is bounded, for some large constant M

$$\begin{aligned}\mathbb{E}\left(\frac{1}{p}x_t^T x_t - \frac{\operatorname{tr}(\Sigma)}{p}\right)^{2+\iota/2} =& \mathbb{E}\left(\frac{1}{p}f_t^T \Sigma f_t - \frac{\operatorname{tr}(\Sigma)}{p}\right)^{2+\iota/2} \\ \leq& M p^{-(1+\iota/4)}\left(\frac{\operatorname{tr}(\Sigma^2)}{p}\right)^{1+\iota/4} = O(n^{-(1+\iota/4)}).\end{aligned}$$

By Markov inequality and Borel–Cantelli lemma, it is easy show that

$$\frac{1}{p}x_t^T x_t - \frac{\operatorname{tr}(\Sigma)}{p} \xrightarrow{a.s.} 0.$$

The proof is complete by checking that $\frac{\operatorname{tr}(\Sigma)}{p} \to \int x dH(x)$ using the dominated convergence theorem. ◻

Lemma 12. *Let $\mathbb{S}_n(t) = \mathbb{S}_n - \frac{1}{n}x_t x_t^T$ be the sample covariance matrix dropping x_t.*

$$\frac{1}{n}x_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t \xrightarrow{a.s.} -1 - \frac{1}{\underline{m}(z)z}.$$

Proof. Note that x_t is independent of $\mathbb{S}_n(t)$. From the proof of Theorem 1 in Bai et al. (2007), e.g equation (2.9) therein, we know that

$$\frac{1}{x_t^T x_t}x_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t - \frac{1}{x_t^T x_t}x_t^T\left(-z\underline{m}(z)\Sigma_n - zI\right)^{-1}x_t \xrightarrow{a.s.} 0$$

Combining with Lemma 11 yield that

$$\frac{1}{n}x_t^T(\mathbb{S}_n(t)-zI)^{-1}x_t-\frac{1}{n}x_t^T\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}x_t\xrightarrow{a.s.}0.$$

Applying Lemma 10 with $\alpha=2+\iota/2$ and noting that $E|f_{1,1}|^{4+\iota}$ is bounded, for some large constant M

$$\begin{aligned}&E\left|\frac{1}{n}x_t^T\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}x_t-\frac{1}{n}\operatorname{tr}\Sigma\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}\right|^{2+\iota/2}\\ \le&Mn^{-(2+\iota/2)}\left\|\Sigma\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}\right\|^{2+\iota/2}\\ =&Mn^{-(2+\iota/2)}\left|z\right|^{-2}\left\|\Sigma\left(\underline{m}(z)\Sigma_n+I\right)^{-1}\right\|^{2+\iota/2}\\ \le&Mn^{-(2+\iota/2)}\left|z\right|^{-2}\left\|\Sigma\right\|^{2+\iota/2}\left\|\left(\underline{m}(z)\Sigma_n+I\right)^{-1}\right\|_{sp}^{2}\end{aligned}$$

Recall from Silverstein (1995), the last paragraph in page 338, that $\left\|\left(\underline{m}(z)\Sigma_n+I\right)^{-1}\right\|_{sp}$ is bounded away from 0, and $|z|>0$ by construction. Hence, for some possibly different constant M, the last bound is further bounded by

$$Mn^{-(2+\iota/2)}\left\|\Sigma\right\|^{2+\iota/2}=Mn^{-(1+\iota/4)}\left(\frac{p}{n}\right)^{1+\iota/4}\left(\frac{\operatorname{tr}(\Sigma^2)}{p}\right)^{1+\iota/4}=O(n^{-(1+\iota/4)}).$$

Then it is elementary to show that, using Markov inequality and Borel–Cantelli lemma,

$$\frac{1}{n}x_t^T(\mathbb{S}_n(t)-zI)^{-1}x_t-\frac{1}{n}\operatorname{tr}\Sigma\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}\xrightarrow{a.s.}0.$$

Finally,

$$\begin{aligned}\frac{1}{n}\operatorname{tr}\Sigma\left(-z\underline{m}(z)\Sigma_n-zI\right)^{-1}=&-\frac{p}{n}\frac{1}{z}\int\frac{\lambda}{1+\underline{m}\lambda}dH_n(\lambda)\\ \xrightarrow{a.s.}&-\frac{1}{cz}\int\frac{\lambda}{1+\underline{m}\lambda}dH(\lambda)=-1-\frac{1}{\underline{m}(z)z}\end{aligned}$$

□

Proof of Proposition 2. Let $\delta_1(x)=\delta(x)\cdot x$. It suffices to show that

$$\frac{1}{n}\left\|\operatorname{diag}(W_n(\delta))-\frac{1}{n}\operatorname{tr}(W_n(\delta))I_n\right\|^2=\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t-\frac{1}{n}\operatorname{tr}\delta_1(S_n)\right)^2\xrightarrow{P}0.$$

Note that $\{\widetilde{x}_t\}$ are exchangeable, and thus by Lemma 2 it suffices to show that

$$\begin{aligned}&E\left(\frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{n}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t-\frac{1}{n}\operatorname{tr}\delta_1(S_n)\right)^2|\mathcal{E}_n\right)\\ =&E\left(\left(\frac{1}{n}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t-\frac{1}{n}\operatorname{tr}\delta_1(S_n)\right)^2|\mathcal{E}_n\right)\to0,\end{aligned}$$

for some event $\mathcal{E}_n$ not depending on t. Applying Weierstrass theorem with the continuity of δ yields that

$$\begin{aligned}\max_{1\le t\le n}\frac{1}{n}x_t^T\delta(S_n)x_t \le & \|\delta(S_n)\|_{sp}\cdot\max_{1\le t\le n}\frac{1}{n}\widetilde{x}_t^T\widetilde{x}_t\\ =&O_{\mathbb{P}}(1)\cdot\max_{1\le t\le n}\underline{S}_n(t,t)\le O_{\mathbb{P}}(1)\left\|\underline{S}_n\right\|_{sp}=O_{\mathbb{P}}(1).\end{aligned}$$

Hence by the asymptotic boundedness of $\frac{1}{n}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t$ it suffices to show that

$$\frac{1}{n}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t-\frac{1}{n}\operatorname{tr}\delta_1(S_n)\xrightarrow{P}0,$$

or equivalently, for some ι_n

$$\frac{1}{p}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t\xrightarrow{P}\int\delta_1(x)dF(x).$$

We can decompose the left-hand-side as

$$\frac{1}{p}\widetilde{x}_t^T\delta(S_n)\widetilde{x}_t=\frac{1}{p}x_t^T\delta(S_n)x_t+\frac{1}{p}\bar{x}^T\delta(S_n)\bar{x}-\frac{2}{p}x_t^T\delta(S_n)\bar{x}.$$

Note that, by Lemma 11

$$\frac{1}{p}\bar{x}^T\delta(S_n)\bar{x}\le\|\delta(S_n)\|_{sp}\frac{1}{p}\bar{x}^T\bar{x}\xrightarrow{P}0,$$

and by Cauchy–Schwarz inequality

$$\left|\frac{1}{p}x_t^T\delta(S_n)\bar{x}\right|\le\sqrt{\frac{1}{p}x_t^T\delta(S_n)x_t\cdot\frac{1}{p}\bar{x}^T\delta(S_n)\bar{x}}.$$

It suffices to prove that

$$\frac{1}{p}x_t^T\delta(S_n)x_t\xrightarrow{a.s.}\int X\delta(x)dF(x). \tag{5.10}$$

Now, similar to $F^{S_n}(x;\beta)$, define an unproper weighted empirical spectral distribution

$$F_t^{S_n}(x):=\frac{1}{p}\sum_{i=1}^p(u_i^Tx_t)^2\mathbb{1}(\lambda_i\le x).$$

We can rewrite that

$$\frac{1}{p}x_t^T\delta(S_n)x_t=\int\delta(x)dF_t^{S_n}(x;x_t).$$

Like in Bai et al. (2007), one can verify that the Stieltjes transform of $F^{S_n}(x;x_t)$

$$m_{F_t^{S_n}}(z)=\int\frac{1}{\lambda-z}dF_t^{S_n}(x)=\frac{1}{p}x_t^T\left(S_n-zI\right)^{-1}x_t$$

for $z\in\mathbb{C}^+=\{z\in\mathbb{C}:\Im z>0\}$. Using the expansions that $S_n=\mathbb{S}_n-\bar{x}\bar{x}^T$ and the Sherman–Morrison formula,

$$\begin{aligned}m_{F_t^{S_n}}(z)=&\frac{1}{p}x_t^T(\mathbb{S}_n-zI-\bar{x}\bar{x}^T)^{-1}x_t\\ =&\frac{1}{p}x_t^T(\mathbb{S}_n-zI)^{-1}x_t+\frac{1}{p}\frac{(\bar{x}^T(\mathbb{S}_n-zI)^{-1}x_t)^2}{1-\bar{x}^T(\mathbb{S}_n-zI)^{-1}\bar{x}}=:T_1+T_2.\end{aligned}$$

Recall that $\mathbb{S}_n = \mathbb{S}_n(t) + \frac{1}{n}x_t x_t^T$ and applying Sherman–Morrison formula,

$$T_1 = \frac{n}{p}\left(1 - \frac{1}{1 + \frac{1}{n}x_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t}\right).$$

Applying Lemma 12, it follows that

$$T_1 \xrightarrow{P} \frac{1}{c}\left(1 + z\underline{m}(z)\right) = 1 + zm(z).$$

We shall show later that $T_2 \xrightarrow{a.s.} 0$, and therefore $m_{F_t^{S_n}}(z) \xrightarrow{a.s.} 1 + zm(z)$, where the limit does not depend on t. By the equivalence between Stieltjes transform and the associated measure, e.g., Theorem B.9 in Bai and Silverstein (2010), we know $\frac{1}{p}x_t^T\delta(S_n)x_t$ converges almost surely to some limit, which are the same for all t. The limit is therefore the same as that of their average $\frac{1}{n}\sum_{t=1}^n \frac{1}{p}x_t^T\delta(S_n)x_t = \frac{1}{p}\operatorname{tr}(\delta_1(S_n)) \xrightarrow{a.s.} \int \delta_1(x)dF(x)$. This completes the proof for (5.10).

It remains to show that $T_2 \xrightarrow{a.s.} 0$. Let $\bar{x}_t = \bar{x} - \frac{1}{n}x_t$ be the sample average dropping x_t and recall that $\mathbb{S}_n(t) = \mathbb{S}_n - \frac{1}{n}x_t x_t^T$. Using Sherman–Morrison formula again and by a direct calculation yields that

$$\begin{aligned}\bar{x}^T(\mathbb{S}_n - zI)^{-1}x_t =& \left(\bar{x}_t + \frac{x_t}{n}\right)^T\left(\mathbb{S}_n(t) + \frac{1}{n}x_t x_t^T - zI\right)^{-1}x_t \\ =& \frac{1}{1 + \frac{1}{n}x_t(\mathbb{S}_n(t) - zI)^{-1}x_t}\left\{\bar{x}_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t + \frac{1}{n}x_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t\right\}.\end{aligned}$$

Note that x_t is independent of $\bar{x}_t$ and $\mathbb{S}_n(t)$. From the proof of Theorem 2 in Pan (2014), by substituting $\frac{x_t}{\|x_t\|}$ for the unit vector $\mathbf{x}_n$ therein, we know that

$$\bar{x}_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t = o(\|x_t\|),$$

Recall from Lemma 12 that the reminder term $\frac{1}{n}x_t^T(\mathbb{S}_n(t) - zI)^{-1}x_t = o(1)$ in the numerator, and furthermore the denominator

$$1 + \frac{1}{n}x_t(\mathbb{S}_n(t) - zI)^{-1}x_t \xrightarrow{a.s.} -\frac{1}{z\underline{m}(z)},$$

where the limit is bounded away from 0. Therefore,

$$\bar{x}^T(\mathbb{S}_n - zI)^{-1}x_t = o(\|x_t\|).$$

Let $z = a + b\mathrm{i}$, where i denotes the imaginary unit and $b > 0$. Recall in the Theorem 2 in Pan (2014), using equation (2.27) therein,

$$\left|\frac{1}{1 - \bar{x}^T(\mathbb{S}_n - zI)^{-1}\bar{x}}\right| = \left|1 + \bar{x}^T\left(S_n - zI\right)^{-1}\bar{x}\right| \le 1 + \bar{x}^T\bar{x}\left\|\left(S_n - zI\right)^{-1}\right\|_{sp} \le 1 + \bar{x}^T\bar{x}\frac{1}{b}.$$

On the other hand, for some large constant M, $\bar{x}^T\bar{x} = \bar{f}^T\Sigma\bar{f} \le M\bar{f}^T\bar{f} \xrightarrow{a.s.} Mc$. It follows that

$$\left|\frac{1}{1 - \bar{x}^T(\mathbb{S}_n - zI)^{-1}\bar{x}}\right| = O(1).$$

Hence,

$$T_2 = \frac{1}{p} \cdot o(\|\bar{x}_t\|^2) \cdot O(1) \xrightarrow{a.s.} 0,$$

by using Lemma 11. Now the proof is complete. □

5.6 Proof of Theorem 4

Proof of Theorem 4. The proof is the same as in the subsection 5.1, after mapping $\psi_j L_n^{i+j}$ to $\psi_i(l) L_n^l$ in (5.4). We omit the details. □

Proof of Theorem 5. We first generalize the results in Lemma 9, by showing that

$$\beta^T \widetilde{X} P_Z \widetilde{X} \beta = \lambda_{\min}^{-1}(\widehat{\Omega}) \cdot \|\beta\|^2 \cdot O_p \left(\sum_{i=1}^{d} \lambda_{\max} \left(\mathbb{E}[\mathbf{v}_i \mathbf{v}_i^T | X] \right) + \lambda_{\max} \left(\Psi_i \Psi_i^T \right) \right). \tag{5.11}$$

The proofs is similar, but we show the differences for completeness. Recall from (5.6) that it suffices to prove that

$$\beta^T \widetilde{X}^T Z Z^T \widetilde{X} \beta = n \cdot \|\beta\|^2 \cdot O_p \left(\sum_{i=1}^{d} \lambda_{\max} \left(\mathbb{E}[\mathbf{v}_i \mathbf{v}_i^T | X] \right) + \lambda_{\max} \left(\Psi_i \Psi_i^T \right) \right).$$

Rewrite

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

where $\Psi_i = \sum_{l=1}^{t-1} \psi_i(l) L_n^l$ like in (5.4). A direct expansion and applying the Cauchy–Schwarz inequality,

$$\begin{aligned}
\beta^T \widetilde{X}^T Z Z^T \widetilde{X} \beta =& \sum_{i=1}^{d} \beta^T \widetilde{X}^T \mathbf{z}_i \mathbf{z}_i^T \widetilde{X} \beta \\
=& \sum_{i=1}^{d} \left((\Psi_i \epsilon + \mathbf{v}_i)^T \widetilde{X} \beta \right)^2 \le 2 \sum_{i=1}^{d} \left(\left(\epsilon^T \Psi_i^T \widetilde{X} \beta \right)^2 + \left(\mathbf{v}_i^T \widetilde{X} \beta \right)^2 \right).
\end{aligned}$$

Furthermore, it follows from Lemma 2 that

$$\begin{aligned}
\left(\epsilon^T \Psi_i^T \widetilde{X} \beta \right)^2 =& O_{\mathbb{P}} \left(\mathbb{E} \left[\left(\epsilon^T \Psi_i^T \widetilde{X} \beta \right)^2 | X \right] \right) \\
=& O_{\mathbb{P}} \left(\beta^T \widetilde{X}^T \Psi_i \Psi_i^T \widetilde{X} \beta \right) = O_{\mathbb{P}} \left(\lambda_{\max}(\Psi_i \Psi_i^T) \cdot \beta^T \widetilde{X}^T \widetilde{X} \beta \right),
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\left(\mathbf{v}_i^T \widetilde{X} \beta \right)^2 =& O_{\mathbb{P}} \left(\mathbb{E} \left[\left(\mathbf{v}_i^T \widetilde{X} \beta \right)^2 | X \right] \right) \\
=& O_{\mathbb{P}} \left(\beta^T \widetilde{X}^T \mathbb{E}[\mathbf{v}_i \mathbf{v}_i^T | X] \widetilde{X} \beta \right) = O_{\mathbb{P}} \left(\lambda_{\max}(\mathbb{E}[\mathbf{v}_i \mathbf{v}_i^T | X]) \cdot \beta^T \widetilde{X}^T \widetilde{X} \beta \right)
\end{aligned}$$

Summing up the upper bounds and noting that $\beta \widetilde{X}^T \widetilde{X} \beta = n \|\beta\|^2 \int x dF^{S_n}(x;\beta) = O_{\mathbb{P}}(n \|\beta\|^2)$ completes the proof of (5.11).

Now, using the condition (2.10) and noting that $\lambda_{\max}(\Psi_i \Psi_i^T) \le \left(\sum_{l=1}^{n-1} |\psi_i(l)|\right)^2 < \infty$, we have that $\beta^T \widetilde{X} P_Z \widetilde{X} \beta = o_{\mathbb{P}}(n \|\beta\|^2)$. Substituting this bound, rather than that from Lemma 9, into the arguments in Section 5.2 immediately completes the proof, and we omit the details to save space. □

References

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices* (2nd ed. ed.). Springer.

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.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab. 1*, 19–42.

Chen, B. and G. 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 (2018). Inference on Causal and Structural Parameters using Many Moment Inequalities. *The Review of Economic Studies 86*, 1867–1900.

Chudik, A., M. H. Pesaran, and E. Tosetti (2011). Weak and strong crosssection dependence and estimation of large panels. *The Econometrics Journal 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. *Probability Theory and 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.

Fan, J., S. Guo, and N. Hao (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 37–65.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in highdimensional crosssectional tests. *Econometrica 83*, 1497–1541.

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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*, 1079–1102.

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

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 University Press.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis.* Cambridge University 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. *Journal of Multivariate Analysis 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. *Probability Theory and Related Fields 151*, 233–264.

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

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.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*, 574–589.

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

Pan, G. (2014). Comparison between two types of large sample covariance matrices. *Ann. Inst. H. Poincar Probab. Statist. 50*, 655–677.

Pan, G., J. Gao, and Y. Yang (2014). Testing independence among a large number of high-dimensional random vectors. *Journal of the American Statistical Association 109*, 600–612.

Phillips, P. C. B. and S. Jin (2014). Testing the martingale hypothesis. *Journal of Business & Economic Statistics 32*, 537–554.

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

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

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

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

Wu, W. B. and X. Shao (2007). A limit theorem for quadratic forms and its applications. *Econometric Theory 23*, 930951.

Xi, H., F. Yang, and J. Yin (2019). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *The Annals of Statistics*.

Xia, N., Y. Qin, and Z. 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. (1986). Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis 20*, 50 – 68.

Zhang, L. (2006). Spectral analysis of large dimensional random matrices. *National University of Singapore PHD Thesis*.

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. *Journal of the American Statistical Association 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. *Journal of the American Statistical Association 113*, 1583–1600.