Unbounded largest eigenvalue of large sample correlation matrices: Asymptotics and applications

Yohji Akama (Tohoku University)*

November 4, 2022

Datasets in economics and wireless communication networks, the leading eigenvalues of the population covariance matrices may tend to infinity [26]. Unbounded population covariance matrices have been studied in [3, 28, 26, 38] to cite a few. On the other hand, the limiting distributions of eigenvalues of sample covariance/correlation matrices with released independence condition have been studied in [9, 4, 14, 29] with Marčenko-Pastur distribution [36, 6], Tracy-Widom distribution [20, 34] and so on. Motivated by [37, 25, 19, 14, 31], we have studied the limiting spectral distribution of sample correlation matrices formed from *equi-correlated normal population*, of which population covariance matrix is unbounded. We discuss our results [2, 1] with large datasets from molecular biology [31, 33], S&P500 stock returns datasets [25], and household datasets [27].

1 Significant dimension of dataset

Let $[x_{ij}]$ be a p by n data matrix, for a random sample of size n from a certain p-dimensional population distribution. Let $\mathbf{S} \in \mathbb{R}^{p\times p}$ be the sample covariance matrix $[\sum_{k=1}^{n} x_{ik}x_{jk}/n]$ and $\mathbf{C} \in \mathbb{R}^{p\times p}$ be the sample correlation matrix.The *empirical spectral distribution* (ESD) of a real symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ is, by definition, $F^{\mathbf{M}}(x) = \frac{1}{p}\#\{\, 1 \le i \le p \mid \lambda_i(\mathbf{M}) \le x \,\} \quad (x \in \mathbb{R})$ where $\lambda_1(\mathbf{M}) \ge \lambda_2(\mathbf{M}) \ge \cdots \ge \lambda_p(\mathbf{M})$ are the eigenvalues of $\mathbf{M}$. In [19], Jiang studied the limiting behaviors of the empirical spectral distribution of $\mathbf{C}$ and the distribution of the extreme eigenvalues of $\mathbf{C}$, assuming that x_{ij} are i.i.d. Hereafter, we suppose

$$p, n \to \infty, \quad \frac{p}{n} \to c \in (0, \infty). \tag{1}$$

This limiting regime is common in random matrix theory [5, 7, 19, 8, 6, 36].

Proposition 1 ([19, Theorem 1.2]). $\mathrm{E}\,|x_{ij}|^2 < \infty$. *Then, almost surely, the empirical spectral distribution $F^{\mathbf{C}}$ weakly converges to a deterministic probability distribution $F_c(x)$ which is* Marčenko-Pastur distribution [36, 6] such that the dimension to sample size ratio index is c.

An *equi-correlation matrix* is, by definition,

$$\mathbf{\Sigma}_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{J} \in \mathbb{R}^{p\times p} \qquad (0 \le \rho < 1),$$

[*] Department of Mathematics, Graduate School of Science, Tohoku University, Aramaki, Aoba, Sendai, 980-8578, Japan.
2022年度科学研究費シンポジウム 大規模複雑データの理論と方法論 新たな発展と関連分野への応用 (2022年11月4,5日)講演予稿
e-mail: `yoji.akama.e8@tohoku.ac.jp`

2010 Mathematics Subject Classification: 60F15, 62H25.
Keywords: Equi-correlated normal population, Marčenko-Pastur distribution, sample correlation matrices, limiting spectral distribution.

where $\mathbf{J}$ is the matrix of 1 of order p and $\lambda_1(\mathbf{\Sigma}_\rho) = (p-1)\rho + 1 \geq \lambda_2(\mathbf{\Sigma}_\rho) = \cdots = \lambda_p(\mathbf{\Sigma}_\rho) = 1 - \rho$. We consider a normal population having the population correlation matrix $\mathbf{\Sigma}_\rho$, as Engle and Kelly [12] proposed *Dynamic Equicorrelation* model for a time series analysis of stock market. Let us call the constant ρ an *equi-correlation coefficient.*

Proposition 2 (A. and Husnaqilati [2])**.** *Let* $\mathbf{C}$ *be a sample correlation matrix formed from a p-dimensional normal population with the equi-correlation coefficient* $0 \leq \rho < 1$*. Then, almost surely,* $F^{\mathbf{C}}(x)$ *weakly converges to* $F_c\left(x/(1-\rho)\right)$.

A default rule to determine the number of significant components [21] (factors [13]) to take, in statistical software SPSS and SAS is *Guttman-Kaiser criterion* [17, 22, 39], which suggests us to retain the number of eigenvalues greater than the average of all eigenvalues of a sample correlation matrix. Nearly four decades ago, for p independent standard normal random variables, Yeomans-Golder [37] showed that Guttman-Kaiser criterion suggests at most $p/2$ variables, by a simulation study. Moreover, for dependent variables, H. F. Kaiser, an American psychologist who introduced Guttman-Kaiser criterion, observed a dichotomous behavior of the criterion by reporting the experience of specialists of factor analysis:

Quotation 3 ([24])**.** ... Humphreys (personal communication, 1984) asserts that, when the number p of attributes is large and the "average" intercorrelation is small, the Kaiser-Guttman rule will overfactor. Tucker (personal communication, 1984) asserts that, when the number of attributes p is small and the structure of the attributes is particularly clear, the Kaiser-Guttman rule will underfactor. ...

Here, "overfactor" ("underfactor", resp.) intends "overestimate" ("underestimate", resp.) the number of factors in the factor model. According to Kaiser [23], 'the "average" intercorrelation' corresponds to a positive constant ρ of $\mathbf{\Sigma}_\rho$.

Yeomans-Golder's simulation study and Kaiser's observation are partially explained by the following:

Theorem 4. *Suppose that* $\mathbf{Z} = n^{-1/2}[z_{ij}] \in \mathbb{R}^{p\times n}$ *where* z_{ij} $(1 \leq i \leq p,\ 1 \leq j \leq n) \overset{i.i.d.}{\sim} N(0,1)$*. Then, it holds almost surely that*

$$\lim_{c\to 0} \lim_{\substack{n,p\to\infty \\ p/n\to c}} F^{\mathbf{Z}\mathbf{Z}^\top \mathbf{\Sigma}_\rho}\left(\frac{1}{p}\operatorname{Tr}\mathbf{Z}\mathbf{Z}^\top\mathbf{\Sigma}_\rho\right) = \begin{cases} \frac{1}{2} & (\rho = 0); \\ 1 & (\rho > 0). \end{cases}$$

We will discuss a free probability theoretic approach for this theorem with "$N(0,1)$" replaced.

For a real symmetric matrix $\mathbf{M}$ of order p, we define the portion $GK^{\mathbf{M}}$ of eigenvalues of $\mathbf{M}$ that Guttman-Kaiser criterion retains, by $p^{-1}\#\left\{\, i \leq p \mid \lambda_i(\mathbf{M}) \geq \operatorname{Tr}(\mathbf{M})/p \,\right\}$. For $GK_{c,\rho} := 1 - F_c\left((1-\rho)^{-1}\right)$, we have the following:

Theorem 5 ([2])**.** *Suppose* $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(\boldsymbol{\mu},\, \mathbf{D}\mathbf{\Sigma}_\rho\mathbf{D})$ *for a deterministic vector* $\boldsymbol{\mu} \in \mathbb{R}^p$*, a deterministic nonsingular diagonal matrix* $\mathbf{D} \in \mathbb{R}^{p\times p}$*, and* $0 \leq \rho < 1$*. Then,* $GK^{\mathbf{S}} \overset{a.s.}{\to} GK_{c,\rho}$ *for* $\boldsymbol{\mu} = \mathbf{0}$ *and* $\mathbf{D} = \sigma\mathbf{I}$ *with* $\sigma > 0$*; and* $GK^{\mathbf{C}} \overset{a.s.}{\to} GK_{c,\rho}$.

Theorem 6 ([2])**.**

1. For any $c > 0$, $GK_{c,\rho}$ is nonincreasing in $\rho \in [0,\, 1)$.
2. If $0 \leq \rho < 1$, then $c \geq (1/\sqrt{1-\rho} + 1)^2 \iff GK_{c,\rho} = 1/c$.

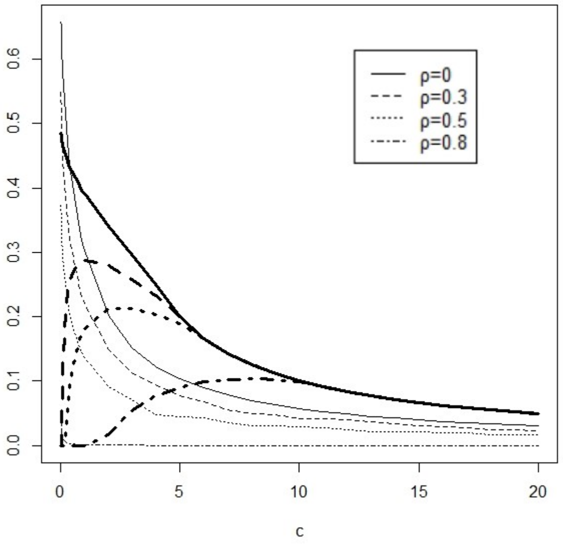


Figure 1: $GK_{c,\rho}$ (thick curves) and $CP_{c,\rho}$ (thin curves). $CP^{\mathbf{M}}(t)$ is the portion of eigenvalues of $\mathbf{M}$ that Jolliffe's rule [21] retains.

2 The largest eigenvalue of sample covariance/correlation matrix

Psychologists have been concerned with the largest eigenvalues of correlation matrices, since Kaiser studied "average" intercorrelation [23] among variables.

Quotation 7 ([15])**.** ... The first eigenvalue of a correlation matrix indicates the maximum amount of the variance of the variables which can be accounted for with a linear model by a single underlying factor. When all correlations are positive, this first eigenvalue is approximately a linear function of the average correlation among the variables. While that is not true when not all the correlations are positive, in the general case the first eigenvalue is approximately equal to a lower bound derived in the paper. That lower bound is based on the maximum average correlation over reversals of variables and over subsets of the variables. Regression tests show these linear approximations are very accurate. The first eigenvalue measures the primary cluster in the matrix, its number of variables and average correlation. ...

By random matrix theory, we get:

Theorem 8 ([1])**.** $\lambda_1(\mathbf{C})/p \overset{a.s.}{\to} \rho, \quad (\lambda_1(\mathbf{S})-\mu)/\sigma \overset{d}{\to} \mathrm{N}(0,1) \qquad (0<\rho<1)$ *where*

$$\mu := \frac{((p-1)\rho+1)\left((1+(n-1)p)\,\rho+p-1\right)}{pn\rho}, \quad \sigma := \frac{(p-1)\rho+1}{\sqrt{2n}}.$$

By the first assertion, $GK_{p/n,\lambda_1(\mathbf{C})/p}$ estimates $GK^{\mathbf{C}}$.

The first (second, resp.) assertion of Theorem 8 is due to the first (second, resp.) assertion of the following proposition. A sequence $(F_p)_p$ of distribution functions is called *tight* [35, p. 8] on $\mathbb{R}^+$, if for every $\varepsilon \in (0,1]$, $\sup_p F_p(M) > 1-\varepsilon$ for some $M \geq 0$.

Proposition 9 ([28, Proposition 2.1]). *Let* $\mathbf{S} = n^{-1}\mathbf{\Sigma}^{1/2}\mathbf{Z}\mathbf{Z}^\top\mathbf{\Sigma}^{1/2} \in \mathbb{R}^{p\times p}$ *with the entries of* $\mathbf{Z} \in \mathbb{R}^{p\times n}$ *being independent, standard normal. Suppose* $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ *is positive semidefinite and deterministic,* $\left(F^{\mathbf{\Sigma}}\right)_p$ *of ESDs is tight on* $\mathbb{R}^+$, *and* $\lim_{p\to\infty} \lambda_1(\mathbf{\Sigma}) = \infty$. *Then,*

1. $\lambda_1(\mathbf{S})/\lambda_1(\mathbf{\Sigma}) \overset{a.s.}{\to} 1.$

2. *If moreover* spectral gap condition *on* $\mathbf{\Sigma}$*:* $\limsup_{\substack{p,n\to\infty \\ p/n\to c}} \lambda_2(\mathbf{\Sigma})/\lambda_1(\mathbf{\Sigma}) < 1$, *then*

$$\sqrt{n}\left(\frac{\lambda_1(\mathbf{S})}{\lambda_1(\mathbf{\Sigma})} - 1 - \beta\right) \overset{d}{\to} \mathrm{N}\left(0,\ \mathrm{E}\left|z_{11}\right|^4 - 1\right) \quad \textit{for} \quad \beta := \frac{1}{n}\sum_{k=2}^{p} \frac{\lambda_k(\mathbf{\Sigma})}{\lambda_1(\mathbf{\Sigma}) - \lambda_k(\mathbf{\Sigma})}.$$

Our first result $\lambda_1(\mathbf{C})/p \overset{a.s.}{\to} \rho$ agrees with the psychologists' work (Quotation 7) and probabilist's work:

Proposition 10 ([19, Theorem 1.1]). *If* $\{\, x_{ij} \mid i, j \geq 1 \,\}$ *are* i.i.d. *and* $\mathrm{E}\left|x_{11}\right|^4 < \infty$, *then* $\lambda_1(\mathbf{C}) \overset{a.s.}{\to} (1+\sqrt{c})^2$.

Theorm 8 and the following yield the phase transition of limiting distribution of the largest eigenvalue, depending on whether $\rho > 0$ or $\rho = 0$.

Proposition 11 ([20, 34]). *Let* $\{\, x_{ij} \mid i, j \geq 1 \,\}$ *be i.i.d. standard normal. Then*

$$\frac{\lambda_1(\mathbf{S}) - \left(\sqrt{p} + \sqrt{n-1}\right)^2}{\left(\sqrt{p} + \sqrt{n-1}\right)\left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{1/3}} \overset{d}{\to} \textit{Tracy-Widom distribution.}$$

It is curious to know whether the limiting distribution of the largest eigenvalue is still dichotomous even when the equi-correlation coefficient $\rho = \rho_n$ decays with respect to n. We are planned to study this when $\rho_n\sqrt{\log p} \to 0, \lambda$, or ∞ in an "ultra-high-dimensional case" (Fan-Jiang [14]) $p = p_n \to \infty$ and $\log p = o(n^{1/3})$ as $n \to \infty$.

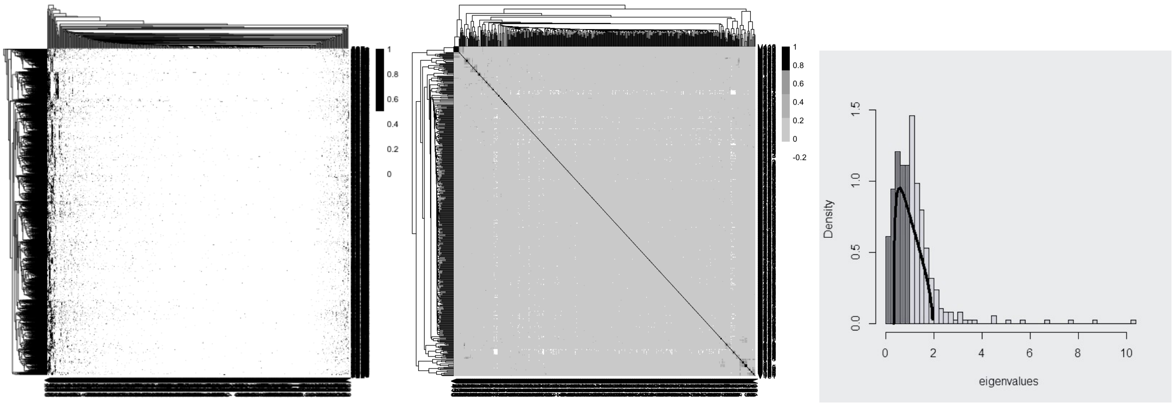
3 Empirical Study

Our theoretical study on equi-correlated normal population is unexpectedly related to psychological datasets, maybe because their datasets have strong correlation structures. To see the mass effect of the number p of variables and the sample size n, we examine our theoretical study with datasets from molecular biology and economics.

3.1 Datasets from molecular biology

To design vaccines, Quadeer et al. [31] considered a multiple sequence alignment (MSA) of a p-residue (site) protein with n sequences where $p = 475$ and $n = 2815$. They represented the MSA with a 0-1 code following [11, 18], and then considered the correlation matrix $\mathbf{C}$. From this, Quadeer et al. [31, 32] detected nine signal eigenvectors from $\mathbf{C}$, by clever randomization for the MSA. At the same time, they introduced an alternative method that employs Marčenko-Pastur distribution. We also examine our study of Guttman-Kaiser criterion with Marčenko-Pastur distribution, by using their MSA dataset of Quadeer et al. [31].

Figure 2 (a)((b), resp.) is the heat map of the binary MSA dataset (the correlation matrix $\mathbf{C}$, resp.) applied by an hierarchical clustering algorithm on columns and rows. The binary MSA dataset of Quadeer et al. [31] is sparse. As for the $p^2 = 225625$ entries



(a) (b) (c)

Figure 2: The binary MSA dataset. (a) The data matrix, (b) the correlation matrix $\mathbf{C}$, and (c) the eigenvalues of $\mathbf{C}$.

of $\mathbf{C}$, the minimum is $-.28922$, the first quantile is $-.00437$, the median is $-.00174$, the mean is .00682, the third quantile is $-.00061$, and the maximum is 1.

Figure 2 (c) is the histogram of the eigenvalues of $\mathbf{C}$. The light gray bins are for the eigenvalues more than 1 and the dark gray bins are for the eigenvalues less than 1. The black solid curve is the density of Marčenko-Pastur distribution with index $p/n = 0.16$ and scale $1 - \lambda_1(\mathbf{C})/p = 0.98$. The $GK^{\mathbf{C}} = 0.39$ has 10% error from the estimated value $GK_{0.16,0.02} = 0.44$, which is close to the theoretical value 1/2 of Theorem 4.

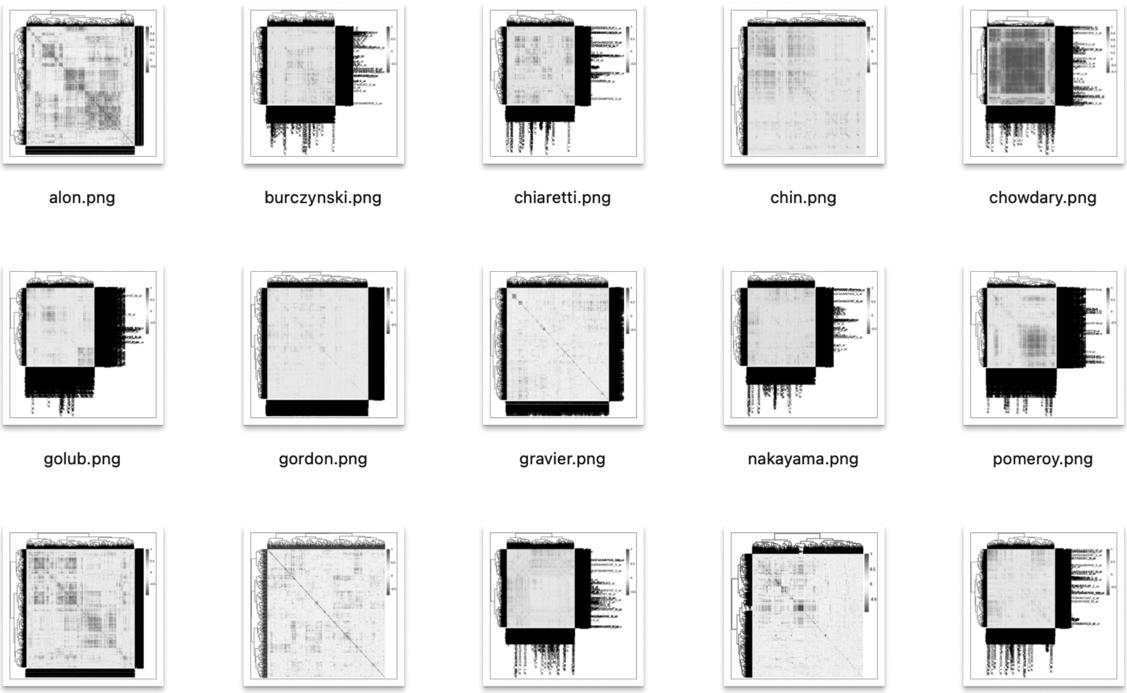
Next, for 16 microarray datasets from [33], Table 1 lists the name of a microarray dataset, p/n, $\lambda_1(\mathbf{C})/p$, $GK^{\mathbf{C}}$, $GK_{p/n,\lambda_1(\mathbf{C})/p}$, $CP^{\mathbf{C}}(.7)$ and p, in the increasing order of p/n. Here p is the number of features and n is the number of observations.

No.	Name	p/n	$\lambda_1(\mathbf{C})/p$	$GK^{\mathbf{C}}$	$GK_{p/n,\lambda_1(\mathbf{C})/p}$	p
1	Sorlie	5.4	.110	.162	.186	456
2	Gravier	17.2	.083	.054	.057	2905
3	Alon	32.2	.450	.030	.031	2000
4	Yeoh	50.9	.145	.020	.020	12625
5	Gordon	69.2	.087	.014	.014	12533
6	Tian	72.9	.089	.014	.014	12625
7	Shipp	92.5	.213	.011	.011	7129
8	Chiaretti	98.6	.181	.01	.01	12625
9	Golub	99.0	.149	.01	.01	7129
10	Pomeroy	118.8	.266	.008	.008	7128
11	West	145.4	.162	.006	.006	7129
12	Burczynski	175.4	.115	.005	.005	22283
13	Chin	188.2	.164	.005	.005	22215
14	Nakayama	212.2	.073	.004	.004	22283
15	Chowdary	214.2	.699	.004	.004	22283
16	Borovecki	718.8	.173	.0013	.0013	22283

Table 1: DNA microarray datasets.

By Table 1, $GK^{\mathbf{C}} > CP^{\mathbf{C}}(.7)$ and $GK^{\mathbf{C}}$ is nonincreasing in p/n as Figure 1 (left). Moreover, all $GK_{p/n,\lambda_1(\mathbf{C})/p}$ are n/p by Theorem 6 (2), by $p/n > (1/\sqrt{1-\lambda_1(\mathbf{C})/p}+1)^2$.

For the correlation matrices of the 15 datasets of Table 1, Figure 3 shows the heat maps applied by a hierarchical clustering to columns and rows. Despite of the various structures of $\mathbf{C}$'s, all the empirical values $GK^{\mathbf{C}}$ are, remarkably, around n/p.



alon.png burczynski.png chiaretti.png chin.png chowdary.png

golub.png gordon.png gravier.png nakayama.png pomeroy.png

shipp.png sorlie.png tian.png west.jpeg yeoh.png

Figure 3: Heat maps of the correlation matrices of DNA microarray datasets.

3.2 Datasets from economics

For a dataset of returns of p S&P500 stocks (or other major markets) for n trading days, Laloux et al. [25] fitted the histogram of the eigenvalues of the correlation matrix $\mathbf{C}$, to the density function of a scaled Marčenko-Pastur distribution.

Table 2 is the list of p/n, $\lambda_1(\mathbf{C})/p$, $GK^{\mathbf{C}}$, and $GK_{p/n,\lambda_1(\mathbf{C})/p}$, for $p = 212$ S&P500 stocks of various periods. The lines of Table 2 are in the increasing order of $\lambda_1(\mathbf{C})/p$. $GK^{\mathbf{C}}$ is decreasing in $\lambda_1(\mathbf{C})/p$ in Table 2, as Theorem 6 (1). The estimators $GK_{p/n,\lambda_1(\mathbf{C})/p}$ of $GK^{\mathbf{C}}$ are mostly 0.

No	Period	p/n	$\lambda_1(\mathbf{C})/p$	$GK^{\mathbf{C}}$	$GK_{p/n,\lambda_1(\mathbf{C})/p}$	p
1	1993-01-04-1995-12-29	.280	.110	.330	.37	212
2	1993-01-04-2022-08-01	.028	.313	.103	0	212
3	2012-08-01-2022-08-01	.084	.399	.099	0	212
4	2005-01-04-2022-08-01	.047	.422	.084	0	212
5	2005-01-04-2013-12-30	.093	.450	.084	0	212

Table 2: The returns of S&P500 datasets.

Next, we consider similar but more categorized datasets. Naturally, the return of a stock is more correlated with the return of a stock of the same industry classification sector, than with the return of a stock of a different industry classification sector. Table 3 is the list of *global industry classification standard* (GICS) sector of S&P500, p/n, $\lambda_1(\mathbf{C})/p$, $GK^{\mathbf{C}}$, $GK_{p/n,\lambda_1(\mathbf{C})/p}$, and p, for the period 2012-2022. Table 3 is ordered in the increasing order of $\lambda_1(\mathbf{C})/p$. All the p of Table 3 are less than the $p = 212$ of Table 2, but some GICS sectors have larger $\lambda_1(\mathbf{C})/p$. In Table 3, $GK^{\mathbf{C}}$ is decreasing if we leave out the first line (Communication services, $p = 19$), the fourth line (Consumer Staples, $p = 23$), the seventh line (Material, $p = 24$), the tenth line (Energy, $p = 16$) from all the lines. Theorem 6 (1) for Guttman-Kaiser criterion holds for $p \geq 28$.

No	GICS	p/n	$\lambda_1(\mathbf{C})/p$	$GK^{\mathbf{C}}$	$GK_{p/n,\lambda_1(\mathbf{C})/p}$	p
1	Communication Services	.007	.357	.210	0	19
2	Consumer Discretionary	.020	.384	.153	0	52
3	Health Care	.018	.394	.149	0	47
4	Consumer Staples	.009	.430	.174	0	23
5	Information Technology	.025	.465	.097	0	62
6	Industrials	.026	.498	.092	0	65
7	Materials	.009	.499	.167	0	24
8	Real estate	.011	.582	.100	0	30
9	Financials	.025	.607	.079	0	63
10	Energy	.006	.687	.062	0	16
11	Utilities	.011	.689	.071	0	28

Table 3: The returns of S&P500 stocks per GICS.

The estimator $GK_{p/n,\lambda_1(\mathbf{C})/p}$ of $GK^{\mathbf{C}}$ in Table 2 and Table 3 are all 0 for $\lambda_1(\mathbf{C})/p > 0.110$. The estimator $GK_{p/n,\lambda_1(\mathbf{C})/p} = 1 - F_{p/n}\left(1/(1 - \lambda_1(\mathbf{C})/p)\right)$ is 0 if $p/n < 1$ but $\lambda_1(\mathbf{C})/p < 1$ is sufficiently large. Figure 4 is the time series of equi-correlations for the datasets, computed by GJR GARCH [16] with correlation structure being *dynamic equicorrelation* [12]. Then, $\lambda_1(\mathbf{C})/p$ is always larger than the time averages $\overline{\rho}$ of the time series of the equi-correlation coefficient.

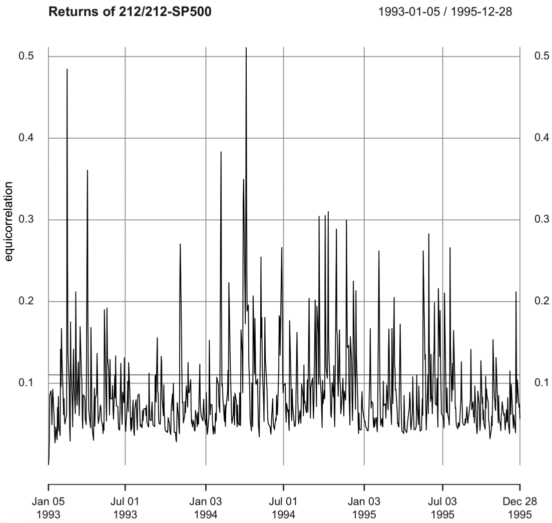


Figure 4: $\lambda_1/p = .110$, average $\overline{\rho}$=.088

For the stock return datasets (Table 2 and Table 3), we computed the heat maps with hierarchical clustering algorithm for the sample correlation matrices $\mathbf{C}$ of the datasets. The heat maps are very different from the heat map (Figure 2(b)) of the correlation matrix of the binary MSA dataset. The correlation matrices of the stock returns have diagonal block structures. Moreover, the sizes of the eleven correlation matrices of Table 3 are small. Since $GK^{\mathbf{C}}$ of the stock returns datasets may be discrepant from our estimator $GK_{p/n,\lambda_1(\mathbf{C})/p}$, we think that Laloux et al. [25]'s fitting of a scaled Marčenko-Pastur distribution to the stock returns dataset is too naive.

Finally, we discuss two household datasets in 2019 by area classification, from [27]. One is for the amount of assets per households and the other is for the average yearly income. These datasets have response variables: the amount of assets and liabilities per household and the average yearly income from the whole of Japan. Meanwhile, the explanatory variables are the amount of assets per households and the average yearly income from each of 66 regions in Japan. The two datasets suffer from *multicollinearity* [10, 21], in view of *variance inflation factors* (VIFs) [10, 21] of the explanatory variables. If all the explanatory variables are uncorrelated, then all the VIFs are 1, but if severe multicollinearities exist, then the VIFs of explanatory variables are large [10, 21].

Name	min VIF	R^2	$\lambda_1(\mathbf{C})/p$	$GK^{\mathbf{C}}$	p	p/n
The amount of assets 2019	1×10^3	1	.98	.03	66	.05
The average yearly income 2019	4×10^4	1	.99	.01	66	.04

Table 4: The household datasets of 2019.

Table 4 is the list of name, min VIF, the (adjusted) coefficient R^2 of determination, $\lambda_1(\mathbf{C})/p$, $GK^{\mathbf{C}}$, $p = 66$ variables, and p/n, in the increasing order of $\lambda_1(\mathbf{C})/p$. Since $R^2 = 1$ for both datasets, we can assume that all the explanatory variables have the equi-correlation coefficient $\rho = 1$. This corresponds to $\lambda_1(\mathbf{C})/p = 1$ by the first assertion of Theorem 8. One of future work is to discuss the multicollinearity and other correlation structures [4, 9, 12, 14, 29, 30] among variables, with the extreme eigenvalues and the bulk eigenvalues of the correlation matrices of datasets.

4 Conclusion

For an equi-correlated normal population with the equi-correlation coefficient ρ ($0 \leq \rho < 1$), we have shown that the limiting distribution of eigenvalues of the sample correlation matrix is Marčenko-Pastur distribution scaled with $1 - \rho$. This scaling of Marčenko-Pastur distribution explains the "phase transitions" of Guttman-Kaiser criterion depending on whether $\rho = 0$ or not as $n, p \to \infty$, $p/n \to c > 0$. In high-dimensional statistics of various fields, when the number of variables are smaller than the size of a sample, a global correlation among the variables causes a perceptible global impact, even if the correlation is minute.

A correlation matrix of finance and biology is often the addition of a background constant correlation and a matrix with block diagonal structure, and the most of block sizes are unchanged but the number of blocks grows. This would be an combination of a spiked population model and our block equi-correlation model, and may be next target of study.

Acknowledgment

The MSA dataset is kindly provided from Prof. A. A. Quadeer through the email correspondence with Prof. M. R. McKay on 22 September, 2022.

References

[1] Y. Akama, *Correlation matrix of equi-correlated normal population: fluctuation of the largest eigenvalue, scaling of the bulk eigenvalues, and stock market*, preprint, 2022.

[2] Y. Akama and A. Husnaqilati, *A dichotomous behavior of Guttman-Kaiser criterion from equi-correlated normal population*, 2022, To appear in JIMS. arXiv:2210.12580v2

[3] M. Aoshima and K. Yata, *Two-sample tests for high-dimension, strongly spiked eigenvalue models*, Stat Sin (2018), 43–62.

[4] Z. Bai and W. Zhou, *Large sample covariance matrices without independence structures in columns*, Stat Sin (2008), 425–442.

[5] Z. D. Bai, *Methodologies in spectral analysis of large dimensional random matrices, a review*, Stat Sin **9** (1999), no. 3, 611–662.

[6] Z. D. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, 2nd ed., Springer, New York, 2010.

[7] Z. D. Bai and Y. Q. Yin, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, Ann. Probab. **21** (1993), no. 3, 1275–1294.

[8] J. Baik and J. W. Silverstein, *Eigenvalues of large sample covariance matrices of spiked population models*, J. Multivar. Anal. **97** (2006), no. 6, 1382–1408.

[9] J. Bryson, R. Vershynin, and H. Zhao, *Marchenko–Pastur law with relaxed independence conditions*, Random Matrices Theory Appl (2021), 2150040.

[10] S. Chatterjee and A. S. Hadi, *Regression analysis by example*, 4th ed., John Wiley & Sons, Hoboken, 2006.

[11] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chkraborty, *Coordinate linkage of HIV evolution reveals regions of immunological vulnerability*, Proc. Natl. Acad. Sci. U.S.A. **108** (2011), no. 28, 11530–11535.

[12] R. Engle and B. Kelly, *Dynamic equicorrelation*, J. Bus. Econ. Stat. **30** (2012), no. 2, 212–228.

[13] L. R. Fabrigar and D. T. Wegener, *Exploratory factor analysis*, Oxford UP, UK, 2011.

[14] J. Fan and T. Jiang, *Largest entries of sample correlation matrices from equi-correlated normal populations*, Ann. Probab. **47** (2019), no. 5, 3321–3374.

[15] S. Friedman and H. F. Weisberg, *Interpreting the first eigenvalue of a correlation matrix*, Educ. Psychol. Meas. **41** (1981), no. 1, 11–21.

[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** (1993), no. 5, 1779–1801.

[17] L. Guttman, *Some necessary conditions for common-factor analysis*, Psychometrika **19** (1954), no. 2, 149–161.

[18] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, *Protein sectors: evolutionary units of three-dimensional structure*, Cell **138** (2009), no. 4, 774–786.

[19] T. Jiang, *The limiting distributions of eigenvalues of sample correlation matrices*, Sankhyā: The Indian Journal of Statistics (2003-2007) (2004), 35–48.

[20] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann Stat **29** (2001), no. 2, 295–327.

[21] I. T. Jolliffe, *Principal component analysis*, 2nd ed., Springer, New York, 2002.

[22] H. F. Kaiser, *The application of electronic computers to factor analysis*, Educ. Psychol. Meas. **20** (1960), no. 1, 141–151.

[23] ———, *A measure of the average intercorrelation*, Educ. Psychol. Meas. **28** (1968),

no. 2, 245–247.

[24] ———, *On Cliff's formula, the Kaiser-Guttman rule, and the number of factors*, Percept. Mot. Ski. **74** (1992), no. 2, 595–598.

[25] L. Laloux, P. Cizeau, M. Potters, and J. Bouchaud, *Random matrix theory and financial correlations*, Int. J. Theor. Appl. Finance **3** (2000), no. 03, 391–397.

[26] Z. Liu, Z. Bai, J. Hu, and H. Song, *CLT for LSS of sample covariance matrices with unbounded dispersions*, 2021, arXiv:2106.10135.

[27] Statistics Bureau, Ministry of Internal Affairs and Communications, *Statistics bureau home page/national survey of family income, consumption and wealth*, `https://www.stat.go.jp/english/data/zenkokukakei/index.htm`.

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

[29] 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** (2021), no. 2, 571.

[30] 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** (2005), no. 4, 974–997.

[31] 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** (2014), no. 13, 7628–7644.

[32] 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 Comput. Biol. **14** (2018), no. 9, 1–29.

[33] J. Ramey, *Datamicroarray*, R package version 1.14.4., 2013.

[34] A. Soshnikov, *A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices*, J. Stat. Phys. **108** (2002), no. 5, 1033–1056.

[35] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge UP, Cambridge, 2000.

[36] J. Yao, S. Zheng, and Z. D. Bai, *Sample covariance matrices and high-dimensional data analysis*, Cambridge UP, New York, 2015.

[37] K. A. Yeomans and P. A. Golder, *The Guttman-Kaiser criterion as a predictor of the number of common factors*, J. R. Stat. Soc. **31** (1982), no. 3, 221–229.

[38] Y. Q. Yin, *Spectral statistics of high dimensional sample covariance matrix with unbounded population spectral norm*, Bernoulli **28** (2022), no. 3, 1729–1756.

[39] W. R. Zwick and W. F. Velicer, *Comparison of five rules for determining the number of components to retain*, Psychol. Bull. **99** (1986), no. 3, 432–442.