Equi-correlated random matrices and high-dimensional statistics

（別紙様式5）

論文内容要旨

(NO. 1)

氏　名	Atina Husnaqilati	提出年	令和　5 年
学位論文の 題　　目	Equi-correlated random matrices and high-dimensional statistics (等相関ランダム行列と高次元統計学)		

論文目次

In a wide array of disciplines, the high-dimensional dataset is being generated. However, when the data dimension p is large, a number of well-known multivariate analysis techniques are known to become ineffective or even inaccurate (Bai-Silverstein, 2010). For datasets where p and the sample size n are large, it is common to consider the limiting regime of random matrix theory: $n, p \to \infty$ and $p/n \to c > 0$. In this asymptotic framework, Bai-Silverstein (2010) and Jiang (2004) studied the limiting spectral distributions (LSDs) of a sample covariance matrix and a sample correlation matrix. Here, the LSD of the sample correlation matrix $\mathbf{R}$ is the limit of empirical spectral distribution (ESD) F R in $n, p \to \infty, p/n \to c > 0$, where $F^{\boldsymbol{R}}(x) = p^{-1}\#\{1 \le i \le p \mid \lambda_i \le x\}$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ being the eigenvalues of $\mathbf{R}$. A sample correlation matrix $\mathbf{R}$ is difficult and important in econometrics and finance because it is invariant under the dilation and the shifting of data. However, $\mathbf{R}$ is complicated because of its definition. Hence, we focus on the LSDs of the sample correlation matrices.

A correlation coefficient may appear in a 1-dimensional sample: standard normal random variables $x_1, x_2, \ldots, x_n$ mutually correlated with a positive constant r enjoy a decomposition

$$x_i = \sqrt{r}\eta + \sqrt{1-r}\,\xi_i \tag{1}$$

where η and ξ_i $(1 \le i \le n)$ are independent, standard normal random variables. Walsh (1947) employed (1) to prove that hypothesis tests worsen nonasymptotically, when a sample $x_1, x_2, \ldots, x_n$ drawn from a normal population is actually mutually correlated with a positive constant r. Moreover, let two samples be $x_1, x_2, \ldots, x_{n_1}$ and $x_1, x_2, \ldots, x_{n_2}$ mutually correlated with positive constants r_1 and r_2, respectively, drawn from two independent normal populations. By the decomposition (1), Walsh (1947) also showed that the proportion of the two variances of the two samples obeys the Snedecor F-statistics scaled by $(1 - r_1)/(1 - r_2)$.

We are also concerned with correlation coefficients in multivariate datasets e.g., high dimensional datasets of econometrics and finance. These datasets are often generated from factor models having many factors. By averaging the effect of the factors, we get factor models with only one factor. For example, we can consider an independent sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ drawn from a p-dimensional normal population such that all the p components of j $(1 \leq j \leq n)$, all the p components of $\boldsymbol{x}_j = [x_{ij}]_{1 \leq i \leq p}$ are mutually correlated by $r \in [0, 1)$ (equi-correlated normal population). In this case, we have the following decomposition:

$$x_{\mathrm{ij}} = \sqrt{r}\eta_j + \sqrt{1-r}\,\xi_{\mathrm{ij}} \qquad (2)$$

We will explore an asymptotic deterioration of statistical inference by using the LSD of the sample correlation matrix formed from an equi-correlated normal population. For this, we combine the decomposition (2) and a simple linear algebraic technique, rank inequality (Bai-Silverstein, 2010).

Theorem 1 (Akama-Husnaqilati, 2022)
Suppose both n and p go to infinity with $\frac{p}{n}$ going to positive c . Then, almost surely, the empirical spectral distribution $F^{\boldsymbol{R}}(x)$ converges weakly to $F_c\left(\frac{x}{1-r}\right)$. Here, $F_c(x)$ is the Marčenko-Pastur distribution of index c .

Theorem 1 answers a question from Fan-Jiang (2019) about the impacts of equi-correlation coefficient r on the LSD of $\mathbf{R}$.

The application of Theorem 1 can be found in principal component analysis (Jolliffe, 2002) which reduces the dimensionality of high-dimensional large samples by retaining the number of new significant uncorrelated variables (principal components (Jolliffe, 2002)) that successively maximize the variance of a dataset. The number of significant principal components is based on some statistical inferences. One of these is Guttman-Kaiser criterion (Kaiser, 1992). It suggests that the number of significant principal components is equal to the number k eigenvalues of $\mathbf{R}$ greater than 1 (Jolliffe, 2002) . By Theorem 1, we prove the following phase transition of the limit theorem.

$$\lim_{c \to 0} \lim_{\substack{p,n \to \infty \\ p/n \to c}} \frac{k}{p} = \begin{cases} \frac{1}{2} & (r = 0); \\ 0 & (r > 0). \end{cases}$$

This elucidates mathematically the deterioration of Guttman-Kaiser criterion in high- dimensional large samples by constant positive correlation among variables, and the convergence of k/p to 1/2 for $r = 0$ suggested by a simulation study of Yeomans-Golder (1982).

Finally, we provide the LSDs of various random matrices formed from equi-correlated normal populations, by using the rank inequality with the decomposition (1) or the decomposition (2). We first consider the product of the sample covariance matrix formed from an equi-correlatted normal population with correlation coefficient $r_1 \in [0, 1)$ and the inverse of the other sample covariance matrix formed from an equi-correlatted normal population with correlation coefficient $r_2 \in [0, 1)$. This matrix has the LSD given by (Bai et al., 1988, Silverstein, 1995) but scaled by $(1 - r_2)/(1 - r_1)$ (Husnaqilati, 2022). This result is a counterpart of the finding from

Walsh (1947) about the Snedecor F-statistics for univariate statistical analysis. Moreover, the combination of (1) and the rank inequality (Bai-Silverstein, 2010) establishes that a Wigner matrix (a symmetric Toeplitz matrix, and Hankel matrix, resp.) with all entries mutually correlated by a nonnegative $r < 1$ has the LSD given by (Bryc et al., 2006) but scaled by $\sqrt{1-r}$ (Husnaqilati, 2022).

別　紙

論文審査の結果の要旨

昨今の大規模高次元データの数理統計学ではデータの次元pと標本の大きさnが比例的に無限大にいく設定（Kolmogorov regime）が有用で、この設定における研究としては、標本分散行列の固有値の分布に関する研究（Z. D. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer, 2nd edition, 2010）や、Jiang（The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, 2004）や近年のHeinyによる標本相関行列**R**の固有値の極限分布の結果がある。

経済学・ファイナンスの大規模高次元データの基本的なモデルは、変数の間の一定の相関係数（同相関係数）rが非負である多次元正規母集団であり、相関行列の逆行列を高速に計算する必要がある時系列解析（R. Engle and B. Kelly. Dynamic equicorrelation. *J Bus Econ Stat*, 2012）において採用されている。データの倍化・移動に関して不変である標本相関行列**R**も経済学・ファイナンス重要である。同相関係数が非負である多次元正規母集団の**R**の固有値のバルク分布の、Kolmogorov regimeにおける極限を、プリンストン大のFanとミネソタ大のJiangは問いかけた（J. Fan and T. Jiang. Largest entries of sample correlation matrices from equi-correlated normal populations. *Ann. Probab.*, 2019)。

フスナキラティ アティナ氏の博士論文はFanとJiangのこの問に解答した。**R**の固有値のバルク分布の極限が、標準的な自由ポアソン分布（F. Hiai and D. Petz. *The semicircle law, free random variables and entropy*, American Mathematical Society, 2000）（Marchenko=Pastur分布（Z. D. Bai and J. W. Silverstein. *ibid.*)）の1−r倍に収縮することを証明した。

関連研究だが、プリンストン大のWalsh（Concerning the effect of intraclass correlation on certain significance tests. *Ann. Math. Stat.*, 1947)は、1次元の標本の変数の間に互いに一定の正の相関係数rがある場合、カイ二乗分布などの標本分布による仮説検定の質の低下を、標本分布の収縮として解明している。Walshは、一定の正の相関係数を持つ1次元正規分布に従う各々の変数は、各々に固有な独立な正規変数とある共通の正規変数の和であることを用いた。この正規変数の分解原理により、フスナキラティ アティナ氏の博士論文は、遥かに困難な標本相関行列に取り組み、この原理が種々のランダム行列（Z. D. Bai and J. W. Silverstein. *ibid.*）に適用可能であることを確認している。

Marchenko=Pastur分布の同相関係数による収縮定理の応用として、高次元大規模データの統計学における有名な推論で現代的統計ソフトウェアに実装されているGuttman-Kaiser基準の質が、変数の間に互いに正の相関係数が存在する場合に著しく低下するという経験則（H. F. Kaiser. On Cliff's formula, the Kaiser-Guttman rule, and the number of factors. *Percept. Mot. Ski.*, 1992)は、同相関係数r>0の正規母集団の標本相関行列**R**の固有値分布の極限定理のrに関する相転移であると数学的に解明した。

このように、フスナキラティ アティナ氏の業績は、(1) 正の同相関係数を持つ多次元正規母集団の標本相関行列の極限固有値分布の収縮定理、(2) (1)の応用として、Guttman-Kaiser基準の、同相関係数に関する相転移定理、(3) (1)が基づく正の同相関係数を持つ多次元正規変数たちの分解原理について、その発展と応用の研究、に及ぶ。

自立して研究活動を行うに必要な高度の研究能力と学識を有することを示している。したがって、フスナキラティ アティナ氏提出の博士論文は、博士（理学）の学位論文として合格と認める。