SELF-INTRODUCTION

Name :
Atina Husnaqilati

Educational Background:

- 2012-2016
  S.Si in Statistik, Universitas Gadjah Mada, Indonesia.
- 2017-2019
  M.Sc in Mathematics, Tohoku University, Japan.
- 2019-now
  Ph.D student in Mathematics, Tohoku University, Japan.

Publication:

- Husnaqilati, Atina. (2016). *Combining parametric, semi-parametric, and nonparametrik survival model with stacked method* http://etd.repository.ugm.ac.id/home/detail_pencarian/98064
- Salmahaminati, Salmahaminati & Husnaqilati, Atina & Yahya, Amri. (2017). Statistical t Analysis for the Solution of Prediction Trash Management in Dusun Tanjung Sari Kec. Ngaglik Kab Sleman, Yogyakarta. *Journal of Physics: Conference Series*. 795. 012046. 10.1088/1742-6596/795/1/012046.
- Husnaqilati, Atina & Utami, Herni & Danardono. (2018). Survival Analysis for Cancer Patient with Stacked Method. *Advanced Science Letters*. 24. 678-681. 10.1166/asl.2018.11786.

A dichotomous behavior of Guttman-Kaiser rule from equi-correlated normal population and the limiting spectral distributions of random matrices

Akama, Y.[1] Husnaqilati, H.[1]

[1] Department of Mathematics, Tohoku University

数学・数理科学専攻若手研究者のための異分野・異業種研究交流会2021 13 November

H.F. Kaiser, an American psychologist who worked in psychometrics and statistical psychology, introduced ***Guttman-Kaiser rule, or eigenvalue-greater-than-one rule* for sample correlation matrix *R*** in order to assess the number ***q*** of significant components or factors in *principal component analysis* (PCA) and *explanatory factor analysis* (EFA) (See the histogram on the right).

As for the number of components or factors Guttman-Kaiser rule retains, the following ***dichotomy*** is reported:

(1) p/2, when the variables are independent, the sample size (*n*) is large and the number variables (*p*) is small, as in a simulation study Yeomans-Golder (*J. R. Stat. Soc. D.* 1992); but

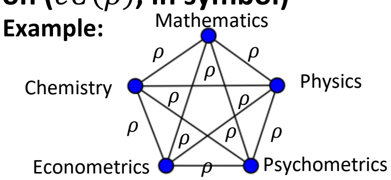
(2) **very small number**, when the number *p* of variables is small and the **"average" intercorrelation** among variables are large, as suggested by Kaiser, Humphreys and Tucker (Kaiser, *Percept. Mot. Ski.* 1992).

According to Kaiser (*Educ. Psychol. Meas.* 1962), the "average" intercorrelation is a positive constant ρ in a so-called **equi-correlation matrix**.

The equi-correlation matrix by definition ($eC(\rho)$, in symbol)

$$eC(\rho) = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{p \times p} \ (\rho \geq 0)$$

Example:



This structure is assumption to estimate a new covariance matrix estimator in various areas of finance (Engle-Kelly, *J. Bus. Econ. Stat*. 2008).

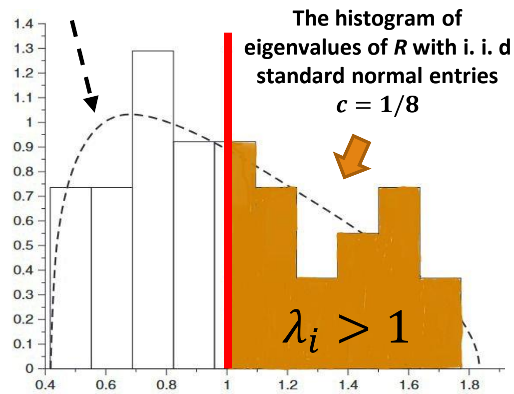
We precisely compute *q/p* in $n, p \to \infty$ with $\frac{p}{n} \to c > 0$ (high-dimensional and large sample sizes) and then the limit in $c \to 0$.

The *empirical spectral distribution* (ESD) of the sample correlation matrix *R of order p* is, by definition, a function

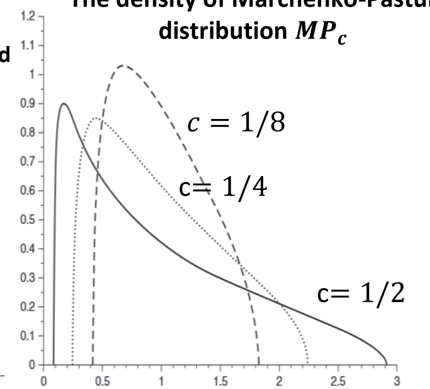
$$F^R(x) = \frac{1}{p} \#\{1 \leq i \leq p : \lambda_i \leq x\} \quad (x \in \mathbb{R})$$

where $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of R. The ESD of the sample covariance matrix S is similarly defined. ***The limiting spectral distribution*** (LSD) is the limit of ESD.

The density of LSD of matrix R



The density of Marchenko-Pastur distribution MP_c

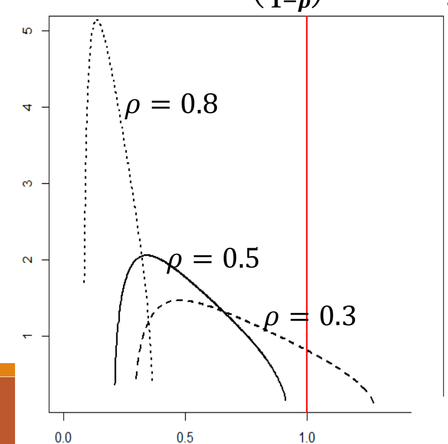


The LSDs of *S* and *R* for equi-correlated normal population

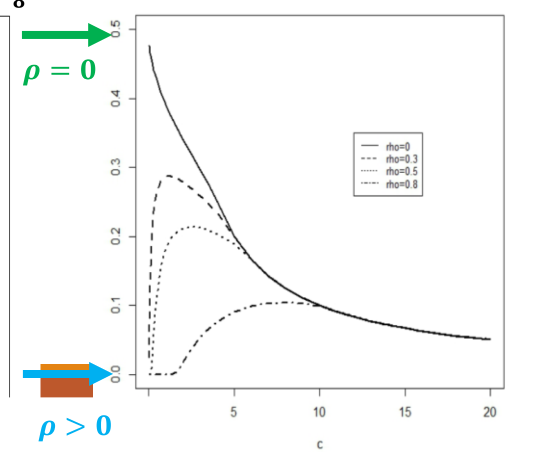
Theorem 1. Suppose that the samples $X_1, \ldots, X_n \sim N_p(0, \sigma^2 eC(\rho))$ are i.i.d. with $\sigma^2 > 0, \rho \geq 0$, and $n, p \to \infty$ with $\frac{p}{n} \to c > 0$. Then, a.s.,

$$\boldsymbol{F^S(x) \to MP_c\left(\frac{x}{\sigma^2(1-\rho)}\right)} \text{ and } \boldsymbol{F^R(x) \to MP_c\left(\frac{x}{1-\rho}\right)}, (x \in \mathbb{R}).$$

The density of $MP_c\left(\frac{x}{1-\rho}\right)$ with $c = \frac{1}{8}$



The rate of GK rule retains depend on c



Outlines

Eigenvalue-greater-than-one rule (GK rule) **for sample correlation matrix *R***

Rejection rate by GK rule is $F^R(1)$

All entries of the samples $X_1, \dots, X_n$ are i.i.d. with $\sigma^2 > 0$, $n, p \to \infty$, $\frac{p}{n} \to c > 0$ and finite second moment.

The LSD of *R* is $MP_c(x)$ (Jiang, *Sankhyā. 2004*)

$$\lim_{c \downarrow 0} MP_c(1) = \frac{1}{2}$$

As in Yeomans-Golder (*J. R. Stat. Soc. D.* 1992)

The samples $X_1, \dots, X_n \sim N_p\left(0, \sigma^2 eC(\rho)\right)$ are i.i.d. with $\sigma^2 > 0, \rho \geq 0$, and $n, p \to \infty$ with $\frac{p}{n} \to c > 0$.

The LSD of *R* is $MP_c\left(\frac{x}{1-\rho}\right)$ $(\rho \geq 0)$

$\rho = 0$:

$$\lim_{c \downarrow 0} MP_c\left(\frac{1}{1-0}\right) = \frac{1}{2}$$

As in Yeomans-Golder (*J. R. Stat. Soc. D.* 1992)

$0 < \rho < 1$:

$$\lim_{c \downarrow 0} MP_c\left(\frac{1}{1-\rho}\right) = 1$$

As in Kaiser (*Percept. Mot. Ski.* 1992)

Proof sketch of Theorem 1 (The LSDs of *S* and *R* for equi-correlated normal population)

1. Assume that the sample $X_1, \dots, X_n \sim N_p\big(0, \sigma^2 eC(\rho)\big)$ is i.i.d. with $\sigma^2 > 0, \rho \geq 0$.

2. Decompose $X = [X_1, \dots, X_n] = \sqrt{\rho}\begin{bmatrix} \xi_1 & \cdots & \xi_n \\ \vdots & \ddots & \vdots \\ \xi_1 & \cdots & \xi_n \end{bmatrix} + \sqrt{1-\rho}\begin{bmatrix} \xi_{11} & \cdots & \xi_{1n} \\ \vdots & \ddots & \vdots \\ \xi_{p1} & \cdots & \xi_{pn} \end{bmatrix}$ $(\xi_j, \xi_{ij} \sim N(0,1)$ are i.i.d. $(1 \leq i \leq p, 1 \leq j \leq n))$.

This matrix has rank at most 1

A

3. The ESD of the sample covariance matrix $S = \frac{1}{n}XX^T$ is close to the ESD of the sample covariance matrix of $\frac{1}{n}AA^T$. By the rank-1 matrix error, the difference between the two ESD is at most $\frac{1}{p}$.

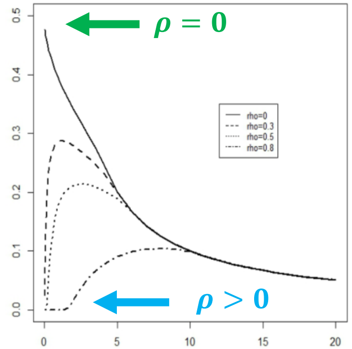
4. For $n, p \to \infty$ with $\frac{p}{n} \to c > 0$,

- LSD of *S* is MP_c scaled by $\sigma^2(1-\rho)$.
- LSD of *R* is the same as LSD of *S* for $\sigma^2 = 1$ (because *R* is invariant under scaling of variables), by Bai-Yin's uniform approximation of the sample mean of p variables (Lemma 2, Bai-Yin, *Ann. Probab*. 1993).

Derivation of dichotomy of GK rule w. r. t. ρ

Corollary 2

1. $\lim_{c \downarrow 0} MP_c(1) = \frac{1}{2} (\rho = 0).$
2. $\lim_{c \downarrow 0} MP_c\left(\frac{1}{1-\rho}\right) = 1 \ (0 < \rho < 1)$



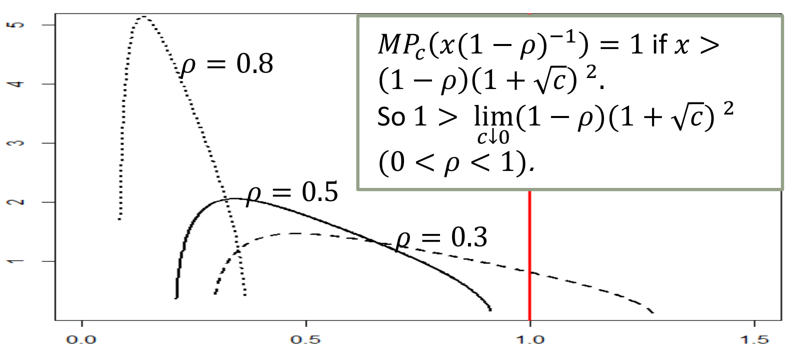
Proof of Corollary 2 (1)

If $X_c \sim MP_c$, then $X_c' := \frac{1}{2\sqrt{c}}(X_c - 1)$ converges in distribution to a random variable X′ that follows Wigner's semi-circle law as $c \to 0$. Thus,

$$\lim_{c \downarrow 0} 1 - MP_c(1) = \lim_{c \downarrow 0} \mathrm{P}(X_c \geq 1) = \lim_{c \downarrow 0} \mathrm{P}(X_c' \geq 0) = \mathrm{P}(X' \geq 0) = \frac{1}{2}$$

Proof of Corollary 2 (2)

The density of $MP_c\left(\frac{x}{1-\rho}\right)$ with $c = \frac{1}{8}$



Conclusion

The random matrices from multinormal population with equi-correlation structure ρ has the same LSDs but scaled by $1 - \rho$

Random matrix	LSD	Scalling of LSD (depending on population)	
		Multinormal with equi-correlation structure	All entries are i.i.d. random variables
$\frac{1}{2}\sqrt{\frac{n}{p}}(S - I_p)$	Semicircle	$1-\rho$	1 [Bai-Yin, *Ann. Probab*. 1988.]
$\frac{1}{2}\sqrt{\frac{n}{p}}(R - I_p)$		????	1 [Jiang, *Sankhyā. 2004*.]
S	Marcenko-Pastur	$1-\rho$	1 [Yao et al., Cambridge UP. 2015.]
R		$1-\rho$	1 [Jiang, *Sankhyā. 2004*.]
$S_1S_2^{-1}$	Fisher LSD	$\frac{1-\rho_2}{1-\rho_1}$	1 [Bai, *Stat. Sin. 1999*.]
$R_1R_2^{-1}$		????	???

S_i, R_i are formed from the i-th population (with equi-correlation structure ρ_i) and I_p is identity matrix order p.