Broken-stick components retention rule for equi-correlated normal population

Yohji Akama[*]

December 7, 2023

Abstract

We consider a p-dimensional normal population such that the different variables have a common correlation coefficient $\rho \in [0, 1)$. We call such a population an *equi-correlated normal population* (ENP). We draw a sample of size n from an ENP, and form a sample covariance matrix $\mathbf{S}$ and a sample correlation matrix $\mathbf{C}$. For a proportional limiting regime, we deduce that the second largest eigenvalue of $\mathbf{S}$ converges in distribution to the type-1 Tracy-Widom distribution. From this, the number of principal components (PCs) the Frontier's broken-stick rule takes from $\mathbf{S}$ converges almost surely to 0 for $\rho = 0$, and 1 for $0 < \rho < 1$. For the limiting distributions, we compare the second largest eigenvalue of $\mathbf{C}$ of ENP, to the eigenvalues of the sample correlation matrix of a bounded spiked eigenvalues model.

1 Introduction

In multivariate statistical analysis, we often employ principal component analysis (PCA) ([20],[24]) and factor analysis (FA) ([22],[33]) to find principal components (PCs) and factors to illustrate a great part of the variance or the correlation in the p-dimensional data by using the sample covariance matrix $\mathbf{S}$ or the sample correlation matrix $\mathbf{C}$ of that data. The PCs and the factors are "significant" if the corresponding eigenvalues of $\mathbf{S}$ or $\mathbf{C}$ of the data are large.

Estimating the number of significant PCs and factors in PCA and FA are important in finance ([11, 12],[30],[13]), biology ([35]), engineering ([39]), psychometrics [25], and so on. *Components retention rules* are methods to determine the number q of the most important PCs. Components retention rules not only balance the accuracy (or fit) of the model with ease of analysis and the potential loss of information, but also identify the cause of phenomena. Many components retention rules have been proposed and compared in [20],[24],[34],[8],[15], to cite a few.

In [2], the author and Husnaqilati handled Kaiser's *intercorrelation* [26], by assuming an *equi-correlated normal population* (ENP), that is, the multinormal population such that the variables have a common correlation coefficient $\rho \geq 0$. Let us call ρ a *population equi-correlation coefficient* (PECC). [2] assumed that both the data dimension p and the sample size n tend to infinity with $p/n \to c > 0$, and identified the limiting spectral distribution of the sample correlation matrix $\mathbf{C}$. Then, they studied the limiting behavior of Kaiser's rule in the limit $c \to 0$, and explained Kaiser's observation [27] as the discontinuity of the limiting value at $\rho = 0$. The phase transitions of various components retention rules concerning the PECC ρ could be interesting from the

[*] Mathematical Institute, Graduate School of Science, Tohoku University, Sendai, 980-8578, Japan. yoji.akama.e8@tohoku.ac.jp

2023 年度科学研究費シンポジウム International Symposium on Recent Advances in Theories and Methodologies for Large Complex Data (2023年12月7–9日)

Keywords: Equi-correlated normal population, sample covariance matrix, the 2nd largest eigenvalue.

viewpoint of random matrix theory in relation to the recent study ([31],[7]) on a sample covariance matrix $\mathbf{S}$ having the *unbounded* largest eigenvalue (See [4] for bounded largest eigenvalue case). Above all, an ENP is related to intraclass correlation [23].

The *broken-stick rule* (Frontier [14], Jackson [19]) is a components retention rule motivated by the study of MacArthur [29] on animal community structure. Let $\mathrm{BS}(\mathbf{C})$ ($\mathrm{BS}(\mathbf{S})$, resp.) be the number of PCs of a sample correlation matrix $\mathbf{C} \in \mathbb{R}^{p\times p}$ (sample covariance matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$, resp.) that the broken-stick rule retains. The i-th ($1 \le i \le p$) largest eigenvalue of a symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ is written by $\lambda_i(\mathbf{M})$.

$$\mathrm{BS}(\mathbf{C}) := \inf\left\{ i \in [1,\, p] \;\middle|\; \lambda_i(\mathbf{C}) \le \sum_{k=i}^{p} k^{-1} \right\} - 1. \tag{1}$$

By [18, Theorem 2.2], for i ($1 \le i \le p$), $p^{-1}\sum_{k=i}^{p} k^{-1}$ is the expectation of the length of the i-th longest subinterval of $I = [0,\, 1]$ divided by $p-1$ points independently obeying the uniform distribution on I. Thus $\mathrm{BS}(\mathbf{C}) < \infty$, because otherwise $\lambda_i(\mathbf{C}) > \sum_{k=i}^{p} k^{-1}$ ($1 \le i \le p$) which implies $p = \operatorname{trace}\mathbf{C} = \sum_{i=1}^{p}\lambda_i(\mathbf{C}) > \sum_{i=1}^{p}\sum_{k=i}^{p} k^{-1} = p$. For $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\mathbf{0},\, \mathbf{\Sigma}_p(\rho))$ and $\rho \in [0,\, 1)$, letting $\inf\emptyset = \infty$, we define $\mathrm{BS}(\mathbf{S})$ to be (1) with $\mathbf{C}$ replaced by $\mathbf{S}$.

Let $\overset{a.s.}{\to}$ and $\overset{P}{\to}$ be the almost sure convergence and the convergence in probability, respectively. For an ENP and the limiting regime $n, p \to \infty, p/n \to c > 0$, we will prove $\mathrm{BS}(\mathbf{C}) \overset{a.s.}{\to} 0$ for $\rho = 0$, $0 < \liminf \mathrm{BS}(\mathbf{C})$ (a.s.) for $0 < \rho < 1$, $\mathrm{BS}(\mathbf{S}) \overset{a.s.}{\to} 0$ for $\rho = 0$, and $\mathrm{BS}(\mathbf{S}) \overset{P}{\to} 1$ for $0 < \rho < 1$.

This paper is organized as follows: The next section proves above results by studying the limiting distribution of $\lambda_2(\mathbf{S})$ of an ENP. Section 3 analyzes the behaviors of the broken-stick rule and various components retention rules, for real datasets of S&P 500 return time-series, Fama-French 100 portfolios time-series, and binary multiple sequence alignment [35]. Section 4 does simulation study for the limiting distribution of $\lambda_2(\mathbf{C})$ of an ENP, in comparison to the theoretical study [32] of a bounded spiked eigenvalues model.

2 Limiting behavior of broken-stick rule for sample covariance matrix of equi-correlated normal populations

Below $\boldsymbol{\mu} \in \mathbb{R}^p$, $\sigma^2 > 0$, $\rho \in [0,\, 1)$ and $c > 0$ are deterministic, and $\mathbf{\Sigma}_p(\rho) = (1-\rho)\mathbf{I}_p + [\rho] \in \mathbb{R}^{p\times p}$. We derive $\lim\limits_{\substack{p,n\to\infty \\ p/n\to c}} \lambda_1(\mathbf{S})/p = \rho$ (a.s.) for the population $\mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}_p(\rho))$ from [41], [31, Theorem 2.1], [7, Theorem 2.1] and so on.

Proposition 2.1 ([1])**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\boldsymbol{\mu},\, \mathbf{D}\mathbf{\Sigma}_p(\rho)\mathbf{D})$ with $\mathbf{D} \in \mathrm{GL}_p(\mathbb{R})$ being deterministic and diagonal. Then $\lim\limits_{\substack{p,n\to\infty \\ p/n\to c}} \lambda_1(\mathbf{C})/p = \rho$ (a.s.).*

Guttman-Kaiser rule ([17],[25],[42]) is a components retention rule that retains the eigenvalues $\lambda_i(\mathbf{M})$ where $\mathbf{M} \in \mathbb{R}^{p\times p}$ is a sample covariance matrix $\mathbf{S}$ or a sample correlation matrix $\mathbf{C}$. Let $\mathrm{GK}(\mathbf{M}) := \#\{\, i \in [1,\, p] \mid \lambda_i(\mathbf{M}) \ge \operatorname{trace}\mathbf{M}/p \,\}$.

Proposition 2.2 ([2])**.** *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\boldsymbol{\mu},\, \mathbf{D}\mathbf{\Sigma}_p(\rho)\mathbf{D})$ with $\mathbf{D} \in \mathrm{GL}_p(\mathbb{R})$ being deterministic and diagonal. If ($\mathbf{M} = \mathbf{S}$, $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{D} = \mathbf{I}_p$) or $\mathbf{M} = \mathbf{C}$, then $\lim_{\substack{p,n\to\infty \\ p/n\to c}} \mathrm{GK}(\mathbf{M})/p$ almost surely converges as $c \to 0$ to $1/2$ ($\rho = 0$); 0 ($\rho > 0$).*

In contrast, $\lim_{\substack{p,n\to\infty \\ p/n\to c}} \mathrm{BS}(\mathbf{C})$ increases from 0 as ρ increases from 0.

Theorem 2.3. *Suppose* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\mathbf{0},\, \sigma^2 \boldsymbol{\Sigma}_p(\rho))$ *and* $n, p \to \infty, p/n \to c$. *Then* $\mathrm{BS}(\mathbf{S}) \overset{a.s.}{\to} 0$ $(\rho = 0)$ *and* $\mathrm{BS}(\mathbf{S}) \overset{P}{\to} 1$ $(\rho > 0)$.

The proof of the case $\rho = 0$ is by

Proposition 2.4 ([5, Theorem 3.1])**.** *If* $\{\, x_{ij} \mid i, j \geq 1 \,\}$ *are* i.i.d., $\mathrm{E}\, x_{11} = 0$, $\mathrm{E}\, x_{11}^2 = 1$, *and* $\mathrm{E}\,|x_{11}|^4 < \infty$, *then* $\lim_{\substack{p,n\to\infty \\ p/n\to c}} \lambda_1(\mathbf{S}) = (1 + \sqrt{c})^2$ *(a.s.)*.

We demonstrate Theorem 2.3 for the case $\rho > 0$ by establishing that the limiting distribution of the *second* largest eigenvalue of $\mathbf{S}$ is the type-1 Tracy-Widom distribution scaled by $1 - \rho$. For this, we use [7, Theorem 2.5] of which setting and the assumptions are reviewed below. Let the data matrix be

$$\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\Xi} \text{ where } \boldsymbol{\Xi} = [\xi_{ij}]_{1 \leq i \leq p+l,\ 1 \leq j \leq n} = [\boldsymbol{\xi}_1 \cdots \boldsymbol{\xi}_n] \in \mathbb{R}^{(p+l)\times n} \tag{2}$$

satisfying the following two conditions:

Assumption 1 ([7, Assumption 1])**.** $\{\, \boldsymbol{\xi}_j = [\xi_{1j} \cdots \xi_{p+l,j}]^\top \mid 1 \leq j \leq n \,\}$ *are i.i.d. random vectors.* $\{\, \xi_{ij} \mid 1 \leq i \leq p + l,\ 1 \leq j \leq n \,\}$ *are independent random variables such that* $\mathrm{E}(\xi_{ij}) = 0$, $\mathrm{E}\,|\xi_{ij}|^2 = 1$, *and* $\sup_{i,j} \mathrm{E}\,|\xi_{ij}|^4 < C$ *for some constant* C*.*

Assumption 2 ([7, Assumption 7])**.** $\sup_{i,j} \mathrm{E}\,|\xi_{ij}|^k < c_k$ *for some constant* c_k $(k \in \mathbb{N})$*.*

Let $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times(p+l)}$ be a deterministic matrix with $l/p \to 0$, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \in \mathbb{R}^{p\times p}$, $\ell_i = \lambda_i(\boldsymbol{\Sigma})$ $(1 \leq i \leq p)$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\ell_1, \ldots, \ell_p) \in \mathbb{R}^{p\times p}$. Let $\boldsymbol{\Gamma} = \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{U}$ be a singular value decomposition with $\mathbf{V} \in \mathbb{R}^{p\times p}$, $\mathbf{U} \in \mathbb{R}^{p\times(p+l)}$, and $\mathbf{V}\mathbf{V}^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_p$. For two sequences of positive numbers a_n and b_n, $a_n \gtrsim b_n$ means $a_n \geq c b_n$ for some constant $c > 0$. Let $a_n \sim b_n$ be $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$. Let $a_n \ll b_n$ be $a_n/b_n \to 0$.

Let a deterministic integer $K > 0$ satisfy the following two assumptions:

Assumption 3 ([7, Assumption 2])**.** $p \gtrsim n$, $d_i := p/(n\ell_i) \to 0$ $(n, p \to \infty)$ *for* $i = 1, 2, \ldots, K$, $\max_{K+1 \leq i \leq p} \ell_i$ *is bounded by a constant* C, $Kn^{-1/6} \to 0$, *and* $K^2 d_K \to 0$.

Assumption 4 ([7, Assumption 3])**.** *There is a constant* s *not depending on* n *such that* $\ell_{i-1}/\ell_i \geq s > 1$, $i = 2, \ldots, K$.

Let $\boldsymbol{\Lambda}_S = \mathrm{diag}(\ell_1, \ldots, \ell_K)$, $\boldsymbol{\Lambda}_P = \mathrm{diag}(\ell_{K+1}, \ldots, \ell_p)$, and

$$\boldsymbol{\Sigma}_1 = \mathbf{U}_2^\top \boldsymbol{\Lambda}_P \mathbf{U}_2 \in \mathbb{R}^{(p+l)\times(p+l)}$$

where $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} \in \mathbb{R}^{p\times(p+l)}$ with $\mathbf{U}_1 \in \mathbb{R}^{K\times(p+l)}$ and $\mathbf{U}_2 \in \mathbb{R}^{(p-K)\times(p+l)}$. The *Stieltjes transform* of a finite measure μ on $\mathbb{R}$ is, by definition, $s_\mu(z) := \int \frac{d\mu(x)}{x-z}$ $(z \in \mathbb{C}\backslash \mathrm{supp}(\mu))$. Let $\mathbb{C}^+$ be the complex upper half plane. Finally, we consider the following assumption:

Assumption 5 ([7, Assumption 8])**.** *There are probability measures* $(\mu_n)_{n\in\mathbb{N}}$ *on* $\mathbb{R}$ *such that* $s_{\mu_n}(z) = -\left(z - n^{-1}\,\mathrm{trace}\left((\mathbf{I}_p + s_{\mu_n}(z)\boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\Sigma}_1\right)\right)^{-1}$ $(z \in \mathbb{C}^+,\ n \in \mathbb{N})$ *and*

$$\limsup_n \ell_{K+1} \cdot \left(-\lim_{z \in \mathbb{C}^+ \to t_n} s_{\mu_n}(z)\right) < 1 \quad \textit{where } t_n := \inf\{\, t \in \mathbb{R} \mid \mu_n((-\infty,\, t]) = 1 \,\}.$$

Let E be an event depending on n. Following [38], we say E *holds with high probability* if $\mathrm{P}(E) \geq 1 - O(n^{-c})$ for some constant $c > 0$ (independent of n).

Proposition 2.5 ([7, Theorem 2.5])**.** *Suppose Assumptions 1-5, $l \ll n^{1/6}$, and $p \sim n$. Let $\epsilon > 0$ be sufficiently small. Assume $1 \leq i - K \leq \log n$. Then, with high probability,*

$$\left| \lambda_i(\mathbf{S}) - \lambda_{i-K}\left(\frac{1}{n}\mathbf{\Xi}^\top \mathbf{\Sigma}_1 \mathbf{\Xi}\right) \right| \leq n^{-2/3-\epsilon}.$$

In particular, λ_{K+1} has limiting Type-1 Tracy-Widom distribution.

We will verify all the assumptions of Proposition 2.5 are satisfied by Theorem 2.3 with $l = 0$, $K = 1$, $i = 2$, $\mathbf{\Lambda}_S = (p-1)\rho + 1$, and $\mathbf{\Sigma}_1 = \mathbf{\Lambda}_P = (1-\rho)\mathbf{I}_p$. Then, the limiting distribution of the second largest eigenvalue $\lambda_2(\mathbf{S})$ of $\mathbf{S}$ will be that of $\lambda_1((1-\rho)n^{-1}\mathbf{\Xi}\mathbf{\Xi}^\top)$ for $\mathbf{\Xi}$ introduced in (2), so we will apply the following results of Soshnikov to $\mathbf{\Xi}$, where the convergence in distribution is denoted by $\xrightarrow{D}$ and

$$\mu_{n,p} := (n^{1/2} + p^{1/2})^2, \quad \sigma_{n,p} := (n^{1/2} + p^{1/2})(n^{-1/2} + p^{-1/2})^{1/3}.$$

Proposition 2.6 ([37, Theorem 1, Corollary 1])**.** *Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{p \times n}$ have independent random entries such that $\mathrm{E}\, x_{ij} = 0$, $\mathrm{E}\, x_{ij}^2 = 1$, $\mathrm{E}\, x_{ij}^k < c_k$ for some constant c_k $(k \in \mathbb{N})$, $\mathrm{E}(x_{ij})^{2m} \leq (\mathrm{const}\ m)^m$, and x_{ij} follow symmetric laws of distribution $(1 \leq i \leq p,\ 1 \leq j \leq n)$. Suppose $n, p \to \infty$, $p/n \to c$. Then $\frac{\lambda_1(\mathbf{X}\mathbf{X}^\top) - \mu_{n,p}}{\sigma_{n,p}} \xrightarrow{D} TW_1$. Moreover, $\lambda_1(\mathbf{X}\mathbf{X}^\top) \leq \mu_{n,p} + O\left(p^{1/2}\log p\right)$ (a.s.).*

Our case satisfies all the assumptions of Proposition 2.5. Assumptions 1 and 2 follow from the normality of the population. Assumption 3 is seen as follows: $K = 1$, $d_K = p/(n((p-1)\rho + 1)) \to 0$ $(n, p \to \infty,\ p/n \to c)$. Besides, $\ell_i = 1 - \rho$ $(i = 2, \ldots, p)$ are bounded, $Kn^{-1/6} \to 0$, and $K^2 d_K \to 0$. Assumption 4 is by $K = 1$. Finally, we check Assumption 5. By $\mathbf{\Sigma}_1 = (1-\rho)\mathbf{I}_p$, the equation of $s_{\mu_n}(z)$ is quadratic. The Marčenko-Pastur probability measure $\nu_{y,t}$ of the dimension-to-sample-size ratio $y > 0$ and the scale parameter $t > 0$, has a point mass $1 - y^{-1}$ at 0 for $y > 1$, and $s_{\nu_{y,t}}(z)$ is given in [40]. Then $\mu_n = (1 - p/n)\,\delta_0 + (p/n)\nu_{p/n,1-\rho}$, so $t_n = (1-\rho)(1 + \sqrt{p/n})^2$. Thus $\ell_{K+1} \cdot (-\lim_{z \in \mathbb{C}^+ \to t_n} s_{\mu_n}(z)) = (1 + \sqrt{p/n})^{-1}$, which assures all the assumptions of Proposition 2.5 for our case. Hence, by Proposition 2.5, the limiting distribution of the second largest eigenvalue $\lambda_2(\mathbf{S})$ of $\mathbf{S}$ is that of the largest eigenvalue $\lambda_1((1-\rho)n^{-1}\mathbf{\Xi}\mathbf{\Xi}^\top)$.

For $\xi \sim \mathrm{N}(0,\ \sigma^2)$, $\mathrm{E}\,\xi^{2k} \leq (2\sigma^2 k)^k$ and $\mathrm{E}\,\xi^{2k-1} = 0$ $(k = 1, 2, \ldots)$, so Proposition 2.6 implies:

Theorem 2.7. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\mathbf{0},\ \sigma^2\mathbf{\Sigma}_p(\rho))$, $\rho \in (0,\ 1)$, $\mathbf{X} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_n]$, and $n, p \to \infty$, $p/n \to c$. Then $\left(\frac{\lambda_2(\mathbf{X}\mathbf{X}^\top)}{1-\rho} - \mu_{n,p-1}\right)/\sigma_{n,p-1} \xrightarrow{D} TW_1$. Moreover*

$$\frac{\lambda_2(\mathbf{X}\mathbf{X}^\top)}{1-\rho} \leq \mu_{n,p-1} + O\left(\sqrt{p-1}\log(p-1)\right) \qquad (\textit{with high probability}).$$

Recall $\sum_{k=1}^{p} k^{-1} = \log p + \gamma + \epsilon_p$ where γ is the Euler constant and $\lim_{p\to\infty} \epsilon_p = 0$. From the last inequality of Theorem 2.7 and Proposition 2.1, Theorem 2.3 follows.

By the same proof argument,

Theorem 2.8. *In the setting of Proposition 2.5,* $\mathrm{BS}(\mathbf{S}) \leq K$ *with high probability.*

3 Broken-stick rule for real datasets

We consider the correlation matrices $\mathbf{C}$ of the following:

1. the dataset [1] obtained from S&P 500 return stock price time-series by removing many quotes having missing values during the period. The dataset is available via `https://zenodo.org/record/8253821`.

2. the datasets (Fan et al. [13]) obtained by cleaning outliers from the dataset of the daily excess returns [21] of Fama-French 100 portfolios [11, 12]. The latter dataset is from Prof. French's data library `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`.

3. binary multiple sequence alignment [35] by the courtesy of Prof. Quadeer.

We compute the n, p/n, $\lambda_1(\mathbf{C})/p$, p, $\mathrm{BS}(\mathbf{C})$, and the numbers $\mathrm{GK}(\mathbf{C})$, $\mathrm{CPV}(\mathbf{C})$, $\mathrm{ACT}(\mathbf{C})$ of PCs the Guttman-Kaiser rule, *Cumulative-Percentage-of-Variance rule* (Jolliffe [24]), and *Adjusted Correlation Thresholding* (Fan et al. [13]) retain from $\mathbf{C}$.

$$\mathrm{CPV}(\mathbf{C}) = \min\left\{ i \in [1,\, p] \,\middle|\, \sum_{k=1}^{i} \lambda_k(\mathbf{C})/p \ge .7 \right\}.$$

$$\mathrm{ACT}(\mathbf{C}) = \max\left\{ i \in [1,\, p] \,\middle|\, \left[\frac{1 - c_{i,n-1}}{\lambda_i(\mathbf{C})} - c_{i,n-1} m_{n,i}\right]^{-1} > 1 + \sqrt{c_{0,n-1}} \right\} \tag{3}$$

$$c_{i,n-1} = \frac{p-i}{n-1},\ m_{n,i} = \frac{1}{p-i}\left[\sum_{k=i+1}^{p} \frac{1}{\lambda_k(\mathbf{C}) - \lambda_i(\mathbf{C})} + \frac{4}{\lambda_{i+1}(\mathbf{C}) - \lambda_i(\mathbf{C})}\right] \quad (1 \le i \le p).$$

The left side of the inequality in (3) is due to Bai-Ding [3], and the ratio against the i-th largest spiked eigenvalue of the population correlation matrix converges in probability to 1 under a mild condition [13]. Besides, the threshold $1+\sqrt{c_{0,n-1}}$ in (3) is optimal [13].

3.1 S&P 500 return time-series

In [10], Engle and Kelly employed their *Dynamic Equicorrelation* to forecast time-series of economics. As in [1], we considered datasets of S&P 500 stock returns for the period 2012-01-04/2021-12-31 (n = 2516=trading days−1), for the 11 sectors (from each sector we choose p stocks without missing data) and the totality of the 11 sectors. We do not clean "outliers" for any days, as "outliers" may contribute to the unboundedness of $\lambda_1(\mathbf{C})$. The rows of Table 1 are listed in the increasing order of $\lambda_1(\mathbf{C})/p$. The 11 sectors have $\mathrm{BS}(\mathbf{C}) = 1$ probably because each of them is relatively homogeneous.

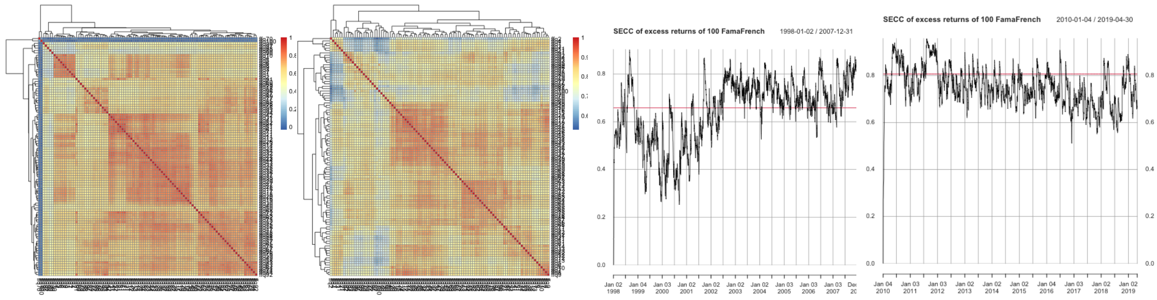
3.2 Fama-French 100 portfolios time-series

In asset pricing and portfolio management, Fama and French designed statistical models, 3-factor model [11] and 5-factor model [12], to describe stock returns. For the datasets of the daily excess returns of 100 companies French chose, Fan et al. [13] computed the estimator ACT and confirmed the three Fama-French factors [11] but not the momentum factor [9].

(A) Fan et al. [13] used the daily excess returns of 100 industrial portfolios formed on the basis of size and book-to-market ratio from January 2, 1998, to December 31, 2007. ("Before 2007-2008 financial crisis")

Sector	n	p/n	$\lambda_1(\mathbf{C})/p$	p	BS	ACT	GK	CPV
Communication Services	2516	.0076	.3571	19	1	3	4	7
Consumer Discretionary	2516	.0207	.3843	52	1	5	8	15
Health Care	2516	.0187	.3948	47	1	5	7	14
Consumer Staples	2516	.0091	.4302	23	1	3	4	7
Information Technology	2516	.0246	.4648	62	1	4	6	14
Industrials	2516	.0258	.4985	65	1	5	6	12
Materials	2516	.0095	.499	24	1	2	4	6
Real estate	2516	.0119	.5819	30	1	3	3	3
Financials	2516	.025	.6086	63	1	4	5	4
Energy	2516	.0064	.6872	16	1	1	1	2
Utilities	2516	.0111	.6897	28	1	2	2	2
The totality	2516	.1705	.3819	429	4	10	46	54

Table 1: The stock return datasets of 11 sectors and the totality (2012-01-04/2021-12-31).



(a) (A) **(b)** (B) **(c)** (A) **(d)** (B)

Figure 1: Heat maps of the correlation matrices and the SECC time-series of the daily excess returns of Fama-French 100 portfolios of (A) 1998-01-02/2007-12-31 and (B) 2010-01-04/2019-04-30. The red lines are $\lambda_1(\mathbf{C})/p$. See Subsection 3.2.

(B) The dataset for the daily excess returns of Fama-French 100 portfolios, from January 4, 2010, to April 30, 2019. ("After 2007-2008 financial crisis")

By Figure 1, the correlation matrix $\mathbf{C}_A$ of (A) has submatrices of larger positive correlation coefficients and/or larger sizes than the correlation matrix $\mathbf{C}_B$ of (B). So $\text{BS}(\mathbf{C}_A) = 2$ and $\text{BS}(\mathbf{C}_B) = 1$. These two datasets are clearly compared, by fitting them to Glosten-Jagannathan-Runkle's Generalized Autoregressive Conditional Heteroskedasticity model (GJR GARCH) [16], the distribution for the univariate estimation being normal, and the correlation model being the Dynamic Equicorrelation [10]. The resulting time-series of sample equi-correlation coefficient (SECC) of the daily excess returns are in Figure 1. The time-series (A) is lower than the time-series (B), as the SECC time-series of the daily returns before February 2020 [1] is lower than that of the daily returns after then; the worldwide stock markets crashed on February 2020 by growing uncertainty due to the Coronavirus disease 2019 pandemic. In Figure 1 (c), the time-series (A) is lower for 1999-2001 ("dot-com bubble") than for the other periods.

The time-series (A) consists of two time-series, and thus has $\text{BS}(\mathbf{C}) = 2$. the time-series (B) is less changing, so $\text{BS}(\mathbf{C}) = 1$. Table 2 shows that the broken-stick rule is smaller than the ACT and the Guttman-Kaiser rule is almost the same as the ACT.

For SECC time-series, the time-average $\overline{\rho}$ satisfies $\overline{\rho} \sim 1.02833\frac{\lambda_1(\mathbf{C})}{p} - 0.09152$ with adjusted R^2 being 0.9924, by the linear regression analysis for the dataset combining

Period	n	p/n	$\lambda_1(\mathbf{C})/p$	p	BS	ACT	GK	CPV	$\overline{\rho}$
(A)	2514	.0398	.6582	100	2	4	5	2	.6016
(B)	2346	.0426	.8062		1	3	3	2	.7511

Table 2: The daily excess returns of Fama-French 100 portfolios. (A) 1998-01-02/2007-12-31 and (B) 2010-01-04/2019-04-30. $\overline{\rho}$ is the time-average of the SECC time-series.

Table 2 and [1, Table 1].

We conjecture $\lim_{\substack{p,n\to\infty \\ p/n\to c}}(\overline{\rho}-\lambda_1(\mathbf{C})/p)=0$ (a.s.) for GJR GARCH+DECO model.

3.3 Binary multiple sequence alignment

We consider the binary multiple sequence alignment (MSA) [35] of a p-residue (site) protein with n sequences where $p = 475$ and $n = 2815$. In [35, p. 7628], Quadeer et al. did "identify groups of coevolving residues within HCV nonstructural protein 3 (NS3) by analyzing diverse sequences of this protein using ideas from random matrix theory and associated methods." They also found "Sequence analysis reveals three sectors of collectively evolving sites in NS3. ..., there remained $\alpha = 9$ eigenvalues greater than $\lambda_{\text{max}}^{\text{rnd}}$, presumably representing intrinsic correlations."[35, Section Results, p. 7631] They detected signals by a randomization from the data.

On the statistical model of [35], Morales-Jimenez, one of the authors of Quadeer [36], commented "the majority of variables (protein positions in the genome) are essentially independent, and there are just some small groups of variables which are correlated, giving rise to the different spikes. These group of variables can be modelled with equi-correlation, but the size of these groups is modelled as fixed, i.e., not growing with the dimension of the protein. That leads to a non-divergent spiked model, like the one considered in our Stat Sinica paper." [32].

Nonetheless, for the dataset of MSA [35], the broken-stick rule and the adjusted correlation thresholding work well. The number 3 of the sectors is detected by the broken-stick rule $\text{BS}(\mathbf{C}) = 3$. The number $\alpha = 9$ of eigenvalues greater than $\lambda_{\text{max}}^{\text{rnd}}$ is nearly $\text{ACT}(\mathbf{C}) = 10$.

	n	p/n	$\lambda_1(\mathbf{C})/p$	p	BS	ACT	GK	CPV
MSA	2815	0.1687	0.0216	475	3	10	188	193

Table 3: The multiple sequence alignment dataset.

4 Comparison of ENP to a bounded spiked eigenvalues model

As for the asymptotic behaviors of the eigenvalues of $\mathbf{S}$ and $\mathbf{C}$, the first-order behavior (i.e., location) are the same under a certain condition (El Karoui [28]), but the second-order behavior (i.e., fluctuation) can be different [32].

For the null case, a sample covariance matrix $\mathbf{S}$ and a sample correlation matrix $\mathbf{C}$ have the same limiting distribution of the largest eigenvalue:

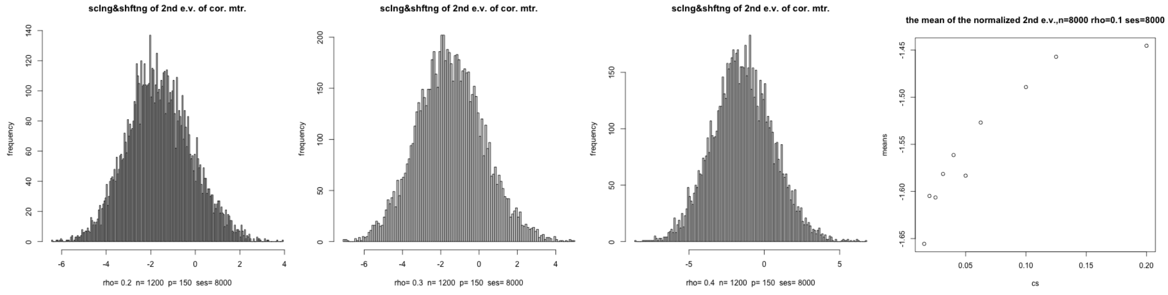
Proposition 4.1 ([6, Theorem 1.6]). *Let the population be* $\mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$. *Suppose* $p, n \to \infty$, $p/n \to c \in (0, 1)$. *Then* $(n\lambda_1(\mathbf{C}) - \mu_{n,p})/\sigma_{n,p} \xrightarrow{D} TW_1$.

An ENP with the population covariance matrix $\boldsymbol{\Sigma}_p(\rho)$ $(\rho > 0)$ is a divergent spiked eigenvalues model. For such an ENP, $\lambda_2(\mathbf{C})$ ($\lambda_2(\mathbf{S})$, resp.) is effectively, so to say, the largest eigenvalue of a sample correlation matrix (a sample covariance matrix, resp.) of a population having covariance matrix $(1-\rho)\mathbf{I}_{p-1}$. Table 4 and Figure 2 (a∼c) show that the dependence of the (variance of the) fluctuation of $\lambda_2(\mathbf{C})$ on the equi-correlation coefficient $\rho > 0$ is stronger than the dependence (Proposition 2.5) of that of $\lambda_2(\mathbf{S})$ on ρ. Let $\tau(\rho)$ be the histogram obtained by 8000 replications of $(n\lambda_2(\mathbf{C})/(1-\rho) - \mu_{n,p-1})/\sigma_{n,p-1}$ from the population $\mathrm{N}_p(\mathbf{0},\ \mathbf{\Sigma}_p(\rho))$ for $n = 1200$ and $p = 150$. Then, the means of $\tau(0.2)$, $\tau(0.3)$, and $\tau(0.4)$ are *smaller* than $\mathrm{E}[TW_1]$.

Figure 2 (d) is the plot of $c = 1/64, 1/50, 1/40, 1/32, 1/25, 1/20, 1/16, 1/10, 1/8, 1/5$ vs. the means of the histogram obtained by 8000 replications of $(8000\lambda_2(\mathbf{C})/(1 - 0.1) - \mu_{8000,8000c-1})/\sigma_{8000,8000c-1}$ from the population $\mathrm{N}_{8000c}(\mathbf{0},\ \mathbf{\Sigma}_{8000c}(0.1))$. The means of the histograms seem nonlinear and bounded in c. It may be $\lim_{\substack{p,n\to\infty\\ p/n\to c}} \mathrm{E}\,\lambda_2(\mathbf{C}) = (1-\rho)(1+\sqrt{c})^2$ for any $\rho \in [0,\ 1)$.

	Mean	Variance	Skewness	Excess kurtosis
$\tau(0.2)$	-1.599	2.050	0.207	0.061
$\tau(0.3)$	-1.512	2.857	0.169	0.059
$\tau(0.4)$	-1.470	3.970	0.125	0.011
TW_1	-1.206	1.607	0.293	0.165

Table 4



(a) $\tau(0.2)$ **(b)** $\tau(0.25)$ **(c)** $\tau(0.4)$ **(d)** $\rho = 0.1$

Figure 2

However, for a bounded spiked eigenvalue model, $\lambda_2(\mathbf{C})$ converges almost surely to a value *larger* than the corresponding eigenvalue λ of the population correlation matrix, if λ is simple [1] and larger than $1+\sqrt{c}$:

Proposition 4.2 ([32, Theorem 1]). *Let* $\boldsymbol{x} = \begin{bmatrix}\xi\\ \eta\end{bmatrix} \in \mathbb{R}^{m+p}$ *be a random vector with* $(4+\delta)$*-th moment for some* $\delta > 0$. *Assume that* $\xi \in \mathbb{R}^m$ *has mean zero, and is independent of* $\eta \in \mathbb{R}^p$*, which has i.i.d. components* η_i *with mean zero and unit variance. Let* $\mathbf{R}$ *be the correlation matrix of* $\boldsymbol{x}$ *such that* $\lambda_i(\mathbf{R}) = 1$ $(m+1 \le i \le m+p)$. *Then,*

$$\lambda_i(\mathbf{R}) > 1+\sqrt{c} \textit{ is simple} \implies \lim_{\substack{p,n\to\infty\\ p/n\to c}} \lambda_i(\mathbf{C}) = \lambda_i(\mathbf{R}) + \frac{c\lambda_i(\mathbf{R})}{\lambda_i(\mathbf{R})-1} \ (a.s.).$$

[1] In an ENP with $\rho > 0$ and $p > 2$, $\mathbf{R} = \mathbf{\Sigma}_p(\rho)$ but $\lambda_i(\mathbf{R}) = 1-\rho$ $(2 \le i \le p)$ is not simple.

References

[1] Y. Akama. Correlation matrix of equi-correlated normal population: fluctuation of the largest eigenvalue, scaling of the bulk eigenvalues, and stock market. *Int. J. Theor. Appl. Finance*, 26:2350006, 2023.

[2] Y. Akama and A. Husnaqilati. A dichotomous behavior of Guttman-Kaiser criterion from equi-correlated normal population. *J. Indones. Math. Soc.*, 28(3):272–303, 2022.

[3] Z. Bai and X. Ding. Estimation of spiked eigenvalues in spiked models. *Random Matrices: Theory and Applications*, 01(02):1150011, 2012.

[4] Z. D. Bai and J. W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.*, 26(1):316–345, 1998.

[5] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 1993.

[6] Z. Bao, G. Pan, and W. Zhou. Tracy-Widom law for the extreme eigenvalues of sample correlation matrices. *Electron. J. Probab.*, 17(none):1–32, 2012.

[7] T. Cai, X. Han, and G. Pan. Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Ann. Stat.*, 48(3):1255–1280, 2020.

[8] R. Cangelosi and A. Goriely. Component retention in principal component analysis with application to cDNA microarray data. *Biol. Direct.*, 2(2), 2007.

[9] M. M. Carhart. On persistence in mutual fund performance. *J. Finance*, 52:57–82, 1997.

[10] R. Engle and B. Kelly. Dynamic equicorrelation. *J. Bus. Econ. Stat.*, 30(2):212–228, 2012.

[11] E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.*, 33(1):3–56, 1993.

[12] E. F. Fama and K. R. French. A five-factor asset pricing model. *J. Financ. Econ.*, 116(1):1–22, 2015.

[13] J. Fan, J. Guo, and S. Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *J. Am. Stat. Assoc.*, 117(538):852–861, 2022.

[14] S. Frontier. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. *Biol. Ecol.*, 25:67–75, 1976.

[15] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory*, 60(8):5040–5053, 2014.

[16] L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance*, 48(5):1779–1801, 1993.

[17] L. Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, 1954.

[18] L. Holst. On the lengths of the pieces of a stick broken at random. *J. Appl. Prob.*, 17:623–634, 1980.

[19] D. A. Jackson. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology*, 74(8):2204–2214, 1993.

[20] J. E. Jackson. *A user's guide to principal components*, volume 587. Wiley, 1991.

[21] M. C. Jensen. The performance of mutual funds in the period 1945–1964. *J. Finance*, 23(2):389–416, 1968.

[22] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall, 6th edition, 2007.

[23] W. D. Johnson and G. G. Koch. Intraclass correlation coefficient. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 685–687. Springer, 2011.

[24] I. T. Jolliffe. *Principal Component Analysis.* Springer, 2nd edition, 2002.

[25] H. F. Kaiser. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.*, 20(1):141–151, 1960.

[26] H. F. Kaiser. A measure of the average intercorrelation. *Educ. Psychol. Meas.*, 28(2):245–247, 1968.

[27] H. F. Kaiser. On Cliff's formula, the Kaiser-Guttman rule, and the number of factors. *Percept. Mot. Ski.*, 74(2):595–598, 1992.

[28] Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.*, 19(6):2362–2405, 2009.

[29] R. MacArthur. On the relative abundance of bird species. *Proc. Natl. Acad. Sci. U.S.A.*, 43:293–295, 1957.

[30] M. W. McCraken and S. Ng. FRED-MD: A monthly database for macroeconomic research. *J. Bus. Econ. Stat.*, 79:677–693, 2017.

[31] F. Merlevède, J. Najim, and P. Tian. Unbounded largest eigenvalue of large sample covariance matrices: Asymptotics, fluctuations and applications. *Linear Algebra Its Appl.*, 577:317–359, Sep 2019.

[32] D. Morales-Jimenez, I. M. Johnstone, M. R. McKay, and J. Yang. Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *Stat. Sin.*, 31(2):571, 2021.

[33] S. A. Mulaik. *Foundations of factor analysis.* CRC press, 2nd edition, 2010.

[34] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.*, 49(4):974–997, 2005.

[35] A. A. Quadeer, R. H. Louie, K. Shekhar, A. K. Chakraborty, I. Hsing, and M. R. McKay. Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design. *J. Virol.*, 88(13):7628–7644, 2014.

[36] A. A. Quadeer, D. Morales-Jimenez, and M. R. McKay. Co-evolution networks of HIV/HCV are modular with direct association to structure and function. *PLoS Computational Biology*, 14:1–29, 2018.

[37] A. Soshnikov. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.*, 108(5):1033–1056, 2002.

[38] T. Tao and V. Vu. Random matrices: Universality of local eigenvalue statistics. *Acta Mathematica*, 206(1):127–204, 2011.

[39] M. O. Ulfarsson and V. Solo. Dimension estimation in noisy PCA with SURE and random matrix theory. *IEEE Trans. Signal Process.*, 56(12):5804–5816, 2008.

[40] J. Yao, S. Zheng, and Z. D. Bai. *Large sample covariance matrices and high-dimensional data analysis.* Cambridge University Press, 2015.

[41] K. Yata and M. Aoshima. PCA consistency for non-gaussian data in high dimension, low sample size context. *Commun. Stat. Theory Methods*, 38(16), 2009.

[42] W. R. Zwick and W. F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychol. Bull.*, 99(3):432–442, 1986.