LARGEVARS: AN R PACKAGE FOR TESTING LARGE VARS FOR THE PRESENCE OF COINTEGRATION

ANNA BYKHOVSKAYA, VADIM GORIN, AND ESZTER KISS

Abstract. **Largevars** R package conducts a cointegration test for high-dimensional vector autoregressions of order k based on the large N, T asymptotics of Bykhovskaya and Gorin [2022, 2024]. The implemented test is a modification of the Johansen likelihood ratio test. In the absence of cointegration the test converges to the partial sum of the Airy_1 point process.

The package and this article contain simulated quantiles of the first ten partial sums of the Airy_1 point process that are precise up to the first 3 digits. An empirical example using **Largevars** on S&P100 stocks that can be found in Bykhovskaya and Gorin [2022, 2024] is also included. The package **Largevars** can be freely downloaded from CRAN.

Keywords: R, cointegration test, high-dimensional VAR.

1. Introduction

Vector Autoregressions (VARs) are a fundamental tool in econometrics and time series analysis, providing a framework for modeling the dynamic interrelationships among multiple time series. However, as the number of variables in a VAR increases, the complexity of the model grows significantly, posing challenges for both estimation and inference. One critical aspect of analyzing VAR models is testing for the presence of cointegration, which can inform whether a set of non-stationary series share a long-run equilibrium relationship. That is, whether a set of non-stationary time series has a stationary linear combination, see, e.g., Johansen [1995].

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). Yet traditional tests for the presence of cointegration (e.g., likelihood ratio of Johansen [1988, 1991]) are not suitable for analyzing large systems, as they tend to significantly over-reject the null hypothesis, see, for example, Ho and Sørensen [1996], Gonzalo and Pitarakis [1999]. To address this issue, Bykhovskaya and Gorin [2022, 2024] propose an approach based on alternative asymptotics, where both the number of coordinates N and the length T of time series are large, and tailored for high-dimensional time series. **Largevars** package implements this approach.

Date: December 4, 2024.

The test in **Largevars** package is based on the squared sample canonical correlations between transformed past levels (lags) and changes (first differences) of the data, see Procedure 2 in Bykhovskaya and Gorin [2024] for precise details. The asymptotic distribution (derived under $N, T \to \infty$ jointly and proportionally) of the test is given by the partial sums of the Airy$_1$ point process. Their quantiles, which are tabulated in Section 5 are of independent interest and have not appeared before, with the exception of Table 1, which corresponds to the quantiles of Tracy-Widom distribution, and was also tabulated in Bejan [2005]. The non-trivial algorithm used to obtain the quantiles is also discussed in detail in Section 5.

2. Getting started

Largevars is available at CRAN and can be installed from CRAN via

```
install.packages("Largevars")
```

Help for using the functions in the package can be called by running ?? `function name` . The empirical example in Section 4.1 of this paper can provide further guidance.

3. Commands

3.1. **Function largevar.** This is the main function in the package that implements the cointegration test for high-dimensional VARs.

```
largevar( data, k=1,r=1 , fin_sample_corr = FALSE, plot_output = TRUE,
significance_level = 0.05)
```

`data`	A numeric matrix where the columns contain individual time series that will be examined for the presence of cointegrating relationships. The rows are indexed by $t = 0, 1, \ldots, T$ and the columns by $i = 1, \ldots, N$.
`k`	The number of lags that we wish to employ in the vector autoregression. The default value is $k = 1$.
`r`	The number of largest eigenvalues used in the test. The default value is $r = 1$.
`fin_sample_corr`	A boolean variable indicating whether we wish to employ finite sample correction on our test statistic. The default value is `fin_sample_corr=FALSE`.

`plot_output`	A boolean variable indicating whether we wish to generate a plot of the empirical distribution of eigenvalues. The default value is `plot_output = TRUE`.
`significance_level`	Specify the significance level at which the decision about the H_0 should be made. The default value is `significance_level = 0.05`.

The function `largevar()` operates according to the steps laid out in Bykhovskaya and Gorin [2024]: it first detrends the data and regresses the detrended data on a constant and appropriately modified first differences and then calculates the squared sample canonical correlations between the residuals obtained from those regressions. The squared sample canonical correlations are equal to the N eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ of the matrix $\tilde{S}_{10}\tilde{S}_{00}^{-1}\tilde{S}_{01}\tilde{S}_{11}^{-1}$ from equation (16) in Bykhovskaya and Gorin [2024]. The test statistic is formed based on the r largest eigenvalues. Any value of r can be used to reject the hypothesis H_0 of no cointegration, and the user can try different options. We recommend small values such as $r = 1, 2, 3$; see Section 3.2 of Bykhovskaya and Gorin [2022] for the discussion.

`largevar()` returns a list object that contains the test statistic, a statistical table with a subset of theoretical quantiles ($q = 0.90, 0.95, 0.97, 0.99$) presented for $r = 1$ to $r = 10$, the decision about H_0 at the significance level specified by the user, and the p-value. These can be accessed by `list$statistic` (numeric value) `list$significance_test$significance_table` (numeric matrix), `list$significance_test$boolean_decision` (numeric value of 0 or 1), and `list$significance_test$p_value` (numeric value), respectively.

The simulations for the quantiles of the limiting distribution were conducted for $r = 1$ to $r = 10$ values. For this reason, p-values are accessible at inputs $r = 1$ to $r = 10$ only. For larger r inputs, the function returns the test statistic but not the p-value and not the decision about H_0 at the significance level specified by the user.

3.2. **Function quantile_tables.** To access the test quantile tables (partial sums of the Airy_1 random sequence presented in Section 4) in R, the user can call the `quantile_tables()` function. Quantile tables are available for $r = 1$ to $r = 10$. The function returns a numeric matrix, where the $0.ab$ quantile corresponds to the row $0.a$ and the column b.

```
> quantile_tables(r=1)
        0     1     2     3     4     5     6     7     8     9
0.0  -Inf -3.90 -3.61 -3.43 -3.30 -3.18 -3.08 -3.00 -2.92 -2.85
0.1 -2.78 -2.72 -2.67 -2.61 -2.56 -2.51 -2.46 -2.41 -2.37 -2.33
```

```
0.2 -2.29 -2.24 -2.20 -2.17 -2.13 -2.09 -2.05 -2.02 -1.98 -1.95
0.3 -1.91 -1.88 -1.84 -1.81 -1.78 -1.74 -1.71 -1.68 -1.65 -1.62
0.4 -1.58 -1.55 -1.52 -1.49 -1.46 -1.43 -1.40 -1.36 -1.33 -1.30
0.5 -1.27 -1.24 -1.21 -1.17 -1.14 -1.11 -1.08 -1.05 -1.01 -0.98
0.6 -0.95 -0.91 -0.88 -0.85 -0.81 -0.78 -0.74 -0.71 -0.67 -0.63
0.7 -0.59 -0.56 -0.52 -0.48 -0.44 -0.39 -0.35 -0.31 -0.26 -0.22
0.8 -0.17 -0.12 -0.07 -0.01  0.04  0.10  0.16  0.23  0.30  0.37
0.9  0.45  0.53  0.63  0.73  0.85  0.98  1.14  1.33  1.60  2.02
```

3.3. **Function sim_function.** This is an auxiliary function that allows the user to calculate an empirical p-value based on a simulation of the data generating process $\widehat{H}_0$ stated in equation (10) of Bykhovskaya and Gorin [2024]. This function should be used only for *quick approximate assessments*, as precise computations of the statistics require much larger numbers of simulations.

```
sim_function(N=NULL, tau=NULL, stat_value=NULL, k = 1, r = 1,
fin_sample_corr = FALSE, sim_num = 1000)
```

`N`	The number of time series used in simulations.
`tau`	The length of the time series used in simulations. If time is indexed as $t = 0, 1, \ldots, T$, then $\tau = T + 1$.
`stat_value`	The test statistic value for which the p-value is calculated.
`k`	The number of lags that we wish to employ in the vector autoregression. The default value is $k = 1$.
`r`	The number of largest eigenvalues used in the test. The default value is $r = 1$.
`fin_sample_corr`	A boolean variable indicating whether we wish to employ finite sample correction on our test statistics. The default value is `fin_sample_corr=FALSE`.
`sim_num`	The number of simulations that the function conducts for H_0. The default value is `sim_num = 1000`.

The function `sim_function()` runs the cointegration test on simulated data generated under $\widehat{H}_0$ and calculates the empirical p-value based for the test statistic specified by the user. For comparison purposes, it is advised to specify the same parameters k and r as employed for `largevar()`.

`sim_function()` returns a list object that contains the simulation values, the empirical p-value (fraction of realizations larger than the test statistic for the original data) and a histogram of the distribution of simulated test statistic values.

4. Examples

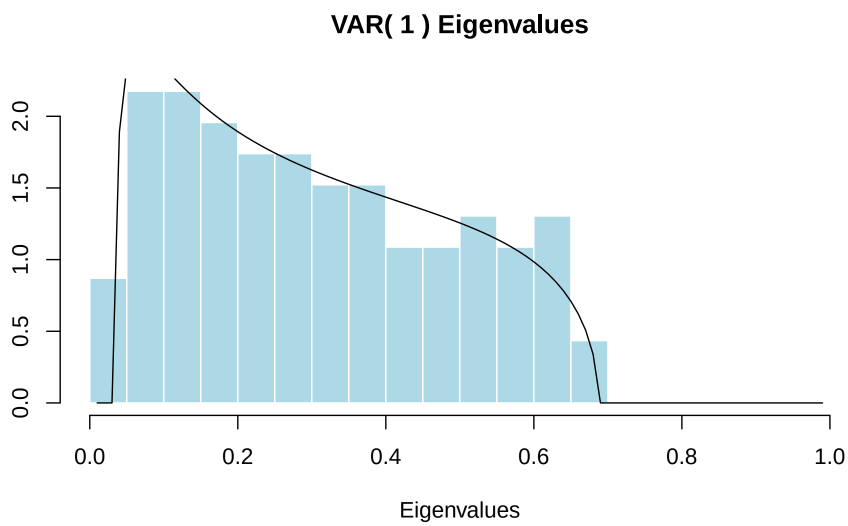
This section provides two examples of the usage of the package. Section 4.1 replicates the S&P100 example from Bykhovskaya and Gorin [2022, 2024], while Section 4.2 uses simulated data. Both examples include the code, which can be copied into R.

4.1. **S&P100.** We use logarithms of weekly adjusted closing prices of assets in the S&P100 over ten years (01.01.2010–01.01.2020), which gives us $\tau = 522$ observations across time. The S&P100 includes 101 stocks, with Google having two classes of stocks. We use 92 of those stocks, those for which data were available for our chosen time period. Only one of Google's two listed stocks is kept in the sample. Therefore, $N = 92$, $T = 521$ and $T/N \approx 5.66$. The data that we use are accessible from the "data" folder in the package.

```
library(Largevars)

## load data
library(readr)
s_p100_price <- read_csv("s_p100_price_adj.csv",show_col_types = FALSE)

## Transform data according to researcher needs
dataSP <- log(s_p100_price[,seq(2,dim(s_p100_price)[2])])

## Turn data frame into numeric matrix to match function requirements
dataSP <- as.matrix(dataSP)

## Use the package documentation by calling help
?largevar

## Use largevar function
### Save the function output (list)
```

```
result <- largevar(data=dataSP,k=1,r=1,fin_sample_corr = FALSE,
     plot_output=TRUE,significance_level=0.05)

### Display the result
result
```

Since we set `plot_output=TRUE`, we obtain a histogram of eigenvalues for the matrix in equation (16) in Bykhovskaya and Gorin [2024]:



The output of `largevar()` is displayed in the Console as:

```
Output for the largevars function
===================================
Cointegration test for high-dimensional VAR(k)                    T=521 , N=92

 10% Critical value    5% Critical value    1% Critical value    Test stat.
 0.45                  0.98                 2.02                 -0.28

If the test statistic is larger than the quantile, reject H0 at the chosen level.
=============================================================================
Test statistic: -0.2777314
The p-value is 0.23
Decision about H0: 0
```

If we want to individually access certain values from the output list, we can do it in the usual way, by referencing the elements of the list:

```
> result$statistic
[1] -0.2777314
> result$significance_test$p_value
[1] 0.23
> result$significance_test$boolean_decision
[1] 0
> result$significance_test$significance_table

       0.90   0.95   0.97   0.99  Test stat.
r=1    0.45   0.98   1.33   2.02 -0.2777314
r=2   -1.87  -1.09  -0.57   0.42 -1.4995879
r=3   -5.90  -4.90  -4.24  -2.99 -5.4154889
r=4  -11.35 -10.15  -9.37  -7.87 -10.5527603
r=5  -18.07 -16.69 -15.79 -14.07 -16.7460847
r=6  -25.95 -24.40 -23.38 -21.45 -23.2178976
r=7  -34.90 -33.19 -32.07 -29.95 -31.1080001
r=8  -44.88 -43.01 -41.79 -39.47 -39.3197363
r=9  -55.82 -53.80 -52.48 -49.99 -49.8419822
r=10 -67.70 -65.53 -64.12 -61.45 -60.4894485
```

```
## sim_function for empirical p-values
result2 <- sim_function(N=92,tau=522,stat_value=-0.2777,k=1,r=1,
                        fin_sample_corr = FALSE,sim_num=1000)
> result2

Output for the sim_function function
===================================
The empirical p-value is 0.245
```

4.2. **Simulation example.** We present an example based on simulated data that users can replicate. The code below generates VAR(2) with $N = 100$, $T = 700$, and

(1)

$$\begin{aligned}
\begin{pmatrix}\Delta X_{1t}\\ \Delta X_{2t}\end{pmatrix} &= \begin{pmatrix}-0.9 & 0.8\\ 0 & 0\end{pmatrix}\begin{pmatrix}X_{1t-2}\\ X_{2t-2}\end{pmatrix} + \begin{pmatrix}-0.7 & 0.8\\ 0 & 0.3\end{pmatrix}\begin{pmatrix}\Delta X_{1t-1}\\ \Delta X_{2t-1}\end{pmatrix} + \begin{pmatrix}\varepsilon_{1t}\\ \varepsilon_{2t}\end{pmatrix},\ t = 1,\dots,T,\\
\begin{pmatrix}\Delta X_{4t}\\ \Delta X_{5t}\end{pmatrix} &= \begin{pmatrix}-0.9 & 0.8\\ 0 & 0\end{pmatrix}\begin{pmatrix}X_{4t-2}\\ X_{5t-2}\end{pmatrix} + \begin{pmatrix}-1.2 & 0.8\\ 0 & 0.25\end{pmatrix}\begin{pmatrix}\Delta X_{4t-1}\\ \Delta X_{5t-1}\end{pmatrix} + \begin{pmatrix}\varepsilon_{4t}\\ \varepsilon_{5t}\end{pmatrix},\ t = 1,\dots,T,\\
\Delta X_{it} &= \varepsilon_{it},\ i \neq 1,2,4,5,\ t = 1,\dots,T,
\end{aligned}$$

where $\Delta X_{it} := X_{it} - X_{it-1}$. The process is initialized by vectors X_0, X_{-1} with independent standard normal coordinates. The data generating process (1) corresponds to a matrix Π of rank 2: Π has two nonzero and linearly independent rows. To be more precise, the coefficient matrices in Eq. (1) correspond to $N-2$ unit root and 2 stationary components.

```
## simulated data
T <- 700
N <- 100
k <- 2
Pi <- matrix(0, N, N)
Pi[1:5,1:5] <- matrix(c(-0.9,rep(0,4),0.8,rep(0,12),-0.9,rep(0,4),0.8,0),5,5)
Gamma <- matrix(0, N, N)
Gamma[1:5,1:5] <- matrix(c(-0.7,rep(0,4),0.8,0.3,rep(0,11),
                        -1.2,rep(0,4),0.8,0.25),5,5)

dX <- matrix(0, N, T)
Xminus1 <- matrix(rnorm(N),N,1)
X0 <- matrix(rnorm(N),N,1)
dX0 <- X0-Xminus1
```

```
epsilon <- matrix(rnorm(N * T), N, T)

dX[,1] <- Pi %*% Xminus1 + Gamma %*% dX0+epsilon[,1]
dX[,2] <- Pi %*% X0 + Gamma %*% dX[,1] + epsilon[,2]
dX[,3] <- Pi %*% (X0+dX[,1]) + Gamma %*% dX[,2] + epsilon[,3]
for (t in 4:T) {
  dX[,t] <- Pi %*% (X0+rowSums(dX[,1:(t-2)])) + Gamma %*% dX[,t-1] + epsilon[,t]
}

data_sim <- matrix(0, N, T+1)
data_sim[,1] <- X0
for (t in 2:(T+1)) {
  data_sim[,t] <- data_sim[,t-1]+dX[,t-1]
}
data_sim <- t(data_sim)

## apply cointegration test
result <- largevar(data=data_sim, k=2, r=2, fin_sample_corr = FALSE,
                   plot_output=TRUE, significance_level=0.05)
> result

Output for the largevars function
===================================
Cointegration test for high-dimensional VAR(k)                 T=700 , N=100

 10% Critical value    5% Critical value    1% Critical value    Test stat.
 -1.87                 -1.09                 0.42                1.15

If the test statistic is larger than the quantile, reject H0 at the chosen level.
=============================================================================
Test statistic: 1.148357
The p-value is less than 0.01
Decision about H0: 1
```

If we want to take a look at how the significance of our test statistics vary across different choices of r, we can call the simulation table. The p-values for our test statistics always stay below the 0.1 and are below 0.01 for $r > 1$.

```
> result$significance_test$significance_table
       0.90    0.95    0.99    Test stat.
r=1    0.45    0.98    2.02    0.6026894
r=2   -1.87   -1.09    0.42    1.1483570
r=3   -5.90   -4.90   -2.99   -1.6981928
r=4  -11.35  -10.15   -7.87   -5.9794140
r=5  -18.07  -16.69  -14.07  -11.1455191
r=6  -25.95  -24.40  -21.45  -18.0542371
r=7  -34.90  -33.19  -29.95  -26.2418961
r=8  -44.88  -43.01  -39.47  -36.2312085
r=9  -55.82  -53.80  -49.99  -46.8432531
r=10 -67.70  -65.53  -61.45  -58.6279645
```

5. Simulation of the Airy$_1$ process and tables

In this section we describe the algorithm used to compute the quantiles for the function `largevar()`. The algorithm takes significant time to carry out, and rather than performing it in each run of the code, we include the output tables inside the package and in this section. These tables are used inside `largevar()` for obtaining the quantiles.

The Airy$_1$ point process is a random infinite sequence of reals

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

that can be defined through the following proposition, where $*$ is the matrix transposition.

Proposition 1 (Forrester [1993], Tracy and Widom [1996]). *Let X_N be an $N \times N$ 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,*

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

The eigenvalues of $\frac{1}{2}(X_N + X_N^*)$ coincide with those of a real symmetric $N \times N$ tridiagonal matrix, as discussed in Dumitriu and Edelman [2002]:

$$\frac{1}{\sqrt{2}}\begin{pmatrix} \mathcal{N}(0,2) & \chi_{n-1} & 0 & & & 0 \\ \chi_{n-1} & \mathcal{N}(0,2) & \chi_{n-2} & & & \\ 0 & \chi_{n-2} & \mathcal{N}(0,2) & & & \\ & & & \ddots & & \\ & & & & \mathcal{N}(0,2) & \chi_1 \\ 0 & & & & \chi_1 & \mathcal{N}(0,2) \end{pmatrix}, \tag{3}$$

where all matrix elements on or above diagonal are independent, $\mathcal{N}(0,2)$ is a normal distribution with mean 0 and variance 2, and χ_ℓ is a square root of a chi-squared distribution with ℓ degrees of freedom.

Moreover, instead of looking at the eigenvalues of a large $N \times N$ symmetric tridiagonal matrix (3), one can look at the eigenvalues of its top-left $\sqrt{N} \times \sqrt{N}$ submatrix. The largest eigenvalues of these two matrices have the same asymptotic distribution; see Edelman and Persson [2005, Section 1.1] and Johnstone et al. [2021, Lemma 5.2].

In our simulations we take advantage of this result and run 10^7 Monte Carlo simulations for $10^4 \times 10^4$ symmetric tridiagonal random matrices (in the form that corresponds to the top-left corner of the matrix in (1) of size 10^8). This is asymptotically equivalent to having run simulations on matrices of size 10^8.

The standard deviations of our results suggest that the error is at most ± 1 in the third digit of the elements of the Airy_1 sequence, meaning that the error is ± 0.01 for $r = 1$ and ± 0.1 for $r = 10$.

The tables below present our simulation results. The $0.ab$ quantile in each table corresponds to the row $0.a$ and the column b.

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-3.90	-3.61	-3.43	-3.30	-3.18	-3.08	-3.00	-2.92	-2.85
0.1	-2.78	-2.72	-2.67	-2.61	-2.56	-2.51	-2.46	-2.41	-2.37	-2.33
0.2	-2.29	-2.24	-2.20	-2.17	-2.13	-2.09	-2.05	-2.02	-1.98	-1.95
0.3	-1.91	-1.88	-1.84	-1.81	-1.78	-1.74	-1.71	-1.68	-1.65	-1.62
0.4	-1.58	-1.55	-1.52	-1.49	-1.46	-1.43	-1.40	-1.36	-1.33	-1.30
0.5	-1.27	-1.24	-1.21	-1.17	-1.14	-1.11	-1.08	-1.05	-1.01	-0.98
0.6	-0.95	-0.91	-0.88	-0.85	-0.81	-0.78	-0.74	-0.71	-0.67	-0.63
0.7	-0.59	-0.56	-0.52	-0.48	-0.44	-0.39	-0.35	-0.31	-0.26	-0.22
0.8	-0.17	-0.12	-0.07	-0.01	0.04	0.10	0.16	0.23	0.30	0.37
0.9	0.45	0.53	0.63	0.73	0.85	0.98	1.14	1.33	1.60	2.02

TABLE 1. Quantiles of $\mathfrak{a}_1$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-8.93	-8.44	-8.12	-7.88	-7.69	-7.52	-7.37	-7.24	-7.12
0.1	-7.01	-6.91	-6.81	-6.72	-6.63	-6.54	-6.46	-6.39	-6.31	-6.24
0.2	-6.17	-6.10	-6.04	-5.97	-5.91	-5.85	-5.79	-5.73	-5.67	-5.61
0.3	-5.56	-5.50	-5.45	-5.39	-5.34	-5.29	-5.23	-5.18	-5.13	-5.08
0.4	-5.03	-4.97	-4.92	-4.87	-4.82	-4.77	-4.72	-4.67	-4.62	-4.57
0.5	-4.52	-4.47	-4.42	-4.37	-4.32	-4.27	-4.22	-4.17	-4.11	-4.06
0.6	-4.01	-3.96	-3.91	-3.85	-3.80	-3.74	-3.69	-3.63	-3.57	-3.52
0.7	-3.46	-3.40	-3.34	-3.27	-3.21	-3.15	-3.08	-3.01	-2.94	-2.87
0.8	-2.80	-2.72	-2.65	-2.57	-2.48	-2.39	-2.30	-2.20	-2.10	-1.99
0.9	-1.87	-1.75	-1.61	-1.46	-1.29	-1.09	-0.86	-0.57	-0.19	0.42

TABLE 2. Quantiles of $\mathfrak{a}_1 + \mathfrak{a}_2$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-15.2	-14.6	-14.1	-13.8	-13.5	-13.3	-13.1	-12.9	-12.8
0.1	-12.6	-12.5	-12.4	-12.2	-12.1	-12.0	-11.9	-11.8	-11.7	-11.6
0.2	-11.5	-11.4	-11.3	-11.3	-11.2	-11.1	-11.0	-10.9	-10.9	-10.8
0.3	-10.7	-10.6	-10.6	-10.5	-10.4	-10.3	-10.3	-10.2	-10.1	-10.1
0.4	-10.0	-9.93	-9.87	-9.80	-9.73	-9.67	-9.60	-9.54	-9.47	-9.40
0.5	-9.34	-9.27	-9.21	-9.14	-9.07	-9.01	-8.94	-8.87	-8.80	-8.74
0.6	-8.67	-8.60	-8.53	-8.46	-8.39	-8.32	-8.25	-8.17	-8.10	-8.02
0.7	-7.95	-7.87	-7.79	-7.71	-7.63	-7.55	-7.46	-7.37	-7.28	-7.19
0.8	-7.10	-7.00	-6.90	-6.79	-6.68	-6.57	-6.45	-6.33	-6.19	-6.05
0.9	-5.90	-5.74	-5.56	-5.37	-5.15	-4.90	-4.60	-4.24	-3.76	-2.99

TABLE 3. Quantiles of $\sum_{i=1}^{3} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-22.7	-21.9	-21.4	-21.0	-20.7	-20.4	-20.1	-19.9	-19.7
0.1	-19.5	-19.3	-19.2	-19.0	-18.9	-18.8	-18.6	-18.5	-18.4	-18.3
0.2	-18.2	-18.0	-17.9	-17.8	-17.7	-17.6	-17.5	-17.4	-17.3	-17.3
0.3	-17.2	-17.1	-17.0	-16.9	-16.8	-16.7	-16.6	-16.6	-16.5	-16.4
0.4	-16.3	-16.2	-16.1	-16.1	-16.0	-15.9	-15.8	-15.7	-15.7	-15.6
0.5	-15.5	-15.4	-15.3	-15.3	-15.2	-15.1	-15.0	-14.9	-14.8	-14.8
0.6	-14.7	-14.6	-14.5	-14.4	-14.4	-14.3	-14.2	-14.1	-14.0	-13.9
0.7	-13.8	-13.7	-13.6	-13.5	-13.4	-13.3	-13.2	-13.1	-13.0	-12.9
0.8	-12.8	-12.7	-12.6	-12.4	-12.3	-12.2	-12.0	-11.9	-11.7	-11.5
0.9	-11.4	-11.2	-11.0	-10.7	-10.5	-10.2	-9.80	-9.37	-8.79	-7.87

TABLE 4. Quantiles of $\sum_{i=1}^{4} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-31.3	-30.3	-29.7	-29.2	-28.9	-28.5	-28.2	-28.0	-27.8
0.1	-27.5	-27.4	-27.2	-27.0	-26.8	-26.7	-26.5	-26.4	-26.2	-26.1
0.2	-26.0	-25.8	-25.7	-25.6	-25.5	-25.3	-25.2	-25.1	-25.0	-24.9
0.3	-24.8	-24.7	-24.6	-24.5	-24.4	-24.3	-24.2	-24.1	-24.0	-23.9
0.4	-23.8	-23.7	-23.6	-23.5	-23.4	-23.3	-23.2	-23.1	-23.1	-23.0
0.5	-22.9	-22.8	-22.7	-22.6	-22.5	-22.4	-22.3	-22.2	-22.1	-22.0
0.6	-21.9	-21.8	-21.7	-21.6	-21.5	-21.4	-21.3	-21.2	-21.1	-21.0
0.7	-20.9	-20.8	-20.7	-20.6	-20.5	-20.4	-20.2	-20.1	-20.0	-19.9
0.8	-19.7	-19.6	-19.5	-19.3	-19.2	-19.0	-18.8	-18.7	-18.5	-18.3
0.9	-18.1	-17.9	-17.6	-17.3	-17.0	-16.7	-16.3	-15.8	-15.1	-14.1

TABLE 5. Quantiles of $\sum_{i=1}^{5} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-40.9	-39.8	-39.1	-38.6	-38.1	-37.8	-37.4	-37.2	-36.9
0.1	-36.7	-36.4	-36.2	-36.0	-35.8	-35.6	-35.5	-35.3	-35.1	-35.0
0.2	-34.8	-34.7	-34.6	-34.4	-34.3	-34.2	-34.0	-33.9	-33.8	-33.7
0.3	-33.5	-33.4	-33.3	-33.2	-33.1	-33.0	-32.9	-32.7	-32.6	-32.5
0.4	-32.4	-32.3	-32.2	-32.1	-32.0	-31.9	-31.8	-31.7	-31.6	-31.5
0.5	-31.4	-31.3	-31.1	-31.0	-30.9	-30.8	-30.7	-30.6	-30.5	-30.4
0.6	-30.3	-30.2	-30.1	-30.0	-29.9	-29.7	-29.6	-29.5	-29.4	-29.3
0.7	-29.1	-29.0	-28.9	-28.8	-28.7	-28.5	-28.4	-28.3	-28.1	-28.0
0.8	-27.8	-27.7	-27.5	-27.3	-27.2	-27.0	-26.8	-26.6	-26.4	-26.2
0.9	-29.0	-25.7	-25.4	-25.1	-24.8	-24.4	-23.9	-23.4	-22.6	-21.5

TABLE 6. Quantiles of $\sum_{i=1}^{6} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-51.5	-50.3	-49.5	-48.9	-48.4	-48.0	-47.6	-47.3	-47.0
0.1	-46.8	-46.5	-46.3	-46.1	-45.8	-45.6	-45.5	-45.3	-45.1	-44.9
0.2	-44.8	-44.6	-44.4	-44.3	-44.1	-44.0	-43.9	-43.7	-43.6	-43.4
0.3	-43.3	-43.2	-43.0	-42.9	-42.8	-42.7	-42.5	-42.4	-42.3	-42.2
0.4	-42.1	-41.9	-41.8	-41.7	-41.6	-41.5	-41.4	-41.2	-41.1	-41.0
0.5	-40.9	-40.8	-40.7	-40.5	-40.4	-40.3	-40.2	-40.1	-40.0	-39.8
0.6	-39.7	-39.6	-39.5	-39.4	-39.2	-39.1	-39.0	-38.8	-38.7	-38.6
0.7	-38.5	-38.3	-38.2	-38.0	-37.9	-37.8	-37.6	-37.5	-37.3	-37.3
0.8	-37.0	-36.8	-36.6	-36.4	-36.3	-36.1	-35.9	-35.6	-35.4	-35.2
0.9	-34.9	-34.6	-34.3	-34.0	-33.6	-33.2	-32.7	-32.1	-31.2	-30.0

TABLE 7. Quantiles of $\sum_{i=1}^{7} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-63.0	-61.7	-60.8	-60.2	-59.7	-59.2	-58.8	-58.5	-58.1
0.1	-57.8	-57.6	-57.3	-57.1	-56.8	-56.6	-56.4	-56.2	-56.0	-55.8
0.2	-55.6	-55.5	-55.3	-55.1	-55.0	-54.8	-54.7	-54.5	-54.4	-54.2
0.3	-54.1	-53.9	-53.8	-53.6	-53.5	-53.4	-53.2	-53.1	-53.0	-52.8
0.4	-52.7	-52.6	-52.4	-52.3	-52.2	-52.0	-51.9	-51.8	-51.7	-51.5
0.5	-51.4	-51.3	-51.2	-51.0	-50.9	-50.8	-50.6	-50.5	-50.4	-50.3
0.6	-50.1	-50.0	-49.9	-49.7	-49.6	-49.5	-49.3	-49.2	-49.0	-48.9
0.7	-48.8	-48.6	-48.5	-48.3	-48.1	-48.0	-47.8	-47.7	-47.5	-47.3
0.8	-47.1	-47.0	-46.8	-46.6	-46.4	-46.1	-45.9	-45.7	-45.4	-45.2
0.9	-44.9	-44.6	-44.2	-43.9	-43.5	-43.0	-42.5	-41.8	-40.9	-39.5

TABLE 8. Quantiles of $\sum_{i=1}^{8} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-75.5	-74.0	-73.1	-72.4	-71.8	-71.3	-70.9	-70.5	-70.2
0.1	-69.9	-69.6	-69.3	-69.0	-68.8	-68.5	-68.3	-68.1	-67.9	-67.7
0.2	-67.5	-67.3	-67.1	-66.9	-66.7	-66.6	-66.4	-66.2	-66.1	-65.9
0.3	-65.7	-65.6	-65.4	-65.3	-65.1	-65.0	-64.8	-64.7	-64.6	-64.4
0.4	-64.3	-64.1	-64.0	-63.8	-63.7	-63.6	-63.4	-63.3	-63.2	-63.0
0.5	-62.9	-62.7	-62.6	-62.5	-62.3	-62.2	-62.1	-61.9	-61.8	-61.6
0.6	-61.5	-61.4	-61.2	-61.1	-60.9	-60.8	-60.6	-60.5	-60.3	-60.2
0.7	-60.0	-59.9	-59.7	-59.5	-59.4	-59.2	-59.0	-58.8	-58.6	-58.5
0.8	-58.3	-58.1	-57.9	-57.6	-57.4	-57.2	-56.9	-56.7	-56.4	-56.1
0.9	-55.8	-55.5	-55.1	-54.7	-54.3	-53.8	-53.2	-52.5	-51.5	-50.0

TABLE 9. Quantiles of $\sum_{i=1}^{9} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

q	0	1	2	3	4	5	6	7	8	9
0.0	$-\infty$	-88.8	-87.2	-86.2	-85.5	-84.9	-84.3	-83.9	-83.5	-83.1
0.1	-82.8	-82.4	-82.1	-81.8	-81.6	-81.3	-81.1	-80.8	-80.6	-80.4
0.2	-80.2	-80.0	-79.8	-79.6	-79.4	-79.2	-79.0	-78.9	-78.7	-78.5
0.3	-78.3	-78.2	-78.0	-77.8	-77.7	-77.5	-77.4	-77.2	-77.1	-76.9
0.4	-76.8	-76.6	-76.5	-76.3	-76.2	-76.0	-75.9	-75.7	-75.6	-75.4
0.5	-75.3	-75.1	-75.0	-74.8	-74.7	-74.5	-74.4	-74.2	-74.1	-73.9
0.6	-73.8	-73.6	-73.5	-73.3	-73.2	-73.0	-72.8	-72.7	-72.5	-72.4
0.7	-72.2	-72.0	-71.8	-71.7	-71.5	-71.3	-71.1	-70.9	-70.7	-70.5
0.8	-70.3	-70.1	-69.9	-69.6	-69.4	-69.2	-68.9	-68.6	-68.3	-68.0
0.9	-67.7	-67.3	-67.0	-66.5	-66.1	-65.5	-64.9	-64.1	-63.1	-61.5

TABLE 10. Quantiles of $\sum_{i=1}^{10} \mathfrak{a}_i$ (based on 10^7 Monte Carlo simulations of $10^8 \times 10^8$ tridiagonal matrices of Dumitriu and Edelman [2002]).

References

A. Bejan. Largest eigenvalues and sample covariance matrices. tracy-widom and painlevé ii: computational aspects and realization in s-plus with applications. *Preprint:* `http://users.stat.umn.edu/~jiang040/downloadpapers/largesteigen/largesteigen.pdf`, 2005.

A. Bykhovskaya and V. Gorin. Cointegration in large VARs. *The Annals of Statistics*, 50 (3):1593–1617, 2022.

A. Bykhovskaya and V. Gorin. Asymptotics of cointegration tests for high-dimensional VAR(k). *Review of Economics and Statistics*, 2024.

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

A. Edelman and P.-O. Persson. Numerical methods for eigenvalue distributions of random matrices. *arXiv preprint math-ph/0501068*, 2005.

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

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

M. Ho 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*, 78(4):726–732, 1996.

S. Johansen. Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control*, 12(2–3):231–254, 1988.

S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

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

I. M. Johnstone, Y. Klochkov, A. Onatski, and D. Pavlyshyn. Spin glass to paramagnetic transition in spherical Sherrington-Kirkpatrick model with ferromagnetic interaction. *arXiv preprint arXiv:2104.07629*, 2021.

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

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

(Anna Bykhovskaya) Duke University

Email address: anna.bykhovskaya@duke.edu

(Vadim Gorin) University of California at Berkeley

Email address: vadicgor@gmail.com

(Eszter Kiss) Duke University

Email address: ekiss2803@gmail.com