Submitted to the Annals of Statistics

COINTEGRATION IN LARGE VARS

BY ANNA BYKHOVSKAYA[*] AND VADIM GORIN[†]

University of Wisconsin-Madison, [*]anna.bykhovskaya@wisc.edu; [†]vadicgor@gmail.com

The paper analyses cointegration in vector autoregressive processes (VARs) for the cases when both the number of coordinates, N, and the number of time periods, T, are large and of the same order. We propose a way to examine a VAR of order 1 for the presence of cointegration based on a modification of the Johansen likelihood ratio test. The advantage of our procedure over the original Johansen test and its finite sample corrections is that our test does not suffer from over-rejection. This is achieved through novel asymptotic theorems for eigenvalues of matrices in the test statistic in the regime of proportionally growing N and T. Our theoretical findings are supported by Monte Carlo simulations and an empirical illustration. Moreover, we find a surprising connection with multivariate analysis of variance (MANOVA) and explain why it emerges.

1. Introduction.

1.1. *Motivation.* The importance of cointegration in economics stems from the seminal papers Granger (1981) and Engle and Granger (1987). For example, as they show, monthly rates on 1-month and 20-year treasury bonds are cointegrated, which means that they are both non-stationary, but they have a stationary linear combination. A lot of other variables in macroeconomics and finance such as price level, consumption, output, trade flows, interest rates and so on are non-stationary, and, thus, are potentially subject to cointegration. When dealing with non-stationary time series, it is always a question whether one should work with levels or with differences. For multivariate settings such as vector autoregression (VAR), the choice of model would depend on whether the series is cointegrated or not.

There are several ways to test for the presence of cointegration (see, e.g., Maddala and Kim (1998) for the detailed description of various methods). One popular approach relies on checking whether the residuals from regressing one of the coordinates on the remaining ones are stationary. It is based on Engle and Granger (1987) and was later extended in Phillips and Ouliaris (1990). Another widely used technique is due to Søren Johansen (Johansen, 1988, 1991)[1]. This approach assumes VAR structure and relies on the likelihood ratio. It tests a null hypothesis of at most ρ cointegrating relationships versus an alternative of between ρ and $r > \rho$ cointegrating relationships. The Johansen test turns out to be related to the eigenvalues of some random matrix and has a non-standard asymptotic distribution.

Neither of the approaches is commonly used in the analysis of large systems. However, in many situations the data turns out to have both cross-sectional and time dimensions being large. Natural examples are given, e.g., by financial data (stock prices, exchange rates, etc.) or by monthly data on trade and investments between countries (where the number of pairs of countries is large). The main difficulty with applying the above cointegration testing approaches to high-dimensional settings is that both of them require N, the cross-sectional dimension of a time series, to be fixed and small. For the former approach, the larger is N, the more regressions we need to run and interpreting their results becomes ambiguous. For

Keywords and phrases: High-dimensional VAR, Cointegration, Johansen test, Jacobi ensemble.

[1]A related approach was also proposed in Stock and Watson (1988).

the latter approach, it turns out, the asymptotic theory stops providing a good approximation for moderate values of T. The test starts to over-reject the null of fewer cointegrations in favor of the alternative (see, for example, Ho and Sørensen (1996), Gonzalo and Pitarakis (1999)). For instance, the simulations reported in Table 1 of Gonzalo and Pitarakis (1999) indicate that the empirical rejection rate based on the 95% asymptotic critical value for the no-cointegration hypothesis (constructed using the asymptotics of Johansen (1988, 1991)) is 20% at $N = 5$, $T = 30$. It eventually decreases to the desired 5% as T grows, but even at $T = 400$ the empirical size is still 6%.

After size distortions became clear econometricians developed various procedures for correcting over-rejection. Popular methods include finite sample correction (e.g., Reinsel and Ahn (1992), Johansen (2002)) and bootstrap (e.g., Swensen (2006), Cavaliere et al. (2012)). Such modifications help to restore correct size for moderate values of T (e.g., $T = 50$), when N is of a smaller order (e.g., $N = 5$). Yet, the question of larger N remained open.

A recent ground-breaking paper Onatski and Wang (2018) shows that the over-rejection when testing the null of no cointegration can be mathematically explained by considering an alternative asymptotics in which both N and T go to infinity jointly and proportionally. When N is large such asymptotic regime turns out to better suit the finite sample properties of the data. While Onatski and Wang (2018) point this out, they do not provide alternative testing procedures.

The observation that large VARs might have analyzable joint limits opens a new area of research. That is, one can try to develop various sophisticated joint asymptotics and derive, as a result, appropriate ways to test for cointegration in settings where both N and T are large. This, however, requires new tools rooted in the random matrix theory. In our paper we propose such cointegration tests and introduce asymptotic theorems which make testing possible. Let us describe our main results.

1.2. *Results.* By modifying the Johansen likelihood ratio (LR) test, we come up with a way to examine the presence of cointegration in a time series when the cross-sectional dimension, N, and the time dimension, T, are of the same order. We consider a vector autoregression (VAR) of order 1 in the error correction representation

$$\Delta X_t = \mu + \Pi X_{t-1} + \varepsilon_t, \qquad t = 1, \dots, T, \tag{1}$$

where $\Delta X_t := X_t - X_{t-1}$, errors $\{\varepsilon_t\}$ are Gaussian with $N \times N$ covariance matrix Λ, and Π, μ are unknown parameters.[2] We do not restrict Λ to be diagonal, which means that any cross-sectional correlation structure is allowed. Such cross-sectional heteroscedasticity assumption may be important in applications, where one expects variables, e.g., countries, to be correlated. In contrast, many previous approaches relied on specific forms of covariance Λ, see, e.g., Breitung and Pesaran (2008), Bai and Ng (2008, Section 7) for reviews and Zhang et al. (2018) for recent developments.

Our model allows for an arbitrary linear trend in X_t, which is a desirable feature when one goes to applications. Yet, it also encompasses as a special case a model without a trend ($\mu = 0$). The way we deal with the data turns out to be invariant to μ, and we do not propose any different procedure even if a researcher has a prior knowledge that there is no trend. (We discuss this in more detail in Section 7.2.)

We are interested in whether the N-dimensional non-stationary process X_t is cointegrated or not. That is, whether there is a non-zero vector β such that $\beta' X_t$ is trend stationary. Our main contributions lie in the construction of the appropriate test, analysis of its asymptotic

[2] Section 7.3 discusses possible generalizations of our approach to VAR(k) for general k.

distribution, and computation of critical values (some of them are reported in Table 1 of Section 3).

Our procedure for cointegration testing relies on the following steps. First, we de-trend X_t by subtracting $\frac{t}{T}(X_T - X_0)$ from X_t. Then we follow Johansen's procedure and regress both first differences, ΔX_t, and lagged de-trended X_t on a constant. That is, we de-mean the data. Then we calculate the squared sample canonical correlations between the residuals from those regressions. This corresponds to the eigenvalues from the modification of the Johansen LR statistics obtained by using the lagged de-trended version of X_t and not de-trended first differences ΔX_t.

The de-trending procedure can be interpreted in the following way: we take the first differences of the observed variables, de-mean them, and re-sum back. This is equivalent (up to a constant X_0 which disappears after we further de-mean the sums) to our de-trending. Such interpretation is related to Elliott et al. (1996) which analyzes unit root testing in the presence of a trend d_t. Under the null hypothesis of a unit root and when the trend d_t is linear, Elliott et al. (1996) suggests to de-mean the first differences.

As discussed in Xiao and Phillips (1999), de-trending plays an important role in improving the performance of cointegration tests. In our procedure there is an additional reason why de-trending matters. It leads to an unexpected connection with the Jacobi ensemble, a familiar object in random matrix theory, but a novel one in econometrics.[3] Eventually, this connection is the main technical tool leading to our asymptotic results and construction of the test.

We show that after proper rescaling our test under the null hypothesis of no cointegration converges to the sum of the first r elements of the Airy_1 point process (we formally define that process in Section 3.1). Airy_1 process is a known object in the random matrix theory and its marginal distributions can be computed in various ways. We also present in Section 5 a simulation study of the speed of convergence, which supports the limiting results even for moderate values of N and T. In Table 2 of that section we demonstrate the significant improvement over the finite sample behavior of the Johansen test and its corrections reported in Gonzalo and Pitarakis (1999).

Along with the size we also report power simulations in Section 5. The power depends on the choice of the alternative and we perform several experiments with random rank 1 matrix Π and random initial condition X_0 of varying magnitudes. In many cases the power is very close to 100%. In particular, we found high power even for moderate N, T, e.g., $N \approx 30$, $T \geq 150$. In general, the results are encouraging and we see quite large power envelopes.

We also present a small empirical illustration of our testing procedure on weekly S&P100 log-prices over 10 years. This gives us approximately 500 observations across time and $T/N \approx 5$, corresponding to high power and close to 5% sample rejection rate in simulations. Log-prices are known to have a unit root, and we do not find any strong evidence that they are cointegrated.

1.3. *Techniques.* Let us briefly indicate technical aspects of our proofs and their mathematical novelty. The key observation that we make is that a small perturbation of the matrix arising in the modified Johansen test has an explicit distribution of its eigenvalues. This distribution is called Jacobi ensemble, and its usual appearances in statistics include sample canonical correlations for two sets of independent data (as opposed to highly dependent in our case, see the discussion in Section 4.1) and multivariate analysis of variance (MANOVA), see, e.g., Muirhead (2009). Our method has two main components which are new, as far as we are aware. First, our perturbation of the model is based on a replacement of a certain

[3] We remark that de-trending is implicitly present in the proofs of Onatski and Wang (2018, 2019), yet in our work it becomes a significant ingredient, rather than a technical detail.

permutation in the matrix formulation of the modified Johansen test by a uniformly random orthogonal matrix. The second ingredient is a challenging computation of matrix integrals[4] leading to the identity of the law of the perturbed matrix with the Jacobi ensemble.

Put it otherwise, we discover an exactly-solvable[5] model in a small neighborhood of the (random) matrix of the Johansen cointegration test. Let us center our attention on this exactly-solvable model. It can be used as an initial point for perturbative arguments leading to asymptotic theorems for our and possibly other modifications of the Johansen test. In addition, our exactly-solvable model is not isolated, but rather it is a representative of a whole class of similar cases. We expect that our approach works in several other situation in which various test statistics in the vector autoregression context can be understood through (other) instances of the Jacobi ensemble. Justifications of this point of view are contained in Sections 7.2 ,7.4, and 9.5.

1.4. *Outline of the paper.* Section 2 describes the setting and the main objects of interest. Section 3 presents asymptotic results, while Section 4 gives a sketch of their proofs. Section 5 shows supporting Monte Carlo simulations and Section 6 applies our test to S&P100. Additional results and extensions are presented in Section 7. Finally, Section 8 concludes. All proofs, unless otherwise noted, are in Appendix.

2. Setting.

2.1. *Building block.* We consider an N-dimensional vector autoregressive process of order 1, VAR(1), based on a sequence of i.i.d. mean zero Gaussian errors $\{\varepsilon_t\}$ with non-degenerate covariance matrix Λ. That is,

$$\Delta X_t = \Pi X_{t-1} + \mu + \varepsilon_t, \qquad t = 1, \dots, T. \tag{2}$$

where $\Delta X_t := X_t - X_{t-1}$ and Π, μ are unknown parameters. We do not impose any restrictions on Λ, thus, we allow for arbitrary correlations across coordinates of X_t. The process is initialized at fixed X_0. We outline possible extensions to autoregressions of higher order, VAR(k), in Section 7.3.

We are interested in analyzing whether X_t is cointegrated. That is, whether there exists a non-zero $N \times r_0$ matrix β such that $\beta' X_t$ is (trend) stationary.

If there exists a $N \times r_0$ matrix β of rank r_0 such that $\beta' X_t$ is (trend) stationary, but there does not exist a $N \times (r_0+1)$ matrix $\tilde{\beta}$ of rank r_0+1 such that $\tilde{\beta}' X_t$ is (trend) stationary, then we say that X_t is cointegrated of order r_0. As shown in Engle and Granger (1987, Granger representation theorem), Johansen (1995, Theorem 4.5) the necessary condition for X_t to be cointegrated of order r_0 is $\mathrm{rank}(\Pi) = r_0$ and, thus, there exist two $N \times r_0$ matrices α and β of rank r_0 such that $\Pi = \alpha\beta'$. For the sufficiency one also needs to require an extra non-degeneracy condition involving α, β, and Γ_i. This additional requirement is used to rule out the $I(2)$ processes, i.e., such that their second differences are stationary, while the first differences are not.

[4]Our arguments have some similarities with proofs in Hua (1963).

[5]A random model is called exactly-solvable or integrable (see Borodin and Gorin (2016) for an overview) if there exist explicit formulas for the expectations of non-trivial random variables describing the system. Such formulas provide a basis for asymptotic analysis of the systems of interest. This is in contrast to generic models where explicit formulas are rarely available.

2.2. *Cointegration test.* We test the null hypothesis H_0 of no cointegration, which is equivalent to $\mathrm{rank}(\Pi) = 0$ or $\Pi \equiv 0$. The complement to the null is $\mathrm{rank}(\Pi) > 0$. Yet, in order to design out test we set $\mathrm{rank}(\Pi) \in [1, r]$, where r is a fixed finite number, to be our alternative hypothesis H_1. Thus, our alternative, in line with Granger representation theorem, can be interpreted as having at most r cointegrating relationships. While the test has different power depending on r and the true data-generating process, when working with real data for which the true value of $\mathrm{rank}(\Pi) \equiv r_0$ is unknown one can use any r in order to reject H_0.[6] Formally, we test

$$H_0 : \Pi \equiv 0 \quad \text{vs.} \quad H_1 : \mathrm{rank}(\Pi) \in [1, r], \tag{3}$$

Our testing procedure is a modification of the Johansen likelihood ratio (LR) test (Johansen, 1988, 1991).[7] We focus on the small r regime, which differs from the classical Johansen LR test, where $r = N$ is more commonly used as an alternative. In that sense, our approach is close to the maximum eigenvalue ($\lambda_{\max}$) test. That turns out to be important in the large N, T setting, see Section 3.2. Our approach consists of several steps:

Step 1. De-trend[8] the data and define

$$\tilde{X}_t = X_{t-1} - \frac{t-1}{T}(X_T - X_0). \tag{4}$$

Note that we also do a time shift. This shift is in line with Johansen test, where lags and first differences are regressed on the observables.

Step 2. De-mean the data. That is, regress de-trended lags, $\tilde{X}_t$, and first differences ΔX_t on a constant. Define the residuals from those regressions as

$$R_{0t} = \Delta X_t - \frac{1}{T}\sum_{\tau=1}^{T} \Delta X_\tau, \quad R_{1t} = \tilde{X}_t - \frac{1}{T}\sum_{\tau=1}^{T} \tilde{X}_\tau. \tag{5}$$

Step 3. Calculate the squared sample canonical correlations between R_0 and R_1. That is, define

$$S_{ij} = \sum_{t=1}^{T} R_{it} R_{jt}^*, \quad i, j = 0, 1, \tag{6}$$

where here and below the notation X^* means transpose of the matrix X (transpose-conjugate whenever complex matrices appear). Then calculate N eigenvalues $\lambda_1 \geq \ldots \geq \lambda_N$ of the matrix $S_{10} S_{00}^{-1} S_{01} S_{11}^{-1}$. The eigenvalues solve the equation

$$\det(S_{10} S_{00}^{-1} S_{01} - \lambda S_{11}) = 0. \tag{7}$$

Step 4. Form the test statistic

$$LR_{N,T}(r) = \sum_{i=1}^{r} \ln(1 - \lambda_i). \tag{8}$$

The subscript N, T in (8) indicates that we modify Johansen LR test to develop the large N, T asymptotics. This statistic after centering and rescaling will be compared with appropriate critical values to decide whether one can reject H_0 or not (see Theorem 2).

[6] In practice we recommend using reasonably small values of r, since our asymptotic theorems are valid in the regime of fixed r and large N, T.

[7] Johansen LR test is based on the maximization of the Gaussian likelihood.

[8] We discuss the role of de-trending in Section 7.2.

Throughout the proofs and extensions, we will be using an alternative way to write the residuals and matrices S_{ij}. For this let us define the *demeaning* operator $\mathcal{P}$. It is a linear operator in a T–dimensional space defined by its matrix

$$\mathcal{P} = \begin{pmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \dots & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & \dots & -\frac{1}{T} \\ & & \ddots & \\ -\frac{1}{T} & -\frac{1}{T} & \dots & 1-\frac{1}{T} \end{pmatrix}. \tag{9}$$

$\mathcal{P}$ is an orthogonal projector on the space orthogonal to the vector $(1,1,\dots,1)$. By definition

$$R_{0it} = [\Delta X \mathcal{P}]_{it} = \Delta X_{it} - \frac{1}{T}\sum_{\tau=1}^{T} \Delta X_{i\tau}, \quad R_{1it} = [\tilde{X}\mathcal{P}]_{it} = \tilde{X}_{it} - \frac{1}{T}\sum_{\tau=1}^{T} \tilde{X}_{i\tau}.$$

Using the fact that $\mathcal{P}^2 = \mathcal{P}$,

$$S_{00} = \Delta X \mathcal{P} \Delta X^*, \quad S_{01} = \Delta X \mathcal{P} \tilde{X}^*, \quad S_{10} = S_{01}^* = \tilde{X}\mathcal{P}\Delta X^*, \quad S_{11} = \tilde{X}\mathcal{P}\tilde{X}^*. \tag{10}$$

Let us emphasize that our test differs from the original Johansen test in the fact that we use $\tilde{X}_t$ instead of X_{t-1}. Note that $\tilde{X}_t$ can be viewed as a rank 1 perturbation of X_{t-1}. Hence, our test statistic is a finite rank perturbation of the original Johansen procedure.

3. Asymptotic results. In this section we formulate our main asymptotic results, explain and discuss the role of the precise setting we chose, and indicate directions for generalisations. We provide a sketch of the proof of our asymptotic theorem in Section 4.

3.1. *Large* $(\boldsymbol{N}, \boldsymbol{T})$ *limit of the test.* Our results use the Airy$_1$ point process. Thus, let us introduce it before we formulate the main theorems. The Airy$_1$ point process is a random infinite sequence of reals

$$\mathfrak{a}_1 > \mathfrak{a}_2 > \mathfrak{a}_3 > \dots$$

which can be defined through the following proposition.

PROPOSITION 1 (Forrester (1993),Tracy and Widom (1996)). *Let Y_N be $N \times N$ matrix of i.i.d. $\mathcal{N}(0,2)$ Gaussian random variables and let $\mu_{1;N} \geq \mu_{2;N} \geq \dots \geq \mu_{N;N}$ be eigenvalues of $\frac{1}{2}(Y_N + Y_N^*)$. Then in the sense of convergence of finite-dimensional distributions*

$$\lim_{N\to\infty} \left\{ N^{1/6}\left(\mu_{i;N} - 2\sqrt{N}\right)\right\}_{i=1}^{N} = \{\mathfrak{a}_i\}_{i=1}^{\infty}. \tag{11}$$

The law of the first coordinate $\mathfrak{a}_1$ is known as the Tracy-Widom F_1 distribution; its distribution function can be written in terms of a solution of the Painleve II differential equation.

We remark that from the computational point of view (11) gives an efficient way to access the distribution of various functions of $\{\mathfrak{a}_i\}_{i=1}^{\infty}$.[9] From the theoretical perspective, one would like to have a more structural definition, which can be used for the analysis. Such definitions exist, yet, unfortunately, none of them is particularly simple.[10]

[9] An even faster way uses tridiagonal matrix models (Dumitriu and Edelman, 2002).

[10] There are several equivalent ways to define Airy$_1$ point process: through Pfaffian formulas for the correlation functions (Forrester, 1993; Tracy and Widom, 1996), through combinatorial formulas for the Laplace transform (Sodin, 2014), through eigenvalues of the Stochastic Airy Operator (Ramirez et al., 2011).

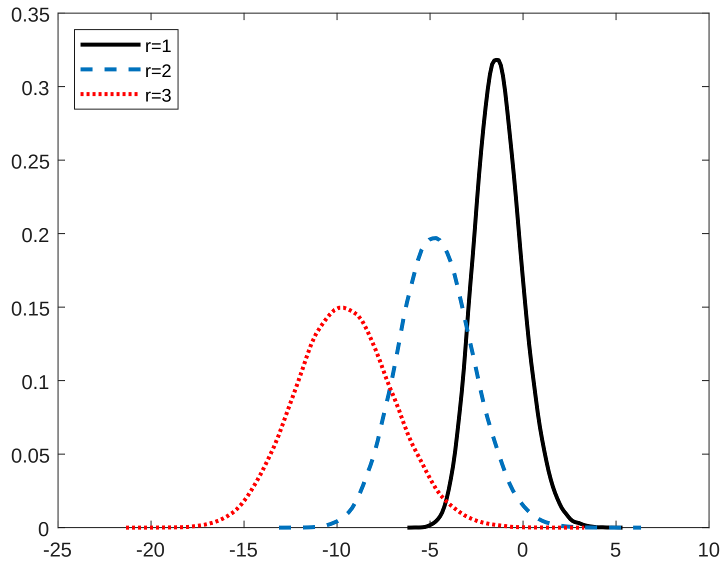


Figure 1: The probability density for the random variables $\sum_{i=1}^{r} \mathfrak{a}_i$. The histogram plots are obtained from 100,000 samples of 200×200 Gaussian matrices following Proposition 1.

THEOREM 2. *Suppose that $T, N \to \infty$ in such a way that the ratio T/N belongs to $[2+\gamma_1, \gamma_2]$ for some $\gamma_1, \gamma_2 > 0$. Suppose that H_0 holds, that is, we have (2) with $\Pi = 0$. Then for each finite $r = 1, 2, \ldots,$ we have convergence in distribution for the largest eigenvalues defined in Eq. (7):*

$$\frac{\sum_{i=1}^{r} \ln(1-\lambda_i) - r \cdot c_1(N,T)}{N^{-2/3} c_2(N,T)} \xrightarrow[T,N\to\infty]{d} \sum_{i=1}^{r} \mathfrak{a}_i, \tag{12}$$

where

(13)

$$c_1(N,T) = \ln(1-\lambda_+), \qquad c_2(N,T) = -\frac{2^{2/3}\lambda_+^{2/3}}{(1-\lambda_+)^{1/3}(\lambda_+ - \lambda_-)^{1/3}} (\mathfrak{p}+\mathfrak{q})^{-2/3} < 0,$$

$$\mathfrak{p} = 2 - \frac{2}{N}, \qquad \mathfrak{q} = \frac{T}{N} - 1 - \frac{2}{N}, \qquad \lambda_\pm = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2}\left[\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}}\right]^2. \tag{14}$$

REMARK 1. *The condition $T/N > 2$ is important: The procedure for constructing λ_i involves computing squared sample canonical correlations between two N–dimensional subspaces in T–dimensional space. If $T/N < 2$, then two subspaces necessary intersect, leading to $\lambda_1 = 1$. Hence, Eq. (12) would need an adjustment in such case.*

Figure 1 shows densities for the random variables $\sum_{i=1}^{r} \mathfrak{a}_i$ for $r = 1, 2, 3$. As r grows, the skewness of $\sum_{i=1}^{r} \mathfrak{a}_i$ decreases, the expectations of $\sum_{i=1}^{r} \mathfrak{a}_i$ tend to $-\infty$ and the variances tend to $+\infty$.

Quantiles of the distribution of $\sum_{i=1}^{r} \mathfrak{a}_i$ serve as critical values for testing the hypothesis H_0 of no cointegrations against the alternative of at most r cointegrations. We report the quantiles for $r = 1, 2, 3$ in Table 1.

$r \backslash \alpha$	0.9	0.95	0.975	0.99
1	0.44	0.97	1.45	2.01
2	-1.88	-1.09	-0.40	0.41
3	-5.91	-4.91	-4.03	-2.99

TABLE 1
Quantiles of $\sum_{i=1}^{r} \mathfrak{a}_i$ *for* $r = 1, 2, 3$ *(based on* 10^6 *Monte Carlo simulations of* $10^8 \times 10^8$ *tridiagonal matrices of Dumitriu and Edelman (2002)).*

Note the shifts by $\frac{2}{N}$ in the definition (14) of $\mathfrak{p}$ and $\mathfrak{q}$. Such finite sample correction is inspired by Johnstone (2008, Discussion before Theorem 1). Theorem 2 remains valid without these shifts, i.e., for the simplified $\mathfrak{p} = 2$, $\mathfrak{q} = \frac{T}{N} - 1$. However, our simulations show that for $T/N < 6$ the shifts improve the speed of convergence and, thus, the finite sample behavior of the test, see Sections 5 and 9.3 for details.

We sketch the proof of Theorem 2 in Section 4 and give full details in Appendix. The two main ideas that we use are:

(I) We show that a small (vanishing as $N, T \to \infty$) perturbation of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ leads to an explicit probability distribution known as the Jacobi ensemble (see Definition 3 below). While this distribution appears in many problems of random matrix theory and multivariate statistics, its connection to the law of the Johansen test statistic remained previously unknown.

(II) Further, we rely on the universality phenomenon from random matrix theory, which says that the particular choice ($\frac{1}{2}(Y_N + Y_N^*)$) for the law of random matrix made in Proposition 1 is not of central importance. Instead, the Airy_1 point process is a universal scaling limit for the largest eigenvalues in various ensembles of symmetric random matrices of growing sizes, see, e.g., Erdos and Yau (2012) and Tao and Vu (2012). In particular, an asymptotic statement similar to Proposition 1 is known for the Jacobi ensemble (see Section 9.3) and, by combining it with the first idea, we eventually arrive at Theorem 2.

It is natural to ask about an extension of Theorem 2 for the case of r growing together with N. We distinguish two subcases here:

- Slow growth: $r = \lfloor N^\theta \rfloor$ for some $0 < \theta < 1$.
- Linear growth: $r = \lfloor \rho N \rfloor$ for some $0 < \rho \leq 1$.

In both cases we expect that (after proper adjustment of c_1 and c_2) the limit in (12) becomes Gaussian. Although we are not going to pursue this direction here, for the slow growth case the proof of asymptotic normality can probably be obtained by the same methods as we use in Theorem 2: we can again use the Jacobi ensemble for the computation. For the Jacobi ensemble individual eigenvalues λ_i (in the regime of growing i and $N - i$) become asymptotically Gaussian, hence, also their sums.[11]

For the linear growth case the situation is more complicated. Our present tools only allow us to prove an asymptotic upper-bound of the form

$$\sum_{i=1}^{\lfloor \rho N \rfloor} \ln(1 - \lambda_i) - N \cdot c_3(T/N, \rho) = o(N^\epsilon), \quad N \to \infty, \tag{15}$$

[11] O'Rourke (2010) proves asymptotic Gaussianity for eigenvalues of Gaussian random matrices and we expect the same proof to work for the Jacobi ensemble case, see also Bao et al. (2013) for more details in the complex setting.

for any $\epsilon > 0$ (for $\epsilon = 1$ this also follows from the results of Onatski and Wang (2018, 2019)),[12] while we expect the expression to be $O(1)$. Yet, there exist very general statements on asymptotic Gaussianity of linear statistics of functions of random matrices (see, e.g., Guionnet and Novak (2015), Mingo and Popa (2013)), and, therefore, there is little doubt in the fact that asymptotic distribution is Gaussian. Note, however, that these methods usually give very limited information about asymptotic variance and it is unclear at this moment how to find a reasonably simple explicit formula for it.

3.2. *Theorem 2 and the role of finite* r*: Discussion.* An important feature of our setting and our asymptotic results is that r is kept finite as $N, T \to \infty$. (As far as the authors know, Eq. (12) in Theorem 2 is the very first statistic for which the computation of the distributional limit became possible in the $N, T \to \infty$, $T/N \in [2+\gamma_1, \gamma_2]$ regime.) In practice, when the dimensions are given to us in the data, this means that r should be much smaller than N and T. Notice that for the purpose of rejecting H_0 any statistic with known asymptotic distribution can be useful. It could be tempting to instead choose $r = N$ and test H_0 of no cointegrations against the alternative of arbitrary number of cointegrations, yet, we believe that in many situations small r allows to infer much more and is a good choice by the following reasons:

3.2.1. *What can be estimated and when?.* The ultimate goal of the cointegration analysis is to find the cointegrating relationships. Turns out, the feasibility of that task crucially depends on their true number. Imagine for a second that the true number of cointegrating relationships is large, say $N/2$, then they span an $N/2$–dimensional subspace of $\mathbb{R}^N$. Thus, the number of unknown parameters which need to be estimated is N^2 up to a multiplicative constant. On the other hand, the total number of observations that we have is NT. Hence, in our asymptotic regime $N, T \to \infty$, $T/N \in [2+\gamma_1, \gamma_2]$, we have only finitely many observations per estimated parameter and consistent estimation is unlikely. The conclusion from this dimension-based heuristics is that in the situation when the number of cointegration relationships is large, our data does not contain enough information for the consistent estimation of all cointegrating relationships and, in a sense, the problem is hopeless. On the other hand, the same heuristics shows that we are in a much better scenario, if we have some a priory reasons to expect that either the number of cointegrating relationships is small, or if this number is in a small neighbourhood of N (in the latter case instead of the space of cointegrating relationships we can estimate its low-dimensional orthogonal complement). The former situation precisely corresponds to our choice of small r in Theorem 2; the first step towards the asymptotic analysis in the latter situation is discussed in Section 7.4.

3.2.2. *Estimation of rank and cointegration relationships.* In the small N cointegration analysis described in detail in Johansen (1988, 1991, 1995), rejection of the hypothesis of no cointegration is the first step. The next step is the estimation of the true rank r_0 of Π in (2) by subsequent rejections of the hypotheses of the rank being smaller than $2, 3, \dots, r_0$. Finally one wants to consistently estimate r_0 cointegrating relationships, closely related to Π itself.

Switching to our limit regime $N, T \to \infty$, $T/N \in [2+\gamma_1, \gamma_2]$, there are two separate cases. If r_0 also grows linearly in N, then the full success of the above program is unlikely: as outlined in Section 3.2.1, we are trying to consistently estimate too many parameters from too few data points.

[12]The value of the constant c_3 can be computed through certain integrals involving the equilibrium measure of the Jacobi ensemble, see (72) and (75) for more details.

On the other hand, if true r_0 is small, then the situation is different (another potentially feasible case is that of small $N - r_0$, which we do not discuss here). We believe that our present results — Theorem 2 and the new approach underlying it — open a path to the full realization of the cointegration analysis for growing N. Let us outline the reasons.

From the technical point of view a key ingredient which needs to be developed is the asymptotics of squared sample canonical correlations $\lambda_1, \dots, \lambda_N$ and corresponding eigenvectors for the model (2) with Π of a fixed and not growing rank r_0. In particular Theorem 2 should be extended from $H_0: \, r_0 = 0$ to arbitrary finite r_0. A natural approach here is by further developing the link to random matrices established in our proof of Theorem 2 via the Jacobi ensemble. This leads to the theory of spiked random matrices. A basic question of this theory is to infer the information on a large-dimensional small rank matrix B from observing $A + B$, where A is a matrix of pure noise (e.g., one can use $A = \frac{1}{2}(Y_N + Y_N^*)$ in the notations of Proposition 1). This is a well-developed theory, see e.g., Johnstone (2001), Baik et al. (2005) for seminal contributions and Bao et al. (2019), Johnstone and Onatski (2020) for the most recent progress. In order to obtain a connection with the VAR-framework (2), we treat the $rank(\Pi) = r_0$ case as a finite rank deformation of the "pure noise case" $\Pi = 0$, which we consider in this text. Hence, by combining our present results with the spiked random matrix theory, we expect to establish the complete asymptotic theory for any finite as $N \to \infty$ value of r_0. (This will require further technical and conceptual efforts and, therefore, is left for future work.)

3.2.3. *Power.* Suppose that the true number of cointegrating relationships is small, say, $r_0 = 1$, yet we are trying to reject the null of no cointegrations using the LR-statistic with $r = N$, i.e.,

$$\sum_{i=1}^{N} \ln(1 - \lambda_i) \tag{16}$$

in the notations of Theorem 2. As we explain at the end of Section 3.1, the standard deviation of (16) is $O(1)$, i.e., after subtracting a constant as in Eq. (15), statistic (16) stays finite as $N, T \to \infty$. Hence, in contrast to (12) there should be no N–dependent rescaling in a test based on (16). This should lead to very different powers of tests based on statistics (12) and (16).

Indeed, we expect the main difference in the joint law of $(\lambda_1, \dots, \lambda_N)$ under the H_0 and under the alternative of one cointegrating relationship to be in the behavior of λ_1 (which is in line with $\ln(1 - \lambda_1)$ being the likelihood ratio test statistic in such setting). Change of λ_1 by $\nu \in \mathbb{R}$, i.e., $\lambda_1 \to \lambda_1 + \nu$, is on the same scale as the standard deviation of (16), and ν needs to be large in order for the test statistic to detect this change and to reject H_0. On the other hand, (12) is changing in the same situation by a number of order $N^{2/3}\nu$, which is much larger, and, hence, rejection of H_0 is more likely. Therefore, one could anticipate that (12) leads to a test of higher power than (16) in this situation.[13]

3.2.4. *Checking model validity.* The fact that we assume r is finite and only use first r largest eigenvalues allows us to use the remaining ones to verify the plausibility of our model specification. That is, we can check whether the statistical properties of $\lambda_{r+1}, \dots, \lambda_N$ agree with the predictions of our model: for any finite r we expect the empirical distribution function $\frac{1}{N-r} \sum_{i=r+1}^{N} \mathbf{1}_{\lambda_i \le x}$ to converge to the CDF of the Wachter distribution (under H_0 this

[13] We do not analyze (16) in this text and, thus, we do not provide any rigorous justification of the above. We remark that the existing in the literature similar comparisons, e.g., Lütkepohl et al. (2001), Paruolo (2001), do not apply in our situation, since they only deal with the small N case.

can be deduced from our Theorems 4 and 14, while more general result under H_1 with finite r follows from (Onatski and Wang, 2018, Theorem 1)). Our empirical example in Section 6, indeed, shows the match of the S&P100 data with the Wachter distribution (see Figure 7).

4. Outline of the proofs. The proof of Theorem 2 rests on the notion of the Jacobi ensemble. Let us first define it and then provide a sketch of the proof of Theorem 2.

4.1. *Jacobi ensemble.*

DEFINITION 3. *A (real) Jacobi ensemble* $\mathbf{J}(N;p,q)$ *is a distribution on* $N \times N$ *symmetric matrices* M *of density proportional to*

$$\det(M)^{p-1}\det(I_N - M)^{q-1}, \quad 0 < M < I_N, \tag{17}$$

with respect to the Lebesgue measure, where $p,q > 0$ *are two parameters and* $0 < M < I_N$ *means that both* M *and* $I_N - M$ *are positive definite.*

When $N = 1$, (17) is the Beta distribution. For general N, the eigenvectors of random Jacobi-distributed M are uniformly distributed, while N eigenvalues $x_1 \geq \cdots \geq x_N$ admit an explicit density with respect to the Lebesgue measure given by

$$\frac{1}{Z(N;p,q)} \prod_{1\leq i<j\leq N} (x_i - x_j) \prod_{i=1}^{N} x_i^{p-1}(1-x_i)^{q-1}, \tag{18}$$

where $Z(N;p,q)$ is an (explicitly known) normalization constant.

The Jacobi ensemble is widely used in statistics (Muirhead, 2009), theoretical physics (Mehta, 2004), and random matrix theory (Forrester, 2010). There are numerous tools for studying it[14], and as a result its asymptotic properties (usually as $N \to \infty$) are known in great detail.

Two particularly important appearances of the Jacobi ensemble in statistics are in the multivariate analysis of variance (MANOVA) and in the canonical-correlation analysis, see Johnstone (2008, Sections 1 and 2) for more detailed discussions and more examples. Let us explain the latter setting, as it has some resemblance with the Johansen test. We start with two rectangular matrices of data: X of size $N \times T$ and Y of size $K \times T$. One can interpret X_{it} (and Y_{it}) as the ith variable at time t.

If $N = K = 1$, then one can think about data sets X and Y as T observations of two variables. Then one can measure the (linear) dependence between these variables by computing the (squared) sample correlation coefficient:

$$\frac{\left(\sum_{t=1}^{T} X_t Y_t\right)^2}{\sum_{t=1}^{T}(X_t)^2 \sum_{t=1}^{T}(Y_t)^2}. \tag{19}$$

A direct computation shows that if the elements of X and Y are i.i.d. mean zero Gaussians, then the law of (19) is given by Beta distribution for any $T = 1, 2, \ldots$.

A generalization of (19) to $N, K \geq 1$ defines the *squared sample canonical correlations* $\{r_i\}$ as solutions to the equation

$$\det\left(S_{XY} S_{YY}^{-1} S_{YX} - r S_{XX}\right) = 0, \tag{20}$$

[14] Just to mention some: Pfaffian point processes (Mehta, 2004), combinatorics of moments (Bai and Silverstein, 2010), variational problems for log-gases (Ben Arous and Guionnet, 1997), Kadell integrals (Kadell, 1997), Schwinger-Dyson/Loop equations (Johansson, 1998), tridiagonal models (Killip and Nenciu, 2004), Macdonald processes (Borodin and Gorin, 2015).

where $S_{XY} = XY^*$, $S_{YY} = YY^*$, $S_{YX} = YX^*$, $S_{XX} = XX^*$ (by X^* we mean transpose of X when X is a real matrix and conjugate-transpose of X when X is complex). Generically the equation (20) has $\min(N,K)$ non-zero solutions. They have a variational meaning. For instance, the maximal solution of Eq. (20) is the maximal squared correlation coefficient between a^*X and b^*Y, where the maximization goes over all N–dimensional vectors a and K–dimensional vectors b.

Now suppose that the columns of X are i.i.d. N–dimensional mean 0 Gaussians with (non-degenerate) covariance matrix Λ and the columns of Y are i.i.d. K–dimensional mean 0 Gaussians with (non-degenerate) covariance matrix Λ'. Assume for simplicity that $N \le K$ and $N + K \le T$. If X and Y are independent, then it can be shown that the (squared) sample canonical correlations $r_1 > r_2 > \cdots > r_N$ have the law of the eigenvalues of the Jacobi ensemble $\mathbf{J}(N; \frac{K-N+1}{2}, \frac{T-N-K+1}{2})$.

At this point we observe both a similarity and a difference between the above instance of the Jacobi ensemble and the matrix appearing in the Johansen test. On one hand, the latter also deals with sample canonical correlations of two data sets. On the other hand, the data sets are no longer independent, instead one is obtained from another by a deterministic linear transformation. So are the roots of Eq. (7) related to the Jacobi ensemble? There is evidence in both directions. First, computer simulations quickly reveal that in one-dimensional case the distribution of the single eigenvalue in the Johansen test is *not* the Beta distribution. Yet, second, recent results of Onatski and Wang (2018, 2019) show that the Law of Large Numbers for the empirical distribution of the eigenvalues appearing in the Johansen test (with roots of Eq. (7) being a particular case) matches the one for the Jacobi ensemble (with shifted dimension parameters) in the limit as $N, T \to \infty$. Those articles were asking for an explanation.

4.2. *Sketch of the proof of Theorem 2.* To prove Theorem 2, we first need to establish the following central statement: a small perturbation of the Johansen test matrix, obtained by replacing the deterministic summation matrix in its definition by a random analogue, exactly matches the Jacobi ensemble. We further show that the perturbation vanishes in the limit as $N, T \to \infty$, thus, allowing us to obtain the asymptotics of the variants of the Johansen test from the known asymptotic results for the Jacobi ensemble.

THEOREM 4. *Suppose that $T, N \to \infty$ in such a way that $T > 2N$ and the ratio T/N remains bounded. Under the hypothesis $\Pi = 0$ for (2), one can couple (i.e. define on the same probability space) the eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$ of the matrix $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ and eigenvalues $x_1 \ge \cdots \ge x_N$ of the Jacobi ensemble $\mathbf{J}(N; \frac{N}{2}, \frac{T-2N}{2})$ in such a way that for each $\epsilon > 0$ we have*

$$\lim_{T,N\to\infty} \mathrm{Prob}\left(\max_{1\le i\le N} |\lambda_i - x_i| < \frac{1}{N^{1-\epsilon}} \right) = 1.$$

The proof of Theorem 4 is based on the following idea. Looking at Eq. (2) when $\Pi = 0$, one can notice that matrices entering into the test and given by Eq. (10) can be expressed in terms of the matrix of data X and the lag operator mapping $X_t \to X_{t-1}$. A computation shows that since we deal with the de-trended and de-meaned data, we can replace the lag operator with its *cyclic version*, which maps X_1 to X_T rather than X_0. The latter is an orthogonal operator whose eigenvalues are roots of unity of order T. Then, the idea is to replace this operator by *uniformly random* orthogonal operator. From that we proceed in two steps:

(I) We show that when the lag operator is replaced by its random counterpart, the eigenvalues of $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ have distribution $\mathbf{J}(N; \frac{N}{2}, \frac{T-2N}{2})$. We remark that this is a new appearance of the Jacobi ensemble, which was not present in the literature before.

(II) We show that replacement of the lag operator by its random counterpart introduces an error, which can be upper-bounded by $N^{\epsilon-1}$ for any $\epsilon > 0$. This part is based on the rigidity results from random matrix theory, which say that eigenvalues of a uniformly random orthogonal matrix can be very closely approximated by equally spaced roots of unity.

The full proof of Theorem 4 is given in Appendix. In addition to the real case, we also simultaneously prove a similar statement for complex matrices, encountering Jacobi ensemble of Hermitian matrices. Generally complex settings are rare guests in economics. Yet, they play a major role in the spectral analysis of time series data and in other areas, such as high energy physics.

By combining Theorem 4 with known asymptotic results for the Jacobi ensemble we can obtain the asymptotics of test statistic (8) in various regimes.

5. Monte Carlo simulations. In this section we illustrate the performance of our test via Monte Carlo simulations. We consider both size (rejection rate) and power.

5.1. *Rejection rate.* First, we compare the finite sample performance of our approach versus Johansen's LR test and one of its corrected versions for (reasonably small) $T = 30$ and $N = 5, \dots, 10$. The finite sample correction takes the form $\frac{T-N}{T}$ and was suggested in Reinsel and Ahn (1992). (Let us note that there are several more advanced empirical finite sample correction procedures for Johansen's LR test and its variations, we refer to Onatski and Wang (2019, Table 2) for a recent comparison of some of those for a wide range of values of N and T.) Table 2 summarizes the results (numbers closer to 5 mean better performance). The $LR_{N,T}$ column is our test and the last two columns are from Gonzalo and Pitarakis (1999). They correspond to the same null of no cointegration, but use a different from ours alternative hypothesis. Our alternative is at most r cointegrating relationships (in Table 2 we use $r = 1$, which means we look at the largest eigenvalue). The LR and RALR tests consider "at most N" cointegrations as H_1, which means that they use the sum of all eigenvalues. We can see that in finite samples when both N and T are of a similar magnitude, our approach significantly outperforms the alternatives based on small N, large T asymptotics.

N	$LR_{N,T}$, $r = 1$	LR	RALR
5	6.60	20.75	3.59
6	5.45	31.66	2.68
7	4.52	47.44	1.98
8	3.80	67.42	2.00
9	3.16	85.00	1.32
10	2.60	96.69	0.96

TABLE 2

Empirical size under no cointegration hypothesis (5% *nominal level). Data generating process:* $\Delta X_{it} = \varepsilon_{it}$*,* $\varepsilon_{it} \sim$ *i.i.d.* $\mathcal{N}(0,1)$*,* $T = 30$*,* $MC = 1{,}000{,}000$ *replications for* $LR_{N,T}$ *and* $MC = 10{,}000$ *for LR and RALR.*

To illustrate the performance of our test as the sample size increases, we fix the ratio $T/N \equiv c$ and plot the empirical size as a function of N. This is shown in Figure 2, where the target is 5% rejection rate. The picture suggests that the test has rejection rate close to 5%. The green solid line corresponding to $c = 4$ achieves 5% rejection rate very fast. Other three curves overshoot 5% by couple percents (e.g., both $c = 5$ and $c = 6$ are always below 7%). Moreover, the larger is c, the higher is the corresponding rejection curve. E.g., $c = 10$ curve (blue, short dash) is strictly above other curves.

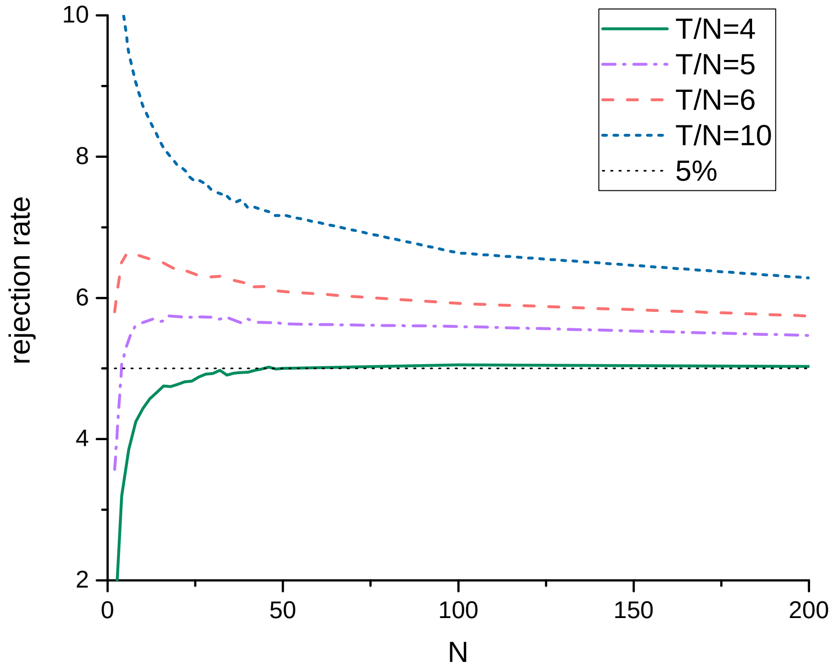


Figure 2: Empirical size when T/N is fixed and N increases.

To improve the finite-sample behavior for large c we suggest to ignore the $\frac{2}{N}$ correction in $\mathfrak{p}$ and $\mathfrak{q}$ in Theorem 2, Eq. (14), when T/N is large. That is, to use simplified formulas instead: $\mathfrak{p} = 2$, $\mathfrak{q} = T/N - 1$. We recalculate the empirical rejection rate under the simplified formulas for $\mathfrak{p}, \mathfrak{q}$ in Figure 3. Under the simplified formulas, the larger is c, the smaller is over-rejection and the closer the curve is to 5% line. As can be seen by comparing Figures 2 and 3, for small c the formula with the correction leads to better rejection rates, while for large c it is the opposite, and we do not gain from finite sample correction. The conclusion is to use the simplified formula when $c = T/N$ is at least 6.

An important feature of Figures 2 and 3 is that as $N, T \to \infty$ with fixed T/N the results improve towards the perfect match to 5%. This is in contrast to various finite sample corrections of Johansen's LR test used earlier. In particular, examination of Table 2 in Onatski and Wang (2019) reveals that each procedure has its own pairs (N, T) where the results are close to perfect (empirical size of the test is very close to the desired 5%) and some others where the results are much worse, yet rejection rates in general do not improve as the sample size increases keeping T/N fixed.

5.2. *Power.* When one analyzes the power of a test, the question is which data generating process (dgp) to use under an alternative H_1. In our setting the space of alternative dgps has growing with N dimension. Thus, there is no clear choice of the proxy alternative hypothesis to use in simulations. Hence, instead we proceed with various random alternatives.[15] Therefore, the corresponding power is also random. For illustrational convenience, we only report averages of those powers.

To analyze the power of our test we conduct several experiments. In all of them the errors ε_{it} are generated as i.i.d. $\mathcal{N}(0,1)$. In the first experiment we randomly sample a matrix Π of

[15] Another approach would be to stick to some ad hoc matrix Π. Yet, we do not expect that to affect our simulations in a qualitatively significant way.

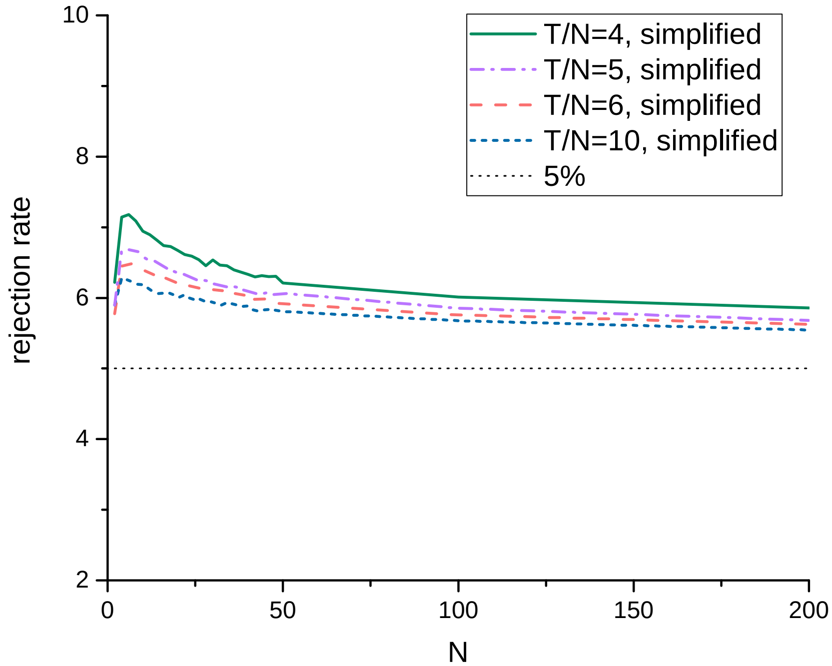


Figure 3: Empirical size when T/N is fixed and N increases. Test based on $\mathfrak{p} = 2$, $\mathfrak{q} = T/N - 1$.

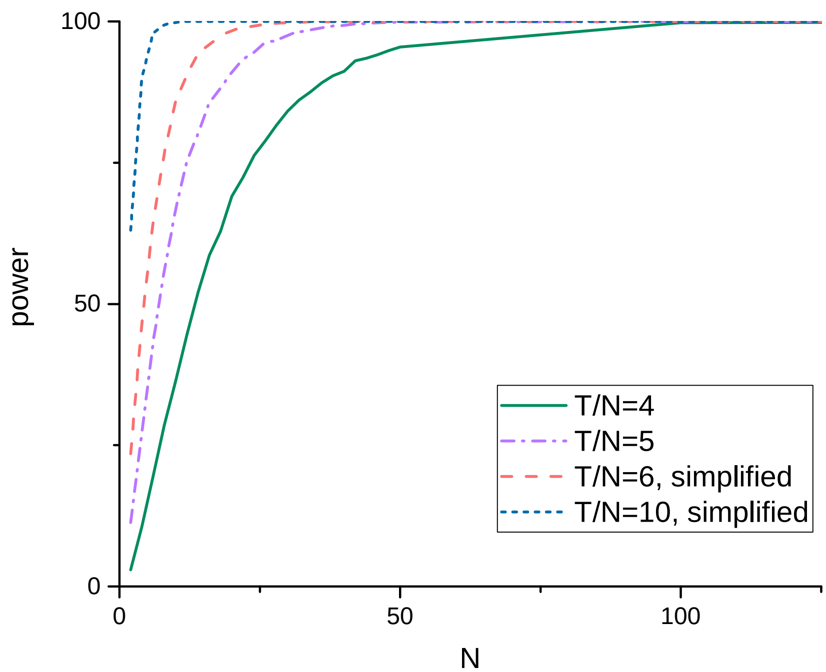


Figure 4: Power against random alternative of 1 cointegrating relationship when T/N is fixed and N increases.

rank 1. We do this by generating two uniformly random N–dimensional unit vectors, u, v, such that the non-zero eigenvalue of uv^* is negative. Then we set $\Pi = uv^*$. By construction,

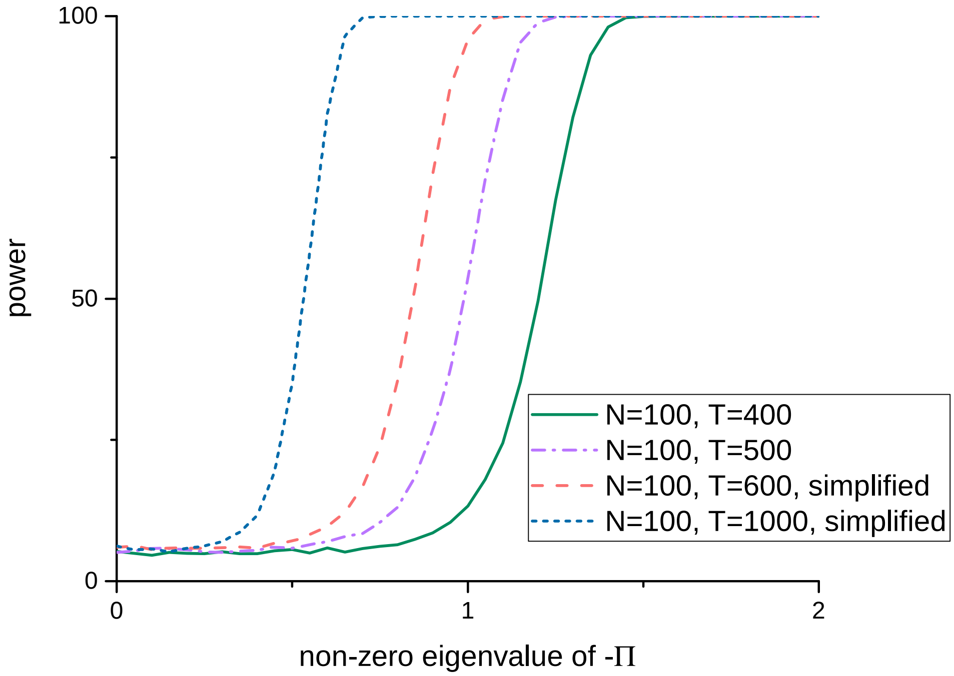


Figure 5: Power against random alternative of 1 cointegrating relationship when Π is symmetric. To ensure stationarity of ΔX a non-zero eigenvalue of $-\Pi$ lies between 0 and 2.

Π has rank 1, singular value 1, and one eigenvalue between -1 and 0 (others are zero),[16] so that X_t is non-stationary, while ΔX_t is stationary. The average power constructed from such random alternative is shown in Figure 4. As in the previous subsection we fix the ratio T/N and plot the average power as a function of N. Following our recommendation, we use simplified formulas for $\mathfrak{p}$ and $\mathfrak{q}$ when $T/N \geq 6$. We can see that the average power quickly reaches 100% for all ratios of T/N. The larger is that ratio, the faster we reach 100%. This is due to the fact that higher ratio means higher time span. This gives the process more chances to accumulate the effects of the presence of cointegration.

In the second experiment, we randomly generate a symmetric matrix Π of rank 1. We do this by generating a uniformly random N–dimensional unit vector v. Then we set $\Pi = -\lambda v v^*$, where λ goes from 0 to 2. The coefficient λ equals to the non-zero eigenvalue of $-\Pi$. The fact that it lies between 0 and 2 guarantees that X_t is non-stationary, while ΔX_t is stationary. Figure 5 shows the power as a function of λ for $N = 100$ and various values of T corresponding to $T/N = \{4, 5, 6, 10\}$ as in the previous experiments. The larger is $-\lambda$, the larger is power, which eventually reaches 100%. When $\lambda = 0$, the dgp corresponds to the null H_0. Thus, all curves start from $\approx 5\%$, which corresponds to the empirical size of the test. We can see that again the larger is T/N, the faster we reach 100%. The reason is the same as in the previous simulation.

Drawing the intuition from the unit root testing literature (e.g., Müller and Elliott (2003), Harvey et al. (2009)), we also analyze the sensitivity of power to the choice of initial condition, X_0.[17] In two previous experiments we set $X_0 = 0$. Figure 6 shows how the power against random alternative of 1 cointegrating relationship for symmetric Π changes for various magnitudes of X_0. To be more specific, we redo the same simulations as in the previous

[16] As N goes to infinity the non-zero eigenvalue is of order $N^{-1/2}$.

[17] Note that X_0 does not affect rejection rates in Section 5.1, because under H_0 it cancels out.

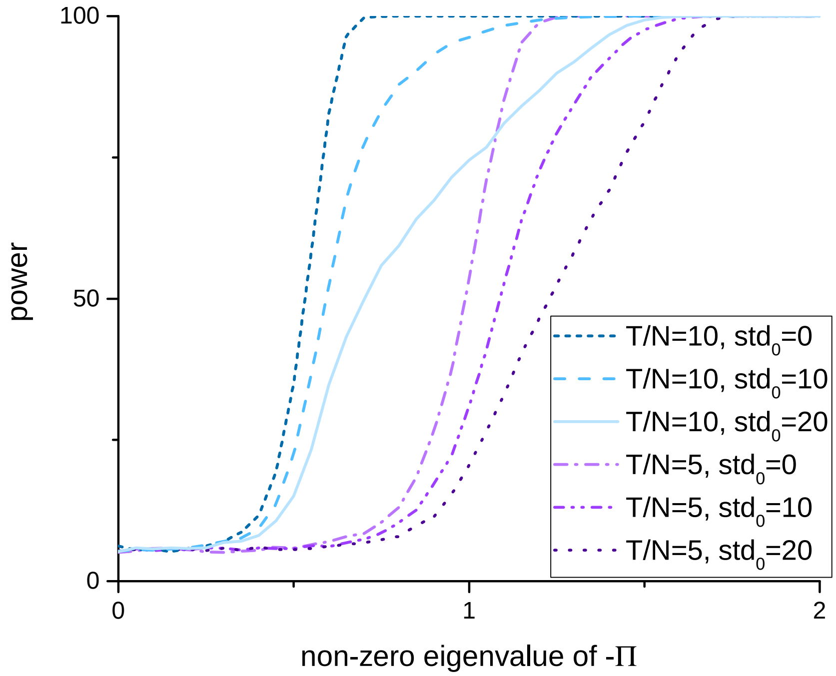


Figure 6: Power against random alternative of 1 cointegrating relationship when Π is symmetric. Initial condition is chosen as $X_0 = \text{std}_0 \cdot \mathcal{N}(0,1)$, $N = 100$.

paragraph, but start with $X_{i0} \sim$ i.i.d. $\text{std}_0 \cdot \mathcal{N}(0,1)$ for each Monte Carlo draw. We consider $T/N = 5$ and $T/N = 10$. Curves with $\text{std}_0 = 0$ are the same as on Figure 5 (they are also represented with the same style and color on both pictures). We can see that the larger is the magnitude of X_0, as measured by its standard deviation std_0, the slower the power reaches 100%.

In contrast to the previous paragraph, the power against random alternative of 1 cointegrating relationship when Π is asymmetric (as in Figure 4) does not exhibit any substantial changes with respect to the magnitude of X_0. Hence, we do not redraw the analogue of Figure 4 for non-zero X_0.

Overall, the simulations suggests the good asymptotic performance of our test procedure both under H_0 and H_1. The theoretical analysis of the power remains an open and challenging question.

6. Empirical illustration. In this section we illustrate our testing strategy by analyzing cointegration in log prices of various stocks. The search for cointegrations is a part of a stock market strategy called pairs trading, see, e.g., Krauss (2017) and references therein. For us this is a convenient testing ground, as both N and T are large.

We use logarithms of weekly S&P100 prices over ten years: $01.01.2010 - 01.01.2020$, which gives us 522 observations across time. The time range is chosen so that we do not need to worry about potential structural breaks due to financial crisis of $2007 - 2008$ and due to COVID-19. S&P100 includes 101 (because one of its component companies, Google, has 2 classes of stock) leading U.S. stocks with exchange-listed options. We use 92 of those stocks (those which were available for the whole ten years span and only one of two Google's stocks). More details on those stocks are in Section 9.6 in Appendix. Therefore, $N = 92$, $T = 521$ and $T/N \approx 5.66$.

Figure 7 shows the histogram of eigenvalues which solve Eq. 7 for the S&P100 data. The key feature of this histogram is that it closely resembles the Wachter distribution, which density is shown by thick orange line in Figure 7. This distribution governs the asymptotics of the eigenvalues of the Jacobi ensemble (see Section 9.3 in Appendix for the details). In particular, we see a precise match between supports of the histogram and of the Wachter distribution. The resemblance is in line with our Theorem 4. Indeed, if we believe that the true data-generating process (2) has $\Pi = 0$, then the theorem combined with the asymptotics of the Jacobi ensemble of Theorem 14 predicts convergence to the Wachter distribution. The conclusion is that Figure 7 shows a close match between theoretical predictions and real data. Simultaneously, this figure is consistent with H_0 of no cointegration.

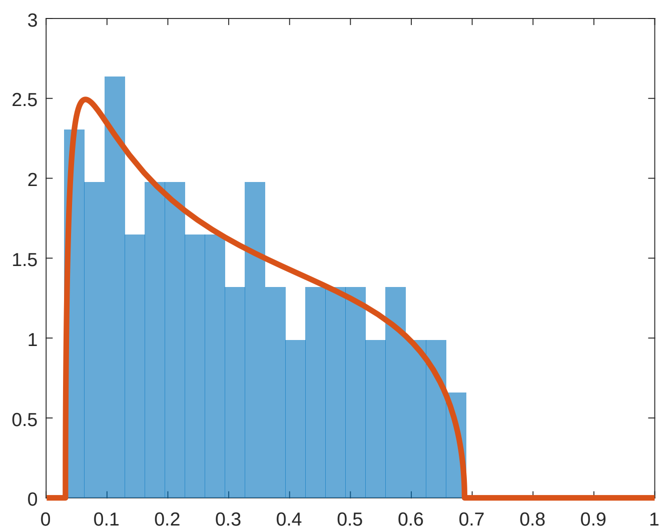


Figure 7: Eigenvalues from S&P data and Wachter distribution.

We compute our test statistic for $r = 1, 2, 3$ using the S&P100 data. In neither of the cases the value of the statistic is large enough for a statistically significant rejection of H_0. If $r = 1$, then the value is -0.27, which corresponds to approximately the 0.78 quantile of the asymptotic distribution shown in Figure 1. For $r = 2, 3$ the values are closer to the right tails of the distribution, but still below the (one-sided) 0.95 quantiles. Hence, we do not see evidence towards the presence of cointegration in S&P100 stock prices for the last 10 years.

7. Extensions. Let us describe possible extensions and modifications of our results. In Subsection 7.1 we consider non-Gaussian errors ε_t. Subsection 7.2 looks at the model without an intercept μ (linear trend in X_t) and considers the effect of de-trending vs. no de-trending in such setting. Subsection 7.3 investigates the performance of our test when the true process follows higher order of autoregression. Finally, in Subsection 7.4 we discuss testing the hypothesis $\Pi = -I_N$ using the same approach as for $\Pi = 0$.

7.1. *Non-gaussian errors.* The result of Theorem 2 is obtained under the assumption that the errors ε_t in Eq. (2) are Gaussian. However, we believe that it should be possible to remove this restriction and it is reasonable to expect that the very same statement should hold for any (independent and identical across time t) distribution of ε_t, as long as it has sufficiently many moments. The underlying reason for this belief is the so-called universality phenomenon in the random matrix theory: asymptotic local spectral characteristics of a random matrix are almost independent of the distributions of the matrix elements, see, e.g., Erdos and Yau (2012), Tao and Vu (2012) for general reviews and Han et al. (2016), Han et al. (2018) for the recent work in contexts of multivariate analysis of variance and canonical correlations. In particular, for the Wigner matrices ($Y + Y^*$, where Y is a square matrix with real i.i.d. entries), it is known that the asymptotic behavior of the largest eigenvalues depends only on the first two moments (expectation and variance) of the distribution of an individual matrix element. Note, however, that our asymptotic result in Theorem 2 holds for any choice of the first two moment of Gaussian noise ε_t: in Eq. (2) the covariance matrix Λ is arbitrary and any shift in expectation can be absorbed into the parameter μ. Hence, we conjecture that the result of Theorem 2 would hold for any distribution of ε_t, as long as it is sufficiently well-behaved.

In order to test this conjecture we made simulations for the case when elements of ε_t are non-gaussian, but i.i.d. across both i and t (corresponding to a diagonal covariance matrix Λ). We ran Monte-Carlo simulations for three different distributions for ε_{it}: uniform on the interval $[0, 1]$, uniform on 3 points $\{1, 2, 3\}$, and the product of two independent $\mathcal{N}(0, 1)$ random variables. In each case for $T = 900$, $N = 300$, and small values of r, we do not see any significant changes in the distribution of $\sum_{i=1}^{r} \ln(1 - \lambda_i)$ from the limit in Theorem 2. However, things go differently when the distribution has heavy tails. In the forth experiment we chose ε_{it} to be Cauchy-distributed, and then the distribution of $\sum_{i=1}^{r} \ln(1 - \lambda_i)$ dramatically changed from what we saw in the Gaussian case. Hence, we conclude, that the existence of at least some number of moments of the errors $\{\varepsilon_t\}$ should be necessary for the validity of the conjecture. Rigorous proof of the conjecture remains an important problem for the future research.

7.2. *Model without trend.* One of the important steps in our testing procedure is de-trending the data. Moreover, the exact form of the de-trending that we use (we subtract the slope of the line which connects X_0 at $t = 0$ and X_T at $t = T$) is an ingredient substantially used in the proof of Theorem 2. Yet, we expect that this is just a technical artifact and statements similar to Theorem 2 should also be true with other forms of de-trending or in models where this step is not needed at all. Let us provide some evidence.

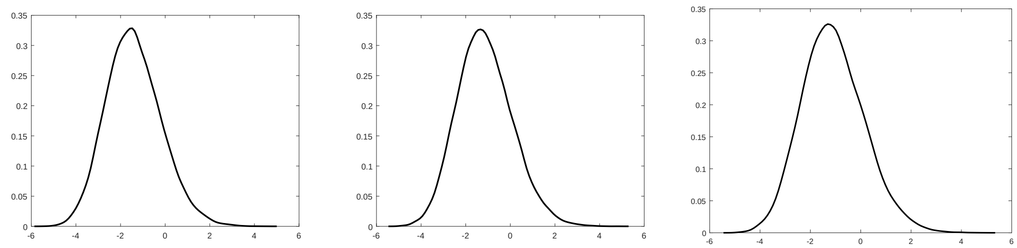
Our model allows for any linear trend and works even if the true value of μ is zero. Yet, when one has such prior knowledge it may seem natural to ignore the de-trending and de-meaning steps (Steps 1 and 2 in Section 2.2) and proceed without them. In this section we compare tests with and without de-trending for the model without μ:

$$\Delta X_t = \Pi X_{t-1} + \varepsilon_t, \qquad t = 1, \dots, T. \tag{21}$$

If both de-trending and de-meaning steps are omitted, we get (the small r version of) the classical Johansen test statistic for the model (21). If only de-trending is omitted, we get the Johansen test statistic corresponding to our original model (2). When both de-trending and de-meaning are implemented, we get our procedure described in Section 2.2. To compare the asymptotic properties of those procedures, we perform a Monte Carlo simulation. Results are reported in Figure 8.

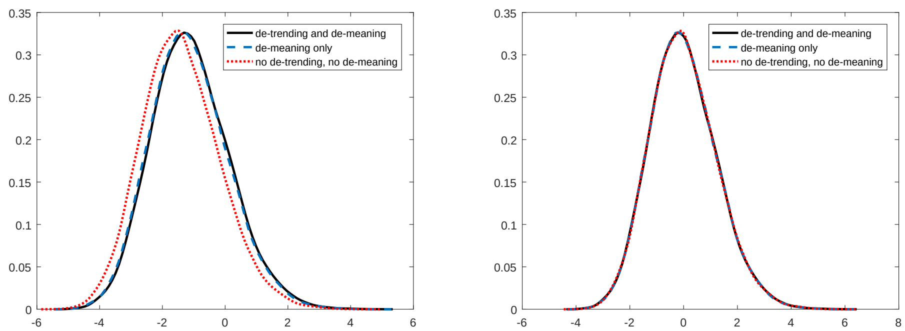
As Figure 8 suggests, the densities have almost identical shape. Figure 9 plots all three of them together as well as their versions with subtracted mean. As we can see, after subtracting the mean, all three densities are identical. This suggests robustness of the Airy_1 point process in our asymptotic results of Theorem 2 (which correspond to Figure 8c) and predicts that some modifications, which preserve the limiting distribution, are possible.

Let us emphasize that our limit theorems currently only apply to Figure 8c, but not to the settings of Figures 8a and 8b. The mismatch of the means in Figure 9a makes one suspect that some modifications of the constants in the asymptotic theorems are needed as soon as we start slightly adjusting the setting.



(a) No de-trending and de-meaning. (b) De-meaning only. (c) De-trending and de-meaning.

Figure 8: Asymptotic distribution of the rescaled $\ln(1-\lambda_1)$ under various testing procedures (Eq. (12) with $r=1$). Data generating process: $\Delta X_{it} = \varepsilon_{it}$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T=500$, $N=100$, $MC = 50,000$ replications.



(a) Three curves together. (b) Mean is subtracted from each of curve.

Figure 9: Asymptotic distribution of the rescaled $\ln(1-\lambda_1)$ under various testing procedures with means normalized to zero (Eq. (12) with $r=1$). Data generating process: $\Delta X_{it} = \varepsilon_{it}$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T=500$, $N=100$, $MC = 50,000$ replications.

7.3. *Higher order of VAR.* In this subsection we discuss the performance of our test when the true data generating process (dgp) is VAR(k), $k>1$. That is

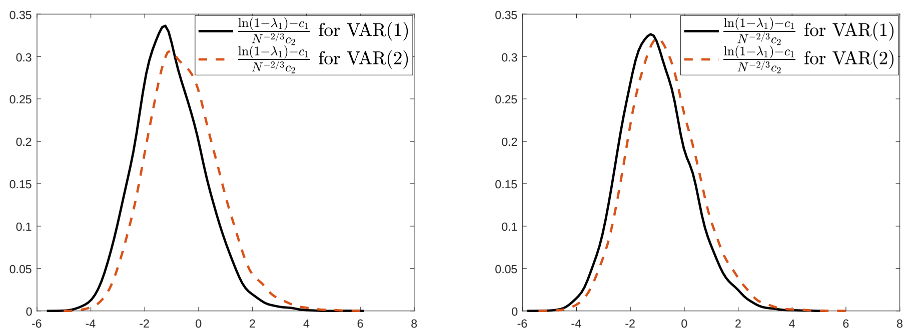
$$\Delta X_t = \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Pi X_{t-k} + \mu + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{22}$$

and the no cointegration situation corresponds to $\Pi \equiv 0$.

Even if $k>1$, we expect to see the Tracy-Widom distribution and marginals of Airy$_1$ process (as in Theorem 2) in the asymptotic behavior of the squared sample canonical correlations from Section 2.2 under mild restrictions on Γ_i. The belief is based on the universality intuition of the random matrix theory. However, we do not expect the scalings (such as the coefficients c_1 and c_2 in Theorem 2) to remain the same. The most plain analogy is the dependence of centering and scaling in the classical Central Limit Theorem on the underlying process. Closer to our context is the asymptotic behavior of sample covariance matrices: when the data is i.i.d., the empirical distribution of the eigenvalues of the sample covariance matrix converges to the Marchenko-Pastur law (and the largest eigenvalues concentrate near the right edge of this distribution), while data with general covariance structure leads to much richer limits, see Bai and Silverstein (2010, Chapter 4) and references therein. Figuring out (even heuristically) any formulas for the scaling coefficients c_1 and c_2 as functions of Γ_i is a challenging open problem for the future research.

There is an important family of cases where one can hope that the formulas for c_1 and c_2 from Theorem 2 remain valid (perhaps, with minor modifications). This is when Γ_i are small and evolution of X_t given by (22) can be treated as a small perturbation of the VAR(1) process. One way to formalize the "smallness' of Γ_i is by requiring them to be of small rank (cf. discussion of rank in Section 3.2) and of small norm.

In order to investigate the above conjecture, we run Monte Carlo simulations for $k = 2$ and Γ_1 of rank 1. We consider the null $\Pi \equiv 0$ and calculate our test statistic based on $r = 1$ under VAR(2) and VAR(1) data generating processes and compare their asymptotic distribution. Under VAR(1) $\Gamma_1 \equiv 0$, while for VAR(2) we use $\Gamma_1 = 0.5E_{11}$ and $\Gamma_1 = 0.5E_{12}$, where E_{ij} is a matrix which has 1 on the intersection of row i and column j and 0 everywhere else; the components of the noise ε_{it} are i.i.d. $\mathcal{N}(0,1)$. The asymptotic distribution of our test statistic with $r = 1$ for various dgps is shown in Figure 10. We see that the distributions on each panel of Figure 10 are close to each other. Hence, testing based on Theorem 2 in such a VAR(2) setting remains valid. Yet, if we consider more general situation of $\Gamma_1 = \theta E_{11}$ or $\Gamma_1 = \theta E_{12}$, then we observe in simulations (not shown) that the quality of approximations significantly deteriorates as θ grows to 1.



(a) $\Gamma_1 = 0.5E_{11}$.

(b) $\Gamma_1 = 0.5E_{12}$.

Figure 10: Asymptotic distribution of the rescaled $\ln(1-\lambda_1)$ (Eq. (12) with $r = 1$) under VAR(1) with $\Pi = 0$ and VAR(2) with $\Pi = 0$ and Γ_1 of rank 1.
Data generating process: $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$, $MC = 10{,}000$ replications.

7.4. *Testing for white noise hypothesis in VAR(1) setting.* The main result of Theorem 2 is a development of a test for the hypothesis $\Pi = 0$ in the VAR(1) model (2). One could also try to understand for which other Π can the testing be possible. The asymptotic distribution depends on the choice of Π, and it is not possible to estimate Π consistently in our regime of T and N growing to infinity proportionally. Thus, for general Π the problem seems infeasible at this point. However, there is another particular choice of Π for which an approach very similar to Theorem 4 still works: $\Pi = -I_N$. Denoting this hypothesis $H_0^{w.n.}$, where $w.n.$ stays for the white noise, the data generating process (Eq. (2)) becomes:

$$H_0^{w.n.}: \qquad X_t = \mu + \varepsilon_t, \qquad t = 1, \dots, T. \tag{23}$$

In other words, we are now testing the hypothesis that the time series X_t is independent across time t against various VAR(1) alternatives. Here is one setup where such testing can be relevant. Suppose that we want to forecast some variable Y and we chose some model for it. After estimating parameters of the model we obtain residuals X_t. If we know that X_t are independent, then they are unforecastable and we cannot further improve our forecasting model. To check the above we can take the residuals X_t and then apply our white noise hypothesis testing procedure.

Let us introduce an adaptation of the Johansen statistic to $H_0^{w.n.}$. As in Eq. (9), we use the notation $\mathcal{P}$ for the de-meaning operator projecting on the hyperplane orthogonal to $(1, 1, \dots, 1)$.

Following Johansen (1988, 1991) we are going to use the de-meaned data $X_t \mathcal{P}$. For the increments ΔX_t in addition to the conventional de-meaning we use an extra modification: we deal with *cyclic increments* $\Delta^c X_t$ defined as:

$$\Delta^c X_t = \begin{cases} X_{t+1} - X_t, & t = 1, 2, \dots, T-1, \\ X_1 - X_T, & t = T. \end{cases}$$

While it might seem bizarre to subtract the last observation from the first one, if we recall that our current hypothesis of interest (23) ignores the time ordering, then this becomes less controversial. Note also a shift of index by 1, as compared to the conventional ΔX_t, which is compensated by the lack of shift $t \to t-1$ in X_t, as compared to Eq. (4). Our choice of definition of Δ^c is important for the following precise asymptotic results. As in Section 2, the conventional Johansen statistic should be thought of as a finite rank perturbation of the modified version that we now introduce.

(24)
$$S_{00}^{w.n.} = \Delta^c X \mathcal{P} (\Delta^c X)^*, \quad S_{01}^{w.n.} = \Delta^c X \mathcal{P} X^*, \quad S_{10}^{w.n.} = X \mathcal{P} (\Delta^c X)^*, \quad S_{11}^{w.n.} = X \mathcal{P} X^*,$$

We further define N numbers $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ as N roots to the equation

$$\det\left(S_{10}^{w.n.} (S_{00}^{w.n.})^{-1} S_{01}^{w.n.} - \lambda S_{11}^{w.n.}\right) = 0. \tag{25}$$

Equivalently, $\{\lambda_i\}$ are eigenvalues of $S_{10}^{w.n.} (S_{00}^{w.n.})^{-1} S_{01}^{w.n.} (S_{11}^{w.n.})^{-1}$.

THEOREM 5. *Suppose that $T, N \to \infty$ in such a way that $T > 2N$ and the ratio T/N remains bounded. Under the hypothesis $H_0^{w.n.}$ one can couple (i.e. define on the same probability space) the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ of the matrix $S_{10}^{w.n.} (S_{00}^{w.n.})^{-1} S_{01}^{w.n.} (S_{11}^{w.n.})^{-1}$ and eigenvalues $x_1 \geq \dots \geq x_N$ of Jacobi ensemble $\mathbf{J}(N; \frac{T-N-1}{2}, \frac{T-2N}{2})$ in such a way that for each $\epsilon > 0$ we have*

$$\lim_{T,N \to \infty} \text{Prob}\left(\max_{1 \leq i \leq N} |\lambda_i - x_i| < \frac{1}{N^{1-\epsilon}} \right) = 1.$$

The proof of Theorem 5 follows a similar strategy as Theorem 4 and we refer to Section 9.5 in Appendix for details; in particular, the proofs rely on yet another novel appearance of the Jacobi ensemble.

The remaining straightforward step to obtain the asymptotics of various statistics built on the eigenvalues $\{\lambda_i\}$ is to combine Theorem 5 with asymptotic results for the Jacobi ensemble presented in Section 9.3. This is in the spirit of Theorem 2.

Note that the hypothesis $H_0^{w.n.}$ implies the maximal amount of cointegrating relationships: each of the N components of X_t is already stationary. A reasonable alternative hypothesis H_1 is the presence of $N - r$ cointegrating relationships. For simplicity, let us concentrate on the case $r = 1$. Then the alternative can be also interpreted as a presence of a single growing factor. In this situation we expect the *smallest* eigenvalue λ_N to be a good test statistic. We see in numerical simulations that λ_N is bounded away from 0 under $H_0^{w.n.}$. It can also be formally proved by combining Theorem 5 with asymptotics of the Jacobi ensemble from Section 9.3. Thus, if λ_N is close to 0, then we are able to reject $H_0^{w.n.}$. The same simulations indicate that λ_N starts to be close to 0 when there are at most $N-1$ cointegrating relationships. Hence, the test based on λ_N should have a good asymptotic power. We leave rigorous justifications of this observation till future research, and for now only mention the following heuristics: the stationary linear combinations of X_t are strongly correlated with the same linear combinations of $\Delta^c X_t$; on the other hand, the growing linear combinations of X_t have very weak correlation with the same linear combinations of $\Delta^c X_t$ (cf. correlations of a one dimensional random walk with its increments). Hence, if the latter are present, the smallest canonical correlations of $X\mathcal{P}$ and $\Delta^c X\mathcal{P}$, which coincide with the eigenvalues of the matrix $S_{10}^{w.n.}(S_{00}^{w.n.})^{-1}S_{01}^{w.n.}(S_{11}^{w.n.})^{-1}$, should become small leading to close to 0 value of λ_N.

8. Conclusion. The paper presents a cointegration test which has desirable empirical size when N and T are of the same magnitude. To our knowledge, this is a first paper which constructs and analyzes asymptotic properties of a test that does not suffer from significant distortions (such as over-rejection) for comparable N and T. The test is based on the Johansen LR test and incorporates some additional steps. First, our procedure reinforces the importance of de-trending in cointegration testing. It turns out, that de-trending is crucial for deriving desirable asymptotic properties. (E.g., only after de-trending one can rewrite the lagged process as a linear function of its first differences.) Second, our asymptotic results reveal and explain an unexpected connection between the Johansen cointegration test and the Jacobi ensemble — a classical ensemble of the random matrix theory whose previous appearances in statistics include multivariave analysis of variance (MANOVA) and sample canonical correlations for independent sets of data.

On the theoretical side the next step would be to go from null hypothesis of zero cointegration to analyzing the behaviour of our test under r cointegrations. This will allow us to calculate the power of the test, reinforcing our simulational findings in Section 5.2, as well as to perform tests of r versus $r+1$ cointegrations.

On the empirical side it would be interesting to apply our test to other data sets beyond what is presented in Section 6. Annual cross-country data provides a natural example of our setting where the number of years and countries is comparable. Another example arises if one considers network-type settings which evolve over time (e.g., as in Bykhovskaya (2021)). Data on trade or on foreign direct investment can potentially be non-stationary, especially if we focus on largest and the most active countries. Moreover, although such monthly data is available, for many countries it only covers 20 years. Thus, we have $T \approx 200$. If we look at directed pairs across 10 largest countries, this gives us $N = 90$ cross-section units, which fits ideally in our setting. Classical cointegration tests are known to perform poorly in the above settings. However, the asymptotic results of our paper open up a possibility of detecting the presence of cointegration in such time-series data.

9. Appendix.

9.1. *New matrix models for the Jacobi ensemble.* Recall that our testing procedure relies on the squared sample canonical correlations between two correlated data sets. As we later show in the proofs of Propositions 11 and 16, an equivalent point of view is that we deal with eigenvalues of a product of two orthogonal projectors P_1 and P_2, where P_1 is a projection on a random N-dimensional subspace $\mathcal{V}$ of a T–dimensional space and P_2 is a projection on $U\mathcal{V}$, where U is a certain deterministic linear operator.

We randomize this problem by replacing U with a *random* operator, whose spectrum is close to U. The randomized problem turns out to be exactly solvable — the eigenvalues of new P_1P_2 coincide with the classical Jacobi ensemble. The goal of this section is to prove this fact. The choices of U and $\mathcal{V}$ depend on the hypothesis that we are testing and, hence, we need several theorems. Throughout this section we are going to deal both with real and complex matrices. According to the customary random matrix theory notation, they are referred to as $\beta = 1$ and $\beta = 2$ cases, respectively.

In what follows I_n means the identity matrix in n–dimensional space. Sometimes we omit n and write simply I when the dimension is clear from the context. For a matrix X we let $[X]_{NN}$ to be its top-left $N \times N$ corner. The parameter $\mathcal{T}$ used in this section is related to T of the main text through $\mathcal{T} = T - 1$.

Throughout this section we repeatedly change the coordinates in various measures, which produces a factor given by the absolute value of the Jacobian of the transformation. We rely on the computation of the Jacobian of the multiplicative change of variables in the space of matrices. We need three forms of it, where in each of them Q stays for a $n \times n$ matrix:

- The map $Z \mapsto QZ$ on $n \times m$ matrices has the Jacobian

$$\left|\frac{\partial(QZ)}{\partial Z}\right| = |\det Q|^{\beta m}, \tag{26}$$

- The map $Z \mapsto QZQ^*$ from the space of $n \times n$ symmetric (Hermitian if $\beta = 2$) matrices to itself has the Jacobian

$$\left|\frac{\partial(QZQ^*)}{\partial Z}\right| = |\det Q|^{\beta n + 2 - \beta} = \begin{cases} |\det Q|^{2n}, & \beta = 2, \\ |\det Q|^{n+1}, & \beta = 1. \end{cases} \tag{27}$$

- The map $Z \mapsto QZQ^*$ from the space of $n \times n$ skew-symmetric (skew-Hermitian if $\beta = 2$) matrices to itself has the Jacobian

$$\left|\frac{\partial(QZQ^*)}{\partial Z}\right| = |\det Q|^{\beta n + \beta - 2} = \begin{cases} |\det Q|^{2n}, & \beta = 2, \\ |\det Q|^{n-1}, & \beta = 1. \end{cases} \tag{28}$$

The first identity (26) follows from the observation that each column of Z is transformed by linear map Q and there are m such columns. The second and third identities are similar and we refer to Forrester (2010, (1.35)) for details.

The first theorem of this section is relevant for the setting of Theorem 4 for VAR(1).

THEOREM 6. *Assume $\mathcal{T} \geq 2N$. Let O be a random $\mathcal{T} \times \mathcal{T}$ matrix chosen from the uniform measure on the group of real orthogonal matrices of determinant 1 if $\beta = 1$ or of complex unitary matrices if $\beta = 2$. Define $U = (I_{\mathcal{T}} + O)^{-1}$. Then the matrix*

$$M = [U]_{NN}([U^*U]_{NN})^{-1}[U^*]_{NN} \tag{29}$$

is distributed as the $N \times N$ real symmetric (if $\beta = 1$) or complex Hermitian (if $\beta = 2$) matrix of density proportional to

$$\det(M)^{\frac{\beta}{2}N + \beta - 2} \det(I_N - M)^{\frac{\beta}{2}(\mathcal{T} - 2N + 1) - 1}, \quad 0 \leq M \leq I_N, \qquad \beta = 1, 2. \tag{30}$$

with respect to the Lebesgue measure on symmetric/Hermitian matrices.

REMARK 2. *Define a $\mathcal{T} \times \mathcal{T}$–dimensional matrix $\mathcal{M}$ by putting M in $N \times N$ corner of $\mathcal{M}$ and filling the rest with zeros. Non-zero eigenvalues of M and $\mathcal{M}$ are the same. Simultaneously, $\mathcal{M}$ can be identified with $P_1P_2P_1$, where P_1 is the orthogonal projector on space $\mathcal{V}$ spanned by the first N coordinate vectors and P_2 is the projector on $U\mathcal{V}$.*

REMARK 3. *If $\mathcal{T} < 2N$, then the spaces $\mathcal{V}$ and $U\mathcal{V}$ necessarily intersect, and therefore M has deterministic eigenvalues which equal 1. This should be taken into account when extending (30) to this case and we will not pursue this direction here.*

PROOF OF THEOREM 6. We parameterize the orthogonal (or unitary if $\beta = 2$) group by the means of the Cayley transform. Namely, we set

$$\mathcal{R} = (I_\mathcal{T} - O)(I_\mathcal{T} + O)^{-1} = \frac{I_\mathcal{T} - O}{I_\mathcal{T} + O}, \quad \text{so that} \quad O = \frac{I_\mathcal{T} - \mathcal{R}}{I_\mathcal{T} + \mathcal{R}}.$$

Since $O^{-1} = O^*$, $\mathcal{R}$ is a skew-symmetric $\mathcal{T} \times \mathcal{T}$ matrix, i.e., $\mathcal{R}^* = -\mathcal{R}$. The uniform (Haar) measure on O leads to the following distribution on $\mathcal{R}$, in which we omitted the irrelevant for us normalization constant:

$$\det(I_\mathcal{T} - \mathcal{R}^2)^{-\frac{\beta}{2}\mathcal{T} + \frac{2-\beta}{2}} d\mathcal{R}, \quad \beta = 1, 2, \tag{31}$$

see, e.g., Forrester (2010, (2.55)). Further, we have

$$U = \frac{I_\mathcal{T} + \mathcal{R}}{2},$$

so that

$$M = [I_\mathcal{T} + \mathcal{R}]_{NN}([I_\mathcal{T} - \mathcal{R}^2]_{NN})^{-1}[I_\mathcal{T} - \mathcal{R}]_{NN}. \tag{32}$$

We partition $\mathcal{T} = N + (\mathcal{T} - N)$ and write $I_\mathcal{T} + \mathcal{R}$ in a block form according to this split:

$$I_\mathcal{T} + \mathcal{R} = \begin{pmatrix} I_N + A & B \\ -B^* & I_{\mathcal{T}-N} + C \end{pmatrix},$$

where A is a $N \times N$ skew-symmetric matrix, C is $(\mathcal{T} - N) \times (\mathcal{T} - N)$ skew-symmetric matrix and B is an arbitrary $N \times (\mathcal{T} - N)$ matrix.

We make a change of variables by introducing $\tilde{B}$ so that

$$B = (I_N - A)\tilde{B}, \quad \text{and} \quad B^* = \tilde{B}^*(I_N + A).$$

Using (26) we compute the Jacobian of the transformation:

$$\left|\frac{\partial B}{\partial \tilde{B}}\right| = |\det(I_N - A)|^{\beta(\mathcal{T}-N)}. \tag{33}$$

Note that

$$\begin{aligned} M &= (I_N + A)(I_N - A^2 + BB^*)^{-1}(I_N - A) \\ &= (I_N + A)\big((I_N - A)(I_N + \tilde{B}\tilde{B}^*)(I_N + A)\big)^{-1}(I_N - A) = \big(I_N + \tilde{B}\tilde{B}^*\big)^{-1} \end{aligned} \tag{34}$$

Using the formula for the determinant of a block matrix

$$\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det \mathbf{A} \cdot \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}),$$

we also have

$$\det(I_\mathcal{T} + \mathcal{R}) = \det\begin{pmatrix} I_N + A & (I_N - A)\tilde{B} \\ -\tilde{B}^*(I_N + A) & I_{\mathcal{T}-N} + C \end{pmatrix} \tag{35}$$

$$= \det(I_N + A)\det\begin{pmatrix} I_N & (I_N - A)\tilde{B} \\ -\tilde{B}^* & I_{\mathcal{T}-N} + C \end{pmatrix} = \det(I_N + A)\det(I_{\mathcal{T}-N} + C + \tilde{B}^*(I_N - A)\tilde{B})$$

$$= \det(I_N + A)\det(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})\det\Big(I_{\mathcal{T}-N} + (I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{-1/2}(C - \tilde{B}^*A\tilde{B})(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{-1/2}\Big),$$

Next, we introduce

$$\tilde{C} = (I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{-1/2}(C - \tilde{B}^*A\tilde{B})(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{-1/2}$$

and notice that the map $C \mapsto \tilde{C}$ preserves the space of all skew-symmetric $(\mathcal{T}-N)\times(\mathcal{T}-N)$ matrices. Using (28) the map has the Jacobian

(36) $$\frac{\partial C}{\partial \tilde{C}} = \det(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{\frac{\beta}{2}(\mathcal{T}-N)+\frac{\beta-2}{2}},$$

Combining (33), (35), and (36), we rewrite the measure (31) as follows:

(37) $$\det(I_{\mathcal{T}} - \mathcal{R}^2)^{-\frac{\beta}{2}\mathcal{T}+\frac{2-\beta}{2}}\, dA\, dB\, dC = \left|\det(I_{\mathcal{T}} + \mathcal{R})^{-\beta\mathcal{T}+2-\beta}\right| dA\, dB\, dC$$
$$= \left|\det(I_N + A)\det(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})\det(I_{\mathcal{T}-N} + \tilde{C})\right|^{-\beta\mathcal{T}+2-\beta}$$
$$\times |\det(I_N - A)|^{\beta(\mathcal{T}-N)} \cdot \det(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{\frac{\beta}{2}(\mathcal{T}-N)+\frac{\beta-2}{2}}\, dA\, d\tilde{B}\, d\tilde{C}.$$

The key property of (37) is that the measure has factorized and projecting onto the $\tilde{B}$–component is straightforward. We conclude that $\tilde{B}$ is distributed according to the measure

(38) $$\det(I_{\mathcal{T}-N} + \tilde{B}^*\tilde{B})^{-\frac{\beta}{2}(\mathcal{T}+N)+\frac{2-\beta}{2}}\, d\tilde{B} = \det(I_N + \tilde{B}\tilde{B}^*)^{-\frac{\beta}{2}(\mathcal{T}+N)+\frac{2-\beta}{2}}\, d\tilde{B},$$

where we used $\det(I + UW) = \det(I + WU)$ in the last equality. According to Forrester (2010, Exercise 3.2.q6), (38) implies that the symmetric (or Hermitian if $\beta = 2$) non-negative definite $N \times N$ matrix $F = \tilde{B}\tilde{B}^*$ has the law

(39) $$\det(F)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\det(I_N + F)^{-\frac{\beta}{2}(\mathcal{T}+N)+\frac{2-\beta}{2}} dF.$$

We have by (34)

$$M = \frac{1}{I_N + F}, \quad F = \frac{1-M}{M}, \quad dM = -\frac{1}{I_N + F} dF \frac{1}{I_N + F}.$$

Using (27) we have

(40) $$\left|\frac{\partial M}{\partial F}\right| = \left|\det\left(\frac{1}{I_N + F}\right)\right|^{\beta N - \beta + 2} = |\det M|^{\beta N - \beta + 2}.$$

Formulas (39) and (40) imply that the matrix M has the distribution

$$\det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\det(M)^{-\frac{\beta}{2}(\mathcal{T}-2N+1)+1}\det(M)^{\frac{\beta}{2}(\mathcal{T}+N)-\frac{2-\beta}{2}}\det(M)^{-\beta N+\beta-2} dM$$
$$= \det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\det(M)^{\frac{\beta}{2}N+\beta-2} dM, \quad 0 \le M \le I_N. \quad \square$$

Our next theorem gives another realisation of the Jacobi ensemble, which is relevant to testing $\Pi = -I_N$ in VAR(1) setting.

THEOREM 7. *Assume $\mathcal{T} \ge 2N$. Let O be a random $\mathcal{T} \times \mathcal{T}$ real matrix chosen from the uniform measure on the group of orthogonal matrices with determinant 1 if $\beta = 1$ or of complex unitary matrices if $\beta = 2$. Let A be $N \times N$ top-left corner of O. Then the matrix*

$$M = (I_N + A)(2I_N + A + A^*)^{-1}(I_N + A^*)$$

is distributed as $N \times N$ real symmetric (if $\beta = 1$) or complex Hermitian (if $\beta = 2$) matrix of density proportional to

(41) $$\det(M)^{\frac{\beta}{2}(\mathcal{T}-N)+\beta-2}\det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}, \quad 0 \le M \le I_N, \qquad \beta = 1, 2.$$

with respect to the Lebesgue measure on symmetric/Hermitian matrices.

PROOF. The computation of the law of A is well-known and Forrester (2010, (3.113)) provides a formula for the density with respect to the Lebesgue measure on all real/complex $N \times N$ matrices. It is

$$\det(I_N - A^*A)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\, dA, \quad 0 \le A^*A \le I_N, \tag{42}$$

where we omit here and below the normalization constant, which makes the total mass of measure equal to 1. The matrix M is a function of A and in the rest of the proof we transform the measure (42) to make the distribution of this function explicit.

Our first step is to rewrite (42) in terms of M. We claim that

$$\det(I_N - A^*A) = \det(I_N - M)\det(2I_N + A + A^*) \tag{43}$$

Indeed, we first define

$$M_1 = (I_N + A^*)(I_N + A)(2I_N + A + A^*)^{-1}$$

and notice that

$$I_N - A^*A = 2I_N + A + A^* - M_1(2I_N + A + A^*) = (I_N - M_1)(2I_N + A + A^*).$$

Hence,

$$\det(I_N - A^*A) = \det(I_N - M_1)\det(2I_N + A + A^*) \tag{44}$$

and (43) is obtained from (44) by using the identity $\det(I - CD) = \det(I - DC)$.

We conclude that the density (42) has the form proportional to

$$\det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\det(2I_N + A + A^*)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}\, dA. \tag{45}$$

Note that the first factor has the desired form from the statement of the theorem.

We now split the Euclidean space of all $N \times N$ real (or complex) matrices into the symmetric (or Hermitian) and skew-symmetric (or skew-Hermitian) parts. For that we define

$$X = I_N + \frac{A + A^*}{2}, \quad Y = \frac{A - A^*}{2}.$$

Since A belongs to a unit matrix ball (because $AA^* + BB^* = I_N$, where B is the top-right $N \times (\mathcal{T} - N)$ corner of O), X is a positive-definite symmetric (or Hermitian) matrix, i.e., $X \ge 0$. The relation

$$(I_N + A)(2I_N + A + A^*)^{-1}(I_N + A^*) = M$$

is now rewritten as

$$(X + Y)(2X)^{-1}(X - Y) = M \quad \text{or} \quad X = 2M + YX^{-1}Y. \tag{46}$$

Note that X^{-1} is positive definite, while Y is skew-Hermitian. This implies that $YX^{-1}Y \le 0$, i.e., $YX^{-1}Y$ is a negative semi-definite Hermitian matrix. Hence, (46) implies

$$X \le 2M,$$

which means that $2M - X$ is a non-negative definite matrix.

Using (46) we make a change of coordinates in the space of matrices $(X, Y) \to (M, Y)$. The Jacobian of this change of coordinates is

$$\det\begin{pmatrix} \frac{\partial M}{\partial X} & \frac{\partial Y}{\partial X} \\ \frac{\partial M}{\partial Y} & \frac{\partial Y}{\partial Y} \end{pmatrix} = \det\begin{pmatrix} \frac{\partial M}{\partial X} & 0 \\ \frac{\partial M}{\partial Y} & I \end{pmatrix} = \det\left(\frac{\partial M}{\partial X}\right)$$

Therefore, we need to compute the Jacobian of the map

$$X \mapsto M = \tfrac{1}{2}\left(X - YX^{-1}Y\right),$$

which maps the Eucledian space of symmetric (Hermitian if $\beta = 2$) matrices to itself. Differentiating, and using $dX^{-1} = X^{-1}dXX^{-1}$, we see the matrix identity

$$2dM = dX - YX^{-1}dXX^{-1}Y = dX + (YX^{-1})dX(YX^{-1})^*, \tag{47}$$

where we used $X = X^*$, $Y = -Y^*$ in the last identity. Hence, the Jacobian of the map $X \mapsto M$ is a function of YX^{-1} and we denote this function $J(YX^{-1})$. Therefore, (45) becomes

$$\sim \det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} \det(X)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} J^{-1}(YX^{-1})\, dM\, dY. \tag{48}$$

We now make another change of variables by setting

$$\tilde{Y} = M^{-1/2} Y M^{-1/2}, \qquad Y = M^{1/2} \tilde{Y} M^{1/2}.$$

The Jacobian of the map $(M,Y) \to (M,\tilde{Y})$ is the same as the Jacobian of the map $Y \to \tilde{Y}$, which is computed by (28) as

$$\left|\frac{\partial \tilde{Y}}{\partial Y}\right| = |\det M|^{-\frac{\beta}{2}N + \frac{2-\beta}{2}} = \begin{cases} (\det M)^{-N}, & \beta = 2, \\ (\det M)^{-(N-1)/2}, & \beta = 1. \end{cases} \tag{49}$$

This converts (48) into

$$\sim \det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} (\det M)^{\frac{\beta}{2}N + \frac{\beta-2}{2}} \det(X)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} J^{-1}(YX^{-1})\, dM\, d\tilde{Y}. \tag{50}$$

It remains to figure out the factors involving $\det X$ and YX^{-1} in the last formula. For that let us introduce the notation

$$\tilde{X} = M^{-1/2} X M^{-1/2},$$

and notice that (46) implies

$$\tilde{X} = 2I_N + \tilde{Y}\tilde{X}^{-1}\tilde{Y}, \qquad 0 \le \tilde{X} \le 2I_N. \tag{51}$$

Solution to (51) gives us a function $\tilde{X} = \tilde{X}(\tilde{Y})$, such that $X = M^{1/2}\tilde{X}M^{1/2}$. Hence, we can transform

$$\det(X)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} = \det(M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} \det(\tilde{X})^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1}. \tag{52}$$

For $J(YX^{-1})$ we notice two properties. First,

$$YX^{-1} = M^{1/2}\tilde{Y}\tilde{X}^{-1}M^{-1/2}.$$

In addition, we have the following statement, which we prove later:

Claim. For the above Jacobian $J(\cdot)$, we have

$$J(A) = J(B^{-1}AB), \qquad \text{whenever } B = B^*. \tag{53}$$

Thus, $J(YX^{-1}) = J(\tilde{Y}\tilde{X}^{-1}) = J(f(\tilde{Y}))$ for a certain function f, and using (52) and (53) we rewrite (50) as

$$\det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} (\det M)^{\frac{\beta}{2}N + \frac{\beta-2}{2}} \det(M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} g(\tilde{Y})\, dM\, d\tilde{Y} \tag{54}$$

for a certain function $g(\tilde{Y})$. Since the parts involving $\tilde{Y}$ and M are now decoupled, we can integrate out $\tilde{Y}$ arriving at the desired expression for the distribution of M:

$$\det(I_N - M)^{\frac{\beta}{2}(\mathcal{T}-2N+1)-1} \det(M)^{\frac{\beta}{2}(\mathcal{T}-N)+\beta-2}\, dM. \tag{55}$$

It remains to prove the claim (53). By definition (47), $J(B^{-1}AB)$ is the Jacobian of the linear map

$$dX \mapsto \tfrac{1}{2}dX + \tfrac{1}{2}(B^{-1}AB)dX(BA^*B^{-1})$$

We split this map into a composition of three:

$$dX \mapsto B\,dX\,B \mapsto \tfrac{1}{2}(B\,dX\,B)+\tfrac{1}{2}A(B\,dX\,B)A^* \mapsto B^{-1}\left[\tfrac{1}{2}(B\,dX\,B)+\tfrac{1}{2}A(B\,dX\,B)A^*\right]B^{-1}.$$

$J(B^{-1}AB)$ is the product of Jacobians of these three maps. The first and the last maps are inverse to each other (above we used the explicit Jacobian of these maps, but this is not even needed) and their Jacobians cancel. The Jacobian of the middle map is the desired $J(A)$. □

9.2. *Proof of Theorem 4.* The proof of Theorem 4 is split into two steps. First, Proposition 11 uses rotation invariance of the Gaussian law to rewrite the matrix $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ as a product of corners of certain matrices, reminiscent of (29). Next, we show in Proposition 12 that replacement of a certain deterministic matrix (in that product) by its random perturbation leads precisely to (29) and simultaneously has controlled effect on the change of eigenvalues.

We need to introduce some deterministic $T\times T$ matrices. The summation matrix Φ has 1's below the diagonal and 0's on the diagonal and everywhere above the diagonal:

$$\Phi = \begin{pmatrix} 0\,0 & 0 & \dots 0 \\ 1\,0 & 0 & \dots 0 \\ 1\,1 & 0 & \dots 0 \\ & \ddots & \\ 1\,1 \dots & 1 & 0 \end{pmatrix}.$$

DEFINITION 8. *V is the $(T-1)$–dimensional hyperplane orthogonal to $(1,1,\dots,1)$.*

We let $\mathcal{P}$ be the orthogonal projector on V (see Eq. (9)). Finally, we set

$$\tilde{\Phi} = \mathcal{P}\Phi\mathcal{P}.$$

We also need the cyclic version of the lead operator F mapping $(x_1,x_2,\dots,x_T)$ to $(x_2,x_3,x_4,\dots,x_T,x_1)$. V is an invariant subspace for F and we denote through F_V the restriction of F onto V.

LEMMA 9. *The operator $\tilde{\Phi}$ preserves the space V. Its restriction onto V coincides with $-(I_V-F_V)^{-1}$, where I_V is the identical operator acting in V.*

PROOF OF LEMMA 9. Let e_i, $i=1,\dots,T$ denote the ith coordinate vector. Then $(e_i - e_{i+1})_{i=1}^{T-1}$ gives a linear basis of V. Since $\Phi e_i = e_{i+1}+e_{i+2}+\dots+e_T$, we have

$$\tilde{\Phi}(e_i-e_{i+1}) = \mathcal{P}\Phi(e_i-e_{i+1}) = \mathcal{P}e_{i+1} = e_{i+1} - \frac{1}{T}\sum_{j=1}^{T} e_j.$$

Applying $I-F$ to the last vector and noticing that $Fe_{i+1}=e_i$, we conclude that

$$(I-F)\tilde{\Phi}(e_i-e_{i+1}) = e_{i+1}-e_i. \quad \square$$

Let us introduce one more notation. Choose some orthonormal basis $\tilde{e}_1,\tilde{e}_2,\dots,\tilde{e}_{T-1}$ of V, e.g., this can be an orthogonalization of e_1-e_2, e_2-e_3, …, $e_{T-1}-e_T$ (but the exact choice is irrelevant for the following theorems).

DEFINITION 10. *For an operator A acting in V we set $[A]_{NN}^V$ to be the $N\times N$ corner of the operator A, taken in the basis $\{\tilde{e}_i\}$.*

Next, we take a uniformly-random orthogonal (or unitary if $\beta = 2$) operator $\tilde{O}$ acting in $(T-1)$–dimensional space V and define an operator $\tilde{U}$ acting in V:

$$\tilde{U} = (I_V - \tilde{O} F_V \tilde{O}^*)^{-1}.$$

Let us introduce a symmetric (or Hermitian if $\beta = 2$) $N \times N$ matrix $\tilde{M}$,

$$\tilde{M} = [\tilde{U}]_{NN}^V ([\tilde{U}^* \tilde{U}]_{NN}^V)^{-1} [\tilde{U}^*]_{NN}^V, \tag{56}$$

PROPOSITION 11. *Choose an arbitrary positive definite covariance matrix Λ. Let ε be $N \times T$ matrix of random variables (real if $\beta = 1$ and complex if $\beta = 2$), such that T columns of ε are i.i.d., and each of them is an N-dimensional mean zero Gaussian vector with covariance Λ. Fix arbitrary N–dimensional vectors X_0 and μ. Define the $N \times T$ matrix $X = (X_1, X_2, \ldots, X_T)$ as in Eq. (2) via recurrence*

$$X_t = X_{t-1} + \mu + \varepsilon_t, \quad t = 1, \ldots, T.$$

Further set $\Delta X_t = X_t - X_{t-1}$ and $\tilde{X}_t = X_{t-1} - \frac{t-1}{T}(X_T - X_0)$, $t = 1, \ldots, T$. Define $N \times N$ matrices:

$$S_{00} = \Delta X \mathcal{P} \Delta X^*, \quad S_{01} = \Delta X \mathcal{P} \tilde{X}^*, \quad S_{10} = S_{01}^* = \tilde{X} \mathcal{P} \Delta X^*, \quad S_{11} = \tilde{X} \mathcal{P} \tilde{X}^*.$$

Then the eigenvalues of the matrix $S_{10} S_{00}^{-1} S_{01} S_{11}^{-1}$ have the same distribution as those of $\tilde{M}$ in (56).

PROOF. We start by expressing the matrices S_{ij} via ε. Clearly, $\Delta X_t = \mu + \varepsilon_t$. Further,

$$(\Delta X \mathcal{P})_t = \mu + \varepsilon_t - \frac{1}{T} \sum_{i=1}^{T} (\mu + \varepsilon_i), \qquad \Delta X \mathcal{P} = \varepsilon \mathcal{P}.$$

Hence, $S_{00} = \varepsilon \mathcal{P} \varepsilon^*$. Next,

$$\tilde{X}_t = X_0 + (t-1)\mu + \sum_{i=1}^{t-1} \varepsilon_i - \frac{t-1}{T} \left(T\mu + \sum_{i=1}^{T} \varepsilon_i \right) = X_0 + \sum_{i=1}^{t-1} \varepsilon_i - \frac{t-1}{T} \sum_{i=1}^{T} \varepsilon_i$$

We claim that $\tilde{X} \mathcal{P} = \varepsilon \tilde{\Phi}^*$. Indeed, $\tilde{X} \mathcal{P}$ coincides with $\tilde{\tilde{X}} \mathcal{P}$, where

$$\tilde{\tilde{X}}_t = \tilde{X}_t - X_0 = \sum_{i=1}^{t-1} \varepsilon_i - \frac{t-1}{T} \sum_{i=1}^{T} \varepsilon_i = \sum_{i=1}^{t-1} \left(\varepsilon_i - \frac{1}{T} \sum_{j=1}^{T} \varepsilon_j \right).$$

Since Φ is the summation operator, we have $\tilde{\tilde{X}} = (\Phi(\varepsilon \mathcal{P})^*)^* = \varepsilon \mathcal{P} \Phi^*$ and the claim is proven as $\tilde{\Phi}^* = \mathcal{P} \Phi^* \mathcal{P}$.

We conclude that

$$S_{00} = \varepsilon \mathcal{P} \varepsilon^*, \quad S_{01} = \varepsilon \tilde{\Phi} \varepsilon^*, \quad S_{10} = S_{01}^* = \varepsilon \tilde{\Phi}^* \varepsilon^*, \quad S_{11} = \varepsilon \tilde{\Phi}^* \tilde{\Phi} \varepsilon^*. \tag{57}$$

Next, note that we can (and will) assume without loss of generality that the covariance matrix Λ is identical, i.e., all matrix elements of ε are i.i.d. random variables $\mathcal{N}(0,1)$ if $\beta = 1$ and $\mathcal{N}(0,1) + \mathbf{i}\mathcal{N}(0,1)$ if $\beta = 2$. Indeed, using (57), if we take any non-degenerate $N \times N$ matrix A and replace ε by $A\varepsilon$, then the matrix $S_{10} S_{00}^{-1} S_{01} S_{11}^{-1}$ is conjugated by A, which keeps eigenvalues unchanged. By choosing appropriate A, we can guarantee that the covariance of columns of $A\varepsilon$ is identical, and then rename $A\varepsilon$ as ε.

In the remaining proof we explicitly couple ε with $\tilde{O}$ (arising in the definition of $\tilde{M}$) by constructing $\tilde{O}$ using randomness coming from ε.

Since for any two rectangular matrices A and B of the same sizes, AB^* and B^*A have the same non-zero eigenvalues, the eigenvalues of $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ can be identified (using also $\mathcal{P}\tilde{\Phi} = \tilde{\Phi}$) with those of

$$(58) \qquad \left(\mathcal{P}\varepsilon^* S_{00}^{-1} \varepsilon \mathcal{P}\right)\left(\tilde{\Phi}\varepsilon^* S_{11}^{-1} \varepsilon \tilde{\Phi}^*\right) = P_1 P_2,$$

where P_1 is the orthogonal projector onto the space spanned by N columns of $\mathcal{P}\varepsilon^*$ and P_2 is the orthogonal projector onto the N columns of $\tilde{\Phi}\varepsilon^*$. Since, the eigenvalues of P_1P_2 are the same as those of $P_1P_2P_1$, and the latter operator acts as 0 on the orthogonal complement of V, we can restrict all the operators to V. Due to invariance of the Gaussian law with respect to rotations by orthogonal (or unitary if $\beta = 2$) matrices, the columns of $\mathcal{P}\varepsilon^*$ span a rotationally-invariant N–dimensional subspace in V. The columns of $\tilde{\Phi}\varepsilon^* = \tilde{\Phi}\mathcal{P}\varepsilon^*$ then span the $\tilde{\Phi}$–image of this subspace.

The eigenvalues of $P_1P_2P_1$ are preserved when we conjugate each of the projectors with an orthogonal transformation O, i.e. replace $P_1P_2P_1$ with $OP_1O^*OP_2O^*OP_1O^* = OP_1P_2P_1O^*$. In this transformation the spaces to where P_1 and P_2 are projecting are replaced by their O–images. We take O to be a random operator satisfying two conditions:

- O maps the span of the columns of $\mathcal{P}\varepsilon^*$ to the subspace spanned by the first N basis vectors, $\tilde{e}_1, \dots, \tilde{e}_N$ in V;
- O is distributed as uniformly random orthogonal operator acting in V, i.e., $O \stackrel{d}{=} \tilde{O}$.

The existence of such O follows from the rotational invariance of the Gaussian law.[18] Hence, OP_1O^* is a projector onto $\langle \tilde{e}_1, \dots, \tilde{e}_N \rangle$, and we conclude that the eigenvalues of $P_1P_2P_1$ coincide with those of

$$[OP_2O^*]_{NN}^V.$$

Since P_2 is the projector on the span of columns of $\tilde{\Phi}\varepsilon^* = O^*(O\tilde{\Phi}O^*)O\varepsilon^*$, OP_2O^* is the projector on the O–image of this span, i.e., OP_2O^* is the projector onto the columns of $(O\tilde{\Phi}O^*)O\varepsilon^*$. Since columns of $O\varepsilon^*$ span $\langle \tilde{e}_1, \dots, \tilde{e}_N \rangle$ and $O\tilde{\Phi}O^* = O(-(I_V - F_V)^{-1})O^* = -\tilde{U}$ on V by Lemma 9, we reach the formula for $\tilde{M}$ as the $N \times N$ corner of the projector on the first N columns of $\tilde{U}$. □

Set $\mathcal{T} = T - 1$ and let O be uniformly random real orthogonal of determinant 1 (if $\beta = 1$) or complex unitary (if $\beta = 2$) $\mathcal{T} \times \mathcal{T}$ matrix. Define $U = (I_{\mathcal{T}} + O)^{-1}$. Then we set, as in (29),

$$(59) \qquad M = [U]_{NN}([U^*U]_{NN})^{-1}[U^*]_{NN}$$

For a matrix A with real spectrum, we set $\lambda_i(A)$ to be the ith largest eigenvalue of A.

PROPOSITION 12. *Assume $\mathcal{T} = T - 1 \geq 2N$. One can couple M from* (59) *with $\tilde{M}$ from* (56) *in such a way that for each $\epsilon > 0$ we have*

$$\text{Prob}\left(\max_{1 \leq i \leq N} |\lambda_i(M) - \lambda_i(\tilde{M})| < \frac{1}{N^{1-\epsilon}}\right) \to 1,$$

as $T, N \to \infty$ in such a way that the ratio T/N remains bounded.

[18] One way to see it is by noticing that the span of the columns of $\mathcal{P}\varepsilon^*$ and the span of N columns of O^{-1} (for uniformly random O) have the same distribution given by the uniform measure on all N–dimensional subspace of T–dimensional space.

REMARK 4. *We believe that the condition $\mathcal{T} \geq 2N$ can be weakened and replaced by T/N bounded away from* 0*, but we stick to it, since that's the situation when Theorem 6 applies.*

Let us compare the definitions of M and $\tilde{M}$. There is only one difference between them: the eigenvalues of $-\tilde{O}F_V\tilde{O}^*$ in the definition of $\tilde{M}$ are deterministic, while the eigenvalues of O are random. Note that both matrices $-\tilde{O}F_V\tilde{O}^*$ and O have uniformly-random eigenvectors. Then the idea of the proof is to rely on the phenomenon of *rigidity* for random matrix eigenvalues, which says that as $T \to \infty$, the eigenvalues of O are very close to their deterministic expected positions, which, in turn, essentially coincide with eigenvalues of $-F_V$. One technical difficulty is that we need to deal with $(I+O)^{-1}$, whose eigenvalues can be large (the absolute value of the largest eigenvalue grows linearly in T), however, as we will see, one can study $(I+O)$ instead of $(I+O)^{-1}$, and the problem of unbounded operators disappears. We proceed to the detailed proof.

PROOF OF PROPOSITION 12. We start by explicitly constructing the desired coupling. The eigenvalues of F_V are all roots of unity of order T different from 1. If $\beta = 2$, then we can diagonalize F_V to turn it into $\mathcal{T} \times \mathcal{T} = (T-1) \times (T-1)$ diagonal matrix with the roots of unity on the diagonal. If $\beta = 1$, the matrix F_V should be block-diagonalized (with blocks of size 2 and one additional block of size 1 corresponding to eigenvalue -1 if $\mathcal{T}$ is even): the pair of complex conjugate roots of unity ω and $\bar{\omega}$ gives rise to the 2×2 matrix of rotation by the angle $|\arg(\omega)|$. Let us denote by D the resulting (block) diagonal matrix multiplied by -1. In order to avoid ambiguity about the order of eigenvalues, we assume that the blocks correspond to the increasing order of $|\arg(-\omega)|$, i.e. the top-left 2×2 corner of D corresponds to the pair of the closest to 1 eigenvalues of D.

The eigenvalues of O also lie on the unit circle and if $\beta = 1$ they come in complex-conjugate pairs. Hence, O can be similarly block-diagonalized (we do not need to multiply by -1 this time) and we denote through D^{rand} the result. The distinction with F_V is that the eigenvalues are *random* and so is D^{rand}. The law of the eigenvalues of O is explicitly known in the random-matrix literature. Both for $\beta = 1$ and $\beta = 2$ they form a determinantal point process on the unit circle with explicit kernel. The *repulsion* between the eigenvalues leads to them being very close to evenly spaced as $T \to \infty$. We summarize this property in the following statement (which is a manifestation of much more general rigidity of eigenvalues, see, e.g., Erdos and Yau (2012)), whose proof can be found in Meckes and Meckes (2013, Lemma 10, $m = 1$, $u = T^\delta$ case, and Section 5).

Claim. There exist constants $c_1(\beta), c_2(\beta) > 0$, such that for $\beta = 1, 2$, every $\delta > 0$, there exists $\mathcal{T}_0(\delta)$ and for every $\mathcal{T} > \mathcal{T}_0(\delta)$ we have [19]

$$\text{Prob}\left(\max_{1 \leq i,j < \mathcal{T}} \left|D - D^{\text{rand}}\right|_{ij} > \frac{1}{\mathcal{T}^{1-\delta}}\right) < c_1(\beta) \cdot \mathcal{T} \cdot \exp\left(-c_2(\beta) \frac{\mathcal{T}^{2\delta}}{\log \mathcal{T}}\right). \tag{60}$$

We remark that since D and D^{rand} are block-diagonal, the bound on the maximum matrix element of their difference is equivalent to a similar bound for any other norm, e.g., for the maximum absolute value among the eigenvalues of $D - D^{\text{rand}}$.

We now choose another $\mathcal{T} \times \mathcal{T}$ uniformly-random orthogonal (or unitary if $\beta = 2$) matrix O_2 (independent from the rest), replace $-\tilde{O}F_V\tilde{O}^*$ with $O_2 D O_2^*$ and replace O with $O_2 D^{\text{rand}} O_2^*$. The invariance of the uniform measure on the orthogonal group $SO(N)$ (or

[19] All the constants can be made explicit, following Meckes and Meckes (2013).

on the unitary group $U(N)$ if $\beta = 2$) with respect to right/left multiplications, implies the distributional identities:

$$-\tilde{O} F_V \tilde{O}^* \stackrel{d}{=} O_2 D O_2^*, \qquad O \stackrel{d}{=} O_2 D^{\text{rand}} O_2^*.$$

The right-hand sides of the identities provide the desired coupling and (60) implies that these two random matrices are close to each other as $\mathcal{T} \to \infty$. Both matrices M and $\tilde{M}$ are obtained from the above two matrices by the following mechanism:

Map $M(Z)$**:** Given an orthogonal (or unitary if $\beta = 2$) matrix Z in $\mathcal{T}$–dimensional space, we define $U(Z) = (I_{\mathcal{T}} + Z)^{-1}$ and an $N \times N$ symmetric (or hermitian) matrix $M(Z)$ through

(61) $$M(Z) = [U(Z)]_{NN}([U^*(Z)U(Z)]_{NN})^{-1}[U^*(Z)]_{NN}.$$

Clearly, under our coupling

$$M = M(O_2 D^{\text{rand}} O_2^*), \qquad \tilde{M} = M(O_2 D O_2^*),$$

and it remains to prove a form of the uniform continuity of the map $Z \mapsto M(Z)$ as $N, \mathcal{T} \to \infty$.

We note that

$$2U(Z) - I_{\mathcal{T}} = \frac{I_{\mathcal{T}} - Z}{I_{\mathcal{T}} + Z}$$

is a skew-symmetric matrix and we use it to write $2U(Z)$ in the block form according to the splitting $\mathcal{T} = N + (\mathcal{T} - N)$.

$$2U(Z) = \begin{pmatrix} I_N + A(Z) & B(Z) \\ -B^*(Z) & I_{\mathcal{T}-N} + C(Z) \end{pmatrix}, \qquad A^*(Z) = -A(Z), \quad C^*(Z) = -C(Z).$$

Then, as in the proof of Theorem 6, we have

(62) $$M(Z) = \left(I_N + \tilde{B}(Z)\tilde{B}^*(Z)\right)^{-1}, \qquad \tilde{B}(Z) = \left(I_N - A(Z)\right)^{-1} B(Z).$$

and our statement boils down to the uniform continuity of the map $Z \mapsto \tilde{B}(Z)$. We remark that in our limit regime the matrix $B(Z)B^*(Z)$ might have large eigenvalues (in fact, of order $\mathcal{T}$) and therefore, $B(Z)$ is potentially an exploding factor in the definition of $\tilde{B}(Z)$. However, the exploding parts would precisely cancel out with similarly growing parts in $(I_N - A(Z))$. In order to see that, we use the formula for the inverse of the block matrix, which reads

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{Q}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{Q} \\ -\mathbf{Q}\mathbf{C}\mathbf{A}^{-1} & \mathbf{Q} \end{pmatrix}, \qquad \mathbf{Q} = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}$$

The advantage of this formula is that when used for $2U(Z)$, the ratio $-\mathbf{C}\mathbf{A}^{-1} = B^*(z)(1 + A(z))^{-1} = \tilde{B}^*(Z)$ gets expressed as $\mathbf{Q}^{-1}(-\mathbf{Q}\mathbf{C}\mathbf{A}^{-1})$, which is the ratio of two blocks in $U(Z)^{-1} = 1 + Z$.[20] Thus, if we write $1 + Z$ itself in the block form according to $\mathcal{T} = N + (\mathcal{T} - N)$ splitting:

$$I_{\mathcal{T}} + Z = \begin{pmatrix} I_N + Z_{11} & Z_{12} \\ Z_{21} & I_{\mathcal{T}-N} + Z_{22} \end{pmatrix},$$

then

$$\tilde{B}^*(Z) = \mathbf{Q}^{-1}(-\mathbf{Q}\mathbf{C}\mathbf{A}^{-1}) = (I_{\mathcal{T}-N} + Z_{22})^{-1} Z_{21},$$

[20] The same observation can be used to link Theorems 6 and 7 to each other.

$$M(Z) = \Big(I_N + Z_{21}^* \big((I_{\mathcal{T}-N} + Z_{22})(I_{\mathcal{T}-N} + Z_{22}^*)\big)^{-1} Z_{21}\Big)^{-1}$$

Hence, the matrix $M(Z)$ has the same (other from 1) eigenvalues as

$$M'(Z) = \Big(I_{\mathcal{T}-N} + \big((I_{\mathcal{T}-N} + Z_{22})(I_{\mathcal{T}-N} + Z_{22}^*)\big)^{-1} Z_{21} Z_{21}^*\Big)^{-1}$$

The latter can be simplified, since Z is orthogonal, implying $Z_{21}Z_{21}^* + Z_{22}Z_{22}^* = I_{\mathcal{T}-N}$:

$$\begin{aligned}(63)\quad M'(Z) &= \Big(I_{\mathcal{T}-N} + \big((I_{\mathcal{T}-N} + Z_{22} + Z_{22}^* + Z_{22}Z_{22}^*)\big)^{-1}(I_{\mathcal{T}-N} - Z_{22}Z_{22}^*)\Big)^{-1} \\ &= \Big(\big((I_{\mathcal{T}-N} + Z_{22} + Z_{22}^* + Z_{22}Z_{22}^*)\big)^{-1}(2\cdot I_{\mathcal{T}-N} + Z_{22} + Z_{22}^*)\Big)^{-1} \\ &= (2\cdot I_{\mathcal{T}-N} + Z_{22} + Z_{22}^*)^{-1}\big(I_{\mathcal{T}-N} + Z_{22} + Z_{22}^* + Z_{22}Z_{22}^*\big)\end{aligned}$$

At this point we rely on the lemma, which is proven below:

LEMMA 13. *Fix any small $\upsilon > 0$. If Z is a uniformly random element of the group $SO(\mathcal{T})$ of orthogonal determinant 1 matrices (or $U(\mathcal{T})$ if $\beta = 2$), then there exists $r > 0$ such that with probability tending to 1 (uniformly) as $\mathcal{T}, N \to \infty$ in such a way that $1 + \upsilon < \mathcal{T}/N < \upsilon^{-1}$, the smallest eigenvalue of $2\,I_{\mathcal{T}-N} + Z_{22} + Z_{22}^*$ is larger than r.*

Now we are ready to compare $M'(Z)$ with $M'(\tilde{Z})$, where $Z = O_2 D^{\text{rand}} O_2^*$, $\tilde{Z} = O_2 D O_2^*$. Fix $\delta > 0$. By (60), with probability tending to 1 as $N, \mathcal{T} \to \infty$, $\|Z - \tilde{Z}\| < \mathcal{T}^{\delta-1}$, where $\|\cdot\|$ is the spectral norm of the matrix (i.e. its largest singular value). Hence, since cutting corners of the matrix only decreases the spectral norm, with probability tending to 1,

$$(64)\qquad \Big\|\big(I_{\mathcal{T}-N} + Z_{22} + Z_{22}^* + Z_{22}Z_{22}^*\big) - \big(I_{\mathcal{T}-N} + \tilde{Z}_{22} + \tilde{Z}_{22}^* + \tilde{Z}_{22}\tilde{Z}_{22}^*\big)\Big\| < 4\mathcal{T}^{\delta-1}.$$

Simultaneously, by Lemma 13, the spectral norm of $(2\,I_{\mathcal{T}-N} + Z_{22} + Z_{22}^*)^{-1}$ is bounded from above by r^{-1} (and, hence, a similar bound holds for Z replaced with $\tilde{Z}$). We conclude that the part of the increment $M'(Z) - M'(\tilde{Z})$ due to the change of the numerator in the right-hand side of (63) is bounded from above (in spectral norm) by

$$(65)\qquad \text{const} \cdot \mathcal{T}^{\delta-1}.$$

For the denominator in (63), we use the matrix identity

$$(66)\qquad (G + \Delta)^{-1} - G^{-1} = -(G + \Delta)^{-1}\Delta\, G^{-1}$$

with $G = 2\,I_{\mathcal{T}-N} + Z_{22} + Z_{22}^*$, $\Delta = \tilde{Z}_{22} + \tilde{Z}_{22}^* - Z_{22} - Z_{22}^*$. By (60), $\|\Delta\| < 2\mathcal{T}^{\delta-1}$ with probability tending to 1. Hence, using Lemma 13 we upper bound the spectral norm of (66) by $\text{const} \cdot r^{-2}\,\mathcal{T}^{\delta-1}$. Since the spectral norm of the numerator in the right-hand side of (63) is at most 4, we conclude that the part of the increment $M'(Z) - M'(\tilde{Z})$ due to the change of the denominator also admits a bound (65). Summing up, we have shown that with probability tending to 1,

$$\|M'(O_2 D^{\text{rand}} O_2^*) - M'(O_2 D O_2^*)\| < \text{const} \cdot \mathcal{T}^{\delta-1}.$$

This finishes the proof of Proposition 12. □

PROOF OF LEMMA 13. We would like to show that for some deterministic $r>0$ the following inequality on the norm holds with probability tending to 1 as $N,\mathcal{T}\to\infty$:

$$\left\| \frac{Z_{22}+Z_{22}^*}{2} \right\| < 1-\tfrac{r}{2}. \tag{67}$$

For that we denote $H=\frac{Z+Z^*}{2}$ and notice that H is a Hermitian matrix with spectrum between -1 and 1. In fact, Forrester (2010, Section 3.7) explains that $\frac{1+H}{2}$ is again a Jacobi ensemble. The law of large numbers for the latter (reviewed in Theorem 14) implies that as $\mathcal{T}\to\infty$, the empirical measure of the eigenvalues $\frac{1}{\mathcal{T}}\sum_{i=1}^{\mathcal{T}}\delta_{\lambda_i(H)}$ converges (weakly in probability[21]) to a certain explicit deterministic measure on $[-1,1]$ with a density $p(x)$.

The matrix $\frac{Z_{22}+Z_{22}^*}{2}$ is a $(\mathcal{T}-N)\times(\mathcal{T}-N)$ corner of H. The largest eigenvalue of this matrix has the following representation:

$$\lambda_1\left(\frac{Z_{22}+Z_{22}^*}{2}\right) = \max_{\substack{u\in\langle e_{N+1},\dots,e_{\mathcal{T}}\rangle\\ |u|=1}} \left(\sum_{i=1}^{\mathcal{T}} \lambda_i(H)\left|(u,v_i(H))\right|^2\right), \tag{68}$$

where $\lambda_i(H)$ is the ith largest eigenvalue of H, $v_i(H)$ is the corresponding eigenvector and $(u,v_i(H))$ is the scalar product with this eigenvector. Since the vectors v_i are orthonormal, $\sum_{i=1}^N\left|(u,v_i(H))\right|^2=1$. On the other hand, all $\lambda_i(H)$ are not greater than 1 and only few of them are close to 1. Hence, (67) would follow, if we manage to show that the sum in (68) is not concentrated on few largest $\lambda_i(H)$. To show that, it suffices to upper-bound $\left|(u,v_i(H))\right|^2$, which we now do.

Let us fix k with $k<\min(N,\mathcal{T}-N)$ and upper bound (68) by

$$\begin{aligned} &\max_{\substack{u\in\langle e_{N+1},\dots,e_{\mathcal{T}}\rangle\\ |u|=1}} \left(\sum_{i=1}^{k}\left|(u,v_i(H))\right|^2 + \lambda_{k+1}(H)\sum_{i=k+1}^{\mathcal{T}}\left|(u,v_i(H))\right|^2\right) \\ &\qquad = \lambda_{k+1}(H)+(1-\lambda_{k+1}(H))\max_{\substack{u\in\langle e_{N+1},\dots,e_{\mathcal{T}}\rangle\\ |u|=1}}\left(\sum_{i=1}^{k}\left|(u,v_i(H))\right|^2\right). \end{aligned} \tag{69}$$

Since the law of the matrix H is invariant under conjugations with orthogonal (or unitary) matrices, the eigenvectors $v_1(H),v_2(H),\dots,v_{\mathcal{T}}(H)$ are uniformly-distributed, i.e. the $\mathcal{T}\times\mathcal{T}$ matrix formed by them is a uniformly-random orthogonal (or unitary) matrix. Hence,

$$\max_{\substack{u\in\langle e_{N+1},\dots,e_{\mathcal{T}-N}\rangle\\ |u|=1}}\left(\sum_{i=1}^{k}\left|(u,v_i(H))\right|^2\right).$$

is a maximum eigenvalue of the $(\mathcal{T}-N)\times(\mathcal{T}-N)$ corner of the projector in $\mathcal{T}$–dimensional space on random k–dimensional subspace chosen uniformly at random. If we denote through P_k and $P_{\mathcal{T}-N}$ the projectors on the first k and $\mathcal{T}-N$ basis vectors, respectively, and take a uniformly random orthogonal (or unitary if $\beta=2$) $\mathcal{T}\times\mathcal{T}$ matrix O, then we deal with the maximal eigenvalue of $P_{\mathcal{T}-N}OP_kO^*P_{\mathcal{T}-N}$. The maximal eigenvalue of this product is the same as the one of $OP_kO^*P_{\mathcal{T}-N}$, or of $P_kO^*P_{\mathcal{T}-N}O$, or of $P_kO^*P_{\mathcal{T}-N}OP_k$. Equivalently, this is the maximal eigenvalue of $k\times k$ corner Y of a projector on uniformly-random $(\mathcal{T}-$

[21] A sequence of *random* probability measures μ_n converges weakly in probability to a measure μ if for each bounded continuous function $f(x)$, the random variables $\int f(x)\mu_n(dx)$ converge in probability to $\int f(x)\mu(dx)$.

N)–dimensional space. Such $k \times k$ matrix Y is distributed as Jacobi ensemble with density proportional to

$$\det(Y)^{\frac{\beta}{2}(\mathcal{T}-N-k+1)-1} \det(1-Y)^{\frac{\beta}{2}(N-k+1)-1} dY, \tag{70}$$

see Forrester (2010, (3.113) and the following formula). We now choose $k = N/2$. Then the largest eigenvalue of the ensemble (70) converges as $N, \mathcal{T} \to \infty$ in probability (see e.g., Johnstone (2008), or Anderson et al. (2010, Section 2.6.2), or Holcomb and Moreno Flores (2012)) to a deterministic number $0 < c < 1$, depending on the ratio $\mathcal{T}/N$.

Simultaneously, $\lambda_{N/2+1}(H)$ converges as $\mathcal{T}, N \to \infty$ to another deterministic number $0 < c' < 1$, due to the aforementioned convergence of the empirical measure of H. Hence, (69) implies that the largest eigenvalue of $\frac{Z_{22}+Z_{22}^*}{2}$ is bounded away from 1. □

REMARK 5. *Although we are not pursuing this direction, it is possible to upgrade the above argument to the case when $N, \mathcal{T} \to \infty$ in such a way that $\mathcal{T}/N \to \infty$. Then r is tending to 0, but we can find the speed of convergence (and then propagate it to the precise bound in Proposition 12). For that we would need to analyze the largest eigenvalue of (70) more precisely (when $\mathcal{T}/N \to \infty$, the Jacobi ensemble concentrates near 1 and degenerates into the Wishart ensemble after proper rescaling), and also use exact asymptotics of $\lambda_{k+1}(H)$ (which can be obtained from the local law of the Jacobi ensemble for H near 1).*

We now have all the ingredients for Theorem 4 (as well as for its $\beta = 2$ version).

PROOF OF THEOREM 4. By Theorem 6, the eigenvalues $x_1 \geq \cdots \geq x_N$ of Jacobi ensemble $\mathbf{J}(N; \frac{N}{2}, \frac{T-2N}{2})$ have the same distribution as those of M defined in (59). On the other hand, by Proposition 11 the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N$ of the matrix $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ have the same distribution as those of $\tilde{M}$ defined in (56). Hence, Proposition 12 gives the desired statement. □

9.3. *Asymptotic of Jacobi ensemble.* In this section we review the asymptotic results for the Jacobi ensemble $\mathbf{J}(N; p, q)$ introduced in the Definition 3 as $N \to \infty$.

We assume that as $N \to \infty$, also $p, q \to \infty$, in such a way that

$$\frac{p-1}{N} = \frac{1}{2}(\mathfrak{p}-1), \quad \mathfrak{p} \geq 1, \qquad \frac{q-1}{N} = \frac{1}{2}(\mathfrak{q}-1), \quad \mathfrak{q} \geq 1, \tag{71}$$

where $\mathfrak{p}$ and $\mathfrak{q}$ are two parameters, which stay bounded away from 1 and from ∞ as $N \to \infty$.[22] We further define the *equilibrium measure* $\mu_{\mathfrak{p},\mathfrak{q}}$ of the Jacobi ensemble through:

$$\mu_{\mathfrak{p},\mathfrak{q}}(x)\,dx = \frac{\mathfrak{p}+\mathfrak{q}}{2\pi} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+-x)}}{x(1-x)} \mathbf{1}_{[\lambda_-,\lambda_+]}\,dx, \tag{72}$$

where the support $[\lambda_-, \lambda_+]$ of the measure is defined via

$$\lambda_\pm = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2}\left(\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}}\right)^2. \tag{73}$$

One can check that $0 < \lambda_- < \lambda_+ < 1$ for every $\mathfrak{p}, \mathfrak{q} > 1$. The law (72) is known as Wachter distribution. Further, define

$$c_\pm = \frac{(\mathfrak{p}+\mathfrak{q})}{2} \frac{\sqrt{\lambda_+-\lambda_-}}{\lambda_\pm(1-\lambda_\pm)}, \tag{74}$$

[22]The use of $p-1$ and $q-1$ instead of p and q is related to the choice of notations in (17). The following theorems would work without this shift as well, however, the shift leads to a faster speed of convergence, cf. Johnstone (2008, Discussion before Theorem 1).

and note that

$$\mu_{\mathfrak{p},\mathfrak{q}}(x) \approx \frac{c_\pm}{\pi}\sqrt{|x-\lambda_\pm|}, \text{ as } x \to \lambda_\pm \quad \text{inside} \quad [\lambda_-, \lambda_+],$$

where the normalization $\frac{1}{\pi}\sqrt{|x-\lambda_\pm|}$ was chosen to match the behavior of the Wigner semicircle law $\frac{1}{2\pi}\sqrt{4-x^2}$ near edges ± 2.

THEOREM 14. *Suppose that $N, p, q \to \infty$ in such a way that $\mathfrak{p} \geq 1$ and $\mathfrak{q} \geq 1$ in (71) stay bounded. For the second conclusions we additionally assume that $\mathfrak{q}$ is bounded away from 1 and for the third conclusion we additionally require $\mathfrak{p}$ to be bounded away from 1. Let $x_1 \geq x_2 \geq \cdots \geq x_N$ be N random eigenvalues of Jacobi ensemble $\mathbf{J}(N; p, q)$. Then*

(I) $\lim_{N\to\infty} \left| \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} - \mu_{\mathfrak{p},\mathfrak{q}} \right| = 0$, *weakly in probability.*
This means that for any continuous function $f(x)$ we have convergence in probability:

$$\lim_{N\to\infty} \left| \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \int_0^1 f(x)\mu_{\mathfrak{p},\mathfrak{q}}(x)dx \right| = 0. \tag{75}$$

(II) *For $\{\mathfrak{a}_i\}_{i=1}^\infty$ as in Proposition 1, we have convergence in finite-dimensional distributions for the largest eigenvalues:*

$$\lim_{N\to\infty} \left\{ N^{2/3} c_+^{2/3} (x_i - \lambda_+) \right\}_{i=1}^{N} \to \{\mathfrak{a}_i\}_{i=1}^\infty. \tag{76}$$

In particular, $N^{2/3}c_+^{2/3}(x_1-\lambda_+)$ converges to the Tracy-Widom distribution F_1.
(III) *We also have convergence in distribution for the smallest eigenvalues*[23]

$$\lim_{N\to\infty} \left\{ N^{2/3} c_-^{2/3} (\lambda_- - x_{N+1-i}) \right\}_{i=1}^{N} \to \{\mathfrak{a}_i\}_{i=1}^\infty. \tag{77}$$

The first proof of (75) appeared in Wachter (1980), for other proofs see Bai and Silverstein (2010) and Dumitriu and Paquette (2012); we follow the notations of the last reference. The asymptotics (76), (77) can be found in Johnstone (2008), and it is a manifestation of the universality for the distributions of largest/smallest eigenvalues of random matrices holding in much wider generality, see, e.g., Deift and Gioev (2007), Erdos and Yau (2012), Tao and Vu (2012). We remark that the original article Johnstone (2008) stated the convergence under an additional parity constraint on the parameters, yet this technical restriction can be removed, see the discussion at the end of Section 2 in Johnstone (2008) and Han et al. (2016).

9.4. *Proof of Theorem 2.* As an intermediate step we prove the following theorem.

THEOREM 15. *Suppose that $T, N \to \infty$ in such a way that $T/N > 2$ remains bounded away from 2 and from ∞. Let $\lambda_1 \geq \cdots \geq \lambda_N$ be eigenvalues of the matrix $S_{10}S_{00}^{-1}S_{01}S_{11}^{-1}$ defined as in Theorem 2 and Proposition 11. Then for $\{\mathfrak{a}_i\}_{i=1}^\infty$ as in Proposition 1, we have convergence in finite-dimensional distributions for the largest eigenvalues:*

$$\lim_{N\to\infty} \left\{ N^{2/3} c_+^{2/3} (\lambda_i - \lambda_+) \right\}_{i=1}^{\infty} \to \{\mathfrak{a}_i\}_{i=1}^\infty, \tag{78}$$

[23] The limiting processes $\{\mathfrak{a}_i\}_{i=1}^\infty$ arising for the largest and smallest eigenvalues are independent.

where

(79)
$$\lambda_\pm = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2}\left[\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}}\right]^2, \qquad c_+ = (\mathfrak{p}+\mathfrak{q})\,\frac{\sqrt{\lambda_+-\lambda_-}}{2\cdot\lambda_+\cdot(1-\lambda_+)},$$

$$\mathfrak{p} = 2 - \frac{2}{N}, \qquad \mathfrak{q} = \frac{T}{N} - 1 - \frac{2}{N}.$$

PROOF. Theorem 4 implies that the asymptotics of largest eigenvalues $x_1, x_2, \dots$ is the same as the one for the largest eigenvalues of $\mathbf{J}(N; \frac{N}{2}, \frac{T-2N}{2})$. For the latter we use Theorem 14 with $\mathfrak{p} = 2 - \frac{2}{N}$, $\mathfrak{q} = \frac{T}{N} - 1 - \frac{2}{N}$. □

PROOF OF THEOREM 2. We use Theorem 15 and the fact that for small x

$$\ln\bigl(1 - (\lambda_+ + x)\bigr) = \ln(1-\lambda_+) - \frac{1}{1-\lambda_+}x + o(x). \quad \square$$

9.5. *Asymptotics under white noise assumption.* In this section we explain how the proof of Theorem 5 is obtained. We follow the notations of Section 9.2 for the deterministic operators and spaces. We take a uniformly-random orthogonal (or unitary if $\beta = 2$) operator $\tilde{O}$ acting in $(T-1)$–dimensional space V and define an operator $\tilde{W}$ acting in V:

$$\tilde{W} = I_V - \tilde{O}(F_V)^{-1}\tilde{O}^*,$$

Let us introduce a symmetric (or Hermitian if $\beta = 2$) $N \times N$ matrix $\widehat{M}$ through

(80)
$$\widehat{M} = [\tilde{W}]_{NN}^V([\tilde{W}^*\tilde{W}]_{NN}^V)^{-1}[\tilde{W}^*]_{NN}^V.$$

PROPOSITION 16. *Choose an arbitrary positive definite covariance matrix Λ. Let ε be $N \times T$ matrix of random variables (real if $\beta = 1$ and complex if $\beta = 2$), such that T columns of ε are i.i.d., and each of them is an N-dimensional mean zero Gaussian vector with covariance Λ. Fix an arbitrary N–dimensional vector μ. Define the $N \times T$ matrix $X = (X_1, X_2, \dots, X_T)$ as in Eq. (23) via*

$$X_t = \mu + \varepsilon_t, \quad t = 1, \dots, T.$$

Further set $\Delta^c X_t = X_{t+1} - X_t$, $t = 1, \dots, T-1$ and $\Delta^c X_T = X_1 - X_t$. Define $N \times N$ matrices:

(81)
$$S_{00}^{w.n.} = \Delta^c X \mathcal{P}(\Delta^c X)^*, \quad S_{01}^{w.n.} = \Delta^c X \mathcal{P} X^*, \quad S_{10}^{w.n.} = X\mathcal{P}(\Delta^c X)^*, \quad S_{11}^{w.n.} = X\mathcal{P}X^*,$$

Then the eigenvalues of the matrix $S_{10}^{w.n.}(S_{00}^{w.n.})^{-1}S_{01}^{w.n.}(S_{11}^{w.n.})^{-1}$ have the same distribution as those of $\widehat{M}$ in (80).

PROOF. The proof is similar to that of Proposition 11 and we omit many details. First, note that μ cancels out both in $X\mathcal{P}$ and in $\Delta^c X$. Hence, without loss of generality we can assume $\mu = 0$.

Next, we define

$$G = \mathcal{P}(F^{-1} - I_T)\mathcal{P}.$$

Note that $\mathcal{P}$ and $(F^{-1} - I_T)$ commute with each other. Using additionally $\mathcal{P}^2 = \mathcal{P}$, we get:

$$S_{00}^{w.n.} = \varepsilon G^* G \varepsilon^*, \quad S_{01}^{w.n.} = \varepsilon G^* \varepsilon^*, \quad S_{10}^{w.n.} = \varepsilon G \varepsilon^*, \quad S_{11}^{w.n.} = \varepsilon \mathcal{P} \varepsilon^*.$$

At this point, arguing as in the proof of Proposition 11, we can assume that the covariance matrix Λ is identical, since any other covariance matrix can be obtained at the cost of conjugation of $S_{10}^{w.n.}(S_{00}^{w.n.})^{-1}S_{01}^{w.n.}(S_{11}^{w.n.})^{-1}$, which leaves the eigenvalues unchanged.

We can further identify non-zero eigenvalues of $S_{10}^{w.n.}(S_{00}^{w.n.})^{-1}S_{01}^{w.n.}(S_{11}^{w.n.})^{-1}$ with those of the product of two projectors $P_2P_1P_2$ (similarly to (58)), where P_1 is the orthogonal projector onto the space spanned by N columns of $G\varepsilon^*$ and P_2 is the orthogonal projector onto the space spanned by N columns of $\mathcal{P}\varepsilon^*$. Next, we rotate the space V, so that the span of N columns of $\mathcal{P}\varepsilon^*$ turns into the span of the first N basis vectors $\tilde{e}_1, \dots, \tilde{e}_N$ in V. At this point we arrive at (80) with $\tilde{W}$ replaced by $-\tilde{W}$. It remains to notice that the introduction of prefactor -1 in front of $\tilde{W}$ does not change the matrix (80). □

Let $\mathcal{T} = T - 1$ and let O be uniformly random real orthogonal of determinant 1 (if $\beta = 1$) or complex unitary (if $\beta = 2$) $\mathcal{T} \times \mathcal{T}$ matrix. Define $W = I_{\mathcal{T}} + O$. Then we set

$$M = [W]_{NN}([W^*W]_{NN})^{-1}[W^*]_{NN}. \tag{82}$$

Recall that for a matrix A with real spectrum, $\lambda_i(A)$ is the ith largest eigenvalue of A.

PROPOSITION 17. *Assume $\mathcal{T} = T - 1 \geq 2N$. One can couple M from (82) with $\widehat{M}$ from (80) in such a way that for each $\epsilon > 0$ we have*

$$\mathrm{Prob}\left(\max_{1\le i\le N} |\lambda_i(M) - \lambda_i(\widehat{M})| < \frac{1}{N^{1-\epsilon}}\right) \to 1,$$

as $T, N \to \infty$ in such a way that the ratio T/N remains bounded.

The proof is almost identical to that of Proposition 12 and we omit it: the key idea is to notice that the only difference between M and $\widehat{M}$ is in the replacement of O by $-\tilde{O}(F_V)^{-1}\tilde{O}^*$; however, by the results of Meckes and Meckes (2013) the latter two matrices can be coupled so that their eigenvectors are the same, while eigenvalues are very close to each other.

Combining Proposition 16, Proposition 17, and Theorem 7, we arrive at the statement of Theorem 5.

9.6. *Data.* As of May 12, 2020, S&P100 consists of the following companies: Apple Inc., AbbVie Inc., Abbott Laboratories Accenture, Adobe Inc., American International Group Allstate, Amgen Inc., American Tower, Amazon.com, American Express Boeing Co., Bank of America Corp, Biogen, The Bank of New York Mellon, Booking Holdings, BlackRock Inc, Bristol-Myers Squibb, Berkshire Hathaway, Citigroup Inc, Caterpillar Inc., Charter Communications, Colgate-Palmolive, Comcast Corp., Capital One Financial Corp., ConocoPhillips, Costco Wholesale Corp., salesforce.com, Cisco Systems, CVS Health, Chevron Corporation, DuPont de Nemours Inc., Danaher Corporation, The Walt Disney Company, Dow Inc., Duke Energy, Emerson Electric Co., Exelon, Ford Motor Company, Facebook Inc., FedEx, General Dynamics, General Electric, Gilead Sciences, General Motors, Alphabet Inc. (Class C), Alphabet Inc. (Class A), Goldman Sachs, Home Depot, Honeywell, International Business Machines, Intel Corp., Johnson & Johnson, JPMorgan Chase & Co., Kraft Heinz, Kinder Morgan, The Coca-Cola Company, Eli Lilly and Company, Lockheed Martin, Lowe's, MasterCard Inc, McDonald's Corp, Mondelēz International, Medtronic plc, MetLife Inc., 3M Company, Altria Group, Merck & Co., Morgan Stanley, Microsoft, NextEra Energy, Netflix, Nike Inc., NVIDIA Corp., Oracle Corporation, Occidental Petroleum Corp., PepsiCo, Pfizer Inc, Procter & Gamble Co, Philip Morris International, PayPal Holdings, Qualcomm Inc., Raytheon Technologies, Starbucks Corp., Schlumberger, Southern Company, Simon Property Group, Inc., AT&T Inc, Target Corporation, Thermo Fisher Scientific, Texas Instruments, UnitedHealth Group, Union Pacific Corporation, United Parcel Service, U.S. Bancorp, Visa Inc., Verizon Communications, Walgreens Boots Alliance, Wells Fargo, Walmart, Exxon Mobil Corp.

Eight of the above companies are not available for the entire period under consideration 01.01.2010 – 01.01.2020. Those companies are AbbVie Inc. (founded in 2013), Charter Communications (was bancrupt in 2009 and got released on NASDAQ in the middle of 2010), Dow Inc. (was spun off of DowDuPont on April 1, 2019), Facebook Inc. (went on IPO on February 1, 2012), General Motors (approached bankruptcy in 2009 and returned to NASDAQ as a new company at the end of 2010), Kraft Heinz (Kraft Foods and H.J. Heinz merged into Kraft Heinz in 2015), Kinder Morgan (was taken in a buyout and began trading again on the NYSE on February 11, 2011), PayPal Holdings (was part of eBay until 2015).

Acknowledgements. The authors are grateful to Donald Andrews, Alexei Borodin, Giuseppe Cavaliere, Alice Guionnet, Bruce Hansen, Søren Johansen, Grigori Olshanski, and anonymous referees for valuable suggestions. V.G. acknowledges support from the NSF Grants DMS-1664619 and DMS-1949820.

REFERENCES

Anderson, G., A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge University Press, 2010.

Bai, J. and S. Ng, "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics*, 2008, *3* (2), 890163.

Bai, Z. and J.W. Silverstein, *Spectral analysis of large dimensional random matrices*, Vol. 20, New York: Springer, 2010.

Baik, J., G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, 2005, *33* (5), 1643–1697.

Bao, Z., G. Pan, and W. Zhou, "Central limit theorem for partial linear eigenvalue statistics of Wigner matrices," *Journal of Statistical Physics*, 2013, *150*, 88–129.

Bao, Z., J. Hu, G. Pan, and W. Zhou, "Canonical correlation coefficients of high-dimensional Gaussian vectors: finite rank case," *The Annals of Statistics*, 2019, *47* (1), 612–640.

Ben Arous, G. and A. Guionnet, "Large deviations for Wigner's law and Voiculescu's non-commutative entropy," *Probability theory and related fields*, 1997, *108* (4), 517–542.

Borodin, A. and V. Gorin, "General $\beta-$Jacobi Corners Process and the Gaussian Free Field," *Communications on Pure and Applied Mathematics*, 2015, *68* (10), 1774–1844.

Borodin, A. and V. Gorin, "Lectures on integrable probability," in "Probability and statistical physics in St. Petersburg," Vol. 91, American Mathematical Soc., 2016, pp. 155–214.

Breitung, J. and M.H. Pesaran, "Unit Roots and Cointegration in Panels," in "Mátyás L., Sevestre P. (eds) The Econometrics of Panel Data. Advanced Studies in Theoretical and Applied Econometrics, vol. 46," Berlin, Heidelberg: Springer, 2008, pp. 279–322.

Bykhovskaya, A., "Time Series Approach to the Evolution of Networks: Prediction and Estimation," *Journal of Business & Economic Statistics*, 2021.

Cavaliere, G., A. Rahbek, and A.R. Taylor, "Bootstrap determination of the co-integration rank in vector autoregressive models," *Econometrica*, 2012, *80* (4), 1721–1740.

Deift, P. and D. Gioev, "Universality at the edge of the spectrum for unitary, orthogonal, and symplectic ensembles of random matrices," *Communications on Pure and Applied Mathematics*, 2007, *69* (6), 867—-910.

Dumitriu, I. and A. Edelman, "Matrix models for beta ensembles," *Journal of Mathematical Physics*, 2002, *43* (11), 5830–5847.

Dumitriu, I. and E. Paquette, "Global fluctuations for linear statistics of β-Jacobi ensembles," *Random Matrices: Theory and Applications*, 2012, *1* (04), 1250013.

Elliott, G., T.J. Rothenberg, and J.H. Stock, "Efficient tests for an autoregressive unit root," *Econometrica*, 1996, *64* (4), 813–836.

Engle, R.F. and C.W. Granger, "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, 1987, *55* (2), 251–276.

Erdos, L. and H.T. Yau, "Universality of local spectral statistics of random matrices," *Bulletin of the American Mathematical Society*, 2012, *49* (3), 377—-414.

Forrester, P.J., "The spectrum edge of random matrix ensembles," *Nuclear Physics B*, 1993, *402* (3), 709–728.

Forrester, P.J., *Log-gases and random matrices*, Princeton University Press, 2010.

Gonzalo, J. and J.Y. Pitarakis, "Dimensionality effect in cointegration analysis," in "Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive WJ Granger," Oxford: Oxford University Press, 1999, chapter 9, pp. 212–229.

Granger, C.W., "Some properties of time series data and their use in econometric model specification," *Journal of Econometrics*, 1981, *16* (1), 121–130.

Guionnet, A. and J. Novak, "Asymptotics of unitary multimatrix models: The Schwinger–Dyson lattice and topological recursion," *The Schwinger–Dyson lattice and topological recursion*, 2015, *268* (10), 2851–2905.

Han, X., G. M. Pan, and B. Zhang, "The Tracy-Widom law for the Largest Eigenvalue of F Type Matrix," *The Annals of Statistics*, 2016, *44* (4), 1564–1592.

Han, X., G.M. Pan, and Q. Yang, "A unified matrix model including both CCA and F matrices in multivariate analysis: the largest eigenvalue and its applications," *Bernoulli*, 2018, *24* (4B), 3447–3468.

Harvey, D.I., S.J. Leybourne, and A.R. Taylor, "Unit root testing in practice: dealing with uncertainty over the trend and initial condition," *Econometric theory*, 2009, *25* (3), 587–636.

Ho, M.S. and B.E. Sørensen, "Finding cointegration rank in high dimensional systems using the Johansen test: an illustration using data based Monte Carlo simulations," *The Review of Economics and Statistics*, 1996, *78* (4), 726–732.

Holcomb, D. and G.R. Moreno Flores, "Edge scaling of the β-Jacobi ensemble," *Journal of Statistical Physics*, 2012, *149* (9), 1136–1160.

Hua, L., *Harmonic analysis of functions of several complex variables in the classical domains*, Vol. 6, American Mathematical Soc., 1963.

Johansen, S., "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, 1988, *12* (2–3), 231–254.

Johansen, S., "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 1991, *59*, 1551–1580.

Johansen, S., *Likelihood-based inference in cointegrated vector autoregressive models*, Oxford University Press, 1995.

Johansen, S., "A small sample correction for the test of cointegrating rank in the vector autoregressive model," *Econometrics*, 2002, *70* (5), 1929–1961.

Johansson, K., "On fluctuations of eigenvalues of random Hermitian matrices," *Duke mathematical journal*, 1998, *91* (1), 151–204.

Johnstone, I., "Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy-Widom limits and rates of convergence," *The Annals of Statistics*, 2008, *36* (6), 2638–2716.

Johnstone, I., "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, 2001, *29* (2), 295–327.

Johnstone, I. and A. Onatski, "Testing in high-dimensional spiked models," *The Annals of Statistics*, 2020, *48* (3), 1231–1254.

Kadell, K.W., "The Selberg–Jack symmetric functions," *Advances in mathematics*, 1997, *130* (1), 33–102.

Killip, R. and I. Nenciu, "Matrix models for circular ensembles," *International Mathematics Research Notices*, 2004, *2004* (50), 2665–2701.

Krauss, C., "Statistical arbitrage pairs trading strategies: Review and outlook.," *Journal of Economic Surveys*, 2017, *31* (2), 513–545.

Lütkepohl, H., P. Saikkonen, and C. Trenkler, "Maximum eigenvalue versus trace tests for the cointegrating rank of a VAR process," *The Econometrics Journal*, 2001, *4* (2), 287–310.

Maddala, G.S. and In-Moo Kim, *Unit Roots, Cointegration, and Structural Change*, Cambridge University Press, 1998.

Meckes, E. and M. Meckes, "Spectral measures of powers of random matrices," *Electronic communications in probability*, 2013, *18.*

Mehta, M.L., *Random matrices*, Elsevier, 2004.

Mingo, J.A. and M. Popa, "Real second order freeness and Haar orthogonal matrices," *Journal of Mathematical Physics*, 2013, *54* (5), 051701.

Muirhead, R.J., *Aspects of multivariate statistical theory*, John Wiley & Sons, 2009.

Müller, U.K. and G. Elliott, "Tests for unit roots and the initial condition," *Econometrica*, 2003, *71* (4), 1269–1286.

Onatski, A. and C. Wang, "Alternative asymptotics for cointegration tests in large vars," *Econometrica*, 2018, *86* (4), 1465–1478.

Onatski, A. and C. Wang, "Extreme canonical correlations and high-dimensional cointegration analysis," *Journal of Econometrics*, 2019.

O'Rourke, S., "Gaussian Fluctuations of Eigenvalues in Wigner Random Matrices," *Journal of Statistical Physics*, 2010, *138* (6), 1045–1066.

Paruolo, P., "The power of lambda max," *Oxford Bulletin of Economics and Statistics*, 2001, *63* (3), 395–403.

Phillips, P.C.B. and S. Ouliaris, "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 1990, *58* (1), 165–193.

Ramirez, J., B. Rider, and B. Virág, "Beta ensembles, stochastic Airy spectrum, and a diffusion," *Journal of the American Mathematical Society*, 2011, *24* (4), 919–944.

Reinsel, G.C. and S.K. Ahn, "Vector autoregressive models with unit roots and reduced rank structure: Estimation. Likelihood ratio test, and forecasting," *Journal of time series analysis*, 1992, *13* (4), 353–375.

Sodin, S., "Several applications of the moment method in random matrix theory," *Proceedings of the International Congress of Mathematicians*, 2014.

Stock, J.H. and M.W. Watson, "Testing for common trends," *Journal of the American statistical Association*, 1988, *83* (404), 1097–1107.

Swensen, A.R., "Bootstrap Algorithms for Testing and Determining the Cointegration Rank in VAR Models," *Econometrica*, 2006, *74* (6), 1699–1714.

Tao, T. and V. Vu, "Random matrices: the universality phenomenon for Wigner ensembles," *Modern aspects of random matrix theory*, 2012, *72*, 121—-172.

Tracy, C.A. and H. Widom, "On orthogonal and symplectic matrix ensembles," *Communications in Mathematical Physics*, 1996, *177* (3), 727–754.

Wachter, K.W., "The limiting empirical measure of multiple discriminant ratios," *The Annals of Statistics*, 1980, *8* (5), 937–957.

Xiao, Z. and P.C.B. Phillips, "Efficient detrending in cointegrating regression," *Econometric Theory*, 1999, *15* (4), 519–548.

Zhang, B., G. M. Pan, and J. T. Gao, "CLT for largest eigenvalues and unit root testing for high-dimensional nonstationary time series," *The Annals of Statistics*, 2018, *46* (5), 2186–2215.

ASYMPTOTICS OF COINTEGRATION TESTS FOR HIGH-DIMENSIONAL VAR(k)

ANNA BYKHOVSKAYA AND VADIM GORIN

Abstract. The paper studies non-stationary high-dimensional vector autoregressions of order k, VAR(k). Additional deterministic terms such as trend or seasonality are allowed. The number of time periods, T, and the number of coordinates, N, are assumed to be large and of the same order. Under such regime the first-order asymptotics of the Johansen likelihood ratio (LR), Pillai-Barlett, and Hotelling-Lawley tests for cointegration is derived: Test statistics converge to non-random integrals. For more refined analysis, the paper proposes and analyzes a modification of the Johansen test. The new test for the absence of cointegration converges to the partial sum of the Airy$_1$ point process. Supporting Monte Carlo simulations indicate that the same behavior persists universally in many situations beyond our theorems.

The paper presents an empirical implementation of the approach to the analysis of stocks in S&P100 and of cryptocurrencies. The latter example has strong presence of multiple cointegrating relationships, while the former is consistent with the null of no cointegration.

Date: March 22, 2022.

The authors would like to thank Bruce Hansen and Alexei Onatski for valuable comments and suggestions. The authors are grateful to Victor Kleptsyn for his help with the proof of Lemma 23. V.G. acknowledges support from the NSF Grants DMS-1664619 and DMS-1949820, BSF grant 2018248.

1. Introduction

Starting from the pioneering work Sims (1980), vector autoregressions (VARs) became a workhorse model in macroeconomics and other fields. A lot of data in macroeconomics and finance is non-stationary (e.g., consumption and output), while properties of VARs can be very different depending on whether one deals with stationary or non-stationary time series. There is a further subdivision in the non-stationary case: It is important to understand if the data is cointegrated or not. That is, whether there exists a stationary non-trivial linear combination of considered series (e.g., log of consumption minus log of output is stationary, while the series per se are unit roots).

Classical tools for testing cointegration (see, e.g., Johansen (1995), Maddala and Kim (1998), Juselius (2006)) failed to achieve desired finite sample performance when the number of time series, N, is large. Thus, they are not commonly used in such settings and designing proper tools to handle the cointegration for large N remained an open problem for years (see, e.g., Choi (2015, Sections 2.3.3, 2.4)). Recently Onatski and Wang (2018, 2019) and Bykhovskaya and Gorin (2021) have opened a new avenue based on "T/N converging to a constant" asymptotic regime. Yet, the testing procedures of these texts covered only VAR(1), while in practice researchers rarely stick to VARs of order 1, instead usually at least two lags are considered. Indeed, as noted already in Pagan (1987), "most applications of Sims' methodology have put number of lags between four and ten." Since Pagan (1987) the lengths of available time-series and computing power only increased, thus, allowing to work with even more complex models. Hence, it is important to generalize and extend the above papers to a VAR(k) setting, which is the main topic of our text.

Our paper analyses a family of tests for the absence of cointegration for non-stationary VAR(k), such as the Johansen likelihood ratio (LR) test (Johansen (1988, 1991)) and related Hotelling-Lawley and Pillai-Bartlett tests (see, e.g., Gonzalo and Pitarakis (1995) and references therein) as N and T jointly and proportionally go to infinity. The shared feature of these tests is that their statistics are based on the squared sample canonical correlations between certain transformations of current changes and past levels of the data. The main contribution of our paper is in the asymptotic analysis of these canonical correlations. First, we show that for VAR(k) with general k, under the null of no cointegration (and some additional technical conditions) the empirical distribution of the squared sample canonical correlations converges to the Wachter distribution. As a corollary, we deduce the first-order deterministic limits of the above test statistics. Second, we introduce a modification of the testing procedure and prove much more refined results in the modified setting. By computing the exact asymptotic behavior of the probability distributions of individual canonical correlations after proper recentering and rescaling, we are able to compute the critical values for the test of no cointegration with correct asymptotic size as N and T jointly and proportionally go to infinity.

We remark that there is a wide scope of literature devoted to the corrections of Johanssen's LR test and its relatives (originally developed based on fixed N, large T asymptotics) for large values of N (see, e.g., Reinsel and Ahn (1992), Johansen (2002), Swensen (2006), Cavaliere et al. (2012), Onatski and Wang (2019)). The distinguishing feature of our work is that we are not trying to correct the finite N asymptotic statements, which stop working for large N, by introducing various empirical adjustments. Instead, we develop a theoretical framework for working with the large N case directly. One advantage is that our approach explains the general phenomenology and predicts the asymptotic behavior, so that the empirical or simulational adjustments for particular values of the parameters of the model are no longer needed.

To achieve the above, in our proofs we use the VAR(1) results of Bykhovskaya and Gorin (2021) as a cornerstone. The main technical work is devoted to producing recursive arguments, which reduce the VAR(k) behavior to that of VAR($k-1$) and eventually to VAR(1). The central role is played by highly non-trivial projections from the group of orthogonal $T \times T$ matrices to the smaller subgroup of orthogonal $(T-N) \times (T-N)$ matrices. While such projections were previously used in the asymptotic representation theory, to the authors' knowledge, this is their first appearance in econometrics or statistics context. Thus, many new properties of those projections need to be developed in our framework.

The rest of the paper is organized as follows. Section 2 describes our setting and provides first asymptotic results. Section 3 constructs our modified test and computes its asymptotics, while Section 4 presents supporting Monte Carlo simulations. Section 5 illustrates our theoretical findings on S&P100 data and on the prices of cryptocurrencies. Finally, Section 6 concludes. All proofs are in Sections 7–10.

2. First order asymptotics of sample canonical correlations

We consider an N-dimensional vector autoregressive process of order k, VAR(k), based on a sequence of i.i.d. mean zero Gaussian errors $\{\varepsilon_t\}$ with non-degenerate covariance matrix Λ. That is, written in the error correction form,

$$\Delta X_t = \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Pi X_{t-k} + \Phi D_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1}$$

where $\Delta X_t := X_t - X_{t-1}$, D_t is a d_D-dimensional vector of deterministic terms, such as a constant, a trend or seasonality (extra explanatory variables are also allowed as long as they are observed), and $\Gamma_1, \ldots, \Gamma_{k-1}$, Π, Φ are unknown parameters. The process is initialized at fixed $X_{1-k}, \ldots, X_0$. We do not impose any restrictions on Λ, thus, we allow for arbitrary correlations across coordinates of X_t. In contrast, many previous approaches relied on specific properties of the covariance matrix Λ, see, e.g., Breitung and Pesaran (2008), Bai and Ng (2008, Section 7) and Zhang et al. (2018).

Remark 1. *Alternatively error correction form can be written as*

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{k-1} \tilde{\Gamma}_i \Delta X_{t-i} + \Phi D_t + \varepsilon_t, \qquad t = 1, \dots, T,$$

so that $\tilde{\Gamma}_i = \Gamma_i - \Pi$. *Whether we use the former (Eq. (1)) or the latter form does not affect our results. The testing procedures of our interest are based on residuals from regressing* X_{t-k} *on* $\Delta X_{t-1}, \dots, \Delta X_{t-k+1}$*, which are the same as residuals from regressing* $X_{t-1} = X_{t-k} + \Delta X_{t-1} + \ldots + \Delta X_{t-k+1}$ *on* $\Delta X_{t-1}, \dots, \Delta X_{t-k+1}$.

We are interested in the behavior of the squared sample canonical correlations between transformed past levels (lags) and changes (first differences) of the data X_t. As shown in (Johansen, 1988, 1991), see also Anderson (1951), they are related to whether the process is cointegrated or not. To be more specific, they appear in the likelihood ratio test for the presence and rank of the cointegration. Let us formally define these correlations. Here and below * means matrix transposition.

Procedure 1 (Johansen (1991))**.** Let $Z_{0t} = \Delta X_t$, $Z_{1t} = (\Delta X_{t-1}^*, \dots, \Delta X_{t-k+1}^*, D_t^*)^*$, and $Z_{kt} = X_{t-k}$. We regress lags Z_{kt} and changes Z_{0t} on regressors Z_{1t} (lagged changes and deterministic terms) and define residuals

$$R_{it} = Z_{it} - \left(\sum_{\tau=1}^{T} Z_{i\tau} Z_{1\tau}^*\right) \left(\sum_{\tau=1}^{T} Z_{1\tau} Z_{1\tau}^*\right)^{-1} Z_{1t}, \quad i = 0, k. \tag{2}$$

Define further $N \times N$ matrices $S_{ij} := \sum_{t=1}^{T} R_{it} R_{jt}^*$, $i, j = 0, k$ and finally set

$$\mathcal{C} = S_{kk}^{-1} S_{k0} S_{00}^{-1} S_{0k}.$$

The N eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ of $\mathcal{C}$ are squared sample canonical correlations of R_0 and R_k, where R_i is $N \times T$ matrix composed of columns R_{it}, $i = 0, k$.

The Johansen's likelihood ratio (LR) statistic for testing the hypothesis $\operatorname{rank}(\Pi) \leq r_1$ (at most r_1 cointegrating relationships) versus the alternative $\operatorname{rank}(\Pi) \in (r_1, r_2]$ (between r_1 and r_2 cointegrating relationships) with $r_2 > r_1$ has the form

$$\sum_{i=r_1+1}^{r_2} \ln(1 - \lambda_i), \tag{3}$$

the Pillai-Barlett statistic is

$$\sum_{i=r_1+1}^{r_2} \lambda_i,$$

and Hotelling-Lawley statistic is

$$\sum_{i=r_1+1}^{r_2} \frac{\lambda_i}{1-\lambda_i},$$

see Gonzalo and Pitarakis (1995) for discussion and many references about these statistics.

In Theorem 3 we show that the empirical measure of eigenvalues of $\mathcal{C}$ converges (weakly in probability) to the Wachter distribution. The theorem generalizes results of Onatski and Wang (2018) from VAR(1) to VAR(k) setting[1].

Definition 2. Wachter distribution is a probability distribution on $[0,1]$ which depends on two parameters $\mathfrak{p} > 1$ and $\mathfrak{q} > 1$ and has density

$$\mu_{\mathfrak{p},\mathfrak{q}}(x) = \frac{\mathfrak{p}+\mathfrak{q}}{2\pi} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{x(1-x)} \mathbf{1}_{[\lambda_-,\lambda_+]}\,, \tag{4}$$

where the support $[\lambda_-, \lambda_+] \subset (0,1)$ of the measure is defined via

$$\lambda_{\pm} = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2}\left(\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}}\right)^2. \tag{5}$$

Theorem 3. *Let X_t follow Eq.* (1)*. Suppose that k is fixed and as $N \to \infty$,*

$$\lim_{N\to\infty} \frac{T}{N} = \tau > (k+1) \qquad \text{and} \tag{6}$$

$$\lim_{N\to\infty} \frac{1}{N}\Big(\operatorname{rank}(\Pi) + \operatorname{rank}(\Gamma_1) + \operatorname{rank}(\Gamma_2) + \cdots + \operatorname{rank}(\Gamma_{k-1}) + d_D\Big) = 0. \tag{7}$$

Then for the empirical measure of eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N$ of $\mathcal{C}$ we have:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i} = \mu_{2,\tau-k}(x)\,\mathrm{d}x, \tag{8}$$

weakly in probability, meaning that for each continuous function $f(x)$ on $x \in [0,1]$, we have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(\lambda_i) = \int_0^1 f(x)\mu_{2,\tau-k}(x)\,\mathrm{d}x, \qquad \text{in probability.} \tag{9}$$

Imposing assumption (7) can be viewed as a dimension reduction (cf. sparsity assumption). Approximating data with low rank matrices is a widely used and powerful technique in data sciences: in machine learning, recommender systems (e.g., movie preference recognition), computational mathematics, etc. We refer to Udell and Townsend (2019) for theoretical

[1]While Onatski and Wang (2018, Theorem 1) allows the data to be VAR(k) under the restriction (7), they construct the matrix $\mathcal{C}$ involved in the statistical testing procedures as if the data was VAR(1). That is, the formal procedure is based on the misspecified VAR(1) setting, while we use the true VAR(k) procedure.

explanations of the suitability of low rank models, as well as many references to the situations where they are very efficient.

Figure 1 illustrates Theorem 3 for independent standard normal errors and $k = 2$, $N = 150$, $T = 1500$. Parameters are $\Phi D_t = 1_N$, $\Gamma_1 = 0.95E_{12}$, $\Pi = -0.1E_{\cdot 1}$, where 1_N is $N \times 1$ column-matrix of ones, E_{12} is an $N \times N$ matrix with one at the intersection of the 1st row and the 2nd column and zeros everywhere else, and $E_{\cdot 1}$ is an $N \times N$ matrix with the first column of ones and zeros everywhere else. Thus, all matrices have rank one. The parameters of the Wachter distribution are $\mathfrak{p} = 2$, $\mathfrak{q} = T/N - k = 8$. The single separated (rightmost) eigenvalue corresponds to rank$(\Pi) = 1$, i.e. one cointegrating relationship.

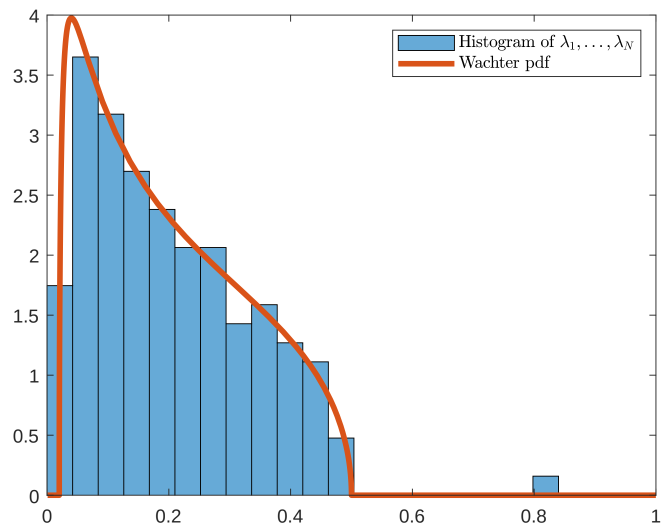


FIGURE 1. Illustration of Theorem 3: Eigenvalues and Wachter distribution. Data generating process: $\Delta X_t = 1_N + 0.95E_{12}\Delta X_{t-1} - 0.1E_{\cdot 1}X_{t-2} + \varepsilon_t$, $T = 1500$, $N = 150$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0, 1)$.

The proof of Theorem 3 is based on treating the setting of (1) as a small rank perturbation of a more restrictive setting analyzed in Section 3. The small rank assumption (7) is crucial for the validity of the theorem and we expect that the asymptotic behavior changes in situations when (7) fails. In contrast, the assumption that $T > (k + 1)N$ can potentially be relaxed. When $T < (k + 1)N$, the matrix $\mathcal{C}$ has deterministic eigenvalues equal to 1, which should be taken into account in $N \to \infty$ asymptotics. This case can be also addressed by our methods, but we do not pursue this direction.

An important corollary of Theorem 3 is that it provides asymptotic behavior of various tests constructed from eigenvalues of $\mathcal{C}$.

Corollary 4. *Under the assumptions of Theorem 3, suppose that the ranks* $r_1 = r_1(N)$ *and* $r_2 = r_2(N)$ *are such that*

$$\lim_{N\to\infty} \frac{r_1}{N} = \rho_1, \quad \lim_{N\to\infty} \frac{r_2}{N} = \rho_2.$$

Let $F(x) = \int_x^1 \mu_{2,\tau-k}(x)\,\mathrm{d}x$*. Then we have convergence in probability for the test statistics:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=r_1+1}^{r_2} \ln(1-\lambda_i) = \int_{F^{-1}(\rho_2)}^{F^{-1}(\rho_1)} \ln(1-x)\mu_{2,\tau-k}(x)\,\mathrm{d}x, \qquad \textit{if} \;\; \rho_1 > 0;$$

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{i=r_1+1}^{r_2} \ln(1-\lambda_i) \le \int_{F^{-1}(\rho_2)}^{1} \ln(1-x)\mu_{2,\tau-k}(x)\,\mathrm{d}x, \qquad \textit{if} \;\; \rho_1 = 0;$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=r_1+1}^{r_2} \lambda_i = \int_{F^{-1}(\rho_2)}^{F^{-1}(\rho_1)} x\mu_{2,\tau-k}(x)\,\mathrm{d}x;$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=r_1+1}^{r_2} \frac{\lambda_i}{1-\lambda_i} = \int_{F^{-1}(\rho_2)}^{F^{-1}(\rho_1)} \frac{x}{1-x}\mu_{2,\tau-k}(x)\,\mathrm{d}x, \qquad \textit{if} \;\; \rho_1 > 0;$$

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{i=r_1+1}^{r_2} \frac{\lambda_i}{1-\lambda_i} \ge \int_{F^{-1}(\rho_2)}^{1} \frac{x}{1-x}\mu_{2,\tau-k}(x)\,\mathrm{d}x, \qquad \textit{if} \;\; \rho_1 = 0.$$

Note that when $\rho_1 = 0$, for the Johansen LR and Hotelling-Lawley statistics we get inequalities, rather than equalities. This is because of the singularity of $\ln(1-\lambda)$ and $\frac{\lambda}{1-\lambda}$ at $\lambda = 1$. However, for similar statistics in the modified setting of the next section the inequalities turn into equalities (as can be proven by combining Theorem 8 and Proposition 10).

For commonly used tests, one often takes $r_2 = N$ (i.e., $H_1 : \mathrm{rank}(\Pi) \le N$), in which case $F^{-1}(\rho_2) = 0$. We also remark that the integrals in Corollary 4 can be explicitly computed in many situations. For instance, the one appearing in the asymptotic of the Pillai-Bartlett statistic for $\rho_1 = 0$, $\rho_2 = 1$ is

$$\int_0^1 x\,\mu_{2,\tau-k}(x)\,\mathrm{d}x = \frac{2}{\tau+2-k}.$$

There are several applications of Theorem 3 and Corollary 4:

- They can be used for validation of the applicability of model (1) to a given dataset. Namely, if VAR(k) model with low rank matrices Γ_i and Π agrees with data, then irrespective of the true values of these parameters we expect to see the Wachter distribution in the histogram of λ_i, $1 \le i \le N$. In Section 5 we perform such validation on S&P100 and cryptocurrencies data sets for VAR(k) with $1 \le k \le 4$ and observe a remarkable match. (VAR(1) for S&P100 is also reported in (Bykhovskaya and Gorin, 2021, Figure 7).)

- They can be used as a screening device for preliminary conclusions about the rank of Π: If the rank is finite, then for any r_1 and r_2 we should be in the ε-neighborhood of the limits in Corollary 4.
- As explained in Onatski and Wang (2018) such results can be used to explain over-rejection in some of the widely used tests for the rank of Π.

In order to draw further economical and statistical conclusions and to develop precise statistical tests and their critical values, one needs to know the corrections to the first order asymptotic results of Theorem 3 and Corollary 4. In the next section we develop such corrections and introduce relevant modifications.

3. Cointegration test: Second order asymptotics

In the regime of N and T growing simultaneously and proportionally, the first order asymptotics of tests based on the squared sample canonical correlations is given in Corollary 4. To perform testing and be able to reject at a given significance level, we need to be more precise and find a centered limit, which would be a random variable rather than a constant. To do this, we need to impose additional conditions on Γ_i, D_t, and Φ in Eq. (1). Let us first describe the modified procedure and then state the asymtotic results.

3.1. Test. We restrict our attention to the case $D_t = 1$, i.e.,

$$\Delta X_t = \mu + \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Pi X_{t-k} + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{10}$$

The null hypothesis of no cointegration is $H_0:$ rank$(\Pi) = 0$ or $\Pi \equiv 0$. Yet, this is not a point hypothesis, as it does not specify Γ_i, $i = 1, \ldots, k-1$. We are going to work with a "proxy" point null approximation and specify the values of Γ_i, $i = 1, \ldots, k-1$. Thus, our model is going to be fully specified (up to a constant μ, which will disappear in the testing procedure).[2] Let us introduce our "proxy" null:

$$\widehat{H}_0 : \ \Pi = \Gamma_1 = \Gamma_2 = \cdots = \Gamma_{k-1} = 0. \tag{11}$$

In other words, under $\widehat{H}_0$ the data-generating process turns into

$$\Delta X_t = \mu + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{12}$$

where μ is an (unknown) N–dimensional vector.

The complement to H_0 is rank$(\Pi) > 0$. Yet, in order to design our test we use an alternative hypothesis

$$H(r): \quad \text{rank}(\Pi) \in [1, r].$$

[2]In Section 4 we present Monte Carlo simulations which indicate that (12) can be replaced by less restrictive (and, thus, closer to H_0) hypothesis, for which asymptotic results of Section 3.2 remain valid.

As in Bykhovskaya and Gorin (2021), our test is based on a modification of the Johansen LR test. The Johansen LR test for the original H_0 (i.e., $\Pi \equiv 0$) versus $H(r)$ is

$$\sum_{i=1}^{r} \ln(1-\lambda_i), \tag{13}$$

where $\lambda_1, \lambda_2, \dots$ are defined in Procedure 1.

Let us describe how our modified test proceeds.

Procedure 2. Step 1. De-trend the data and define

$$\tilde{X}_t = X_{t-1} - \frac{t-1}{T}(X_T - X_0). \tag{14}$$

Note that we do a time shift in line with the notation in Bykhovskaya and Gorin (2021).

Step 2. Define regressors and dependent variables: For any $a \in \mathbb{Z}$, set

$$a \mid T = a + kT, \quad \text{where } k \in \mathbb{Z} \text{ is such that } a + kT \in \{1, 2, \dots, T\}.$$

Define

$$\tilde{Z}_{0t} = \Delta X_{t|T} \equiv \Delta X_t, \quad \tilde{Z}_{kt} = \tilde{X}_{t-k+1|T}, \quad \tilde{Z}_{1t} = (\Delta X^*_{t-1|T}, \dots, \Delta X^*_{t-k+1|T}, 1)^*.$$

The main difference between $\tilde{Z}_{it}$ and Z_{it} from Procedure 1 is the usage of cyclic indices: values at $t = 0, -1, \dots$ are replaced by values at $t = T, T-1, \dots$.

Step 3. Calculate residuals from regressions $\tilde{Z}_{0t}$ on $\tilde{Z}_{1t}$ and $\tilde{Z}_{kt}$ on $\tilde{Z}_{1t}$:

$$\tilde{R}_{it} = \tilde{Z}_{it} - \left(\sum_{\tau=1}^{T} \tilde{Z}_{i\tau}\tilde{Z}^*_{1\tau}\right)\left(\sum_{\tau=1}^{T} \tilde{Z}_{1\tau}\tilde{Z}^*_{1\tau}\right)^{-1} \tilde{Z}_{1t}, \quad i = 0, k. \tag{15}$$

Step 4. Calculate the squared sample canonical correlations between $\tilde{R}_0$ and $\tilde{R}_1$, where $\tilde{R}_i$ is $N \times T$ matrix composed of columns $\tilde{R}_{it}$, $i = 0, k$. That is, define

$$\tilde{S}_{ij} = \sum_{t=1}^{T} \tilde{R}_{it}\tilde{R}^*_{jt}, \quad i, j = 0, 1, \qquad \text{and} \tag{16}$$

$$\tilde{\mathcal{C}} = \tilde{S}_{10}\tilde{S}_{00}^{-1}\tilde{S}_{01}\tilde{S}_{11}^{-1}. \tag{17}$$

Then calculate N eigenvalues $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_N$ of the matrix $\tilde{\mathcal{C}}$. The eigenvalues solve the equation

$$\det(\tilde{S}_{k0}\tilde{S}_{00}^{-1}\tilde{S}_{0k} - \tilde{\lambda}\tilde{S}_{kk}) = 0. \tag{18}$$

Step 5. Form the test statistic

$$LR_{N,T}(r) = \sum_{i=1}^{r} \ln(1-\tilde{\lambda}_i). \tag{19}$$

The subscript N, T in (19) indicates that we modify Johansen LR test to develop the large N, T asymptotics. This statistic after centering and rescaling will be compared with appropriate critical values to decide whether one can reject $\widehat{H}_0$ or not (see Theorem 9). Visually rejections correspond to the case when largest eigenvalues are separated from the rest (as in Figure 1).

One can also consider other functions of largest eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots$ such as Pillai-Barlett or Hotelling-Lawley statistics. The asymptotic behavior in those cases can be derived in the same way as we treat statistic (19) in Theorem 9.

An alternative way to write residuals $\tilde{R}_i$, $i = 0, k$ is via an orthogonal projector: Let $\mathcal{W}$ be a linear subspace of dimension $N(k-1)+1$ in T–dimensional vector space, spanned by vector $(1, 1, \dots, 1)$ and all rows of matrices $(\Delta X)(L_c^i)^*$, $1 \le i \le (k-1)$, where L_c is a cyclic version of the conventional lag operator. I.e., the cyclic lag operator L_c maps a vector $(x_1, x_2, \dots, x_T)$ to $(x_T, x_1, x_2, \dots, x_{T-1})$. Let $P_{\perp\mathcal{W}}$ denote the projector on orthogonal complement to $\mathcal{W}$. Then

$$\tilde{R}_0 = (\Delta X) P_{\perp\mathcal{W}}, \qquad \tilde{R}_k = \tilde{X}(L_c^{k-1})^* P_{\perp\mathcal{W}}. \tag{20}$$

3.2. Second order asymptotics. In this section we show that the eigenvalues $\tilde{\lambda}_i$, $i = 1, \dots, N$ are very close (up to $N^{-1+\epsilon}$ for arbitrary $\epsilon > 0$) to a known random matrix distribution. From this result we deduce our main theorem (Theorem 9), which gives the large N, T limit of the test statistic $LR_{N,T}(r)$ in Eq. (19). Before formally stating the results, let us define the relevant random matrix distributions.

3.2.1. *Definitions.*

Definition 5. The (real) Jacobi ensemble $\mathbf{J}(N; p, q)$ is a distribution on $N \times N$ real symmetric matrices $\mathcal{M}$ of density proportional to

$$\det(\mathcal{M})^{p-1} \det(I_N - \mathcal{M})^{q-1}\, d\mathcal{M}, \qquad 0 < \mathcal{M} < I_N, \tag{21}$$

with respect to the Lebesgue measure, where $p, q > 0$ are two parameters, I_N is the $N \times N$ identity matrix, and $0 < \mathcal{M} < I_N$ means that both $\mathcal{M}$ and $I_N - \mathcal{M}$ are positive definite.

The Jacobi ensemble is a generalization of the Beta distribution to the space of square matrices (when $N = 1$ we get the Beta distribution).

Definition 6. The Airy_1 point process is a random infinite sequence of reals

$$\mathfrak{a}_1 > \mathfrak{a}_2 > \mathfrak{a}_3 > \dots$$

which can be defined through the following proposition.

Proposition 7 (Forrester (1993),Tracy and Widom (1996))**.** *Let X_N be $N \times N$ matrix of i.i.d. $\mathcal{N}(0,2)$ Gaussian random variables and let $\mu_{1;N} \geq \mu_{2;N} \geq \dots \mu_{N;N}$ be eigenvalues of $\frac{1}{2}(X_N + X_N^*)$. Then in the sense of convergence of finite-dimensional distributions*

$$\lim_{N\to\infty} \left\{ N^{1/6} \left(\mu_{i;N} - 2\sqrt{N} \right) \right\}_{i=1}^{N} = \{\mathfrak{a}_i\}_{i=1}^{\infty}. \tag{22}$$

The marginals of the Airy$_1$ point process can be calculated via various methods, see, e.g., Forrester (2010) for more details.

3.2.2. *Theorems.*

Theorem 8. *Fix $C > 0$ and suppose that $T, N \to \infty$ in such a way that $\frac{T}{N} \in [k+1+C^{-1}, C]$. Under the hypothesis $\widehat{H}_0$ given by (11) for (1), one can couple (i.e. define on the same probability space) the eigenvalues $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \ldots \geq \tilde{\lambda}_N$ of the matrix $\tilde{S}_{k0}\tilde{S}_{00}^{-1}\tilde{S}_{0k}\tilde{S}_{kk}^{-1}$ and eigenvalues $x_1 \geq \cdots \geq x_N$ of the Jacobi ensemble $\mathbf{J}(N; \frac{N}{2}, \frac{T-(k+1)N}{2})$ in such a way that for each $\epsilon > 0$ we have*

$$\lim_{T,N\to\infty} \mathrm{Prob}\left(\max_{1\leq i\leq N} |\tilde{\lambda}_i - x_i| < \frac{1}{N^{1-\epsilon}} \right) = 1.$$

The proof of Theorem 8 relies on two steps. First, we modify our matrix $\tilde{\mathcal{C}}$ a bit, which leads to a surprising appearance of the Jacobi ensemble, as shown in Section 8. Second, in Section 9 we show that the distance between the original model and the modified one becomes small as $N \to \infty$.

Combining Theorem 8 with known asymptotic results for the Jacobi ensemble, which we recall in Proposition 10 in Section 7, we derive the asymptotics of (19) in the following theorem.

Theorem 9. *Fix $C > 0$ and suppose that $T, N \to \infty$ in such a way that $\frac{T}{N} \in [k+1+C^{-1}, C]$. Under the proxy $\widehat{H}_0$ given by (11) for (1), for each finite $r = 1, 2, \ldots$ we have convergence in distribution for the largest eigenvalues defined in Eq. (18):*

$$\frac{\sum_{i=1}^{r} \ln(1 - \tilde{\lambda}_i) - r \cdot c_1(N,T)}{N^{-2/3} c_2(N,T)} \xrightarrow[T,N\to\infty]{d} \sum_{i=1}^{r} \mathfrak{a}_i, \tag{23}$$

where

$$c_1(N,T) = \ln(1-\lambda_+), \qquad c_2(N,T) = -\frac{2^{2/3}\lambda_+^{2/3}}{(1-\lambda_+)^{1/3}(\lambda_+ - \lambda_-)^{1/3}} (\mathfrak{p}+\mathfrak{q})^{-2/3} < 0, \tag{24}$$

$$\mathfrak{p} = 2, \qquad \mathfrak{q} = \frac{T}{N} - k, \qquad \lambda_\pm = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2} \left[\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}} \right]^2. \tag{25}$$

Theorem 9 gives us the basis of cointegration testing in the large N, T setting. Treating $\widehat{H}_0$ as a proxy for H_0, we can use our asymptotic results to test high-dimensional VARs for the presence of cointegration. Note that although the asymptotic result (23) is shown under the proxy $\widehat{H}_0$, the Monte-Carlo simulations in Section 4 suggest that it extends well beyond $\widehat{H}_0$. We believe that the same asymptotic results and testing procedures continue to hold in many situations with non-zero Γ_i in Eq. (1). To be more precise, we expect that the conclusion of Theorem 9 is still true in the setting (1) with $\Pi = 0$ subject to certain technical conditions in the spirit of small rank restriction (7). The intuition is that addition of small rank matrices is negligible when compared to the rest of the process and, thus, does not change the asymptotics.

Formally, to perform testing one first needs to calculate the statistic $LR_{N,T}(r)$ following the Procedure 2. We recommend using small values of r, such as $r = 1, 2$, or 3. Then one needs to calculate $\frac{LR_{N,T}(r) - r \cdot c_1(N,T)}{N^{-2/3} c_2(N,T)}$, as in Theorem 9, and compare the result with quantiles of the sum of Airy$_1$, $\sum_{i=1}^{r} \mathfrak{a}_i$. If the rescaled statistic is larger than the α quantile, we reject the null of no cointegration at $(1 - \alpha)$ level. We report the quantiles for $r = 1$ in Table 1.

r \ α	0.9	0.95	0.975	0.99
1	0.44	0.97	1.45	2.01
2	-1.88	-1.09	-0.40	0.41
3	-5.91	-4.91	-4.03	-2.99

Table 1. Quantiles of $\sum_{i=1}^{r} \mathfrak{a}_i$ for $r = 1, 2, 3$ (based on 10^6 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman (2002)).

Theorem 9 means that under $\widehat{H}_0$ the largest eigenvalues $\tilde{\lambda}_i$ are close to λ_+, which is the right point of the support of the Wachter distribution in Eq. (4). The relevance of this theorem for cointegration testing stems from the fact that we expect some of the eigenvalues to be much larger than λ_+ when cointegrating relationships are present, cf. Figure 1.

Finally, let us emphasize that r is kept fixed as N and T grow in Theorem 9. The role of this choice and motivations for sticking to it are discussed in detail in (Bykhovskaya and Gorin, 2021, Section 3.2).

4. Monte Carlo simulations

4.1. Size. We refer to Bykhovskaya and Gorin (2021, Section 5) for the finite sample size and power performance of our test (for $k = 1$). Results for VAR(k) are similar and we do not show them in much detail here. For illustration purpose and to see the comparative statics, Table 2 reports the empirical size for $T = 522$, $N = 92$ (those numbers correspond to our empirical example in Section 5.1) for tests based on VAR(k), $k = 1, 2, 3, 4$ procedures.[3] We

[3]Depending on assumed order of autoregression, we have different number of regressors in Procedure 2.

VAR(1)	VAR(2)	VAR(3)	VAR(4)
5.81%	5.92%	6.12%	6.95%

Table 2. Empirical size under no cointegration hypothesis (5% nominal level) based on VAR(k) tests, $k = 1, 2, 3, 4$. Data generating process: $\Delta X_{it} = \varepsilon_{it}$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0, 1)$, $T = 522$, $N = 92$, $MC = 1,000,000$ replications.

can see that the numbers are close to the desired 5%, and for the same N and T lower order of VAR leads to slightly better results.

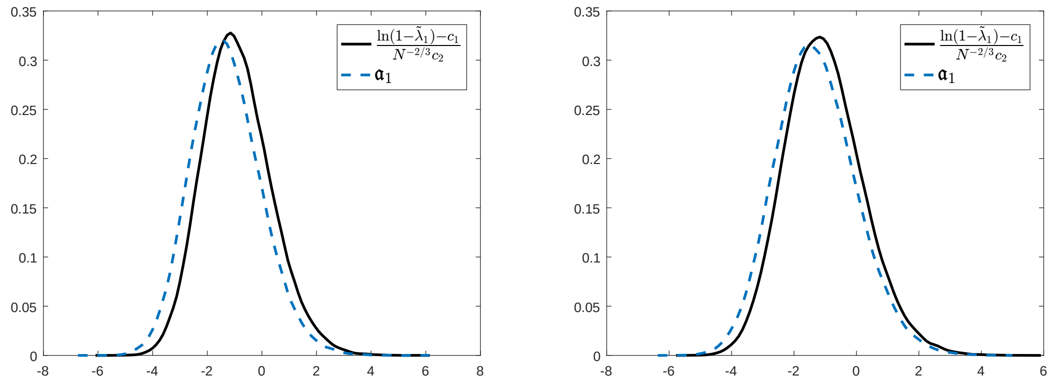
In the rest of this section we focus on the aspects of the test, which are specific to $k > 1$.

4.2. $\mathbf{H_0}$ vs. $\widehat{\mathbf{H}}_\mathbf{0}$. The first aspect is the introduction of the proxy null $\widehat{H}_0$. For N=100, T=500, $k = 2$ we simulate the data based on i.i.d. $\mathcal{N}(0, 1)$ errors ε_{it}, zero Π, and non-zero Γ_1 (i.e., this corresponds to H_0, but not $\widehat{H}_0$). We then compare the density of the test based on the largest eigenvalue $\tilde{\lambda}_1$ ($r = 1$ case of Theorem 9 with statistic $\frac{\ln(1-\tilde{\lambda}_1)-c_1(N,T)}{N^{-2/3}c_2(N,T)}$), with the density of the first coordinate of the Airy$_1$ point process, $\mathfrak{a}_1$. The aim of this simulation is to check whether Theorem 9 can hold under less restrictive H_0 instead of $\widehat{H}_0$. If the densities coincide, then it means that we can still use the asymptotics from Theorem 9 to test the null of no cointegration.

Let E_{ij} be a matrix with 1 at the cell (i, j) and 0s everywhere else and let $E_{\cdot j}$ be a matrix with 1s filling the entire column j and 0s everywhere else. In the first experiment we set $\Gamma_1 = 0.95E_{11}$, while in the second one we consider an asymmetric matrix $\Gamma_1 = 0.1E_{\cdot 1} + 0.95E_{12}$. Thus, the first case corresponds to Γ_1 of rank 1 and the second to rank 2. The results are illustrated in Figure 2. We interpret these figures as a strong argument towards the validity of an analogue of Theorem 9 well-beyond the $\widehat{H}_0$ setting.[4]

4.3. Order of VAR. It is essential for the experiments in the last paragraph that the data generating process is VAR(2) and the procedure we use also corresponds to VAR(2), i.e. $k = 2$ in the notations of Sections 2 and 3. As illustrated in Figure 3, using larger k would lead to similar results, while using incorrect VAR(1) procedure, when the data generating process is VAR(2), implies wrong centering and scaling. Moreover, one can spot in Figure 4 that underestimating order of VAR can be misinterpreted as a presence of cointegration (largest eigenvalue separated from the rest leading to the large value of the test statistic). Yet as we increase the order, the largest eigenvalue becomes inseparable from the rest, and no sign of false cointegration remains present. Thus, practitioners are encouraged to experiment with the order of VAR to make sure they are detecting cointegration and not a wrong model.

[4]The minor mismatches between densities as in Figure 2 should be expected even under $\widehat{H}_0$. Theorem 8 (after multiplication of the result by $N^{2/3}$, as in Eq. (23)) predicts errors of at least const $\cdot N^{-1/3}$ in the approximations under $\widehat{H}_0$.



(A) $\Gamma_1 = 0.95E_{11}$. (B) $\Gamma_1 = 0.1E_{\cdot 1} + 0.95E_{12}$.

FIGURE 2. Airy$_1$ and asymptotic distribution of the rescaled $\ln(1-\tilde{\lambda}_1)$ under H_0. Data generating process: $\Delta X_t = \Gamma_1 \Delta X_{t-1} + \varepsilon_t$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$, $MC = 100,000$ replications.

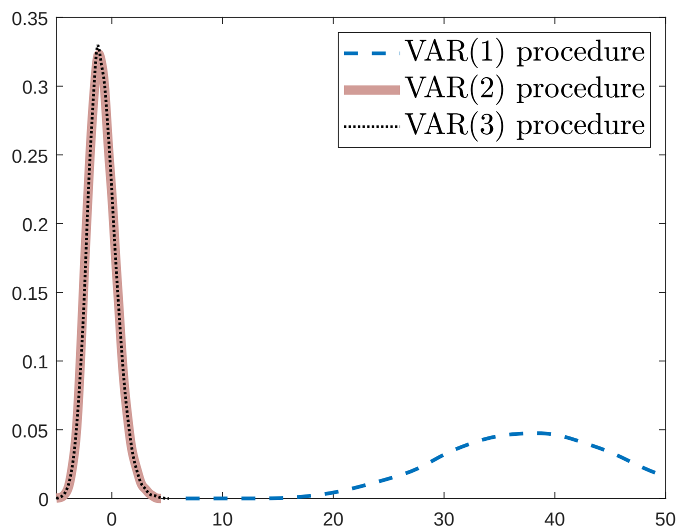
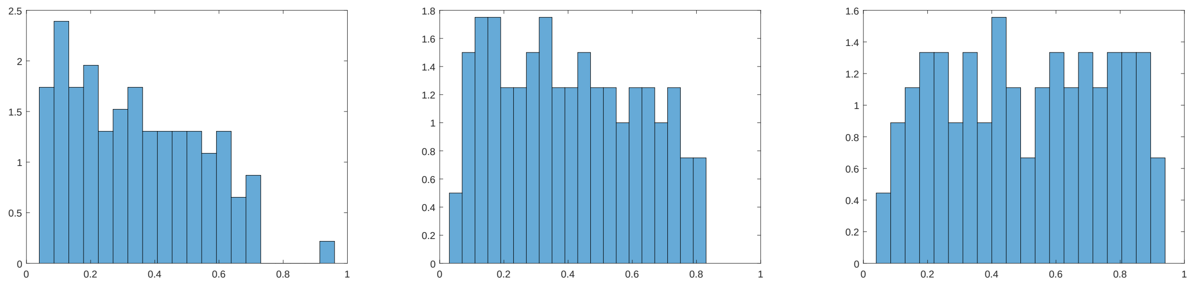


FIGURE 3. Density of rescaled $\ln(1-\tilde{\lambda}_1)$ obtained from various procedures. The correct procedure corresponds to VAR(2), $k = 2$. Data generating process: $\Delta X_t = 0.95E_{11}\Delta X_{t-1} + \varepsilon_t$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$, $MC = 10,000$ replications.

Related simulations for $\Gamma_1 = \theta E_{11}$ are also reported in (Bykhovskaya and Gorin, 2021, Section 7.3): For small θ, such as $\theta = 0.5$, the VAR(1) procedure still performs well, however as θ grows to 1 the performance quickly deteriorates ($\theta = 0.95$ in Figure 3). All of the above reinforces the importance of using VAR(k) rather than VAR(1) procedure.



(A) VAR(1) procedure. (B) VAR(2) procedure. (C) VAR(3) procedure.

FIGURE 4. Eigenvalues obtained from various procedures. The correct procedure corresponds to VAR(2), $k = 2$. Data generating process: $\Delta X_t = 0.95 E_{11} \Delta X_{t-1} + \varepsilon_t$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$.

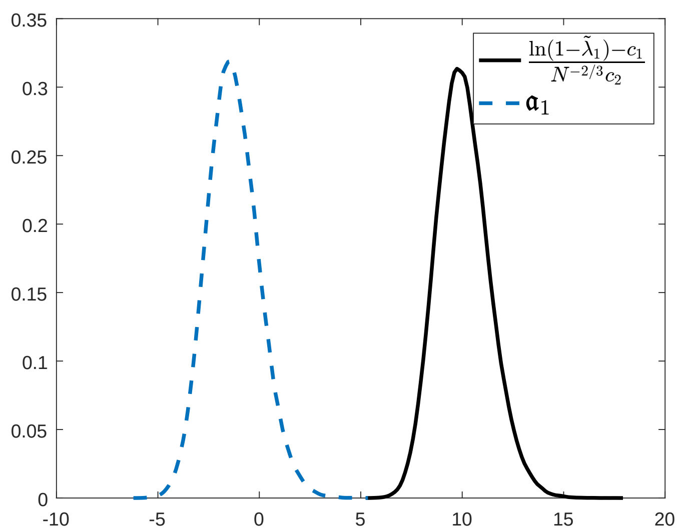


FIGURE 5. Airy$_1$ and asymptotic distribution of the rescaled $\ln(1-\tilde{\lambda}_1)$ under H_0 with Γ_1 of full rank. Data generating process: $\Delta X_{it} = 0.95 \Delta X_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$, $MC = 100{,}000$ replications.

4.4. Small ranks. To illustrate the importance of small ranks (e.g., (7) in Theorem 3) we also redo the same procedure for a matrix Γ_1 of full rank and set Γ_1 to be $0.95 I_N$, where I_N is an $N \times N$ identity matrix. The result is shown in Figure 5. While the shape and the scale (corresponding to $N^{2/3}$ rescaling in Theorem 9) of the distribution remains similar, the location changes. Thus, the small rank restriction of Eq. (7) is important not only in the context of Theorem 3, but also for the correct centering in possible generalizations of Theorem 9.

4.5. Power. Finally, we simulate the process based on $\Pi \neq 0$. We use $\Pi = -0.95 E_{11}$ and $\Gamma_1 = 0$. Figure 6 shows the results of the simulation. Two curves are separated, moreover,

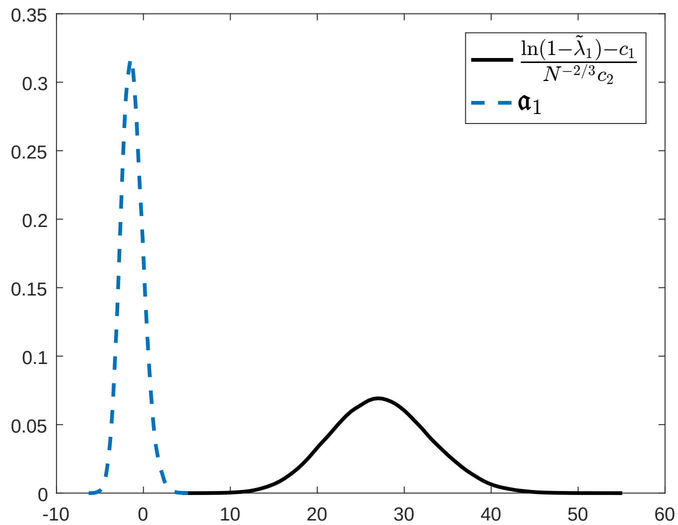


FIGURE 6. Airy$_1$ and asymptotic distribution of the rescaled $\ln(1-\tilde{\lambda}_1)$ under H_1. Data generating process: $\Delta X_t = -0.95 E_{11} X_{t-2} + \varepsilon_t$, $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$, $T = 500$, $N = 100$, $MC = 100,000$ replications.

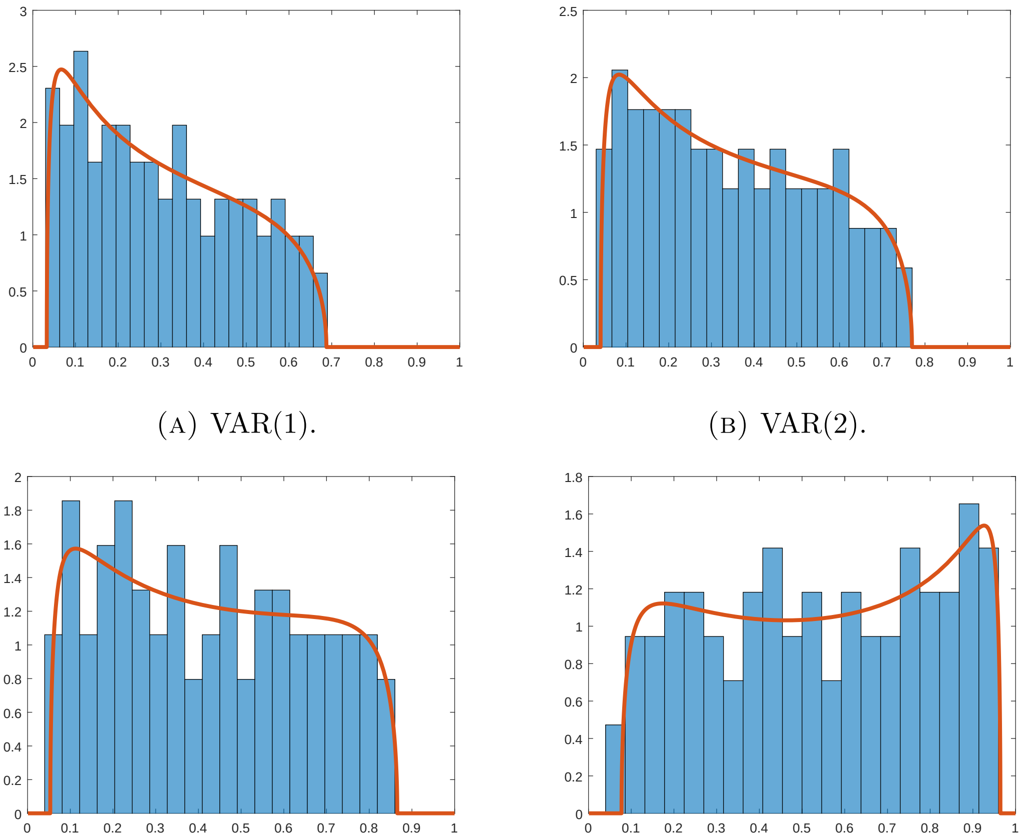
the black straight line (test distribution) is flatter than the blue dashed curve ($\mathfrak{a}_1$). First, the separation of the curves is in line with the usefulness of Theorem 9 in cointegration hypothesis testing, since the test statistic was designed to distinguish between $\Pi = 0$ and $\Pi \neq 0$. Second, distinct variances are due to the fact that (as we expect from comparison with results on spiked random matrices in the literature, see, e.g., Baik et al. (2005)) under the alternative ($\Pi \neq 0$) the test needs to be scaled differently: Instead of $N^{2/3}$ rescaling one should use $N^{1/2}$.

5. Empirical illustrations

5.1. S&P100. We illustrate our asymptotic theorems on the S&P100 data. We use logarithms of weekly prices of assets in S&P100 over ten years: 01.01.2010 – 01.01.2020, which gives us 522 observations across time. The more detailed description of the variables can be found in Bykhovskaya and Gorin (2021, Section 6).

For the S&P100 data set we use the Procedure 2 to calculate $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$. We do this for various choices of k. The case $k = 1$ (VAR(1)) corresponds to Bykhovskaya and Gorin (2021). The results are shown in Figure 7.

We see a striking match between the histograms and Wachter densities for all $k = 1, 2, 3, 4$, which is an indication that the setting of Theorem 3 is a proper modelling for the S&P data. We do not see any outliers in the largest eigenvalues, which would appear if there was a cointegration. Indeed, our test statistics based on Theorem 9 are -0.28, -0.71, -1.07, -3.84



(A) VAR(1). (B) VAR(2).

(C) VAR(3). (D) VAR(4).

FIGURE 7. Eigenvalues from S&P data (blue histogram) and Wachter distribution (orange line) based on various VAR(k) settings.

for $k = 1, 2, 3, 4$, respectively, while the 5% and 10% critical values are 0.97 and 0.44. Because the former numbers are smaller than the latter numbers, we do not reject the "no cointegration" hypothesis.

5.2. Cryptocurrencies. In this subsection we redo the calculations for cryptocurrencies instead of S&P stocks. We use data of Keilbar and Zhang (2021) (25 series from github repository). Logarithms of daily prices for two years (from Oct 5, 2017 to Oct 4, 2019) are shown in Figure 8. The results of our testing Procedure 2 to calculate $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ are shown in Figure 9.

Similarly to the S&P example in the previous subsection, we see a close match between the eigenvalues and the Wachter distribution.[5] Yet, there is a major difference between Figures 7 and 9: the latter has around 3 eigenvalues to the right of the support of the orange curve (Wachter distribution). This is an indication of the presence of approximately 3 cointegrating

[5]The Wachter distribution depends on the order of VAR, k, and on the ratio T/N. Thus, orange curves in Figures 7 and 9 have different shapes and supports.

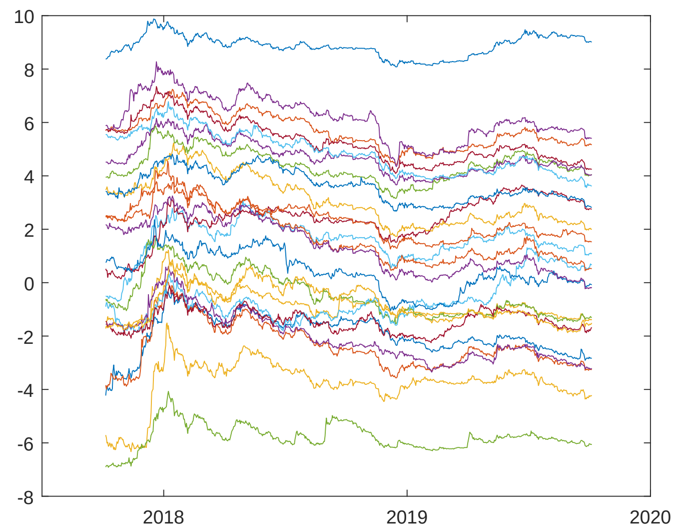


FIGURE 8. Time series of daily log prices for 25 cryptocurrencies.

relationships. This is reinforced by our test, which has p-value below 0.01 for all four choices of order of VAR(k) ($k = 1, 2, 3, 4$).

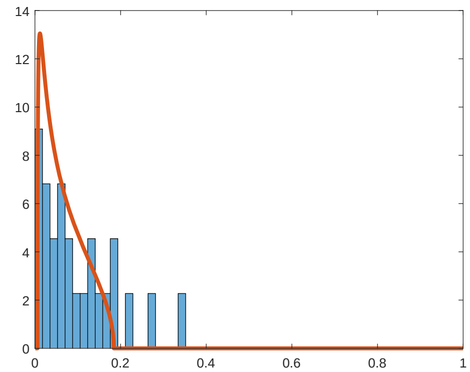
The difference in results for traditional stocks and cryptocurrencies can be explained by the fact that cryptocurrency market is still very inefficient and, thus, has numerous trading possibilities. The presence of cointegration can be one such inefficiency.

6. Conclusion

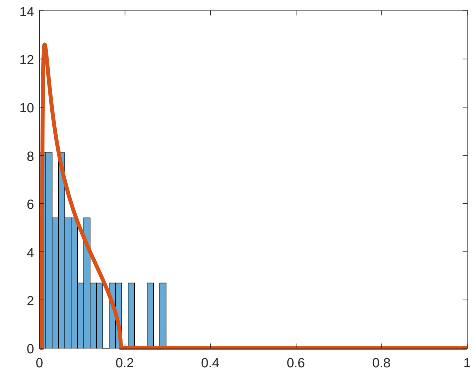
High-dimensional data becomes more and more widespread in economics and other sciences. Thus, the appropriate machinery is needed. We believe that the use of the random matrix theory is inevitable for the development of the area: as soon as dimensions are high, random matrices kick in. Along these lines, in our paper the central role is played by random matrix objects: Wachter distribution, Airy_1 point process, Jacobi ensemble.

The present paper focused on non-stationary high-dimensional VARs and presented the asymptotic limit of the Johansen LR test for cointegration as well as its modifications. Because the limit is non-random, the appropriate second order statistic was derived and the new test for the presence of cointegration was proposed. The new test builds upon the Johansen LR, while having some extra modifications. This new test is suitable for a vector autoregression of order k with an intercept.

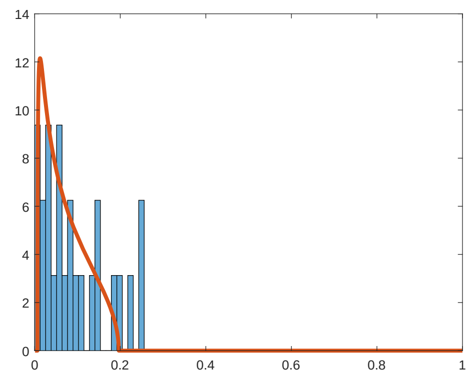
The main focus of the present paper is the null of no cointegration. The next essential step would be to be able to test whether the cointegration rank is r for $r > 0$, i.e. find the true rank of cointegration. Heuristics for finding the correct value of r can be already seen in our simulations and data sets, cf. Figures 1, 7, and 9: when there are no cointegrations, all



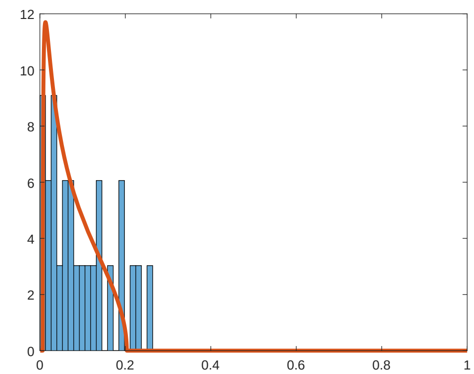
(A) VAR(1).



(B) VAR(2).



(C) VAR(3).



(D) VAR(4).

FIGURE 9. Eigenvalues from cryptocurrencies data (blue histogram) and Wachter distribution (orange line) based on various VAR(k) settings.

eigenvalues (squared canonical correlations) are to the left from the end-point of the support of the Wachter distribution. In contrast, we expect each cointegrating relationship to lead to an eigenvalue between the right end-point of the support of the Wachter distribution and 1. Figuring out the exact conditions under which this heuristics is correct forms an important problem for the future research.

7. Appendix. Proofs 1: Asymptotic of Jacobi ensemble

In this section we review the asymptotic results for the Jacobi ensemble $\mathbf{J}(N; p, q)$ introduced in the Definition 5 as $N \to \infty$.

We assume that as $N \to \infty$, also $p, q \to \infty$, in such a way that

$$\frac{p}{N} = \frac{1}{2}(\mathfrak{p} - 1), \quad \mathfrak{p} \geq 1, \qquad \frac{q}{N} = \frac{1}{2}(\mathfrak{q} - 1), \quad \mathfrak{q} \geq 1, \tag{26}$$

where $\mathfrak{p}$ and $\mathfrak{q}$ are two parameters, which stay bounded away from 1 and from ∞ as $N \to \infty$.[6] We further define the *equilibrium measure* $\mu_{\mathfrak{p},\mathfrak{q}}$ of the Jacobi ensemble through:

$$\mu_{\mathfrak{p},\mathfrak{q}}(x)\,dx = \frac{\mathfrak{p}+\mathfrak{q}}{2\pi} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+-x)}}{x(1-x)} \mathbf{1}_{[\lambda_-,\lambda_+]}\,\mathrm{d}x, \tag{27}$$

where the support $[\lambda_-, \lambda_+]$ of the measure is defined via

$$\lambda_\pm = \frac{1}{(\mathfrak{p}+\mathfrak{q})^2} \left(\sqrt{\mathfrak{p}(\mathfrak{p}+\mathfrak{q}-1)} \pm \sqrt{\mathfrak{q}}\right)^2. \tag{28}$$

One can check that $0 < \lambda_- < \lambda_+ < 1$ for every $\mathfrak{p}, \mathfrak{q} > 1$. Further, define

$$c_\pm = \frac{(\mathfrak{p}+\mathfrak{q})}{2} \frac{\sqrt{\lambda_+ - \lambda_-}}{\lambda_\pm(1-\lambda_\pm)}, \tag{29}$$

and note that

$$\mu_{\mathfrak{p},\mathfrak{q}}(x-\lambda_\pm) \approx \frac{c_\pm}{\pi}\sqrt{|x-\lambda_\pm|}, \text{ as } x \to \lambda_\pm \quad \text{inside} \quad [\lambda_-, \lambda_+],$$

where the normalization $\frac{1}{\pi}\sqrt{|x-\lambda_\pm|}$ was chosen to match the behavior of the Wigner semicircle law $\frac{1}{2\pi}\sqrt{4-x^2}$ near edges ± 2.

Proposition 10 (See Johnstone (2008), Forrester (2010), and Han et al. (2016))**.** *Suppose that $N, p, q \to \infty$ in such a way that $\mathfrak{p} \geq 1$ and $\mathfrak{q} \geq 1$ in* (26) *stay bounded. For the second conclusions we additionally assume that $\mathfrak{q}$ is bounded away from 1 and for the third conclusion we additionally require $\mathfrak{p}$ to be bounded away from 1. Let $x_1 \geq x_2 \geq \cdots \geq x_N$ be N random eigenvalues of Jacobi ensemble $\mathbf{J}(N; p, q)$. Then*

(1) $\lim_{N\to\infty} \left| \frac{1}{N} \sum_{i=1}^N \delta_{x_i} - \mu_{\mathfrak{p},\mathfrak{q}} \right| = 0$, *weakly in probability.*

This means that for any continuous function $f(x)$ we have convergence in probability:

$$\lim_{N\to\infty} \left| \frac{1}{N} \sum_{i=1}^N f(x_i) - \int_0^1 f(x)\mu_{\mathfrak{p},\mathfrak{q}}(x)\mathrm{d}x \right| = 0. \tag{30}$$

(2) *For $\{\mathfrak{a}_i\}_{i=1}^\infty$ as in Proposition 7, we have convergence in finite-dimensional distributions for the largest eigenvalues:*

$$\lim_{N\to\infty} \left\{ N^{2/3} c_+^{2/3} (x_i - \lambda_+) \right\}_{i=1}^\infty \to \{\mathfrak{a}_i\}_{i=1}^\infty. \tag{31}$$

In particular, $N^{2/3} c_+^{2/3} (x_1 - \lambda_+)$ converges to the Tracy-Widom distribution F_1.

[6]Johnstone (2008, Theorem 1) suggests to use $\frac{p-1}{N}$ and $\frac{q-1}{N}$ instead of $\frac{p}{N}$ and $\frac{q}{N}$, respectively, in order to improve the speed of convergence. However, we found in Bykhovskaya and Gorin (2021) that for the tests in VAR(1) case the usefulness of this correction depends on the exact value of the ratio T/N and we are not going to pursue this direction here.

(3) *We also have convergence in distribution for the smallest eigenvalues*[7]

$$\lim_{N\to\infty}\left\{N^{2/3}c_-^{2/3}\left(\lambda_- - x_{N+1-i}\right)\right\}_{i=1}^{\infty} \to \{\mathfrak{a}_i\}_{i=1}^{\infty}. \tag{32}$$

8. Appendix. Proofs 2: A new model for the Jacobi ensemble

The Jacobi ensemble appearing in Theorem 8 originates in the following computation of exact distribution. In addition to real symmetric matrices ($\beta = 1$ in the usual random matrix notations) it also covers the case of complex Hermitian matrices ($\beta = 2$).

Theorem 11. *Fix $k = 1, 2, \dots$ and assume $\mathcal{T} \geq (k+1)N$. Let $\mathcal{V}$ be an N–dimensional subspace in the $\mathcal{T}$–dimensional space and let O be uniformly random orthogonal $\mathcal{T}\times\mathcal{T}$ matrix with determinant 1 if $\beta = 1$ (O is a uniformly random unitary matrix if $\beta = 2$). Let P be an orthogonal projector on the space orthogonal to $O\mathcal{V}$, $O^2\mathcal{V}, \dots, O^{k-1}\mathcal{V}$. Let P_1 be a projector on the subspace $P\mathcal{V}$ and P_2 be a projector on the subspace $PO^{k-1}(I_{\mathcal{T}}+O)^{-1}\mathcal{V}$. Then non-zero eigenvalues of $P_1P_2P_1$ coincide with those of the Jacobi ensemble of $N\times N$ real symmetric if $\beta = 1$ (complex Hermitian if $\beta = 2$) matrices of density proportional to*

$$\det(\mathcal{M})^{\frac{\beta}{2}N+\beta-2}\det(I_N-\mathcal{M})^{\frac{\beta}{2}(\mathcal{T}-(k+1)N+1)-1}\,d\mathcal{M},\quad 0\leq\mathcal{M}\leq I_N,\qquad \beta=1,2. \tag{33}$$

Remark 12. *We can replace $PO^{k-1}(I_{\mathcal{T}}+O)^{-1}$ in the definition of P_2 with $PO^k(I_{\mathcal{T}}+O)^{-1}$. Indeed, if $k>1$, then $PO^k(I_{\mathcal{T}}+O)^{-1}\mathcal{V}+PO^{k-1}(I_{\mathcal{T}}+O)^{-1}\mathcal{V} = PO^{k-1}\mathcal{V} = 0$. If $k=1$, then P disappears (one can say that it becomes an identical operator) and the random operators $(I_{\mathcal{T}}+O)^{-1}$ and $O(I_{\mathcal{T}}+O)^{-1} = (I_{\mathcal{T}}+O^{-1})^{-1}$ has the same law, since the uniform measure on the orthogonal group is invariant under the inversion $O\mapsto O^{-1}$.*

By a similar argument applied inductively we can replace $PO^{k-1}(I_{\mathcal{T}}+O)^{-1}$ with $PO(I_{\mathcal{T}}+O)^{-1}$. Note, however, that for $k>1$ we can not replace it with $P(I_{\mathcal{T}}+O)^{-1}$.

The $k=1$ case of Theorem 11 is established in (Bykhovskaya and Gorin, 2021, Theorem 6 in Appendix). The proof of Theorem 11 uses the following three auxiliary ingredients.

Lemma 13 (Block matrix inversion formula)**.** *For matrices* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, *we have:*

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{Q} & -\mathbf{Q}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{Q} & \mathbf{D}^{-1}+\mathbf{D}^{-1}\mathbf{C}\mathbf{Q}\mathbf{B}\mathbf{D}^{-1}\end{pmatrix},\qquad \mathbf{Q} = (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}, \tag{34}$$

Proof. Direct computation. □

Lemma 14 (Cayley transform)**.** *Suppose that all eigenvalues of $N\times N$ matrix O are different from -1. Then O is an orthogonal matrix with determinant 1, if and only if the matrix $\mathcal{R}$ defined through*

$$\mathcal{R} = (I_N-O)(I_N+O)^{-1} = \frac{I_N-O}{I_N+O},\qquad \textit{so that}\qquad O = \frac{I_N-\mathcal{R}}{I_N+\mathcal{R}} \tag{35}$$

[7]The limiting processes $\{\mathfrak{a}_i\}_{i=1}^{\infty}$ arising for the largest and smallest eigenvalues are independent.

is skew-symmetric, i.e. it satisfies $\mathcal{R}^* = -\mathcal{R}$.

Proof. The formulas (35) imply that $O^* = O^{-1}$ if and only if $\mathcal{R}^* = -\mathcal{R}$. On the other hand, for a skew-symmetric $\mathcal{R}$, we have $\det(I_N + \mathcal{R}) = \det(I_N + \mathcal{R}^*) = \det(I_N - \mathcal{R})$. Hence, $\det(O) = \det\left(\frac{I_N - \mathcal{R}}{I_N + \mathcal{R}}\right) = 1$. □

Lemma 15. *Choose two positive integers* M, N *and set* $\mathcal{T} = M + N$. *Let* O *be a uniformly random* $\mathcal{T} \times \mathcal{T}$ *orthogonal matrix with determinant* 1. *Write* O *in the block form according to* $\mathcal{T} = M + N$ *splitting:*

$$O = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

Then $\tilde{O} := A - B(I_N + D)^{-1}C$ *is a* $M \times M$ *orthogonal matrix of determinant* 1 *uniformly distributed among all such matrices. In addition, the random matrices* $\tilde{O}$ *and* $B(I_N + D)^{-1}$ *are independent.*

Remark 16. *The law of* $\widehat{W} := B(I_N + D)^{-1}$ *is explicit. The computation* (42) *below implies that the density of* $\widehat{W}$ *is proportional to*

$$\det\big(I_N + \widehat{W}^*\widehat{W}\big)^{1/2-M-N/2}\, d\widehat{W}.$$

Remark 17. *The computation of the law of* $\tilde{O}$ *is mentioned in Olshanski (2003) and Neretin (2002) with its roots going back to Hua (1963). However, we could not locate the statements concerning also* $B(I_N + D)^{-1}$ *in the literature.*

Proof of Lemma 15. First, note that the distribution of eigenvalues of D is absolutely continuous and, hence, $I_N + D$ is almost surely invertible and the matrix $\tilde{O}$ is well-defined. Our next task is to show that $\tilde{O}$ is an orthogonal matrix with determinant 1. We use Cayley transform for that. Combining (35) with (34) we have

$$\begin{aligned} \mathcal{R} &= \frac{I_{\mathcal{T}} - O}{I_{\mathcal{T}} + O} = \frac{2}{I_{\mathcal{T}} + O} - I_{\mathcal{T}} \\ &= \begin{pmatrix} 2Q - I_M & -2QB(I_N + D)^{-1} \\ -2(I_N + D)^{-1}CQ & 2(I_N + D)^{-1} + 2(I_N + D)^{-1}CQB(I_N + D)^{-1} - I_N \end{pmatrix}, \\ & \qquad Q = (I_M + A - B(I_N + D)^{-1}C)^{-1}. \end{aligned} \tag{36}$$

Since $\mathcal{R}$ is skew-symmetric, so is its top–left $M \times M$ corner $\tilde{\mathcal{R}} = 2Q - I_M$. We claim that $\tilde{O}$ is the Cayley transform of $\tilde{\mathcal{R}}$, which would imply that $\tilde{O}$ is orthogonal of determinant 1. Indeed,

$$\frac{I_M - \tilde{\mathcal{R}}}{I_M + \tilde{\mathcal{R}}} = \frac{2I_M - 2Q}{2Q} = Q^{-1} - I_M = A - B(I_N + D)^{-1}C = \tilde{O}. \tag{37}$$

It remains to compute the distributions of $\tilde{O}$ and $B(I_N+D)^{-1}$ and show their independence. In terms of $\mathcal{R}$ the distribution of O (as a uniformly random orthogonal matrix of determinant

1) is given by the density proportional to

$$\det(I_{\mathcal{T}} - \mathcal{R}^2)^{-\frac{1}{2}\mathcal{T}+\frac{1}{2}} d\mathcal{R} = \det(I_{\mathcal{T}} - \mathcal{R})^{1-\mathcal{T}} d\mathcal{R} = \det(I_{\mathcal{T}} + \mathcal{R})^{1-\mathcal{T}} d\mathcal{R}, \tag{38}$$

see, e.g., Forrester (2010, (2.55)) and notice that $\det(I_{\mathcal{T}} - \mathcal{R}) = \det(I_T + \mathcal{R})$ for the two equalities. We rewrite the block form (36) of $\mathcal{R}$ as

$$\mathcal{R} = \begin{pmatrix} \tilde{\mathcal{R}} & -W \\ W^* & \mathcal{R}_2 \end{pmatrix},$$

where $\mathcal{R}_2$ is $N \times N$ skew-symmetric and W is an arbitrary $M \times N$ matrix. We further introduce the notation $\widehat{W} := (I_M + \tilde{\mathcal{R}})^{-1} W$. Recalling that $\tilde{\mathcal{R}} = 2Q - I_M$, we transform

$$\widehat{W} = 2(I_M + \tilde{\mathcal{R}})^{-1} QB(I_N + D)^{-1} = B(I_N + D)^{-1}. \tag{39}$$

We also define

$$\widehat{\mathcal{R}}_2 := \big(I_N + \widehat{W}^*\widehat{W}\big)^{-1/2}\big(-\mathcal{R}_2 + \widehat{W}^*\tilde{\mathcal{R}}\widehat{W}\big)\big(I_N + \widehat{W}^*\widehat{W}\big)^{-1/2}.$$

Note that $(\tilde{R}, \widehat{W}, \widehat{R}_2)$ is an alternative parameterization of $\mathcal{R}$, in which $\widehat{W}$ is an arbitrary $M \times N$ matrix and $\widehat{R}_2$ is an arbitrary $N \times N$ skew-symmetric matrix. Using the formula for the determinant of a block matrix

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det \mathbf{A} \cdot \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}), \tag{40}$$

we rewrite (38) as

$$\begin{aligned} \det(I_{\mathcal{T}} - \mathcal{R})^{1-\mathcal{T}}\, d\tilde{\mathcal{R}}\, dW\, d\mathcal{R}_2 &= \det \begin{pmatrix} I_M - \tilde{\mathcal{R}} & W \\ -W^* & I_N - \mathcal{R}_2 \end{pmatrix}^{1-\mathcal{T}} d\tilde{\mathcal{R}}\, dW\, d\mathcal{R}_2 \\ &= \det(I_M - \tilde{\mathcal{R}})^{1-\mathcal{T}} \det(I_N - \mathcal{R}_2 + W^*(I_M - \tilde{\mathcal{R}})^{-1} W)^{1-\mathcal{T}}\, d\tilde{\mathcal{R}}\, dW\, d\mathcal{R}_2 \\ &= \det\big(I_M - \tilde{\mathcal{R}}\big)^{1-\mathcal{T}} \det\big(I_N - \mathcal{R}_2 + \widehat{W}^*(I_M + \tilde{\mathcal{R}})\widehat{W}\big)^{1-\mathcal{T}}\, d\tilde{\mathcal{R}}\, dW\, d\mathcal{R}_2 \end{aligned} \tag{41}$$

Further, notice that

$$\big(I_N + \widehat{W}^*\widehat{W}\big)^{1/2}\big(I_N + \widehat{\mathcal{R}}_2\big)\big(I_N + \widehat{W}^*\widehat{W}\big)^{1/2} = I_N - \mathcal{R}_2 + \widehat{W}^*(I_M + \tilde{\mathcal{R}})\widehat{W}.$$

Hence, the last line of (41) is transformed into

$$\begin{aligned} &\det\big(I_M - \tilde{\mathcal{R}}\big)^{1-\mathcal{T}} \det\big(I_N + \widehat{W}^*\widehat{W}\big)^{1-\mathcal{T}} \det\big(I_N + \widehat{\mathcal{R}}_2\big)^{1-\mathcal{T}}\, d\tilde{\mathcal{R}}\, dW\, d\mathcal{R}_2 \\ &= \det\big(I_M + \tilde{\mathcal{R}}\big)^{1-M} \det\big(I_N + \widehat{W}^*\widehat{W}\big)^{1/2-M-N/2} \det\big(I_N + \widehat{\mathcal{R}}_2\big)^{1-M-N}\, d\tilde{\mathcal{R}}\, d\widehat{W}\, d\widehat{\mathcal{R}}_2, \end{aligned} \tag{42}$$

where in the last line we use $\mathcal{T} = M + N$ and change variables $dW \mapsto d\widehat{W}$ and $d\mathcal{R}_2 \mapsto d\widehat{\mathcal{R}}_2$ using the general Jacobian computations:

- The map $Z \mapsto QZ$ on $n \times m$ matrices has the Jacobian

$$\left|\frac{\partial(QZ)}{\partial Z}\right| = |\det Q|^m, \tag{43}$$

- The map $Z \mapsto QZQ^*$ from the space of $n \times n$ skew-symmetric matrices to itself has the Jacobian

$$\left|\frac{\partial(QZQ^*)}{\partial Z}\right| = |\det Q|^{n-1}. \tag{44}$$

The first identity (43) follows from the observation that each column of Z is transformed by linear map Q and there are m such columns. The second is similar and we refer to Forrester (2010, (1.35)) for details.

The key important feature of the last line of (42) is that it has a product form, which implies the joint independence of $\tilde{\mathcal{R}}$, $\widehat{W}$, and $\widehat{\mathcal{R}}_2$. Hence, the density of $\tilde{\mathcal{R}}$ is proportional to $\det\big(1+\tilde{\mathcal{R}}\big)^{1-M} d\tilde{\mathcal{R}}$. Comparing with (38) and noting that the dimension changed from $\mathcal{T}$ to M, we conclude that $\tilde{O}$ is a uniformly random $M \times M$ orthogonal matrix of determinant 1.

Next, recalling that $\tilde{O}$ is a deterministic function of $\tilde{\mathcal{R}}$ by (37), we conclude that $\tilde{O}$ is independent with $\widehat{W}$, which is precisely $B(I_N + D)^{-1}$ by (39). □

Proof of Theorem 11. We only give a proof for the real case $\beta = 1$; the complex case can be proven by the same argument. The proof is induction in k with base case $k = 1$ being (Bykhovskaya and Gorin, 2021, Theorem 6 in Appendix) and the induction step being based on Lemma 15.

Step 1. We first note that the particular choice of *deterministic* space $\mathcal{V}$ in the statement of the theorem is not important: any other deterministic choice of $\mathcal{V}$ can be achieved by a change of basis of the $\mathcal{T}$–dimensional space, which keeps the probability distribution of O and, hence, entire construction invariant. In particular, the probability distribution of $P_1P_2P_1$ is unchanged. However, we need to be more careful, if we would like to make $\mathcal{V}$ random, as correlations with O might cause issues.

Step 2. Take any $r \in \mathbb{Z}$. We claim that replacement of $\mathcal{V}$ with $O^r\mathcal{V}$ everywhere in the statement of Theorem 11 does not change the eigenvalues of $P_1P_2P_1$. Indeed, the only important feature of O^r here is that it is an orthogonal operator commuting with O. Hence, the change $\mathcal{V} \mapsto O^r\mathcal{V}$ leads to the image of the projector P being multiplied by O^r; in more details, the transformation takes the form $P \mapsto O^rPO^{-r}$. Further, $P\mathcal{V}$ gets transformed to $O^rP\mathcal{V}$ and P_1 undergoes a similar transformation: $P_1 \mapsto O^rP_1O^{-r}$. The same is true for P_2: it undergoes the transformation $P_2 \mapsto O^rP_2O^{-r}$. We conclude that the product $P_1P_2P_1$ is transformed into $O^rP_1P_2P_1O^{-r}$. Since conjugations do not change eigenvalues, we are done.

The arguments of Steps 1 and 2 might give a feeling that we can actually replace $\mathcal{V}$ by any random space. However, this is not the case. Repeating the same arguments, we see that replacement $\mathcal{V} \mapsto A\mathcal{V}$ leads to the same eigenvalues of the projector $P_1P_2P_1$ as if we replaced $O \mapsto A^*OA$. In both Steps 1 and 2 O had the same distribution as A^*OA, hence, the eigenvalues were unchanged. But in general, if A is correlated with O in a non-trivial way, then the distribution might change.[8]

Step 3. We now transform the statement of Theorem 11 by replacing $\mathcal{V}$ with $O^{1-k}\mathcal{V}$ and further replacing O by O^{-1} everywhere. Since the uniform (Haar) measure on the orthogonal matrices is invariant under inversion, the law of eigenvalues of $P_1P_2P_1$ is unchanged and the ingredients of Theorem 11 are now as follows:

- O is a uniformly random $\mathcal{T} \times \mathcal{T}$ orthogonal matrix with determinant 1 and $\mathcal{V}$ is an arbitrary (deterministic) N–dimensional subspace of the $\mathcal{T}$–dimensional space, whose choice is irrelevant for the statement.
- P is the projector on the orthogonal complement of $O^{k-2}\mathcal{V}$, $O^{k-3}\mathcal{V}$, $\dots$, $O\mathcal{V}$, $\mathcal{V}$.
- P_1 is the projector on the subspace $PO^{k-1}\mathcal{V}$ and P_2 is the projector on the subspace $PO(I_T+O)^{-1}\mathcal{V}$. (The latter can be replaced by $P(I_T+O)^{-1}\mathcal{V}$ without changing the outcome. Indeed, for that we need start from $PO^k(I_T+O)^{-1}$ instead of $PO^{k-1}(I_T+O)^{-1}$, which is possible by Remark 12).
- The claim is that the eigenvalues of $P_1P_2P_1$ are distributed as (33).

We are going to prove this last statement by induction in k. For that we choose $\mathcal{V}$ to be the span of the last N coordinate vectors, split $\mathcal{T} = M + N$ with $M = \mathcal{T} - N$ and project everything on the first M coordinate vectors (which are orthogonal complement to $\mathcal{V}$). We rely on Lemma 15 and use A, B, C, D and $\tilde{O}$ notation from that lemma.

Step 4. We claim that the subspace in M–dimensional space spanned by the first M coordinates of $O^{k-2}\mathcal{V}$, $O^{k-3}\mathcal{V}$, $\dots$, $O\mathcal{V}$ (since $\mathcal{V}$ has zero projection on the first M coordinates, we do not need it here) is the same as the subspace spanned by $\tilde{O}^{k-3}\langle B\rangle$, $\tilde{O}^{k-4}\langle B\rangle$, $\dots$, $\langle B\rangle$, where $\langle B\rangle$ is N–dimensional space spanned by columns of the $M \times N$ matrix B.

Indeed, the first M coordinates of $O\mathcal{V}$ are $\langle B\rangle$ by definition of the block structure in Lemma 15. Further, to go from powers of O to powers of $\tilde{O}$ we make the following observation: take a vector w in $\mathcal{T}$–dimensional space and write it as $w = \binom{w_2}{w_1}$, where w_1 is N-dimensional vector (one can think of w_1 being in $\mathcal{V}$) and w_2 is M–dimensional vector (one can think of w_2 being in the orthogonal complement of $\mathcal{V}$) and write

$$O\begin{pmatrix} w_2 \\ w_1 \end{pmatrix} = \begin{pmatrix} u_2 \\ u_1 \end{pmatrix}.$$

Then the M–dimensional vector u_2 takes the form

$$u_2 = Aw_2 + Bw_1 = \tilde{O}w_2 + B((I_N + D)^{-1}Cw_2 + w_1).$$

[8]For instance, if $\mathcal{V}$ is spanned by eigenvectors of O, then the spaces $O\mathcal{V}$, $O^2\mathcal{V}$,…,$O^{k-1}\mathcal{V}$ all coincide, which is a very different behavior from the case of deterministic $\mathcal{V}$.

Since we only care about the linear span of columns and $\langle B\rangle$ already belongs to the desired linear span, the last term can be ignored and we arrive at $\tilde{O}w_2$, which then implies the claim.

Step 5. Next, consider the projection of $O^{k-1}\mathcal{V}$ on the orthogonal complement to $O^{k-2}\mathcal{V}$, $O^{k-3}\mathcal{V}$, …, $O\mathcal{V}$, $\mathcal{V}$. This is the same as the the projection of the first M coordinates of $O^{k-1}\mathcal{V}$ on the orthogonal complement (in M–dimensional space) to first M coordinates of $O^{k-2}\mathcal{V}$, $O^{k-3}\mathcal{V}$, …, $O\mathcal{V}$. Hence, combining with the argument of Step 4, this is the same as the projection of $\tilde{O}^{k-2}\langle B\rangle$ on the orthogonal complement of $\tilde{O}^{k-3}\langle B\rangle$, $\tilde{O}^{k-4}\langle B\rangle$, …, $\langle B\rangle$. It is convenient to note that $\langle B\rangle = \langle B(I_N+D)^{-1}\rangle$.

Step 6. Finally, consider the projection of $(I_T+O)^{-1}O\mathcal{V}$ on the orthogonal complement of $O^{k-2}\mathcal{V}$, $O^{k-3}\mathcal{V}$, …, $O\mathcal{V}$, $\mathcal{V}$. By Steps 4 and 5 this is the same as the projection of the first M coordinates of $(I_T+O)^{-1}O\mathcal{V}$ on the orthogonal complement of $\tilde{O}^{k-3}\langle B(I_N+D)^{-1}\rangle$, $\tilde{O}^{k-4}\langle B(I_N+D)^{-1}\rangle$, …, $\langle B(I_N+D)^{-1}\rangle$. Representing $(I_T+O)^{-1}O\mathcal{V}$ in the block form, the first M coordinates of $(I_T+O)^{-1}O\mathcal{V}$ are the span of the columns of the sum of the top–left corner of $(I_T+O)^{-1}$ multiplied by B plus the top-right corner of $(I_T+O)^{-1}$ multiplied by D. Using Lemma 13, we get the span of the columns of

$$(I_M+A-B(I_N+D)^{-1}C)^{-1}B-(I_M+A-B(I_N+D)^{-1}C)^{-1}B(I_N+D)^{-1}D \\ = (I_M+\tilde{O})^{-1}B(I_N+D)^{-1},$$

which is the same as $(I_M+\tilde{O})^{-1}\langle B(I_N+D)^{-1}\rangle$.

Step 7. Combining the results of Steps 6 and 7 with Lemma 15, we identify the eigenvalues of $P_1P_2P_1$ with the eigenvalues of $\tilde{P}_1\tilde{P}_2\tilde{P}_1$ obtained by the following procedure:

- $\tilde{O}$ is a uniformly random $M\times M$ orthogonal matrix with determinant 1, where $M=\mathcal{T}-N$.
- $\tilde{P}$ is the projector on orthogonal complement of $\tilde{O}^{k-3}\langle B(I_N+D)^{-1}\rangle$, $\tilde{O}^{k-4}\langle B(I_N+D)^{-1}\rangle$, …, $\langle B(I_N+D)^{-1}\rangle$.
- $\tilde{P}_1$ is the projector on the subspace $\tilde{P}\tilde{O}^{k-2}\langle B(I_N+D)^{-1}\rangle$ and $\tilde{P}_2$ is the projector on the subspace $\tilde{P}(I_M+\tilde{O})^{-1}\langle B(I_N+D)^{-1}\rangle$.

Since $\langle B(I_N+D)^{-1}\rangle$ is independent from $\tilde{O}$ by Lemma 15, this is the same form as the one at the end of Step 3, but with k decreased by 1, $\mathcal{T}$ decreased by N, and $\mathcal{V}$ replaced by $\langle B(I_N+D)^{-1}\rangle$. Decreasing k by 1 and $\mathcal{T}$ by N leaves the formula (33) unchanged, hence, we can invoke the induction assumption, thus, finishing the proof. □

9. Appendix. Proofs 3: A perturbation of the Jacobi ensemble.

In this section we use Theorem 11 to prove Theorem 8.

Recall the cyclic shift[9] operator L_c acting in T–dimensional space. Let V be the $(T-1)$–dimensional space orthogonal to the vector $(1, 1, \dots, 1)$, i.e., $V = \{(x_1, \dots, x_T) \mid x_1 + \dots + x_T = 0\}$. Note that V is an invariant space for L_c and let L_V denote the restriction of L_c on the subspace V.

Take a uniformly-random orthogonal (or unitary if $\beta = 2$) operator $\tilde{O}$ acting in $(T-1)$–dimensional space V and define an operator $\tilde{L}$ acting in V:

$$\tilde{L} = -\tilde{O} L_V \tilde{O}^*.$$

Proposition 18. *Assume $T > (k+1)N$ and let $\tilde{L}$ be as above. Take an arbitrary N–dimensional subspace $\mathcal{U}$ in $(T-1)$–dimensional space V. Let P be the orthogonal projector on the space orthogonal to $\tilde{L}\mathcal{U}$, $\tilde{L}^2\mathcal{U}, \dots, \tilde{L}^{k-1}\mathcal{U}$. Let P_1 be the projector on the subspace $P\mathcal{U}$ and P_2 be the projector on the subspace $P\tilde{L}^k(I_V + \tilde{L})^{-1}\mathcal{U}$. Then the distributions of non-zero eigenvalues of $P_1P_2P_1$ coincides with that of the squared sample canonical correlations solving* (18) *under the hypothesis $\widehat{H}_0$.*

Comparing Proposition 18 with Theorem 11 and Remark 12 one notices that the differences are in restricting on the subspace V (hence, decreasing the dimension by 1) and in replacement $O \leftrightarrow \tilde{L}$.

Proof of Proposition 18.

Step 1. We start by transforming the Gaussian noise ε_t. Let ε be $N \times T$ matrix, whose t-th column is ε_t. Take any non-degenerate $N \times N$ matrix A and transform $\varepsilon \mapsto A\varepsilon$. Thus, we leave X_0 unchanged and recalculate X_t, $1 \le t \le T$. We claim that the canonical correlations solving Eq. (18) are unchanged. Indeed, the linear subspace $\mathcal{W}$ stays the same and so does the projector $P_{\perp\mathcal{W}}$. For each $t = 1, 2, \dots, T$, the vector ΔX_t is transformed by $\Delta X_t \mapsto A\Delta X_t + (I_N - A)\mu$ and $\tilde{X}_t$ is transformed by $\tilde{X}_t \mapsto A\tilde{X}_t + (I_N - A)X_0$. Recall that the space $\mathcal{W}$ includes vector $(1, \dots, 1)$, which leads to the projector $P_{\perp\mathcal{W}}$ canceling the additional terms $(I_N - A)\mu$ and $(I_N - A)X_0$ in the last two formulas. Hence, the matrices $\tilde{R}_0$ and $\tilde{R}_k$ are transformed by $\tilde{R}_0 \mapsto A\tilde{R}_0$ and $\tilde{R}_k \mapsto A\tilde{R}_k$. Therefore,

$$\tilde{S}_{k0}\tilde{S}_{00}^{-1}\tilde{S}_{0k} \mapsto A\tilde{S}_{k0}\tilde{S}_{00}^{-1}\tilde{S}_{0k}A^*, \qquad \tilde{S}_{kk} \mapsto A\tilde{S}_{kk}A^*.$$

We conclude that Eq. (18) is multiplied by $\det(A)\det(A^*)$ and, hence, its roots are preserved.

By choosing an $A = \Lambda^{-1/2}$ the covariance matrix Λ becomes identical. Hence, for the rest of the proof we assume without loss of generality that Λ is identical, which means that the matrix elements of ε are i.i.d. standard Gaussians.

Step 2. Let us now reduce the canonical correlations solving Eq. (18) to eigenvalues for a product of projectors. By definition the canonical correlations are eigenvalues of $N \times N$

[9]Note that in Bykhovskaya and Gorin (2021) we expressed all the operators in terms of $F = L_c^{-1}$ rather than L_c.

matrix

$$\tilde{S}_{k0}\tilde{S}_{00}^{-1}\tilde{S}_{0k}\tilde{S}_{kk}^{-1} = \tilde{R}_k\tilde{R}_0^*(\tilde{R}_0\tilde{R}_0^*)^{-1}\tilde{R}_0\tilde{R}_k^*(\tilde{R}_k\tilde{R}_k^*)^{-1}$$

Note that for any two rectangular matrices A and B of the same sizes the non-zero eigenvalues of AB^* and of B^*A coincide. Hence, the desired canonical correlations are also eigenvalues of $T \times T$ matrix

$$\left[\tilde{R}_0^*(\tilde{R}_0\tilde{R}_0^*)^{-1}\tilde{R}_0\right] \cdot \left[\tilde{R}_k^*(\tilde{R}_k\tilde{R}_k^*)^{-1}\tilde{R}_k\right].$$

The last matrix is a product of two projectors:[10] the first one projects on the space spanned by columns of $\tilde{R}_0^*$ and the second one projects on columns of $\tilde{R}_k^*$.

Step 3. The next step is to express via ε various matrices involved in constructing $\tilde{R}_0$ and $\tilde{R}_k$. Let $\mathcal{P}$ be the orthogonal projector on the subspace V. Under $\widehat{H}_0$ we have $\Delta X_t = \mu + \varepsilon_t$. Also

$$(\Delta X\mathcal{P})_t = \mu + \varepsilon_t - \frac{1}{T}\sum_{\tau=1}^{T}(\mu + \varepsilon_\tau) = \varepsilon_t - \frac{1}{T}\sum_{\tau=1}^{T}\varepsilon_\tau, \qquad \Delta X\mathcal{P} = \varepsilon\mathcal{P}.$$

Further, we define the $T \times T$ summation matrix Φ. It has 1's below the diagonal and 0's on the diagonal and everywhere above the diagonal:

$$\Phi = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ & & \ddots & & \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix}.$$

We set

$$\tilde{\Phi} = \mathcal{P}\Phi\mathcal{P}.$$

By a straightforward linear algebra (see (Bykhovskaya and Gorin, 2021, Section 9.2) for some details) one shows that the linear operator $\tilde{\Phi}$ preserves the space V (orthogonal to $(1, 1, \dots, 1)$). In addition, its restriction on the subspace V coincides with $L_V(I_V - L_V)^{-1}$, where I_V is the identical operator acting in V.

We can write

$$\tilde{X}_t = X_{t-1} - \frac{t-1}{T}(X_T - X_0) = X_0 + (t-1)\mu + \sum_{\tau=1}^{t-1}\varepsilon_\tau - \frac{t-1}{T}\left(T\mu + \sum_{\tau=1}^{T}\varepsilon_\tau\right) = X_0 + \sum_{\tau=1}^{t-1}\varepsilon_\tau - \frac{t-1}{T}\sum_{\tau=1}^{T}\varepsilon_\tau. \tag{45}$$

[10]For a closer match to the proposition that we are proving, note also that if P_1 and P_2 are projectors, then eigenvalues of P_1P_2 and $P_1P_2P_1$ are the same.

We claim that $\tilde{X}\mathcal{P} = \varepsilon\tilde{\Phi}^*$. Indeed, $\tilde{X}\mathcal{P}$ coincides with $\tilde{\tilde{X}}\mathcal{P}$, where

$$\tilde{\tilde{X}}_t = \tilde{X}_t - X_0 = \sum_{\tau=1}^{t-1}\varepsilon_\tau - \frac{t-1}{T}\sum_{\tau=1}^{T}\varepsilon_\tau = \sum_{\tau=1}^{t-1}\left(\varepsilon_\tau - \frac{1}{T}\sum_{s=1}^{T}\varepsilon_s\right).$$

Since Φ is the summation operator, we have $\tilde{\tilde{X}} = (\Phi(\varepsilon\mathcal{P})^*)^* = \varepsilon\mathcal{P}\Phi^*$ and the claim is proven because $\tilde{\Phi}^* = \mathcal{P}\Phi^*\mathcal{P}$.

Step 4. Previous steps yield the following expressions for $\tilde{R}_0$ and $\tilde{R}_k$. Take the N–dimensional space $\tilde{\mathcal{U}}$ (belonging to $(T-1)$-dimensional space V) spanned by the columns of $\mathcal{P}\varepsilon^*$. Let $\tilde{P}$ be the orthogonal projector on the space orthogonal to $L_c\tilde{\mathcal{U}}$, $L_c^2\tilde{\mathcal{U}}$, $\dots$, $L_c^{k-1}\tilde{\mathcal{U}}$. (Note that L_c can be replaced by L_V in the last definition without changing $\tilde{P}$). Then the space spanned by N columns of $\tilde{R}_0^*$ is $\tilde{P}\tilde{\mathcal{U}}$. On the other hand, the space spanned by N columns of $\tilde{R}_k^*$ is $\tilde{P}L_c^{k-1}\tilde{\Phi}\tilde{\mathcal{U}} = \tilde{P}L_V^k(I_V - L_V)^{-1}\tilde{\mathcal{U}}$. At this point, we see strong similarities with objects in the statement of Proposition 18 with main difference being in the assignment of randomness: L_c is deterministic and $\tilde{\mathcal{U}}$ is random, but $\tilde{L}$ is random and $\mathcal{U}$ is deterministic. Thus, it remains to relocate the random part.

For that we notice that due to the rotational invariance of the Gaussian law (here it is important that we made the covariance matrix Λ identical on the first step), the space $\tilde{\mathcal{U}}$ spanned by the columns of $\mathcal{P}\varepsilon^*$ has the same law as $\tilde{O}^*\mathcal{U}$. The reason is that both laws give uniformly random N–dimensional subspace of $(T-1)$–dimensional space V.

Since everything was previously expressed through the span of columns of $\mathcal{P}\varepsilon^*$, denote $\tilde{\mathcal{U}}$, we now simply replace those by the columns of $\tilde{O}^*\mathcal{U}$. Then the space orthogonal to $L_c\tilde{\mathcal{U}}$, $L_c^2\tilde{\mathcal{U}}$, $\dots$, $L_c^{k-1}\tilde{\mathcal{U}}$ becomes the space orthogonal to $L_c\tilde{O}^*\mathcal{U}$, $L_c^2\tilde{O}^*\mathcal{U}$, $\dots$, $L_c^{k-1}\tilde{O}^*\mathcal{U}$. Equivalently, this is the space orthogonal to $\tilde{O}^*\tilde{L}\mathcal{U}$, $\tilde{O}^*\tilde{L}^2\mathcal{U}$, $\dots$, $\tilde{O}^*\tilde{L}^{k-1}\mathcal{U}$. $\tilde{P}$ is the projector on this space. We conclude that the law of canonical correlations (18) is the same as the law of non-zero eigenvalues of the product of two projectors: the first one projects on the subspace $\tilde{P}\tilde{O}^*\mathcal{U}$ and the second one projects on the subspace $\tilde{P}L_c^k(I_V - L_c)^{-1}\tilde{O}^*\mathcal{U} = \tilde{P}\tilde{O}^*\tilde{L}^k(I_V + \tilde{L})^{-1}\mathcal{U}$. Up to a change of basis (by matrix $\tilde{O}$), which does not change the eigenvalues, we have arrived precisely at the expression from the statement of the proposition. □

The next proposition explains the effect of the replacement $O \leftrightarrow \tilde{L}$ on the eigenvalues of the product of projectors in Theorem 11 and Proposition 18. We need to introduce some additional notations.

Choose positive integers k, N, and $\mathcal{T}$, such that $\mathcal{T} \geq (k+1)N$ and an arbitrary N–dimensional subspace $\mathcal{V}$ in $\mathcal{T}$–dimensional space. Let

$$f^{k,N,\mathcal{T};\mathcal{V}} : SO(\mathcal{T}) \to \{0 \leq x_1 \leq x_2 \leq \dots \leq x_N \leq 1\}$$

be a map from the group $SO(\mathcal{T})$ of orthogonal $\mathcal{T}\times\mathcal{T}$ matrices of determinant 1 to N–tuples of reals on $[0,1]$ interval, defined by the following procedure: Take $O \in SO(\mathcal{T})$. Let P be the orthogonal projector on the space orthogonal to $O\mathcal{V}$, $O^2\mathcal{V}$,..., $O^{k-1}\mathcal{V}$. Let P_1 be the projector on the subspace $P\mathcal{V}$ and P_2 be the projector on the subspace $PO^k(I_\mathcal{T}+O)^{-1}\mathcal{V}$. Then $f^{k,N,\mathcal{T};\mathcal{V}}$ maps O to N largest eigenvalues of $P_1P_2P_1$.

We also need three norms:

(1) $\|v\|_2$ is the L_2 norm of a vector $v=(v_1,v_2,\dots,v_N)$, defined as $\|v\|_2=\sqrt{\sum_{i=1}^N v_i^2}$.
(2) $\|v\|_\infty$ is the supremum norm of a vector $v=(v_1,\dots,v_N)$, defined as $\|v\|_\infty=\max_i |v_i|$.
(3) $\|A\|_2$ is the spectral norm of a matrix A, defined as the square root of the largest eigenvalue of AA^*. Equivalently, $\|A\|_2=\max_v \frac{\|Av\|_2}{\|v\|_2}$.

Proposition 19. *Suppose that k is fixed, while N is growing and $\mathcal{T}$ depends on N in such a way that $\frac{\mathcal{T}}{N}\in[k+1+C_1,C_2]$ for some $C_1,C_2>0$. Let O_1 and O_2 be two $\mathcal{T}\times\mathcal{T}$ random matrices, such that:*

- *O_1 is a uniformly random $\mathcal{T}\times\mathcal{T}$ orthogonal matrix with determinant 1.*
- *The eigenvalues of O_2 are almost surely different from -1.*
- *For each $\varepsilon>0$ we have*

$$\lim_{N\to\infty}\mathrm{Prob}\left(\|O_1-O_2\|_2<\frac{1}{N^{1-\varepsilon}}\right)=1. \tag{46}$$

Then for each $\varepsilon>0$ we have

$$\lim_{N\to\infty}\mathrm{Prob}\left(\|f^{k,N,\mathcal{T};\mathcal{V}}(O_1)-f^{k,N,\mathcal{T};\mathcal{V}}(O_2)\|_\infty<\frac{1}{N^{1-\varepsilon}}\right)=1. \tag{47}$$

Proposition 19 claims a continuity of map $f^{k,N,\mathcal{T};\mathcal{V}}$. The proof needs care because of the inversions in the definition of the map $f^{k,N,\mathcal{T};\mathcal{V}}$.

Remark 20. *Proposition 19 has a version for complex numbers, in which all orthogonal matrices are replaced by unitary matrices. The proof of the complex version is the same.*

The proof of Proposition 19 relies on three lemmas which we prove later in this section. For these lemmas we write matrices O_1 and O_2 of Proposition 19 in the block forms according to the splitting $\mathcal{T}=(\mathcal{T}-N)+N$:

$$O_1=\begin{pmatrix}A_1 & B_1\\ C_1 & D_1\end{pmatrix},\qquad O_2=\begin{pmatrix}A_2 & B_2\\ C_2 & D_2\end{pmatrix}. \tag{48}$$

Lemma 21. *Let O be a $\mathcal{T}\times T$ orthogonal matrix of determinant 1 written in the block form $O=\begin{pmatrix}A & B\\ C & D\end{pmatrix}$ according to the splitting $\mathcal{T}=(\mathcal{T}-N)+N$. If all eigenvalues of O are different from -1, then so are the eigenvalues of D and of $A-B(I_N+D)^{-1}C$.*

Lemma 22. *Under the assumptions of Proposition 19 we have*

$$\lim_{N\to\infty}\mathrm{Prob}\left(\|(A_1-B_1(I_N+D_1)^{-1}C_1)-(A_2-B_2(I_N+D_2)^{-1}C_2)\|_2<\frac{1}{N^{1-\varepsilon}}\right)=1. \tag{49}$$

Lemma 23. *Under the assumptions of Proposition 19, let $\tilde{\mathcal{V}}_0$ be the N–dimensional subspace of $(\mathcal{T}-N)$–dimensional space spanned by the last N coordinate vectors. There exists an $(\mathcal{T}-N)\times(\mathcal{T}-N)$ orthogonal matrix U_1, depending only on $B_1(I_N+D_1)^{-1}$ and an $(\mathcal{T}-N)\times(\mathcal{T}-N)$ orthogonal matrix, U_2, depending both on $B_1(I_N+D_1)^{-1}$ and on $B_2(I_N+D_2)^{-1}$, such that $U_1\tilde{\mathcal{V}}_0 = \langle B_1\rangle$, $U_2\tilde{\mathcal{V}}_0 = \langle B_2\rangle$ and*

$$\lim_{N\to\infty} \mathrm{Prob}\left(\|U_1-U_2\|_2 < \frac{1}{N^{1-\varepsilon}}\right) = 1. \tag{50}$$

Proof of Proposition 19. The proof is induction in k with the base case $k=1$ proven in Bykhovskaya and Gorin (2021), see the continuity of $M(Z)$ in the proof of Proposition 13 there. For the induction step we recycle the ideas in the proof of Theorem 11.

First, recall a property of function f, which we established in Steps 1 and 2 of Theorem 11: if U is a $\mathcal{T}\times\mathcal{T}$ orthogonal matrix, then

$$f^{k,N,\mathcal{T};U\mathcal{V}}(O) = f^{k,N,\mathcal{T};\mathcal{V}}(U^*OU). \tag{51}$$

Note that conjugations (by the same orthogonal matrix for O_1 and O_2) leave the three conditions of Proposition 19 unchanged, hence, the (51) implies that statement of proposition remains the same for any choice $\mathcal{V}$.

Second, we make the replacements of Steps 2 and 3 of Theorem 11 individually for $f^{k,N,\mathcal{T};\mathcal{V}}(O_1)$ and $f^{k,N,\mathcal{T};\mathcal{V}}(O_2)$. The replacement $\mathcal{V}\mapsto O^{k-1}\mathcal{V}$ does not change the eigenvalues of $P_1P_2P_1$, while inversion of O_1 and O_2 keeps the conditions of Proposition 19 unchanged. Summing up, we replace the map $O\mapsto f^{k,N,\mathcal{T};\mathcal{V}}(O)$ in Proposition 19 by a new map $O\mapsto \tilde{f}^{k,N,\mathcal{T};\mathcal{V}_0}(O)$ defined through: Let $\mathcal{V}_0$ be the subspace spanned by the last N coordinate vectors in $\mathcal{T}$–dimensional space. Let P be the orthogonal projector on the space orthogonal to $\mathcal{V}, O\mathcal{V}, O^2\mathcal{V},\dots, O^{k-2}\mathcal{V}$. Let P_1 be the projector on the subspace $PO^{k-1}\mathcal{V}$ and P_2 be the projector on the subspace $P(I_\mathcal{T}+O)\mathcal{V}$ (or, equivalently, on $PO(I_\mathcal{T}+O)\mathcal{V}$). Then $\tilde{f}^{k,N,\mathcal{T};\mathcal{V}_0}(O)$ is N largest eigenvalues of $P_1P_2P_1$.

Using the block notations (48), steps 4-7 in the proof of Theorem 11 imply the following almost sure identities:

$$\tilde{f}^{k,N,\mathcal{T};\mathcal{V}_0}(O_1) = \tilde{f}^{k-1,N,\mathcal{T}-N;\langle B_1\rangle}(A_1 - B_1(I_N+D_1)^{-1}C_1), \tag{52}$$

$$\tilde{f}^{k,N,\mathcal{T};\mathcal{V}_0}(O_2) = \tilde{f}^{k-1,N,\mathcal{T}-N;\langle B_2\rangle}(A_2 - B_2(I_N+D_2)^{-1}C_2). \tag{53}$$

We would like to check that the right-hand sides of (52) and (53) are close by using the induction assumption.

Using (51) and Lemma 23, we rewrite the right-hand sides of (52) and (53) as:

$$\tilde{f}^{k-1,N,\mathcal{T}-N;\tilde{\mathcal{V}}_0}\big(U_1^*(A_1-B_1(I_N+D_1)^{-1}C_1)U_1\big); \qquad \tilde{f}^{k-1,N,\mathcal{T}-N;\tilde{\mathcal{V}}_0}\big(U_2^*(A_2-B_2(I_N+D_2)^{-1}C_2)U_2\big). \tag{54}$$

Let us check that we can apply the induction assumption to deduce that the expressions of (54) are close to each other:

- By Lemma 15, $A_1 - B_1(I_N + D_1)^{-1}C_1$ is a uniformly random $(\mathcal{T} - N) \times (\mathcal{T} - N)$ orthogonal matrix. By Lemma 23, U_1 is a function of $B_1(I_N + D_1)^{-1}$. Hence, using Lemma 15 again, we conclude that U_1 is independent from $A_1 - B_1(I_N + D_1)^{-1}C_1$. Therefore, $U_1^*(A_1 - B_1(I_N + D_1)^{-1}C_1)U_1$ is a uniformly random orthogonal matrix, as desired.
- By Lemma 21, $(I_N + D_2)^{-1}$ is well-defined and no eigenvalues of $A_2 - B_2(I_N + D_2)^{-1}C_2$ are equal to -1. Hence, the eigenvalues of $U_2^*(A_2 - B_2(I_N + D_2)^{-1}C_2)U_2$ are almost surely different from -1.
- Combining Lemmas 22 and 23 we conclude that

$$\lim_{N\to\infty} \mathrm{Prob}\left(\|U_1^*(A_1 - B_1(I_N + D_1)^{-1}C_1)U_1 - U_2^*(A_2 - B_2(I_N + D_2)^{-1}C_2)U_2\|_2 < \frac{1}{N^{1-\varepsilon}} \right) = 1.$$

Hence, using the $(k-1)$ statement, the expressions in (54) are close to each other as $N \to \infty$ and, therefore, (52) is close to (53). □

Proof of Lemma 21. Let us show that D has no eigenvalues -1. We argue by contradiction and assume that there exists an N–dimensional vector v of length 1 such that $Dv = -v$. Note that $B^*B + D^*D = I_N$ by orthogonality of O. Hence, using the notation $\langle \cdot, \cdot \rangle$ for the scalar product, we have

$$\langle Bv, Bv \rangle = \langle B^*Bv, v \rangle = \langle (I_N - D^*D)v, v \rangle = \langle v, v \rangle - \langle Dv, Dv \rangle = 1 - 1 = 0.$$

Therefore, $Bv = 0$, which readily implies that the $\mathcal{T}$–dimensional vector $\binom{0}{v}$ is an eigenvector of O with eigenvalue -1. Contradiction.

Next, for the matrix $A - B(I_N + D)^{-1}C$, let us use its representation as a Cayley transform developed in (37):

$$A - B(I_N + D)^{-1}C = \frac{I_{\mathcal{T}-N} - \mathcal{R}}{I_{\mathcal{T}-N} + \mathcal{R}},$$

where $\mathcal{R}$ is a $(\mathcal{T} - N) \times (\mathcal{T} - N)$ skew-symmetric matrix. If v was an eigenvector of $A - B(I_N + D)^{-1}C$ with eigenvalue -1, then we would have

$$(I_{\mathcal{T}-N} - \mathcal{R})v = -(I_{\mathcal{T}-N} + \mathcal{R})v,$$

which is impossible for non-zero v. □

Proof of Lemma 22. Note that whenever X is a submatrix of Y, we have $\|X\|_2 \le \|Y\|_2$. Hence, the spectral norms of the differences $A_1 - A_2$, $B_1 - B_2$, $C_1 - C_2$, $D_1 - D_2$ are all small with probability tending to 1 as $N \to \infty$. Addition, multiplication, and inversion of matrices are all Lipschitz operations as long as factors are bounded for the multiplication and singular values are bounded away from 0 for the inversion. Therefore, it remains to show that the norms of the factors B_1, $(I_N + D_1)^{-1}$, and C_1 are uniformly bounded (since B_1, $(I_N + D_1)^{-1}$, and C_1 are close to B_2, $(I_N + D_2)^{-1}$, and C_2, respectively, the norms of the latter are then going to be bounded as well). For B_1 and C_1 the bound on the norm is straightforward, as they are submatrices of O_1, whose norm is 1. Hence, $\|B_1\|_2 \le 1$ and $\|C_1\|_2 \le 1$.

In order to deal with $(I_N + D_1)^{-1}$ we rely on the fact that the distribution of the symmetric $N \times N$ matrix $Y = D_1^* D_1$ is explicit. It has density (see, e.g., (Forrester, 2010, (3.113) and the formula immediately after)) proportional to:

$$\det Y^{-1/2} \det(I_N - Y)^{\frac{\mathcal{T}}{2} - N - 1/2}\, dY, \qquad 0 < Y < I_N. \tag{55}$$

This is a particular case of the Jacobi ensemble of Definition 5 and we can use the large N asymptotic of the latter recorded in Theorem 10. Therefore, there exists a constant $0 < c < 1$, such that all the eigenvalues of Y are smaller than c with probability tending to 1 as $N \to \infty$. Hence, by the triangular inequality

$$\min_{\|v\|_2 = 1} \|(I_N + D_1)v\|_2 \ge 1 - \sqrt{c},$$

with probability tending to 1 as $N \to \infty$. We conclude that

$$\lim_{N \to \infty} \operatorname{Prob}\left(\|(I_N + D_1)^{-1}\| < \frac{1}{1 - \sqrt{c}} \right) = 1. \quad \square.$$

Proof of Lemma 23. We will be proving a slightly different statement, in which $\tilde{\mathcal{V}}_0$ is the span of the first (rather than last) coordinate vectors. The desired statement of the theorem is then obtained by replacing $U_1 \mapsto U_1 \cdot \mathfrak{S}$, and $U_2 \mapsto U_2 \cdot \mathfrak{S}$, where $\mathfrak{S}$ is the (orthogonal matrix) which swaps ith and $(\mathcal{T} - N + 1 - i)$th basis vectors for $i = 1, 2, \dots, \mathcal{T} - N$.

We know that the matrices B_1 and B_2 are close to each other and our aim is to show that the orthonormal bases of $\langle B_1 \rangle = \langle B_1 (I_N + D_1)^{-1} \rangle$ and its orthogonal complement, and $\langle B_2 \rangle = \langle B_2 (I_N + D_2)^{-1} \rangle$ and its orthogonal complement can be chosen to also be close to each other. For that we need to produce some formulas for these bases, which is what we do in the rest of the proof. The delicacy of this argument stems from the fact that given a space, in general, there might be no continuous way to produce an orthogonal matrix, such that the space is spanned by its first columns. (For instance, by the hairy ball theorem one can not continuously complement a unit vector in 3–dimensional space to an orthonormal basis.) Hence, we need to be more careful.

We start by replacing B_1 with

$$X_1 := B_1 (I_N + D_1)^{-1} \big((I_N + D_1^*)^{-1} B_1^* B_1 (I_N + D_1)^{-1} \big)^{-1/2}$$

and replacing B_2 with

$$X_2 := B_2(I_N + D_2)^{-1}\big((I_N + D_2^*)^{-1} B_2^* B_2 (I_N + D_1)^{-1}\big)^{-1/2}.$$

Clearly, $\langle B_1\rangle = \langle X_1\rangle$ and $\langle B_2\rangle = \langle X_2\rangle$. The advantage of X_1 and X_2 is that their columns are orthonormal. Indeed,

$$X_1^* X_1 = \big((I_N + D_1^*)^{-1} B_1^* B_1 (I_N + D_1)^{-1}\big)^{-1/2} (I_N + D_1^*)^{-1} B_1^* B_1 (I_N + D_1)^{-1} \times \big((I_N + D_1^*)^{-1} B_1^* B_1 (I_N + D_1)^{-1}\big)^{-1/2} = I_N$$

and similarly for X_2.

Claim. X_1 and X_2 are asymptotically close to each other:

$$\lim_{N\to\infty} \mathrm{Prob}\left(\|X_1 - X_2\|_2 < \frac{1}{N^{1-\varepsilon}} \right) = 1. \tag{56}$$

Note that X_1 and X_2 are built out of O_1 and O_2 with operations of addition, multiplication, inversion, and square root. The first one is Lipschitz in spectral norm, the second one is Lipschitz as long as the factors are uniformly bounded, and for the last two we additionally need the singular values of the factors to be uniformly bounded away from 0 uniformly[11]. We already explained in the proof of Lemma 22 that B_1 has spectral norm at most 1 and that $(I_N + D_1)$ (and hence also its inverse and its transpose) has singular values bounded away from 0 and ∞. Hence, it remains only to deal with $B_1^* B_1$ in the definition of X_1. Since B_1 is a $(\mathcal{T} - N) \times N$ submatrix of uniformly random $\mathcal{T} \times \mathcal{T}$ matrix, the law of $\Lambda = B_1^* B_1$ is explicit. It has density (see, e.g., (Forrester, 2010, (3.113) and the formula immediately after)) proportional to:

$$\det \Lambda^{\frac{\mathcal{T}}{2} - N - 1/2} \det(I_N - \Lambda)^{-1/2}\, d\Lambda, \qquad 0 < \Lambda < I_N. \tag{57}$$

This is a particular case of the Jacobi ensemble of Definition 5 and we can use the large N asymptotic of the latter recorded in Theorem 10, which implies that the eigenvalues of Λ are bounded away from 0 as $N \to \infty$. The claim is proven.

Next, we produce the desired orthogonal matrix U_1 by the Gramm-Schmidt orthogonalization procedure: letting e_k be the k–th coordinate vector in $(\mathcal{T} - N)$–dimensional space, and X_1^k be the k–th column of X_1, we start from $(\mathcal{T} - N)$ vectors

$$X_1^1, X_1^2, \dots, X_1^N,\ e_{N+1}, e_{N+2}, \dots, e_{\mathcal{T}-N}$$

and orthogonalize them. This is a valid procedure, since the Gramm matrix of the above vectors is almost surely non-degenerate (this is equivalent to the non-degeneracy of the top

[11]For the square root operation on positive-definite matrices $x \mapsto \sqrt{x}$ we can first rescale x so that its spectrum belongs to $[c_0, 1]$ segment for some $c_0 > 0$ and then use Taylor series expansion of the square root: $\sqrt{x} = \sqrt{1 + (x-1)} = 1 + \frac{x-1}{2} - \frac{1}{4}(x-1)^2 + \dots$ to deduce the Lipschitz property.

$N \times N$ corner of X_1, which is true due to absolute continuity of the distribution of this corner with respect to the Lebesgue measure on $N \times N$ matrices that can be deduced from Remark 16).

We set the columns of U_1 to be the vectors from the orthogonalization procedure. Since the vectors are orthonormal, U_1 is orthogonal. Note that since the columns of X_1 are orthonormal, the first N steps of the orthogonalization procedure are trivial and the first N columns of U_1 are $X_1^1, X_1^2, \dots, X_1^N$. In particular, these N columns span $\langle B_1 \rangle$, as desired.

We proceed to the construction of U_2. It is tempting to do exactly the same procedure (with all indices 1 replaced by indices 2), but that is not going to work: the problem is that while the top $N \times N$ corner of X_1 was almost surely non-degenerate, but it can have singular values arbitrary close to 0. Eventually, this leads to unstability of the orthogonalization procedure and, hence, there is no way to guarantee that the results of orthogonalization for X_1 and X_2 are close to each other.

Therefore, we proceed in a different way. Set $F := U_1^{-1} X_2$. Because the first N columns of U_1 are X_1, we have

$$U_1^{-1} X_1 = \begin{pmatrix} I_N \\ 0_{(\mathcal{T}-2N)\times N} \end{pmatrix},$$

where $0_{(\mathcal{T}-2N)\times N}$ stays for the $(\mathcal{T}-2N) \times N$ filled with 0 matrix elements. Hence, since X_1 and X_2 were close, we have

$$\lim_{N\to\infty} \operatorname{Prob}\left(\left\| F - \begin{pmatrix} I_N \\ 0_{(\mathcal{T}-2N)\times N} \end{pmatrix} \right\|_2 < \frac{1}{N^{1-\varepsilon}} \right) = 1. \tag{58}$$

Let F^k, $k = 1, \dots, N$, denote the columns of F and consider $\mathcal{T} - N$ vectors

$$F^1, F^2, \dots, F^N, \; e_{N+1}, e_{N+2}, \dots, e_{\mathcal{T}-N}.$$

We are going to orthogonalize these vectors. The advantage over the procedure we used for X_1 is that now the top $N \times N$ submatrix of F is close to identity, which is going to make the orthogonalization procedure well-behaved. In order to make the orthogonalization procedure explicit, we are going to use a block version of the Cholesky decomposition.

For that set $M = \mathcal{T} - 2N$ and write F in the block form according to the splitting $\mathcal{T} - N = N + M$:

$$F = \begin{pmatrix} Y_2 \\ Z_2 \end{pmatrix}.$$

Let W_2 denote the $(\mathcal{T}-N) \times (\mathcal{T}-N)$ matrix written in the $N + M$ block form as

$$W_2 = \begin{pmatrix} Y_2 & 0 \\ Z_2 & I_M \end{pmatrix}.$$

We would like to perform orthogonalization of the columns of W_2. For that we first compute

$$W_2^* W_2 = \begin{pmatrix} Y_2^* Y_2 + Z_2^* Z_2 & Z_2^* \\ Z_2 & I_M \end{pmatrix}. \tag{59}$$

We further would like to represent $W_2^* W_2$ as

$$W_2^* W_2 = \begin{pmatrix} I_N & 0 \\ Q_2 & I_M \end{pmatrix} \begin{pmatrix} G_2 & 0 \\ 0 & H_2 \end{pmatrix} \begin{pmatrix} I_N & Q_2^* \\ 0 & I_M \end{pmatrix} = \begin{pmatrix} G_2 & G_2 Q_2^* \\ Q_2 G_2 & Q_2 G_2 Q_2^* + H_2 \end{pmatrix}. \tag{60}$$

Comparing with (59) we conclude that

$$G_2 = Y_2^* Y_2 + Z_2^* Z_2 = F^* F, \quad Q_2 = Z_2 (F^* F)^{-1}, \quad H_2 = I_M - Z_2 (F^* F)^{-1} Z_2^*. \tag{61}$$

Since the spectral norm of a submartix is at most the spectral norm of the matrix, (58) implies that

$$\lim_{N\to\infty} \mathrm{Prob}\left(\|Y_2 - I_N\|_2 < \frac{1}{N^{1-\varepsilon}} \right) = 1, \qquad \lim_{N\to\infty} \mathrm{Prob}\left(\|Z_2 - 0_{M\times N}\|_2 < \frac{1}{N^{1-\varepsilon}} \right) = 1. \tag{62}$$

Therefore, W_2 is close to I_{M+N}, G_2 is close to I_N, Q_2 is close to $0_{N\times M}$, H_2 is close to I_M.

Note that G_2 and H_2 are positive-definite symmetric matrices, hence, they have well-defined square roots. In addition,

$$\begin{pmatrix} I_N & Q_2^* \\ 0 & I_M \end{pmatrix}^{-1} = \begin{pmatrix} I_N & -Q_2^* \\ 0 & I_M \end{pmatrix}.$$

The goal of all these manipulations with matrices is to define

$$\tilde U_2 = W_2 \cdot \begin{pmatrix} I_N & -Q_2^* \\ 0 & I_M \end{pmatrix} \cdot \begin{pmatrix} G_2^{-1/2} & 0 \\ 0 & H_2^{-1/2} \end{pmatrix}. \tag{63}$$

The two key properties of $\tilde U_2$ are:

- The span of the first N columns of $\tilde U_2$ coincides with $\langle F \rangle$.
- $\tilde U_2$ is orthogonal. Indeed, using (60) we have

$$\tilde U_2 \tilde U_2^* = W_2 \begin{pmatrix} I_N & -Q_2^* \\ 0 & I_M \end{pmatrix} \begin{pmatrix} G_2^{-1} & 0 \\ 0 & H_2^{-1} \end{pmatrix} \begin{pmatrix} I_N & 0 \\ -Q_2 & I_M \end{pmatrix} W_2^* = W_2 (W_2^* W_2)^{-1} W_2^* = I_{\mathcal{T}-N}.$$

Hence, we can finally set

$$U_2 := U_1 \cdot \tilde U_2,$$

We have

$$U_2 \tilde{\mathcal{V}}_0 = \langle U_1 F \rangle = \langle X_2 \rangle = \langle B_2 \rangle,$$

as desired. It remains to show that the matrix $\tilde U_2$ is very close to identity, as this would imply that U_2 is close to U_1. For that we consider each factor in (63) and see that they are close to identical matrices by (62). Hence,

$$\lim_{N\to\infty} \operatorname{Prob}\left(\left\|\tilde{U}_2 - I_{N+M}\right\|_2 < \frac{1}{N^{1-\varepsilon}}\right) = 1,$$

as desired. □

Proof of Theorem 8. We start by explicitly constructing the desired coupling. For the Jacobi ensemble we use the realization of Theorem 11 and for the matrix of the Johansen test we use the realization of Proposition 18. We set $\mathcal{T} = T - 1$ to match the notations and it remains to couple O of Theorem 11 with $\tilde{L} = -\tilde{O} L_V \tilde{O}^*$ of Proposition 18.

The eigenvalues of L_V are all roots of unity of order T different from 1. In the complex case $\beta = 2$ we can diagonalize L_V to turn it into $\mathcal{T} \times \mathcal{T} = (T-1) \times (T-1)$ diagonal matrix with the roots of unity on the diagonal. In the real case $\beta = 1$, the matrix L_V should be block-diagonalized (with blocks of size 2 and one additional block of size 1 corresponding to eigenvalue -1 if $\mathcal{T}$ is even): the pair of complex conjugate roots of unity ω and $\bar{\omega}$ gives rise to the 2×2 matrix of rotation by the angle $|\arg(\omega)|$. Let us denote by D the resulting (block) diagonal matrix multiplied by -1. In order to avoid ambiguity about the order of eigenvalues, we assume that the blocks correspond to the increasing order of $|\arg(-\omega)|$, i.e. the top-left 2×2 corner of D corresponds to the pair of the closest to 1 eigenvalues of D.

The eigenvalues of O also lie on the unit circle and if $\beta = 1$, then they come in complex-conjugate pairs. Hence, O can be similarly block-diagonalized (we do not need to multiply by -1 this time) and we denote through D^{rand} the result. The distinction with L_V is that the eigenvalues are *random* and so is D^{rand}. The law of the eigenvalues of O is explicitly known in the random-matrix literature. Both for $\beta = 1$ and $\beta = 2$ they form a determinantal point process on the unit circle with explicit kernel. The repulsion between the eigenvalues leads to them being very close to evenly spaced as $T \to \infty$. We summarize this property in the following statement (which is a manifestation of a more general rigidity of eigenvalues, see, e.g., Erdos and Yau (2012)), whose proof can be found in Meckes and Meckes (2013, Lemma 10, $m = 1$, $u = T^{\delta}$ case, and Section 5).

Claim. There exist constants $c_1(\beta), c_2(\beta) > 0$, such that for $\beta = 1, 2$, every $\delta > 0$, there exists $\mathcal{T}_0(\delta)$ and for every $\mathcal{T} > \mathcal{T}_0(\delta)$ we have [12]

$$\operatorname{Prob}\left(\max_{1\le i,j<\mathcal{T}} \left|D - D^{\text{rand}}\right|_{ij} > \frac{1}{\mathcal{T}^{1-\delta}}\right) < c_1(\beta) \cdot \mathcal{T} \cdot \exp\left(-c_2(\beta) \frac{\mathcal{T}^{2\delta}}{\log \mathcal{T}}\right). \tag{64}$$

We remark that since D and D^{rand} are block-diagonal, the bound on the maximum matrix element of their difference is equivalent to a similar bound for any other norm, e.g., for the spectral norm, which we used in Proposition 19.

[12] All the constants can be made explicit, following Meckes and Meckes (2013).

We now choose another $\mathcal{T}\times\mathcal{T}$ uniformly-random orthogonal (or unitary if $\beta=2$) matrix O_2 (independent from the rest), replace $-\tilde{O}L_V\tilde{O}^*$ with $O_2DO_2^*$ and replace O with $O_2D^{\mathrm{rand}}O_2^*$. The invariance of the uniform measure on the orthogonal group $SO(N)$ (or on the unitary group $U(N)$ if $\beta=2$) with respect to right/left multiplications, implies the distributional identities:

$$-\tilde{O}L_V\tilde{O}^* \stackrel{d}{=} O_2DO_2^*, \qquad O \stackrel{d}{=} O_2D^{\mathrm{rand}}O_2^*.$$

The right-hand sides of the identities provide the desired coupling and (64) implies that these two random matrices are close to each other as $\mathcal{T}\to\infty$.

It now remains to apply Proposition 19 (see also Remark 20) with the first matrix being $O_2D^{\mathrm{rand}}O_2^*$ and the second matrix being $O_2DO_2^*$. □

10. Appendix. Proofs 4: Small rank perturbations

In this section we prove Theorem 3 by combining Theorem 8 with Theorem 10 and general statements about small rank perturbations. The key step of the proof is the following observation:

Theorem 24. *Let R_i, $i=0,k$ be as in Section 2 for X_t solving Eq.* (1) *and let $\tilde{R}_i$, $i=0,k$ be as in Section 3 under $\widehat{H}_0$ for X_t solving Eq.* (12)*. Suppose that X_0 and the noises ε_t used in the constructions of R_i and $\tilde{R}_i$ are the same. Introduce $T\times T$ projection matrices:*

(65)
$$P_0=R_0^*(R_0R_0^*)^{-1}R_0, \quad P_k=R_k^*(R_kR_k^*)^{-1}R_k, \quad \tilde{P}_0=\tilde{R}_0^*(\tilde{R}_0\tilde{R}_0^*)^{-1}\tilde{R}_0, \quad \tilde{P}_k=\tilde{R}_k^*(\tilde{R}_k\tilde{R}_k^*)^{-1}\tilde{R}_k.$$

Then under the assumptions (6), (7) *of Theorem 3 we have*

$$\lim_{N\to\infty}\frac{1}{N}\mathrm{rank}\left(P_0P_kP_0-\tilde{P}_0\tilde{P}_k\tilde{P}_0\right)=0. \tag{66}$$

Proof. Throughout the proof we assume that the matrices $R_iR_i^*$ and $\tilde{R}_i\tilde{R}_i^*$ are invertible. In principle, invertibility might fail for some N: in such situation we can still use Moore–Penrose inverse in order for the statements to make sense, and we are not going to detail this.

In the following argument we use various properties of ranks:

- If a matrix A differs from a matrix B only in $\mathfrak{r}$ columns, then $\mathrm{rank}(A-B)\le\mathfrak{r}$;
- $\mathrm{rank}(CA-CB)\le\mathrm{rank}(A-B)$;
- $\mathrm{rank}(A+B)\le\mathrm{rank}(A)+\mathrm{rank}(B)$;
- If matrices A and B are invertible, then $\mathrm{rank}(A^{-1}-B^{-1})=\mathrm{rank}(A-B)$.

We refer to the time series defined by (1) as X_t and to the time series defined by (12) as $\mathcal{X}_t$. We form two $N\times T$ matrix X and $\mathcal{X}$ with columns X_t and $\mathcal{X}_t$, $t=1,\dots,T$, respectively. Our first task is to show that $\frac{1}{N}\mathrm{rank}(X-\mathcal{X})\to 0$ as $N\to\infty$.

For that we subtract (12) from (1) to get:

$$\Delta X_t - \Delta \mathcal{X}_t = \Pi X_{t-k} + \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Phi D_t - \mu, \quad t = 1, 2, \ldots, T. \tag{67}$$

In the matrix form, (67) represents the $N \times T$ matrix $\Delta(X - \mathcal{X})$ with columns $\Delta X_t - \Delta \mathcal{X}_t$ as a sum of $k+2$ low rank matrices, with the total rank (coming from the right-hand side of (67)) at most

$$\operatorname{rank}(\Pi) + \sum_{i=1}^{k-1} \operatorname{rank}(\Gamma_i) + d_D + 1. \tag{68}$$

The matrix $X - \mathcal{X}$ is obtained from $\Delta(X - \mathcal{X})$ by multiplication by the summation matrix

$$\Phi = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ & & \ddots & & \\ 1 & 1 & \ldots & 1 & 1. \end{pmatrix}$$

Hence, the rank of $X - \mathcal{X}$ is at most (68) and $\frac{1}{N}\operatorname{rank}(X - \mathcal{X}) \to 0$ by assumption (7) of Theorem 3.

Next, we should take into account that the procedures for constructing R_0, R_k from X and $\tilde{R}_0$, $\tilde{R}_k$ from $\mathcal{X}$ are slightly different. Namely, the latter involves cyclic shifts of indices, rather than usual shifts, involves regressing over only constants, rather than d_D deterministic terms D_t, and finally involves detrending (14). However, cyclic shifts only affect the first $k-1$ indices t and, hence, lead to bounded difference in ranks, and similarly for detrending. Regressing on d_D terms leads to another $O(d_D)$ difference in ranks, which is negligible after division by N in the limit $N \to \infty$ by (7). The conclusion is that R_0, R_k from one side and $\tilde{R}_0$, $\tilde{R}_k$ on the the other side are constructed from two finite sets of matrices, which differ by small rank perturbations, by finitely many of operations of addition, multiplication, and inversion. Each of these operations preserves the smallness of the rank of perturbations and, hence,

$$\lim_{N\to\infty} \frac{1}{N}\operatorname{rank}(R_0 - \tilde{R}_0) = 0, \quad \lim_{N\to\infty} \frac{1}{N}\operatorname{rank}(R_k - \tilde{R}_k) = 0. \tag{69}$$

Since $P_0 P_k P_0$ and $\tilde{P}_0 \tilde{P}_k \tilde{P}_0$ are obtained from R_0, R_k and $\tilde{R}_0$, $\tilde{R}_k$, respectively, by the same algebraic operations, (66) follows from (69). □

Proof of Theorem 3. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ denote the eigenvalues of $\mathcal{C} = S_{kk}^{-1} S_{k0} S_{00}^{-1} S_{0k}$ and let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_N$ denote the eigenvalues of $\tilde{\mathcal{C}} = \tilde{S}_{kk}^{-1} \tilde{S}_{k0} \tilde{S}_{00}^{-1} \tilde{S}_{0k}$. Combining

Theorem 8 with Theorem 10 we conclude that the empirical measure of $\tilde{\lambda}_i$ converges:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta_{\tilde{\lambda}_i} = \mu_{2,\tau-k}. \tag{70}$$

We would like to show that $\tilde{\lambda}_i$ can be replaced by λ_i in (70). Note that although the spectra of matrices $\mathcal{C}$ and $\tilde{\mathcal{C}}$ are real, but these matrices are not symmetric. Similarly to (66), $\frac{1}{N}\mathrm{rank}(\mathcal{C}-\tilde{\mathcal{C}}) \to 0$, however, in general, for non-symmetric matrices even rank 1 perturbations can lead to significant changes in the spectrum. Hence, we need to be more careful and symmetrize $\mathcal{C}$ and $\tilde{\mathcal{C}}$ by using projectors as in (66).

Note that for any two $K\times M$ matrices A and B, the non-zero eigenvalues of AB^* and of B^*A coincide. Recalling that $S_{ij} = R_i R_j^*$ and using the notations (65), we conclude that the eigenvalues of $\mathcal{C}$ are the same as N largest eigenvalues of $P_k P_0$. Since $P_0^2 = P_0$, they are also the same as N largest eigenvalues of $P_k P_0 P_0$ and the same as those of $P_0 P_k P_0$. Similarly, the eigenvalues of $\tilde{\mathcal{C}}$ are the same as N largest eigenvalues of $\tilde{P}_0 \tilde{P}_k \tilde{P}_0$.

Denote $\mathfrak{r} = \mathrm{rank}(P_0 P_k P_0 - \tilde{P}_0 \tilde{P}_k \tilde{P}_0)$. All the involved matrices are symmetric and we can use classical inequalities between eigenvalues of a Hermitian matrix A and Hermitian matrix $A+B$, where B has rank $\mathfrak{r}$, see, e.g., Horn and Johnson (2013, Corollary 4.3.5). In our situation the inequalities read

$$\lambda_{m-\mathfrak{r}} \ge \tilde{\lambda}_m \ge \lambda_{m+\mathfrak{r}}, \quad 1 \le m-\mathfrak{r} \le m+\mathfrak{r} \le N. \tag{71}$$

Therefore, for any points $0 < a < b < 1$,

$$\Big| \#\{1 \le i \le N \mid \lambda_i \in [a,b]\} - \#\{1 \le i \le N \mid \tilde{\lambda}_i \in [a,b]\} \Big| \le 2\mathfrak{r}.$$

Hence, (70) implies

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i} = \mu_{2,\tau-k}. \tag{72}$$

□

References

Anderson, T. W., "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *Annals of Mathematical Statistics*, 1951, *22* (3), 327–351.

Bai, J. and S. Ng, "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics*, 2008, *3* (2), 89–163.

Baik, J., G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability*, 2005, *33* (5), 1643–1697.

Breitung, J. and M.H. Pesaran, "Unit Roots and Cointegration in Panels," in "Mátyá L., Sevestre P. (eds) The Econometrics of Panel Data. Advanced Studies in Theoretical and Applied Econometrics, vol. 46," Berlin, Heidelberg: Springer, 2008, pp. 279–322.

Bykhovskaya, A. and V. Gorin, "Cointegration in large VARs," *Annals of Statistics*, 2021.

Cavaliere, G., A. Rahbek, and A.R. Taylor, "Bootstrap determination of the cointegration rank in vector autoregressive models," *Econometrica*, 2012, *80* (4), 1721–1740.

Choi, I., "Panel cointegration," in "Baltagi B.H. (eds) The Oxford handbook of panel data," Oxford University Press, 2015.

Dumitriu, I. and A. Edelman, "Matrix models for beta ensembles," *Journal of Mathematical Physics*, 2002, *43* (11), 5830–5847.

Erdos, L. and H. T. Yau, "Universality of local spectral statistics of random matrices," *Bulletin of the American Mathematical Society*, 2012, *49* (3), 377—-414.

Forrester, P. J., "The spectrum edge of random matrix ensembles," *Nuclear Physics B*, 1993, *402* (3), 709–728.

Forrester, P. J., *Log-gases and random matrices*, Princeton University Press, 2010.

Gonzalo, J. and J. Y. Pitarakis, "Comovements in Large Systems," *Statistics and Econometrics Series, Vol. 10. Working Paper 95-38, Universidad Carlos III de Madrid*, 1995.

Han, X., G. M. Pan, and B. Zhang, "The Tracy-Widom law for the Largest Eigenvalue of F Type Matrix," *The Annals of Statistics*, 2016, *44* (4), 1564–1592.

Horn, R. A. and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, 2013.

Hua, L., *Harmonic analysis of functions of several complex variables in the classical domains*, Vol. 6, American Mathematical Soc., 1963.

Johansen, S., "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, 1988, *12* (2–3), 231–254.

Johansen, S., "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 1991, *59*, 1551–1580.

Johansen, S., *Likelihood-based inference in cointegrated vector autoregressive models*, Oxford University Press, 1995.

Johansen, S., "A small sample correction for the test of cointegrating rank in the vector autoregressive model," *Econometrics*, 2002, *70* (5), 1929–1961.

Johnstone, I., "Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy-Widom limits and rates of convergence," *Annals of statistics*, 2008, *36* (6), 2638–2716.

Juselius, Katarina, *The Cointegrated VAR Model: Methodology and Applications*, Oxford University Press, 2006.

Keilbar, G. and Y. Zhang, "On Cointegration and Cryptocurrency Dynamics," *Digital Finance*, 2021, *3* (1), 1–23. `https://github.com/QuantLet/CryptoDynamics/blob/master/CryptoDynamics_Series/price.csv`.

Maddala, G.S. and In-Moo Kim, *Unit Roots, Cointegration, and Structural Change*, Cambridge University Press, 1998.

Meckes, E. and M. Meckes, "Spectral measures of powers of random matrices," *Electronic communications in probability*, 2013, *18.*

Neretin, Y., "Hua-type integrals over unitary groups and over projective limits of unitary groups," *Duke Mathematical Journal*, 2002, *114* (2), 239–266.

Olshanski, G., "The problem of harmonic analysis on the infinite-dimensional unitary group," *Journal of Functional Analysis*, 2003, *205* (2), 464–524.

Onatski, A. and C. Wang, "Alternative asymptotics for cointegration tests in large vars," *Econometrica*, 2018, *86* (4), 1465–1478.

Onatski, A. and C. Wang, "Extreme canonical correlations and high-dimensional cointegration analysis," *Journal of Econometrics*, 2019.

Pagan, A., "Three econometric methodologies: A critical appraisal," *Journal of Economic surveys*, 1987, *1* (1), 3–23.

Reinsel, G.C. and S.K. Ahn, "Vector autoregressive models with unit roots and reduced rank structure: Estimation. Likelihood ratio test, and forecasting," *Journal of time series analysis*, 1992, *13* (4), 353–375.

Sims, C. A., "Macroeconomics and Reality," *Econometrica*, 1980, *48* (1), 1–48.

Swensen, A.R., "Bootstrap Algorithms for Testing and Determining the Cointegration Rank in VAR Models," *Econometrica*, 2006, *74* (6), 1699–1714.

Tracy, C. A. and H. Widom, "On orthogonal and symplectic matrix ensembles," *Communications in Mathematical Physics*, 1996, *177* (3), 727–754.

Udell, M. and A. Townsend, "Why are big data matrices approximately low rank?," *SIAM Journal on Mathematics of Data Science*, 2019, *1* (1), 144–160.

Zhang, B., G. M. Pan, and J. T. Gao, "CLT for largest eigenvalues and unit root testing for high-dimensional nonstationary time series," *The Annals of Statistics*, 2018, *46* (5), 2186–2215.

University of Wisconsin-Madison
Email address, A.B.: `anna.bykhovskaya@wisc.edu`
Email address, V.G.: `vadicgor@gmail.com`