International Symposium on Theories, Methodologies and Applications for Large Complex Data

December 4-6, 2024

Venue:

Conference Room 202, Tsukuba International Congress Center
2-20-3 Takezono, Tsukuba, Ibaraki 305-0032, Japan

Organizers:

Makoto Aoshima (University of Tsukuba)
Kazuyoshi Yata (University of Tsukuba)
Aki Ishii (Tokyo University of Science)
Kento Egashira (Tokyo University of Science)

Supported by

Grant-in-Aid for Scientific Research (A) 20H00576 (Project Period: 2020-2024)
"Innovative developments of theories and methodologies for large complex data"
(Principal Investigator: Makoto Aoshima)

Grant-in-Aid for Challenging Research (Exploratory) 22K19769 (Project Period: 2022-2024)
"Developments of statistical compression technology for massive data having tensor structures"
(Principal Investigator: Makoto Aoshima)

Program (UTC+9)

December 4 (Wednesday)

13:40∼13:50 Opening

13:50∼14:30 Aki Ishii[*,a], Yumu Iwana[b], Kazuyoshi Yata[c] and Makoto Aoshima[c]
[a](Department of Information Sciences, Tokyo University of Science)
[b](Graduate School of Science and Technology, University of Tsukuba)
[c](Institute of Mathematics, University of Tsukuba)
Statistical inference on high-dimensional covariance structures under the SSE models

14:40∼15:20 Yoshikazu Terada
(Graduate School of Engineering Science, Osaka University)
Statistical properties of matrix decomposition factor analysis

15:30∼16:10 Tsutomu T. Takeuchi
(Division of Particle and Astrophysical Science, Nagoya University)
High-dimensional statistics in astrophysics and its perspective

16:30∼17:10 Shao-Hsuan Wang

(Graduate Institute of Statistics, National Central University)

High-dimensional inference on a cross data matrix-based method

17:20∼18:00 Yuan-Tsung Chang[*,a], Nobuo Shinozaki[b] and William E. Strawderman[c]

[a](The Institute of Statistical Mathematics)

[b](Faculty of Science and Technology, Keio University)

[c](Department of Statistics, Rutgers University)

On estimation of a matrix mean under matrix loss

December 5 (Thursday)

9:00∼11:00 **Young Researchers Session**

1. Tetsuya Umino (Graduate School of Science and Technology, University of Tsukuba)
   Automatic sparse estimation of high-dimensional cross-covariance matrix
2. Dongsun Yoon (Department of Statistics, Seoul National University)
   Augmented estimation of principal component subspace in high dimensions
3. Giheon Seong (Department of Statistics, Seoul National University)
   James-Stein estimator of spiked leading eigenvector of high-dimensional covariance matrix
4. Yongjae Kim (Department of Statistics, Seoul National University)
   General measures of attribution disclosure risk for gauging privacy of synthetic data
5. Guan Xin (Graduate School of Engineering Science, Osaka University)
   Regularized k-POD clustering for high-dimensional missing data

11:10∼11:50 Masaaki Imaizumi

(Komaba Institute for Science, University of Tokyo / RIKEN AIP)

Non-sparse high-dimensional statistics: structured model, neural network, and universality

11:50∼13:40 Lunch

13:40∼18:00 **Special Invited and Keynote Sessions**

19:00∼21:00 Dinner

December 6 (Friday)

9:00∼9:40 Shogo Nakakita

(Komaba Institute for Science, University of Tokyo)

On dimension-free concentration of logistic regression

9:50∼10:30 Sangil Han

(Institute for Data Innovation in Science, Seoul National University)

Subspace recovery in winsorized PCA

10:45∼11:25 Yuta Koike

(Graduate School of Mathematical Sciences, University of Tokyo)

High-dimensional bootstrap and asymptotic expansion

11:35∼12:15 Takahiro Nishiyama[*,a], Masashi Hyodo[b] and Shoichi Narita[c]

[a](Department of Business Administration, Senshu University)

[b](Faculty of Economics, Kanagawa University)

[c](Graduate School of Economics, Kanagawa University)

On a test for assessing vector correlation for latent factor models in high-dimensional settings

12:25∼13:05 Yohji Akama

(Mathematical Institute, Tohoku University)

Asymptotic locations of bounded and unbounded eigenvalues of sample correlation matrices of certain factor models – application to a components retention rule

13:05∼13:10 Closing

(∗ Speaker)

Special Invited and Keynote Sessions

December 5 (Thursday)

Special Invited Session

13:40∼14:30 **Difference between large statistical model and medium statistical model**

Speaker: Shurong Zheng
(School of Mathematics and Statistics, Northeast Normal University)

Discussion Leader: Kento Egashira (Department of Information Sciences, Tokyo University of Science)

14:40∼15:30 **Principal component analysis for zero-inflated compositional data**

Speaker: Sungkyu Jung
(Institute for Data Innovation in Science, Seoul National University)

Discussion Leader: Kazuyoshi Yata (Institute of Mathematics, University of Tsukuba)

Keynote Session

15:50∼16:50 **A generalized mean approach for distributed-PCA**

Speaker: Su-Yun Huang
(Institute of Statistical Science, Academia Sinica)

Discussion Leader: Yuan-Tsung Chang (The Institute of Statistical Mathematics)

17:00∼18:00 **Alignment and matching tests for high-dimensional tensor signals via tensor contraction**

Speaker: Jianfeng Yao
(School of Data Science, Chinese University of Hong Kong (Shenzhen))

Discussion Leader: Yuta Koike (Graduate School of Mathematical Sciences, University of Tokyo)

Statistical inference on high-dimensional covariance structures under the SSE models

Aki Ishii[a], Yumu Iwana[b], Kazuyoshi Yata[c] and Makoto Aoshima[c]

[a] Department of Information Sciences, Tokyo University of Science
[b] Graduate School of Science and Technology, University of Tsukuba
[c] Institute of Mathematics, University of Tsukuba

Key words and phrases: ECDM; HDLSS; Noise-reduction method; SSE model

1 Introduction

We consider correlation tests for "high-dimension, low-sample-size (HDLSS)" data. Recently, Aoshima and Yata [2] created the two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. In this talk, we focus on the SSE model. Suppose that we take samples $\boldsymbol{x}_j$, $j = 1, \ldots, n$, of size n (≥ 4), which are independent and identically distributed (i.i.d.) as a p-variate distribution. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^\top, \boldsymbol{x}_{2j}^\top)^\top$ and assume that $\boldsymbol{x}_{ij} \in \mathbf{R}^{p_i}$, $i = 1, 2$, with $p_1 \in [1, p-1]$ and $p_2 = p - p_1$. We also assume that $\boldsymbol{x}_j$ has an unknown mean vector, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$, and unknown covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^\top & \boldsymbol{\Sigma}_2 \end{pmatrix} \ (\geq \boldsymbol{O}),$$

that is, $\mathrm{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i$, $\mathrm{Var}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_i$, $i = 1, 2$, and $\mathrm{Cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \mathrm{E}(\boldsymbol{x}_{1j}\boldsymbol{x}_{2j}^\top) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^\top = \boldsymbol{\Sigma}_*$. Let σ_{ij} be the j-th diagonal element of $\boldsymbol{\Sigma}_i$ for $i = 1, 2$; $j = 1, \ldots, p_i$, and assume $\sigma_{ij} > 0$ for all i, j. We denote the correlation coefficient matrix between $\boldsymbol{x}_{1j}$ and $\boldsymbol{x}_{2j}$ by $\mathrm{Corr}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \boldsymbol{P}$, where $\boldsymbol{P} = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{1p_1})^{-1/2}\boldsymbol{\Sigma}_*\mathrm{diag}(\sigma_{21}, \ldots, \sigma_{2p_2})^{-1/2}$. Here, $\mathrm{diag}(\sigma_{i1}, \ldots, \sigma_{ip_i})$ denotes the diagonal matrix of elements, $\sigma_{i1}, \ldots, \sigma_{ip_i}$. Then, we consider testing the following hypotheses:

$$H_0 : \boldsymbol{P} = \boldsymbol{O} \quad \text{vs.} \quad H_1 : \boldsymbol{P} \neq \boldsymbol{O}. \tag{1}$$

2 Correlation test under the SSE model

We assume that p_1 is fixed. We also assume the following condition:

(A-i) $\dfrac{\lambda_{\max}(\boldsymbol{\Sigma}_2)}{\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}} \to 1, \quad p_2 \to \infty$, where $\lambda_{\max}(\boldsymbol{\Sigma})$ is the largest eigenvalue of $\boldsymbol{\Sigma}_2$

The model (A-i) is one of the SSE models and is called the "uni-SSE model" in Ishii, Yata and Aoshima [3].

Aoshima and Yata [1] gave a test statistic for testing (1) and Yata and Aoshima [4] improved the test statistic by using the *extended cross-data-matrix (ECDM) methodology*. They gave asymptotic normality of the test statistic under one of the NSSE models.

Let $\Delta = \mathrm{tr}(\mathbf{\Sigma}_* \mathbf{\Sigma}_*^\top) (= \|\mathbf{\Sigma}_*\|_F^2)$, where $\|\cdot\|_F$ is the Frobenius norm. We introduce an unbiased estimator of Δ by the ECDM methodology. We define $n_{(1)} = \lceil n/2 \rceil$ and $n_{(2)} = n - n_{(1)}$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. Let

$$\boldsymbol{V}_{n(1)(k)} = \begin{cases} \{\lfloor k/2 \rfloor - n_{(1)} + 1, \ldots, \lfloor k/2 \rfloor\} & \text{if } \lfloor k/2 \rfloor \geq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor\} \cup \{\lfloor k/2 \rfloor + n_{(2)} + 1, \ldots, n\} & \text{otherwise;} \end{cases}$$

$$\boldsymbol{V}_{n(2)(k)} = \begin{cases} \{\lfloor k/2 \rfloor + 1, \ldots, \lfloor k/2 \rfloor + n_{(2)}\} & \text{if } \lfloor k/2 \rfloor \leq n_{(1)}, \\ \{1, \ldots, \lfloor k/2 \rfloor - n_{(1)}\} \cup \{\lfloor k/2 \rfloor + 1, \ldots, n\} & \text{otherwise} \end{cases}$$

for $k = 3, \ldots, 2n-1$, where $\lfloor x \rfloor$ denotes the largest integer $\leq x$. Also, let $\#\boldsymbol{A}$ denote the number of elements in a set $\boldsymbol{A}$. Note that $\#\boldsymbol{V}_{n(l)(k)} = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{n(1)(k)} \cap \boldsymbol{V}_{n(2)(k)} = \emptyset$ and $\boldsymbol{V}_{n(1)(k)} \cup \boldsymbol{V}_{n(2)(k)} = \{1, \ldots, n\}$ for $k = 3, \ldots, 2n-1$. It should be noted that

$$i \in \boldsymbol{V}_{n(1)(i+j)} \quad \text{and} \quad j \in \boldsymbol{V}_{n(2)(i+j)} \quad \text{for } i < j \ (\leq n). \tag{2}$$

Let

$$\overline{\boldsymbol{x}}_{l(1)(k)} = n_{(1)}^{-1} \sum_{j \in \boldsymbol{V}_{n(1)(k)}} \boldsymbol{x}_{lj} \quad \text{and} \quad \overline{\boldsymbol{x}}_{l(2)(k)} = n_{(2)}^{-1} \sum_{j \in \boldsymbol{V}_{n(2)(k)}} \boldsymbol{x}_{lj}, \quad l = 1, 2$$

for $k = 3, \ldots, 2n-1$. We consider the following quantity:

$$\widehat{\Delta}_{ij} = (\boldsymbol{x}_{1i} - \overline{\boldsymbol{x}}_{1(1)(i+j)})^\top (\boldsymbol{x}_{1j} - \overline{\boldsymbol{x}}_{1(2)(i+j)}) (\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_{2(1)(i+j)})^\top (\boldsymbol{x}_{2j} - \overline{\boldsymbol{x}}_{2(2)(i+j)})$$

for all $i < j \ (\leq n)$. Let $u_n = n_{(1)} n_{(2)} \{(n_{(1)} - 1)(n_{(2)} - 1)\}^{-1}$. Yata and Aoshima [4] proposed an unbiased estimator of Δ by

$$\widehat{T}_n = \frac{2u_n}{n(n-1)} \sum_{i<j}^{n} \widehat{\Delta}_{ij}.$$

Theorem 2.1. *Assume (A-i) and some regularity conditions. Then, it holds that as* $m = \min\{p, n\} \to \infty$

$$\frac{n(\widehat{T}_n - \Delta)}{\lambda_{\max}(\mathbf{\Sigma}_2)} + \mathrm{tr}(\mathbf{\Sigma}_1) \Rightarrow \sum_{s=1}^{p_1} \lambda_{1s} \chi_{1s}^2,$$

where λ_{1s} *is the s-th eigenvalue of* $\mathbf{\Sigma}_1$, χ_{1s}^2 *stands for a chi-square random variable with* 1 *degree of freedom and* χ_{1s}^2, $s = 1, ..., p_1$ *are mutually independent.*

[1] M. Aoshima, K. Yata, Two-stage procedures for high-dimensional data, Sequential Anal. (Editor's special invited paper) 30 (2011) 356–399.

[2] M. Aoshima, K. Yata, Two-sample tests for high-dimension, strongly spiked eigenvalue models. Statist. Sinica 28 (2018) 43–62.

[3] A. Ishii, K. Yata, M. Aoshima, Hypothesis tests for high-dimensional covariance structures. Ann. Inst. Statist. Math. 73 (2021) 599–622.

[4] Yata, K., Aoshima, M. (2016). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. J. Multivariate Anal. 151 (2016) 151–166.

Statistical Properties of Matrix Decomposition Factor Analysis

Yoshikazu Terada*
Graduate School of Engineering Science, Osaka University
Center for Advanced Integrated Intelligence Research, RIKEN

Exploratory factor analysis, often referred to as factor analysis, is an important technique of multivariate analysis (Anderson 2003). Factor analysis is a method for exploring the underlying structure of a set of variables and is applied in various fields. In factor analysis, we consider the following model for a p-dimensional observation x:

$$x = \mu + \Lambda f + \epsilon, \tag{1}$$

where $\mu \in \mathbb{R}^p$ is a mean vector, m is the number of factors ($m < p$), $\Lambda \in \mathbb{R}^{p\times m}$ is a factor loading matrix, f be a m-dimensional centered random vector with the identity covariance, ϵ be a p-dimensional uncorrelated centered random vector, which is independent from f, with diagonal covariance matrix $\mathrm{Var}(\epsilon) = \Psi^2 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Each component of f and ϵ are called the common and unique factors, respectively.

For a constant $c_\Lambda > 0$, let $\Theta_\Lambda := \{\Lambda \in \mathbb{R}^{p\times m} \mid |\lambda_{jk}| \leq c_\Lambda \ (j = 1, \ldots, p;\ k = 1, \ldots, m)\}$ be the parameter space for the factor loading matrix Λ. For positive constants $c_L, c_U > 0$, define the parameter space for Ψ as $\Theta_\Psi := \{\mathrm{diag}(\sigma_1, \ldots, \sigma_p) \mid c_L \leq |\sigma_j| \leq c_U \ (j = 1, \ldots, p)\}$. Let $\Phi = [\Lambda, \Psi] \in \mathbb{R}^{p\times(m+p)}$, and define $\Theta_\Phi = \{\Phi = [\Lambda, \Psi] \mid \Lambda \in \Theta_\Lambda \text{ and } \Psi \in \Theta_\Psi\}$. For the factor model (1) with $\Phi = [\Lambda, \Psi]$, the covariance matrix of x is represented as $\Phi\Phi^\top = \Lambda\Lambda^\top + \Psi^2$.

We assume that the factor model (1) is true with some unknown parameter $\Phi_* = [\Lambda_*, \Psi_*] \in \Theta_\Phi$. Let $\Sigma_* = \Phi_*\Phi_*^\top = \Lambda_*\Lambda_*^\top + \Psi_*^2$ denote the true covariance matrix. It should be noted that the statistical properties described later still hold as a minimum contrast estimator even when the factor model (1) is not true. Let $(x_1, f_1, \epsilon_1), \ldots, (x_n, f_n, \epsilon_n)$ be i.i.d. copies of (x, f, ϵ), where $(f_1, \epsilon_1), \ldots, (f_n, \epsilon_n)$ are not observable in practice. Throughout the paper, it is assumed that $n > m+p$. In factor analysis, we aim to estimate (Λ_*, Ψ_*) from the observations $X_n = (x_1, \ldots, x_n)^\top$. Here, we note that the factor model (1) has an indeterminacy. For example, for any $m \times m$ orthogonal matrix R, a rotated loading matrix $\Lambda_* R$ can also serve as a true loading matrix. Thus, let $\Theta_\Phi^* = \{\Phi \in \Theta_\Phi \mid \Sigma_* = \Phi\Phi^\top\}$ be the set of all possible true parameters.

There are several estimation approaches to estimate the parameter $\Phi = [\Lambda, \Psi]$. The theoretical properties of these estimation approaches have been extensively studied. Moreover, most of these estimators can be formulated as minimum discrepancy function estimators.

*This research was supported by JSPS KAKENHI Grant (JP20K19756, JP20H00601, JP23H03355, and JP24K14855).

Thus, we can apply the general theory of minimum discrepancy function estimators to derive the theoretical properties of the estimators (Shapiro 1983, 1985).

In the early 2000s, a novel estimator based on matrix factorization was developed for factor analysis. According to Adachi & Trendafilov (2018), this method was originally developed by Professor Henk A. L. Kiers and first appeared in Socan's dissertation (Socan 2003). This method is called matrix decomposition factor analysis (MDFA for short). The MDFA algorithm always provides proper solutions (i.e., no Heywood cases in MDFA); thus, it is computationally more stable than the maximum likelihood estimator. From the aspect of computational statistics, matrix decomposition factor analysis has been well-studied, and several extensions have been developed.

In matrix decomposition factor analysis, the estimator is obtained by minimizing the following principal component analysis-like loss function:

$$\mathcal{L}_n(\mu, \Lambda, \Psi, F, E) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - (\mu + \Lambda f_i + \Psi e_i)\|^2,$$

where $e_i = (e_{i1}, \ldots, e_{ip})^\top$, $E = (e_1, \ldots, e_n)^\top \in \mathbb{R}^{n \times p}$, and $F = (f_1, \ldots, f_n)^\top \in \mathbb{R}^{n \times m}$. Certain constraints are imposed on the common factor matrix F and the normalized unique factor matrix E. Unlike classical factor analysis, matrix decomposition factor analysis treats the common factors F and normalized unique factors E as parameters that are estimated simultaneously with $\Phi = [\Lambda, \Psi]$. The number of parameters linearly depends on the sample size n, and the standard asymptotic theory of classical M-estimators cannot be directly applied to analyze its theoretical properties. As a result, the statistical properties of the MDFA estimator have yet to be discussed, leading to the open problem: Can matrix decomposition factor analysis truly be regarded as "factor analysis"?

In this talk, we establish the statistical properties of matrix decomposition factor analysis to answer this question. We show that as the sample size n goes to infinity, the MDFA estimator converges to the true parameter $\Phi_* \in \Theta_\Phi^*$. First, we formulate the MDFA estimator as the semiparametric profile likelihood estimator and derive the explicit form of the profile likelihood. Next, we reveal the population-level loss function of matrix decomposition factor analysis and its fundamental properties. Then, we show the statistical properties of matrix decomposition factor analysis.

References

Adachi, K. & Trendafilov, N. T. (2018), 'Some mathematical properties of the matrix decomposition solution in factor analysis', *Psychometrika* **83**, 407–424.

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis*, Wiley.

Shapiro, A. (1983), 'Asymptotic distribution theory in the analysis of covariance structures', *South African Statistical Journal* **17**(1), 33–81.

Shapiro, A. (1985), 'Asymptotic equivalence of minimum discrepancy function estimators to G.L.S. estimators', *South African Statistical Journal* **19**(1), 73–81.

Socan, G. (2003), The incremental value of minimum rank factor analysis., PhD dissertation, University of Groningen.

High-Dimensional Statistics in Astrophysics and its Perspective

Tsutomu T. Takeuchi[1,2]*, Kazuyoshi Yata, Kento Egashira,
Makoto Aoshima, Nanase Harada, Kohji Yoshikawa, Aki Ishii, Hiroma Okubo,
Ryusei R. Kano, Wen E. Shi, Aina May So, Hai-Xia Ma,
Sena A. Matsui, Koichiro Nakanishi, Sucheta Cooray, Kotaro Kohno

1. Division of Particle and Astrophysical Science, Nagoya University, Japan,
2. Research Center for Statistical Machine Learning, Institute of Statistical Mathematics, Japan,

1 Main Result

If we denote the dimension of data as d and the number of samples as n, we often meet a case with $n \ll d$. Traditionally in astronomy, such a situation is regarded as ill-posed, and they thought that there was no choice but to throw away most of the information in data dimension to let $d < n$. The data with $n \ll d$ is referred to as high-dimensional low sample size (HDLSS). To deal with HDLSS problems, a method called high-dimensional statistics has been developed rapidly in the last decade.

In this work, we first introduce the high-dimensional statistical analysis. We apply two representative methods in the high-dimensional statistical analysis methods, the noise-reduction principal component analysis (NRPCA) and automatic sparse principal component analysis (A-SPCA), to a spectroscopic map of a nearby archetype starburst galaxy NGC 253 taken by the Atacama Large Millimeter/Submillimeter Array (ALMA). The ALMA map is a typical HDLSS dataset. First we analyzed the original data including the Doppler shift due to the systemic rotation. The high-dimensional PCA could describe the spatial structure of the rotation precisely. We then applied to the Doppler-shift corrected data to analyze more subtle spectral features. The NRPCA and A-SPCA could quantify the very complicated characteristics of the ALMA spectra. Particularly, we could extract the information of the global outflow from the center of NGC 253. This method can also be applied not only to spectroscopic survey data, but also any type of HDLSS data. The main result is published in Takeuchi et al. (2024) and Takeuchi et al. (2024), Toukei SUuri, in press.

2 Further Development for the Next Generation Data

The original data of this study were recently updated to the one with much higher quality. The new data contains information of very weak spectral lines from molecules or radicals (ionized

*E-mail: tsutomu.takeuchi.ttt@gmail.com.

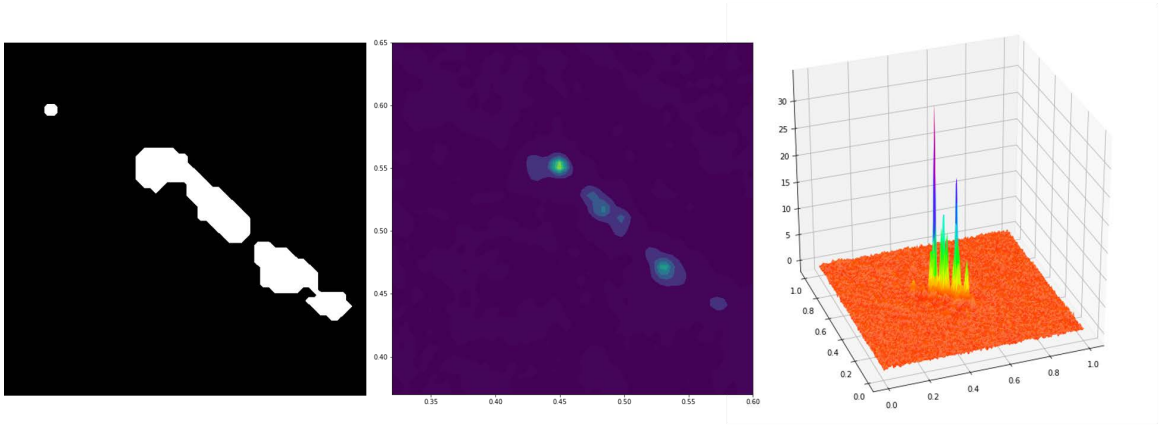

Figure 1: The bright regions of NGC 253 map cut out by the mask. Left: the mask region map. White regions have significant intensity signals. Center: the cut-out region with significantly bright emission. Right: the bird's view of the signal.

molecules) in NGC 253. To analyze such data, it would make sense to apply an analysis method which can deal with nonlinear correlation of data features. Kernel PCA is one of such possibilities. We will develop this study with such methods as our next step.

References

Takeuchi, T. T., Yata, K., Egashira, K., et al. 2024, ApJS, 271, 44

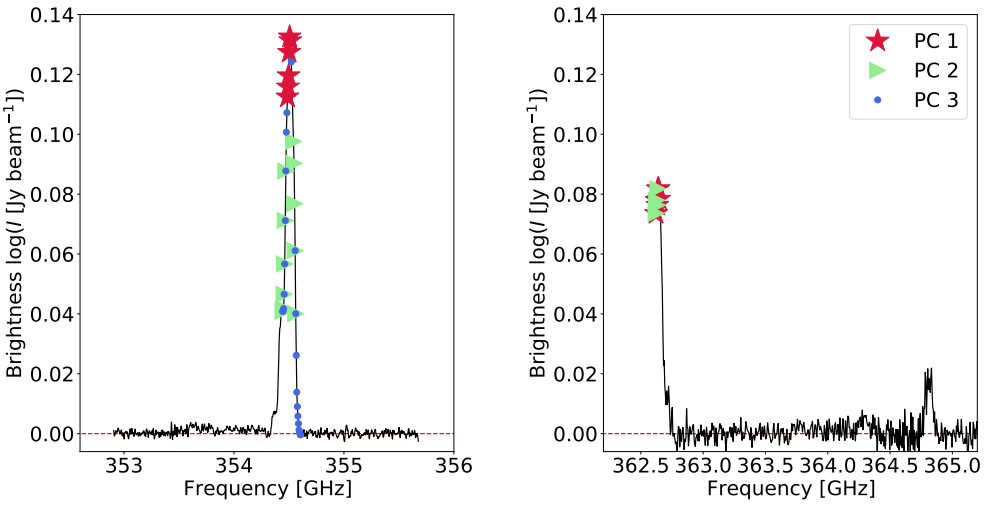

Figure 2: Responsible features to characterize PCs from the RPCA for the ALMA map of NGC 253, after the Doppler shift correction due to the systemic rotation. Information on the details of this figure is found in Takeuchi et al. (2024).

High-dimensional inference on a cross data matrix-based method

Shao-Hsuan Wang (National Central University)

Abstract

The concept of the cross-data matrix originates from the work of Yata and Aoshima (2010), who demonstrated that the cross-data matrix-based principal component analysis (CDM-PCA) method can effectively reduce noise and enhance the performance of principal component analysis (PCA) in high-dimensional, low-sample-size settings. This innovative approach has inspired numerous subsequent studies. For instance, Wang, Huang, and Chen (2020) established the asymptotic normality of estimates for principal component directions, while Wang and Huang (2022) derived finite-sample approximations and explored the asymptotic behavior of CDM-based PCA through matrix perturbation theory. More recently, Hung and Huang (2023) introduced a more stable variant of CDM-PCA, termed product-PCA (PPCA). This formulation offers a more convenient structure for theoretical analysis and has been shown to be more robust than PCA in preserving the correct ordering of leading eigenvalues, even in the presence of outliers.

In this talk, I will discuss recent advances in the cross-data matrix-based methods for high-dimensional data analysis, which will be presented in two parts. First, I will introduce cross-data matrix-based Multilinear Principal Component Analysis (CDM-MPCA) along with its numerical studies. In the second part, I will present the limiting spectral distribution (LSD) for the singular values of large cross-data matrix-based sample covariance matrix. Additionally, I will compare this distribution with the Marchenko–Pastur law (MP law), which characterizes the asymptotic behavior of the singular values of large sample covariance matrix.

On Estimation of a Matrix Mean under Matrix Loss

Yuan-Tsung Chang (The Institute of Statistical Mathematics),
Nobuo Shinozaki (Faculty of Science and Technology, Keio University)
William E. Strawderman (Department of Statistics, Rutgers University)

International Symposium on Theories, Methodologies and Applications for Large Complex Data, Dec. 4-6, 2024, at Tsukuba International Congress Center

Abstract Consider estimating an $n \times p$ matrix means of matrix random variables $X_{n\times p}$ under matrix quadratic error loss function. Abu-Shanab, Kent and Strawderman (2012) studied the independent normal distributions version and proposed a matrix version of shrinkage estimators which is dependent on a tuning constant a. We generalize their results to a broad class of models including estimation of Poisson means, estimating of Binomial samples sizes, estimating of natural parameters of discrete and continuous exponential families.

1 Introduction

Let

$$X_{n\times p} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) = (x_{ij})_{n\times p} \tag{1}$$

be an $n \times p$ matrix of independent random variables such that

$$E(X) = \Theta = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p) = (\theta_{ij})_{n\times p}. \tag{2}$$

The object is to estimate Θ and, in particular, to find an estimator which improves over the unbiased estimator

$$\delta_0(X) = (\boldsymbol{\delta}_{01}(\boldsymbol{X}_1), \ldots, \boldsymbol{\delta}_{0p}(\boldsymbol{X}_p)) = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p) \tag{3}$$

We make the following assumption throughout.

Assumption (1): There exists an estimator of the form

$$\boldsymbol{\delta}(\boldsymbol{X}) = \boldsymbol{X} + \boldsymbol{g}(\boldsymbol{X}),$$

such that for each j, $\boldsymbol{\delta}(\boldsymbol{X}_j) = \boldsymbol{X}_j + \boldsymbol{g}(\boldsymbol{X}_j)$ dominates the estimator $\boldsymbol{\delta}_0(\boldsymbol{X}_j) = \boldsymbol{X}_j$ under scalar quadratic loss

$$L_j(\boldsymbol{\theta}_j, \boldsymbol{d}_j) = \sum_{i=1}^{n} (\theta_{ij} - d_{ij})^2 w_{ij}^2(\theta_{ij}). \tag{4}$$

It follows from Assumption (1) that the difference in risk between $\boldsymbol{\delta}(\boldsymbol{X}_j)$ and $\boldsymbol{X}_j$ satisfies

$$\begin{aligned} \Delta_j(\boldsymbol{\theta}_j, \boldsymbol{\delta}(\boldsymbol{X}_j)) &= E[L_j(\boldsymbol{\theta}_j, \boldsymbol{X}_j + \boldsymbol{g}(\boldsymbol{X}_j)) - L_j(\boldsymbol{\theta}_j, \boldsymbol{X}_j)] \\ &= \sum_{i=1}^{n} w_{ij}^2(\theta_{ij}) E[g_{ij}^2(\boldsymbol{X}_j)] + 2\sum_{i=1}^{n} w_{ij}^2(\theta_{ij}) E[(X_{ij} - \theta_{ij}) g_{ij}(\boldsymbol{X}_j)] \\ &\leq 0. \end{aligned} \tag{5}$$

We consider domination of the matrix estimator (3) by the matrix estimator

$$\delta_a(X)_{n\times p} = (X + G(X)) = (\boldsymbol{X}_1 + a\boldsymbol{g}(\boldsymbol{X}_1), \ldots, \boldsymbol{X}_p + a\boldsymbol{g}(\boldsymbol{X}_p)), \tag{6}$$

where $\boldsymbol{g}(\boldsymbol{X}_j) = (g_{1j}(\boldsymbol{X}_j), \ldots, g_{nj}(\boldsymbol{X}_j))^t$ satisfies Assumption (1). Under the matrix loss $L(\Theta, D)_{p\times p}$ where the jk-th component is given by

$$(L(\Theta, D))_{jk} = \sum_{i=1}^{n} w_{ij}(\theta_{ij}) w_{ik}(\theta_{ik})(d_{ij} - \theta_{ij})(d_{ik} - \theta_{ik}). \tag{7}$$

The risk of $D(X)$ is defined by $R(\Theta, D(X)) = E\{L(\boldsymbol{\Theta}, D)\}$. Let two estimators of Θ be $\widehat{\Theta}^1$ and $\widehat{\Theta}^2$ those are depending on X. $\widehat{\Theta}^1$ dominates $\widehat{\Theta}^2$ if $\Delta R = R(\Theta, \widehat{\Theta}^2) - R(\Theta, \widehat{\Theta}^1)$ is positive semipositive definite for all Θ and for some Θ, ΔR is positive definite.

Abu-Shanab, Kent and Strawderman (2012) studied a version of that problem where $X_{ij} \sim N(\theta_{ij}, \sigma^2)$ and showed that if a shrinkage estimator of the form $\boldsymbol{X}_j + \boldsymbol{g}(\boldsymbol{X}_j)$ satisfies Assumption (1), then the matrix estimator $\delta_a(X)_{n\times p}$ dominates $X_{n\times p}$ for $0 < a \leq 2/p$.
Hence, a shrinkage estimator with a smaller (by a factor of $2/p$) shrinkage constant also dominates X for the matrix loss (7).

Our main result, Theorem 1, generalizes this result to the more general setting of (1) without the restriction to the normality. Hence it applies to a broad class of models studied in the literature including, but not limited to estimation of Poisson means under weighted and unweighted quadratic loss, estimation of Binomial sample sizes under weighted and unweighted loss, estimation of natural parameters of discrete and non-discrete exponential families. Note also that X_{ij} may be interpreted as general unbiased estimator of θ_{ij} and need not be the original observation. Note also that domination in the matrix loss sense implies (and is equivalent to) simultaneous domination for all scalar losses of the form $\boldsymbol{\alpha}' L(\Theta, D)\boldsymbol{\alpha}$ for all $\boldsymbol{\alpha} \in R^p$.

Some motivation for this version of matrix loss is discussed in Abu-Shanab, Kent and Strawderman (2012). The main result is given in Section 2.

2 The main result

Theorem 1 Let $X, \Theta, \delta_a(X), L(\Theta, D)$ be as (1), (2), (6) and (7), respectively. Suppose that Assumption (1) holds. Then the matrix estimator $\delta_a(X)$ dominates $\delta_0(X) = X$ under matrix loss $L(\Theta, D)$ for $0 < a \leq 1/p$.

3 Some illustrative applications

We give some applications of Theorem 1 in the following.
1) Simultaneous estimation of Poisson matrix means
2) Simultaneous estimation of binomial sample sizes
3) Application to the exponential density families

Acknowledgments Professor William Edward Strawderman, Sr., PhD, passed away on October 1st, 2024. We are deeply saddened that this paper becomes a memorial to him. May he rest in peace.

References

[1] Abu-Shanab, R., Kent, J. T. and Strawderman, W. E. (2012). Shrinkage estimation with a matrix loss function. *Electron. J. Statist.* 6: 2347-2355.

Automatic sparse estimation of high-dimensional cross-covariance matrix

Tetsuya Umino[a], **Kazuyoshi Yata**[b], **and Makoto Aoshima**[b]

[a]Graduate School of Science and Technology, University of Tsukuba
[b]Institute of Mathematics, University of Tsukuba

A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively small. This is the so-called " HDLSS " or " large p, small n " data situation where $p/n \to \infty$; here p is the data dimension and n is the sample size. Such data situations occur in many areas of modern science such as genomics, medical imaging, text recognition, finance, chemometrics, and so on.

Suppose we take samples, $\boldsymbol{x}_j,\ j = 1, \ldots, n$, of size $n\ (\geq 4)$, which are independent and identically distributed (i.i.d.) as a p-variate distribution. Here, we consider situations where the data dimension p is very high compared to the sample size n. Let $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^T, \boldsymbol{x}_{2j}^T)^T$ and assume $\boldsymbol{x}_{ij} \in \mathbf{R}^{p_i}$, $i = 1, 2$, with $p_1 \in [1, p-1]$ and $p_2 = p - p_1$. We assume that $\boldsymbol{x}_j$ has an unknown mean vector, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$, and unknown covariance matrix,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_* \\ \boldsymbol{\Sigma}_*^T & \boldsymbol{\Sigma}_2 \end{pmatrix}\ (\geq \boldsymbol{O}),$$

that is, $\mathrm{E}(\boldsymbol{x}_{ij}) = \boldsymbol{\mu}_i$, $\mathrm{Var}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_i$, $i = 1, 2$, and $\mathrm{Cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = \mathrm{E}(\boldsymbol{x}_{1j}\boldsymbol{x}_{2j}^T) - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T = \boldsymbol{\Sigma}_*$.

Aoshima and Yata [1] and Yata and Aoshima [4, 5] considered testing the cross-covariance matrix by

$$H_0 : \boldsymbol{\Sigma}_* = \boldsymbol{O} \quad \text{vs.} \quad H_1 : \boldsymbol{\Sigma}_* \neq \boldsymbol{O} \tag{1}$$

for high-dimensional settings. When $(p_1, p_2) = (p-1, 1)$ or $(1, p-1)$, (1) implies the test of correlation coefficients. Aoshima and Yata [1] gave a test statistic for the test and Yata and Aoshima [4, 5] improved the test statistic by using a method called the *extended cross-data-matrix (ECDM) methodology*.

In this talk, we consider the problem of estimating the cross-covariance matrix, $\boldsymbol{\Sigma}_*$. There have been several studies on sparse estimation of the entire covariance matrix. For example, Bien and Tibshirani [3] proposed a sparse estimator of the covariance matrix based on L1-penalties, and Bickel and Levina [2] proposed a thresholding estimator of the covariance matrix. However, to our knowledge, sparse estimation of the cross-covariance matrix does not seem to have been studied in high-dimensional settings.

Recently, Yata and Aoshima [6] proposed a new sparse PCA (SPCA) method called the automatic SPCA (A-SPCA). A-SPCA does not depend on any threshold (tuning) values. In this talk, by applying the idea of A-SPCA to the estimation of the cross-covariance matrix,

we propose a new sparse estimator of $\mathbf{\Sigma}_*$. We show that the proposed estimator is consistent without any threshold (tuning) values.

Acknowledgements: Research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), JSPS, under Contract Number 22K03412. Research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 22K19769.

References

[1] M. Aoshima and K. Yata. Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper)*, 30: 356–399, 2011.

[2] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36: 2577 – 2604, 2008.

[3] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98: 807–820, 2011.

[4] K. Yata and M. Aoshima. Correlation tests for high-dimensional data using extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 117:313–331, 2013.

[5] K. Yata and M. Aoshima. High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. *Journal of Multivariate Analysis*, 151:151–166, 2016.

[6] K. Yata and M. Aoshima. Automatic sparse PCA for high-dimensional data. *Statistica Sinica*, 2025 (in press).

Augmented Estimation of Principal Component Subspace in High Dimensions

Dongsun Yoon
Seoul National University

In this paper, we introduce a novel estimator, called the Augmented Principal Component Subspace, for estimating the principal component subspace for high-dimensional low-sample size data with spiked covariance structure. Our approach augments the naive sample principal component subspace by incorporating additional information from predefined reference directions. Augmented principal component subspace asymptotically reduces every principal angle between the estimated and the true subspaces, thereby outperforming the naive estimator regardless of the metric used. The estimator ' s efficiency is validated both analytically and through numerical studies, demonstrating significant improvements in accuracy when the reference directions contain substantial information about the true principal component subspace. Additionally, we suggest AugmentedPCA using this estimator, and explore connections between our method and the recently proposed James-Stein estimator for principal component directions.

James-Stein Estimator of Spiked Leading Eigenvector of High-dimensional Covariance Matrix

Giheon Seong
Seoul National University

Recently, a James-Stein shrinkage (JS) estimator has gained attention as a powerful tool for estimating the leading eigenvector of covariance matrices. In a series of seminal works, the efficacy of the JS estimator has been demonstrated under a spiked covariance model, using the high-dimensional, low-sample-size (HDLSS) asymptotic regime, where the number of variables increases while the sample size n remains fixed. We extend the application of the JS shrinkage to the regime of $n, p \to \infty$ with appropriate rate and reveal a key condition involving a signal-to-noise ratio, for the JS estimator to be useful. This approach utilizes geometric representation, a phenomenon that arises in high-dimensional asymptotics, to interpret the structure of parameters and estimators on a sphere within a lower-dimensional space. Furthermore, we develop shrinkage estimators for principal component variance and scores, enabling their application in high-dimensional principal component analysis.

General measures of Attribution Disclosure Risk for gauging privacy of synthetic data

Yongjae Kim
Seoul National University

As the demand for synthetic data continues to grow, there is an increasing need for rigorous measures to assess whether synthetic data is safe or poses significant privacy risks. Correct Attribution Probability (CAP) is a widely used risk measure; however, its theoretical foundation has not been fully established within a solid statistical framework. In this paper, we propose a statistical framework for defining CAP and introduce a modified version to clarify its theoretical meaning. We also demonstrate the limitations of CAP as a comprehensive risk measure and argue why it cannot serve as an all-encompassing solution. Furthermore, we develop a generalized version of CAP, termed Attribution Disclosure Risk (ADR), which provides a more comprehensive and versatile assessment of synthetic data risk, incorporating CAP as a special case at both the population and sample levels. Numerical studies demonstrate that our proposed measure consistently captures the risk inherent in synthetic data and offers flexibility to accommodate various intruder scenarios, applicable to both simulated and real datasets.

Regularized k-POD Clustering for High-Dimensional Missing Data

Xin Guan*[1] and Yoshikazu Terada[1,2]

[1]Graduate School of Engineering Science, Osaka University
[2]RIKEN Center for Advanced Intelligence Project

1 Introduction

Clustering is an important technique that groups data points without labels into several clusters. Notably, the k-means clustering is one of the most popular clustering methods, the main idea of which is to find k cluster centers and then cluster data points by assigning them to their nearest centers. The k-means clustering has been widely used in various fields for its easy and fast implementation based on heuristic algorithms like Lloyd's algorithm, and not relying on specific assumptions of data distribution. However, the issue of clustering for missing data, especially the k-means clustering for missing data receives far less attention, even though the problem of missing data is ubiquitous in real-world applications for imperfect data collection process.

The main challenge is that the classical k-means clustering requires the data matrix to be complete, and thus directly conducting it on an incomplete data matrix is infeasible. The traditional approach is to pre-process the incomplete data matrix by complete-case analysis or multiple imputation to construct a new complete data matrix for conducting k-means clustering, which is not appropriate for large proportions of missingness and high-dimensional data.

Alternatively, the k-POD clustering proposed by Chi et al. (2016) is a natural extension for k-means clustering to missing data and can be applicable for even large missingness proportions and high-dimensional data. Write $\mathrm{X} = (x_{ij})_{n\times p} \in \mathbb{R}^{n\times p}$ for the data matrix with n data points $\mathrm{X}_1, \ldots, \mathrm{X}_n$ in $\mathbb{R}^p$. The k cluster centers $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ are encoded by a matrix $\mathrm{M} = (\mu_{lj})_{k\times p} \in \mathbb{R}^{k\times p}$, where the l-th row represents the l-th cluster center. The membership between data points and cluster centers is denoted by a binary matrix $\mathrm{U} = (u_{il})_{n\times k} \in \{0,1\}^{n\times k}$, where $u_{il} = 1$ if and only if i-th data point X_i is assigned to l-th cluster. Since one data point is assigned to a unique cluster, it must satisfy that $\mathrm{U}\mathbf{1}_k = \mathbf{1}_n$, where $\mathbf{1}$ is the all-one vector. For a complete data matrix X, the k-means clustering can be expressed as

$$\min_{\mathrm{U},\mathrm{M}} \|\mathrm{X} - \mathrm{UM}\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, calculated as $(\sum_{i,j} a_{ij}^2)^{1/2}$ for $\mathrm{A} = (a_{ij})$. If there exist missing entries in X, the loss function cannot be directly calculated. Denoting all observed positions in X by a set $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$, the k-POD clustering introduces a mapping $\mathcal{P}$ onto the set Ω to replace the missing entries with zero. That is, $\mathcal{P}_\Omega : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$, and $(\mathcal{P}_\Omega(\mathrm{X}))_{ij} = x_{ij}$ if $(i,j) \in \Omega$, 0 otherwise. Then, the k-POD clustering is given by

$$\min_{\mathrm{U},\mathrm{M}} \|\mathcal{P}_\Omega(\mathrm{X} - \mathrm{UM})\|_F^2. \tag{2}$$

The optimization procedure consists of filling in missing entries by the corresponding cluster means and conducting k-means clustering on the new data matrix alternatively.

However, the k-POD clustering is not consistent even under the missing completely at random mechanism (Terada & Guan 2024). The estimated cluster centers of the k-POD clustering and k-means clustering converge to different solutions as $n \to \infty$. The direct reason for the bias simply comes from the difference between loss functions of k-means and k-POD. Specifically, all positions of X are used by k-means, while only observed positions, i.e., $(i,j) \in \Omega$, are included by k-POD, and thus in general, one can hardly expect the same solutions based on these two different loss functions.

In this talk, we proposed regularized k-POD clustering for high-dimensional missing data. Specifically, we introduce a regularization function of cluster centers to the loss of k-POD clustering, which shrinks cluster centers feature-wisely. This offers a significant advantage of reducing the bias of estimated cluster centers, in the case when noise features exist that have no contribution to the true cluster structure, which is common in high-dimensional space.

*Corresponding author: xin@sigmath.es.osaka-u.ac.jp (XG)

2 Methodology and optimization

Suppose that the data matrix $\mathrm{X} = (x_{ij})_{n\times p}$ is column-wised centered, that is, $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, \ldots, p$. Write $\mathrm{X}_i \in \mathbb{R}^p$ for the i-th data point ($i = 1, \ldots, n$) and $\mathrm{X}_{(j)} \in \mathbb{R}^n$ for the j-th column of X ($j = 1, \ldots, p$). Denote by Ω the set of observed positions of X and suppose that the number of clusters k is fixed.

We defined the loss function of regularized k-POD clustering with respect to membership $\mathrm{U} \in \{0,1\}^{n\times k}$, $\mathrm{U}\mathbf{1}_k = \mathbf{1}_n$, and cluster centers $\mathrm{M} \in \mathbb{R}^{k\times p}$ to be

$$\widehat{L}_n(\mathrm{U}, \mathrm{M}) = \|\mathcal{P}_\Omega(\mathrm{X} - \mathrm{UM})\|_F^2 + \lambda \cdot J(\mathrm{M}). \tag{3}$$

The first term is the loss of the k-POD clustering, and $J(\mathrm{M})$ is a regularization function with respect to M. To shrink the estimated cluster centers feature-wisely, we consider two types of $J(\mathrm{M})$:

$$\text{The } l_0 \text{ penalty}: \quad J_0(\mathrm{M}) = \sum_{j=1}^{p} \mathbb{1}(\|\mathrm{M}_{(j)}\| > 0)$$

$$\text{The group lasso penalty}: \quad J_1(\mathrm{M}) = \sum_{j=1}^{p} w_j \|\mathrm{M}_{(j)}\|,$$

where $\mathrm{M}_{(j)} = (\mu_{1j}, \ldots, \mu_{kj})^T$ denotes the j-th column of cluster centers M with μ_{lj} being the j-th component of the l-th cluster center ($l = 1, \ldots, k$). The function $\mathbb{1}(\cdot)$ is the indicator function and w_j is the weight for $\mathrm{M}_{(j)}$. Both types of $J(\cdot)$ are column-wised, which means that all elements of $\mathrm{M}_{(j)}$, that is $\{\mu_{1j}, \ldots, \mu_{kj}\}$ would be shrunk together. The l_0 type $J_0(\cdot)$ constrains the number of non-zero columns of M, while the group lasso type $J_1(\cdot)$ constrains the weighted sum of l_2 norms of M in each feature. Therefore, with suitable regularization parameter λ, the estimated cluster centers $\widehat{\mathrm{M}}$ would be sparse in columns. In addition, the group lasso type contains weights. We consider the weights based on the k-POD estimator $\widetilde{\mathrm{M}}$, that is, $w_j = 1/\|\widetilde{\mathrm{M}}_{(j)}\|$. If the estimated cluster centers of the k-POD clustering in a feature are relatively concentrated, the corresponding weight would be relatively large, which makes the group lasso estimator in the corresponding feature more likely to be zero.

We applied the majorization-minimization algorithm (MM algorithm) to minimize the proposed loss function Eq. (3). we propose Algorithm 1 for regularized k-POD clustering. Specifically, given current $\mathrm{U}^{(t)}$ and $\mathrm{M}^{(t)}$, $t \in \mathbb{N}$, the $(t+1)$-th iteration consists of two steps. Step 1 imputes missing entries of X by the corresponding entries of multiplication matrix of current $\mathrm{U}^{(t)}$ and $\mathrm{M}^{(t)}$, so that we can get a new complete data matrix $\widehat{\mathrm{X}}^{(t+1)}$. Step 2 updates $\mathrm{U}^{(t+1)}$ and $\mathrm{M}^{(t+1)}$ by applying regularized k-means clustering on the imputed data matrix $\widehat{\mathrm{X}}^{(t+1)}$. Repeat the iteration until the loss (Eq. (3)) converges. Note that Algorithm 1 is a general framework for any type of $J(\cdot)$, and the difference in results comes from Step 2.

Algorithm 1 Regularized k-POD clustering

Input: incomplete data matrix X, set of observed positions Ω, number of clusters k.
Parameters: regularization parameter λ, weights $\{w_j\}$
Initialize $\mathrm{U}^{(0)}$ and $\mathrm{M}^{(0)}$
while Loss function (3) does not converge **do**
1: Impute $\widehat{\mathrm{X}}^{(t+1)} = \mathcal{P}_\Omega(\mathrm{X}) + \mathcal{P}_{\Omega^c}(\mathrm{U}^{(t)}\mathrm{M}^{(t)})$
2: Update $\mathrm{U}^{(t+1)}$ and $\mathrm{M}^{(t+1)}$ by applying regularized k-means on $\widehat{\mathrm{X}}^{(t+1)}$
end while
Output: $\mathrm{U}^{(t+1)}$ and $\mathrm{M}^{(t+1)}$

The results of numerical experiments have been introduced in this talk, which verified the better performance of the proposed method.

References

Chi, J. T., Chi, E. C. & Baraniuk, R. G. (2016), 'k-pod: A method for k-means clustering of missing data', *The American Statistician* **70**(1), 91–99.

Terada, Y. & Guan, X. (2024), 'Some notes on the k-means clustering for missing data'.
URL: *https://arxiv.org/abs/2410.00546*

NON-SPARSE HIGH-DIMENSIONAL STATISTICS: STRUCTURED MODEL, NEURAL NETWORK, AND UNIVERSALITY

MASAAKI IMAIZUMI[1,2]

[1]*The University of Tokyo,* [2] *RIKEN Advanced Intelligence Project*

Abstract. In this talk, we present several results in high-dimensional statistics. Specifically, we consider the linear regression model with the universality, an estimation problem of the single-index model, and the rigorous learnability of high-dimensional neural networks with many neurons. The analysis in these studies uses the theory of the high-dimensional central limit theorem, the nonlinear component of the proportionally high-dimensional regime, and detailed analysis of macro-level dynamic of a group of neurons.

1. Outline

1.1. **Linear Regression.** We consider a linear regression model with p-dimensional covariates and a parameter. Suppose that we observe i.i.d. n pairs $\{(X_i, Y_i)\}_{i=1}^n$ of a covariate $X_i \in \mathbb{R}^p$ and a target variable $Y_i \in \mathbb{R}$ generated from the following linear model with the true parameter $\theta_0 \in \mathbb{R}^p$:

$$Y_i = X_i^\top \theta_0 + \xi_i,\ i = 1, ..., n,$$

where ξ_i is a centered noise variable. Let $\Sigma = \mathbb{E}[X_i X_i^\top]/n$ be a covariance matrix of the covariate.

We are interested in the statistical inference of an estimator of θ_0 in the model. Rigorously, we define an empirical risk minimizer problem with a loss function ℓ and a regularizer R_0:

$$\widehat{\theta}_{\mathbf{X}} \in \mathop{\mathrm{argmin}}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_0(Y_i - X_i^\top \theta) + R_0(\theta) \right\}.$$

Then, we are interested in a probability law of the estimator $\widehat{\theta}_{\mathbf{X}}$. Here, we are particularly interested in the proportional limit of the coefficient dimension

p and sample size n: $n, p \to \infty$ and $p/n \to \exists\kappa \in (0, \infty)$. When data follow a Gaussian distribution, an asymptotic distribution of $\widehat{\theta}_{\mathbf{X}}$ has been well studied.

Even when the data do not follow a Gaussian distribution, if the asymptotic properties are preserved, this is referred to as the *universality*. For the universality of the asymptotic distribution of an estimator to hold, an existing research has shown that it holds when all p elements of the data vector X_i are independent. In contrast, in [TI24], we demonstrate that universality can also hold when the elements of the data vector X_i exhibit a class of dependence known as block dependence.

1.2. **Single-Index Model.** We next consider the single-index model: for a pair (X, Y) of p-dimensional random features X and random responses Y, we consider the following model

$$\mathbb{E}[Y \mid X] = g(\beta^\top X), \tag{1}$$

where $g : \mathbb{R} \to \mathbb{R}$ is an unknown link function that monotonically increases, and $\beta \in \mathbb{R}^p$ is an unknown deterministic coefficient vector. Suppose that we observe i.i.d. n pairs $\{(X_i, Y_i)\}_{i=1}^n$ of a feature vector $X_i \in \mathbb{R}^p$ and a target variable $Y_i \in \mathbb{R}$ that follow the single-index model (1).

Our goal is to estimate both the coefficient β and the link function $g(\cdot)$. Here, We consider the proportional high-dimensional regime: we are particularly interested in the proportional limit of the coefficient dimension p and sample size n: $n, p \to \infty$ and $p/n \to \exists\kappa \in (0, \infty)$.

[SUI24] develops a methodology for statistical inference for the single-index model, by deriving the asymptotic normality of an estimator. This method is based on an analysis of the first-order method and the deconvolution technique for statistical models.

References

[SUI24] Kazuma Sawaya, Yoshimasa Uematsu, and Masaaki Imaizumi. High-dimensional single-index models: Link estimation and marginal inference. *preprint arXiv:2404.17812*, 2024.

[TI24] Toshiki Tsuda and Masaaki Imaizumi. Universality of estimator for high-dimensional linear models with block dependency. *arXiv preprint arXiv:2410.19244*, 2024.

Title: Difference between Large Statistical Model and Medium Statistical Model

Shurong Zheng
(School of Mathematics and Statistics, Northeast Normal University)

Abstract:

In this talk, we will show that the large statistical model will have a very different performance compared with the medium model. For example, when the sample size is fixed and the dimension of data increases (convergence regime), the power function of the log-likelihood ratio test for the covariance matrix will tend to one. Moreover, under the convergence regime, the estimated number of factors in the factor model will be more accurate. Moreover, we will give some other examples to show the difference between large statistical model and medium statistical model.

Principal component analysis for zero-inflated compositional data

Sungkyu Jung

(Institute for Data Innovation in Science, Seoul National University)

Abstract: Recent advances in DNA sequencing technology have heightened interest in microbiome data, which is often high-dimensional and presents challenges due to its compositional nature and zero-inflation. In this talk, I will introduce new PCA methods for zero-inflated compositional data, based on a framework called principal compositional subspace. These methods aim to identify both the principal compositional subspace and corresponding principal scores that best approximate the data while maintaining its compositional properties. Theoretical properties such as existence and consistency of the principal compositional subspace are investigated. Simulation studies show these methods achieve lower reconstruction errors than existing log-ratio PCA methods in linear patterns and perform comparably in curved patterns. The methods successfully uncover the low-rank structure in four microbiome compositional datasets with excessive zeros.

A Generalized Mean Approach for Distributed-PCA

Zhi-Yu Jou, Su-Yun Huang,
Institute of Statistical Science, Academia Sinica, Taiwan
Hung Hung*,
Institute of Health Data Analytics and Statistics, National Taiwan University, Taiwan
and
Shinto Eguchi
Institute of Statistical Mathematics, Japan

Abstract

Principal component analysis (PCA) is a widely used technique for dimension reduction. As datasets continue to grow in size, distributed-PCA (DPCA) has become an active research area. A key challenge in DPCA lies in efficiently aggregating results across multiple machines or computing nodes due to computational overhead. Fan et al. (2019) introduced a pioneering DPCA method to estimate the leading rank-r eigenspace, aggregating local rank-r projection matrices by averaging. However, their method does not utilize eigenvalue information. In this article, we propose a novel DPCA method that incorporates eigenvalue information to aggregate local results via the matrix β-mean, which we call β-DPCA. The matrix β-mean offers a flexible and robust aggregation method through the adjustable choice of β values. Notably, for $\beta = 1$, it corresponds to the arithmetic mean; for $\beta = -1$, the harmonic mean; and as $\beta \to 0$, the geometric mean. Moreover, the matrix β-mean is shown to associate with the matrix β-divergence, a subclass of the Bregman matrix divergence, to support the robustness of β-DPCA. We also study the stability of eigenvector ordering under eigenvalue perturbation for β-DPCA. The performance of our proposal is evaluated through numerical studies.

Keywords: distributed computing; eigenvalue perturbation; generalized matrix mean; matrix divergence; PCA

*Corresponding author. *Email:* hhung@ntu.edu.tw

Alignment and matching tests for high-dimensional tensor signals via tensor contraction

Jianfeng Yao

School of Data Science
The Chinese University of Hong Kong (Shenzhen)
e-mail: jeffyao@cuhk.edu.cn

This is a joint work with Ruihan Liu and Zhenggang Wang,
Department of Statistics and Actuarial Science,
The University of Hong Kong

MSC2020 subject classifications: Primary 62H15; secondary 60B20,62H10.
Keywords and phrases: High-dimensional tensors, Low-rank tensors, Tensor signal alignment, Tensor signal matching, Tensor contraction, Linear spectral statistics, Random matrix theory.

In the era of "big data", the analysis of high-dimensional tensor data has become increasingly important in various fields, including genomics, economics, image analysis, and machine learning. High-order tensor data often exhibit intrinsic low-rank structures [14, 25]. To capture these low-rank structures, the "signal plus noise" tensor model has been widely adopted [9, 11, 15]. Let $n_1, \ldots, n_d \in \mathbb{N}^+$ denote d dimension numbers, where $d \geq 3$, and let $N = n_1 + \cdots + n_d$. The d-fold rank-R spiked tensor model is defined as:

$$\boldsymbol{T} = \sum_{r=1}^{R} \beta_r \boldsymbol{x}^{(r,1)} \otimes \cdots \otimes \boldsymbol{x}^{(r,d)} + \frac{1}{\sqrt{N}} \boldsymbol{X}, \tag{1}$$

where $\beta_1 \geq \cdots \geq \beta_R > 0$ are the signal-to-noise ratios (SNRs), $\{\boldsymbol{x}^{(1,l)}, \cdots, \boldsymbol{x}^{(R,l)}\}$ are mutually orthogonal unit vectors $\mathbb{R}^{n_l}$ for each $1 \leq l \leq d$ [13], and $\boldsymbol{X} = (X_{i_1 \cdots i_d})_{n_1 \times \cdots \times n_d} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is a noise tensor with independent and identically distributed (i.i.d.) entries, each having zero mean and unit variance. Specifically, the rank-1 spiked tensor model [21] is given by:

$$\boldsymbol{T} = \beta \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(d)} + \frac{1}{\sqrt{N}} \boldsymbol{X}, \tag{2}$$

where $\beta > 0$ is the single SNR of the model.

The primary focus of most existing literature is on recovering the signal vectors $\{\boldsymbol{x}^{(1,l)}, \ldots, \boldsymbol{x}^{(R,l)}\}$, $1 \leq l \leq d$ from the observed tensor $\boldsymbol{T}$, with a particular

emphasis on the computational efficiency of recovery algorithms. In the case of the rank-one model (2) with symmetric and i.i.d. Gaussian noise $\boldsymbol{X}$, [10] showed that computing the maximum likelihood (ML) estimator of $\beta \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(d)}$ is in general NP-hard, and [1] provided a comprehensive discussion on the relationship between the computational complexity of the ML estimator and the value of the SNR β. To reduce the computational complexity, [21] proposed the use of the power iteration method and approximate message passing (AMP) algorithms. These two methods have been extensively investigated by [5, 7, 12, 15, 20] for AMP and by [11] for power iteration. Moreover, [21] introduced the tensor unfolding method, which involves unfolding the tensor data $\boldsymbol{T}$ into matrices, enabling the recovery of signals through Principal Component Analysis (PCA). [6] conducted a comprehensive study of the tensor unfolding method for the general asymmetric model (2) under fairly general noise distribution assumptions.

However, when the SNRs fall below the phase transition threshold, these recovery methods often fail. In such cases, a less ambitious but potentially more achievable goal is to test the alignment of a signal in $\boldsymbol{T}$ with a given directional tensor $\boldsymbol{a}^{(1)} \otimes \cdots \otimes \boldsymbol{a}^{(d)}$, where $\boldsymbol{a}^{(j)}, 1 \le j \le d$ are d given directional unit vectors in $\mathbb{R}^{n_j}$, respectively. This leads to the following *tensor signal alignment test* between two hypotheses:

$$\begin{aligned} &H_0 : \boldsymbol{a}^{(l)} \perp \boldsymbol{x}^{(r,l)} \quad \text{for } 1 \le l \le d,\ 1 \le r \le R. \\ &H_1 : \text{there exists at least one } 1 \le l \le d,\ 1 \le r \le R \text{ such that } \boldsymbol{a}^{(l)} \not\perp \boldsymbol{x}^{(r,l)}. \end{aligned} \tag{3}$$

Despite the tensor signal alignment test appearing more tractable than signal recovery, to the best of our knowledge, there is no established and rigorously justified procedure for addressing this problem. The difficulty stems from the high dimensionality of the tensors and the lack of a meaningful test statistic.

We leverage the tensor contraction operator $\boldsymbol{\Phi}_d$, originally proposed in [22], which maps an arbitrary tensor $\boldsymbol{T}$ and unit vectors $\{\boldsymbol{a}^{(j)}\}$ to a matrix $\boldsymbol{R}$:

$$\begin{aligned} &\boldsymbol{\Phi}_d : \mathbb{R}^{n_1 \times \cdots \times n_d} \times \mathbb{S}^{n_1 - 1} \times \cdots \times \mathbb{S}^{n_d - 1} \longrightarrow \mathbb{R}^{N \times N}, \\ &(\boldsymbol{T}, \boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(d)}) \longmapsto \boldsymbol{R} = \begin{pmatrix} \boldsymbol{0}_{n_1 \times n_1} & \boldsymbol{T}^{12} & \cdots & \boldsymbol{T}^{1d} \\ (\boldsymbol{T}^{12})' & \boldsymbol{0}_{n_2 \times n_2} & \cdots & \boldsymbol{T}^{2d} \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{T}^{1d})' & (\boldsymbol{T}^{2d})' & \cdots & \boldsymbol{0}_{n_d \times n_d} \end{pmatrix}. \end{aligned} \tag{4}$$

Here, for a pair of indices $1 \le j_1 < j_2 \le d$, $\boldsymbol{T}^{j_1 j_2}$ is an $n_{j_1} \times n_{j_2}$ matrix, called *second order contraction matrix of* $\boldsymbol{T}$ *along the directions* $\{\boldsymbol{a}^{(j_1)}, \boldsymbol{a}^{(j_2)}\}$, as introduced in [16]. It is defined by:

$$\boldsymbol{T}^{j_1 j_2} = \left[\sum_{i_j = 1, j \neq j_1, j_2}^{n_j} T_{i_1 \cdots i_d} \prod_{l=1, l \neq j_1, j_2}^{d} a_{i_l}^{(l)} \right]_{n_{j_1} \times n_{j_2}}. \tag{5}$$

From a mathematical perspective, the contraction operator $\boldsymbol{\Phi}_d$ has several advantages. Firstly, $\boldsymbol{\Phi}_d$ is linear in $\boldsymbol{T}$. When applied to the R-rank tensor in (1), we have

$$\begin{aligned}\boldsymbol{R} &= \boldsymbol{\Phi}_d(\boldsymbol{T}, \boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(d)}) \\ &= \sum_{r=1}^{d} \beta_r \boldsymbol{\Phi}_d(\boldsymbol{x}^{(r,1)} \otimes \cdots \otimes \boldsymbol{x}^{(r,d)}, \boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(d)}) + \frac{1}{\sqrt{N}} \boldsymbol{\Phi}_d(\boldsymbol{X}, \boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(d)}), \\ &= \boldsymbol{S} + \boldsymbol{M}. \end{aligned} \tag{6}$$

where $\boldsymbol{S}$ is the contracted signal matrix containing the $\boldsymbol{R}$ tensor signals, and $\boldsymbol{M}$ is the residual matrix representing pure noise. Under the null hypothesis H_0, $\boldsymbol{S} = \boldsymbol{0}$ implying $\boldsymbol{R} = \boldsymbol{M}$. In contrast, under the alternative H_1, $\boldsymbol{S} \neq \boldsymbol{0}$, result in $\boldsymbol{R} \neq \boldsymbol{M}$.

Furthermore, both the contracted signal matrix $\boldsymbol{S}$ and noise matrix $\boldsymbol{M}$ are symmetric, with $\boldsymbol{S}$ having a finite rank. This allows us to analyze the contracted data matrix $\boldsymbol{R}$ using linear spectral statistics (LSS), a powerful tool from random matrix theory. Central limit theorems for LSS of random matrices have received much attention in high-dimensional statistics, see [2, 3, 17, 19, 26] for a few classical references. In our case, by employing an appropriate LSS of $\boldsymbol{R}$ with an established asymptotic distribution, we can effectively distinguish between the two hypotheses.

We first establish that the eigenvalue distribution of $\boldsymbol{R}$ has a limit ν when the d dimensions $\{n_j\}$ grow to infinity in comparable rates. Next, we introduce the following test statistic:

$$\widehat{T}_N^{(d)} = \|\boldsymbol{R}\|_2^2 - N \int_{-\infty}^{\infty} x^2 \nu(dx). \tag{7}$$

Here, $\|\boldsymbol{R}\|_2^2 = \sum_{i,j=1}^{N} R_{i,j}^2$ is a linear spectral statistic of $\boldsymbol{R}$. As one of the main results of this paper, we establish that under the null hypothesis H_0,

$$\frac{\widehat{T}_N^{(d)} - \xi_N^{(d)}}{\sigma_N^{(d)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{8}$$

where $\xi_N^{(d)}$ and $\sigma_N^{(d)}$ are known parameters that can be calculated numerically. Under the alternative hypothesis H_1,

$$\frac{\widehat{T}_N^{(d)} - \xi_N^{(d)}}{\sigma_N^{(d)}} - \mathcal{D}^{(d)}/\sigma_N^{(d)} \xrightarrow{d} \mathcal{N}(0,1), \tag{9}$$

where $\mathcal{D}^{(d)}/\sigma_N^{(d)}$ is a positive mean drift. Consequently, the asymptotic normal distribution in (8) enables us to construct a test for a given significance level α, while the distribution in (9) guarantees a positive power for the test, which depends on the magnitude of $\mathcal{D}^{(d)}/\sigma_N^{(d)}$.

When $d = 2$, the tensor model (1) reduces to a finite-rank perturbed or spiked random matrix. In this context, the signal alignment test in (3) can be seen as a tensor extension of existing tests for the presence of spikes along given directions, as studied by [4, 8, 18, 23, 24].

However, when $d \geq 3$, a fundamental difference emerges: the elements $\boldsymbol{T}^{j_1 j_2}$ in the contracted data matrix $\boldsymbol{R}$ become correlated. This correlation significantly increases the complexity of studying the matrix, making the analysis more challenging compared to the $d = 2$ case. The presence of these correlations necessitates the development of novel techniques to effectively analyze the eigenvalue distribution and establish the asymptotic properties of the test statistic $\widehat{T}_N^{(d)}$ in high dimensions.

The main contributions of this article are as follows.

(i) We conduct an in-depth analysis of the contracted data matrix $\boldsymbol{R}$, whose entries display significant correlations and deviate from traditional random matrix models in which the elements of the noise matrix are typically assumed to be independent of one another, including

   (a) The characterization of its limiting spectral distribution (LSD) through a vector Dyson equation, along with entrywise behaviors of the resolvent.

   (b) The establishment of CLT for a broad class of its LSS.

(ii) We establish a rigorous procedure for the tensor signal alignment test (3) by establishing the normality asymptotic of the test statistic and deriving its power function under a general alternative hypothesis.

(iii) We also address the problem of testing for the matching of two high-dimensional low-rank tensor signals. To tackle this problem, we employ an approach similar to the one established for the tensor signal alignment test.

The contributions presented in this article are novel. One notable innovation is that our tensor signal model in (1) allows for non-Gaussian and non-symmetric signals. This sets our work apart from most existing literature on high-dimensional tensor data models, which typically assumes symmetry or Gaussianity for either the tensor signal, the tensor noise, or both.

References

[1] Arous, G. B., Mei, S., Montanari, A. and Nica, M. (2019). The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics* **72** 2282–2330.

[2] Bai, Z., Jiang, D., Yao, J. F. and Zheng, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *The Annals of Statistics* **37** 3822-3840.

[3] Bai, Z. and Silverstein, J. W. (2004). CLT for Linear Spectral Statistics of Large-Dimensional Sample Covariance Matrices. *Annals of Probability* 553–605.

[4] Bao, Z., Ding, X., Wang, J. and Wang, K. (2022). Statistical inference for principal components of spiked covariance matrices. *The Annals of Statistics* **50** 1144–1169.
[5] Ben Arous, G., Gheissari, R. and Jagannath, A. (2020). Algorithmic thresholds for tensor PCA. *Annals of Probability* **48** 2052–2087.
[6] Ben Arous, G., Huang, D. Z. and Huang, J. (2023). Long Random Matrices and Tensor Unfolding. *The Annals of Applied Probability* **33** 5753–5780.
[7] Chen, W.-K. (2019). Phase transition in the spiked random tensor with Rademacher prior. *The Annals of Statistics* **47** 2734–2756.
[8] Hallin, M., Paindaveine, D. and Verdebout, T. (2010). Optimal rank-based testing for principal components. *Annals of statistics* **38** 3245–3299.
[9] Han, Y. and Zhang, C.-H. (2022). Tensor principal component analysis in high dimensional CP models. *IEEE Transactions on Information Theory* **69** 1147–1167.
[10] Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)* **60** 1–39.
[11] Huang, J., Huang, D. Z., Yang, Q. and Cheng, G. (2022). Power iteration for tensor PCA. *Journal of Machine Learning Research* **23** 1–47.
[12] Jagannath, A., Lopatto, P. and Miolane, L. (2020). Statistical thresholds for tensor PCA. *The Annals of Applied Probability* **30** 1910–1933.
[13] Kolda, T. G. (2001). Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications* **23** 243–255.
[14] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.
[15] Lesieur, T., Miolane, L., Lelarge, M., Krzakala, F. and Zdeborová, L. (2017). Statistical and computational phase transitions in spiked tensor estimation. *2017 IEEE International Symposium on Information Theory (ISIT)* 511–515.
[16] Lim, L.-H. (2005). Singular values and eigenvalues of tensors: a variational approach. In *1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2005.* 129–132. IEEE.
[17] Lytova, A. and Pastur, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *The Annals of Probability* **37** 1778 – 1840.
[18] Naumov, A., Spokoiny, V. and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields* **174** 1091–1132.
[19] Pan, G. M. and Zhou, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* 1232-1270.
[20] Perry, A., Wein, A. S. and Bandeira, A. S. (2020). Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques* **56** 230–264.
[21] Richard, E. and Montanari, A. (2014). A statistical model for tensor

PCA. *Advances in Neural Information Processing Systems* **27**.

[22] Seddik, M. E. A., Guillaud, M. and Couillet, R. (2024). When random tensors meet random matrices. *Annals of Applied Probability* **34** 203 – 248.

[23] Silin, I. and Fan, J. (2020). Hypothesis testing for eigenspaces of covariance matrix. *arXiv preprint arXiv:2002.09810*.

[24] Silin, I. and Spokoiny, V. (2018). Bayesian inference for spectral projectors of the covariance matrix. *Electronic Journal of Statistics* **12** 1948–1987.

[25] Udell, M. and Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science* **1** 144–160.

[26] Zheng, S., Bai, Z. and Yao, J. (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli* **23** 1130-1178.

On dimension-free concentration of logistic regression

Shogo Nakakita*

The logistic regression model is a fundamental binary classification model in statistics and machine learning. Given a sequence of $\mathbb{R}^p \times \{0,1\}$-valued independent and identically distributed (i.i.d.) random variables $\{(\boldsymbol{X}_i, Y_i); i = 1, \dots, n\}$, it supposes Y_i as conditionally Bernoulli-distributed random variables such that for some $\boldsymbol{\theta} \in \mathbb{R}^p$, for all $i = 1, \dots, n$ and $\boldsymbol{x} \in \mathbb{R}^p$,

$$Y_i | \boldsymbol{X}_i = \boldsymbol{x} \sim \mathrm{Ber}(\sigma(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)), \tag{1}$$

where $\sigma(t) = 1/(1 + \exp(-t))$ with $t \in \mathbb{R}$ is the link function. Each component of $\boldsymbol{\theta}$ explains the relationship between the corresponding component of $\boldsymbol{X}_i$ and the conditional probability $\mathbb{P}(Y_i = 1 | \boldsymbol{X}_i)$. The model is widely used for academic and industrial purposes as its interpretations are simple.

Our interest is the estimation of $\boldsymbol{\theta}$ with good prediction performance under high-dimensional settings. To estimate $\boldsymbol{\theta}$, we frequently consider the minimization problem of the following empirical risk function (or the $(-1/n)$-scaled log-likelihood function):

$$\mathcal{R}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \left(-Y_i \log \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta} \rangle) - (1 - Y_i) \log(1 - \sigma(\langle \boldsymbol{X}_i, \boldsymbol{\theta} \rangle)) \right). \tag{2}$$

We expect that the minimization of $\mathcal{R}_n(\boldsymbol{\theta})$ is a good approximation of the minimization of the unknown population risk function $\mathcal{R}(\boldsymbol{\theta}) := \mathbb{E}[\mathcal{R}_n(\boldsymbol{\theta})]$. If it holds true, the minimizers of $\mathcal{R}_n(\boldsymbol{\theta})$ also achieve small population risk and thus good prediction performance. Under the low-dimensional setting where p is fixed in n and $n \to \infty$, the classical argument for maximum likelihood estimation validates this idea. However, this idea becomes difficult to justify in situations where p is large relative to n. For example, Sur and Candès (2019) point out that the minimizers of the empirical risk $\mathcal{R}_n(\boldsymbol{\theta})$ can perform poorly as the estimators of the minimizers of the population risk $\mathcal{R}(\boldsymbol{\theta})$ under high-dimensional settings. In contrast, our study examines when the minimization of $\mathcal{R}_n(\boldsymbol{\theta})$ is a good approximation of the minimization of $\mathcal{R}(\boldsymbol{\theta})$ even with large p.

In particular, we study uniform concentration bounds and a uniform law of large numbers as their corollary for $\mathcal{R}_n(\boldsymbol{\theta})$ around $\mathcal{R}(\boldsymbol{\theta})$ on $B[R]$, where $B[R] = \{\boldsymbol{\theta}' \in \mathbb{R}^p; \|\boldsymbol{\theta}'\|_2 \le R\}$ with $R \ge 0$ is the known bounded parameter space. Let us consider the following ball-constrained minimization problem (ball-constrained logistic regression) instead of the unconstrained minimization on $\mathbb{R}^p$:

$$\text{minimize } \mathcal{R}_n(\boldsymbol{\theta}) \text{ subject to } \|\boldsymbol{\theta}\|_2 \le R. \tag{3}$$

This is a smooth convex optimization problem on a bounded convex set; it has solutions, which we can find efficiently. Note that constraints on balls or spheres are not only mild but also typical

*Komaba Institute for Science, University of Tokyo. Email address: nakakita@g.ecc.u-tokyo.ac.jp

in previous studies (Kuchelmeister and van de Geer, 2024; Hsu and Mazumdar, 2024). If we can conclude that $\mathcal{R}_n(\boldsymbol{\theta})$ is uniformly close to $\mathcal{R}(\boldsymbol{\theta})$ on $B[R]$, then solving the minimization problem (3) (i.e., maximum likelihood estimation with the parameter space $B[R]$) is a good approximation of the minimization of $\mathcal{R}(\boldsymbol{\theta})$ on $B[R]$. If the following uniform law of large numbers holds, then we can support this idea asymptotically:

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in B[R]} |\mathcal{R}_n(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta})| = 0 \text{ almost surely.} \tag{4}$$

The uniform law (4) is one of the most fundamental arguments in large-sample theory. It concludes that the minimization of the empirical risk $\mathcal{R}_n(\boldsymbol{\theta})$ is asymptotically equivalent to the minimization of the population risk $\mathcal{R}(\boldsymbol{\theta})$ (van der Vaart, 2000). To derive sufficient conditions for the uniform law under high-dimensional settings, we analyse non-asymptotic uniform concentration bounds on $\sup_{\boldsymbol{\theta}\in B[R]} |\mathcal{R}_n(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta})|$.

Our study gives a dimension-free uniform concentration bound yielding a mild sufficient condition for the uniform law of large numbers. We derive a bound such that for some explicit $c > 0$ independent of $\boldsymbol{\Sigma}$, n, and p, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\sup_{\boldsymbol{\theta}\in B[R]} |\mathcal{R}_n(\boldsymbol{\theta}) - \mathcal{R}(\boldsymbol{\theta})| \le c\left(\sqrt{\frac{\|\boldsymbol{\Sigma}\|\mathbf{r}(\boldsymbol{\Sigma}) + (1 + \|\boldsymbol{\Sigma}\|)(1 + \log\delta^{-1})}{n}} + \frac{\|\boldsymbol{\Sigma}\|\sqrt{1 + \log\delta^{-1}}}{n}\right). \tag{5}$$

It is noteworthy that this bound gives a mild and natural sufficient condition $\mathbf{r}(\boldsymbol{\Sigma})/n \to 0$ for the uniform law of large numbers.

Acknowledgements

I gratefully acknowledge Pierre Alquier for his enlightening comments. This work was supported by JSPS KAKENHI Grant Number JP24K02904 and JST CREST Grant Numbers JPMJCR21D2 and JPMJCR2115.

References

Hsu, D. and Mazumdar, A. (2024). On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437. PMLR.

Kuchelmeister, F. and van de Geer, S. (2024). Finite sample rates for logistic regression with small noise or few samples. *Sankhya A*. Advance online publication.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Subspace Recovery in Winsorized PCA

Sangil Han
Seoul National University

We explore the theoretical properties of subspace recovery using Winsorized Principal Component Analysis (WPCA), utilizing a common data transformation technique that caps extreme values to mitigate the impact of outliers. Despite the widespread use of winsorization in various tasks of multivariate analysis, its theoretical properties, particularly for subspace recovery, have received limited attention. We provide a detailed analysis of the accuracy of WPCA, showing that increasing the number of samples while decreasing the proportion of outliers guarantees the consistency of the sample subspaces from WPCA with respect to the true population subspace. Furthermore, we establish perturbation bounds that ensure the WPCA subspace obtained from contaminated data remains close to the subspace recovered from pure data. Additionally, we extend the classical notion of breakdown points to subspace-valued statistics and derive lower bounds for the breakdown points of WPCA. Our analysis demonstrates that WPCA exhibits strong robustness to outliers while maintaining consistency under mild assumptions. A toy example is provided to numerically illustrate the behavior of the upper bounds for perturbation bounds and breakdown points, emphasizing winsorization's utility in subspace recovery.

High-dimensional bootstrap and asymptotic expansion

Yuta Koike *†

December 23, 2024

Let $X_1, \ldots, X_n$ be independent centered random vectors in $\mathbb{R}^d$ with finite variance. Set

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i.$$

The aim of this paper is to investigate the accuracy of bootstrap approximation for the maximum type statistic

$$T_n := \max_{1 \le j \le d} S_{n,j}$$

when both n and d tend to infinity. Specifically, we consider the so-called wild bootstrap method: Let $w_1, \ldots, w_n$ be i.i.d. random variables independent of the data $X_1, \ldots, X_n$ such that $\mathrm{E}[w_1] = 0$ and $\mathrm{E}[w_1^2] = 1$. Define the wild bootstrap version of S_n as follows:

$$S_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i (X_i - \bar{X}), \quad \text{where } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i. \tag{0.1}$$

Given a significance level $\alpha \in (0,1)$, let $\hat{c}_{1-\alpha}$ be the $(1-\alpha)$-quantile of the conditional law of $T_n^* := \max_{1 \le j \le d} S_{n,j}^*$ given the data. The main result of this paper is an asymptotic expansion formula for the bootstrap coverage probability $P(T_n \ge \hat{c}_{1-\alpha})$. To state the result, we introduce some notation:

- ϕ_Σ is the density of $N(0, \Sigma)$.
- f_Σ is the density of $Z^\vee := \max_{1 \le j \le d} Z_j$ with $Z \sim N(0, \Sigma)$. Also, $c_{1-\alpha}^G$ is the $(1-\alpha)$-quantile of $Z^\vee$.
- $\mathbf{1}_d$ is the all-ones vector in $\mathbb{R}^d$.
- $\overline{X^3} := n^{-1} \sum_{i=1}^n X_i^{\otimes 3}$.

Theorem (Asymptotic expansion of bootstrap coverage probability)**.** *Under regularity conditions,*

$$P(T_n \ge \hat{c}_{1-\alpha}) = \alpha - (1 - \mathrm{E}[w_1^3]) Q_n(c_{1-\alpha}^G) - \mathrm{E}[R_n(\alpha)] + O\left(\frac{\log^3(dn)}{n} \log n \right)$$

as $d, n \to \infty$, where

$$Q_n(t) := -\frac{1}{6\sqrt{n}} \langle \mathrm{E}[\overline{X^3}], \int_{(-\infty, t]^d} \nabla^3 \phi_\Sigma(z) dz \rangle \quad (t \in \mathbb{R}),$$

*Graduate School of Mathematical Sciences, University of Tokyo
†CREST, Japan Science and Technology Agency

$$R_n(\alpha) := \frac{1}{\sqrt{n}} \frac{\left\langle \overline{X^3} \otimes \mathbf{1}_d, \Psi_\alpha^{\otimes 2} \right\rangle}{2 f_\Sigma(c^G_{1-\alpha})}, \qquad \Psi_\alpha := \int_{(-\infty, c^G_{1-\alpha}]^d} \nabla^2 \phi_\Sigma(z) dz$$

and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of tensors.

As a corollary, we obtain the following blessing-of-dimensionality type phenomenon:

Corollary. *Under the assumptions of the above theorem, if the covariance matrix of S_n has identical diagonal entries and bounded eigenvalues as $d, n \to \infty$, then*

$$P(T_n \geq \hat{c}_{1-\alpha}) = \alpha + O\left(\frac{\log^3(dn)}{n} \log n + \sqrt{\frac{\log^3 d}{dn}} \right),$$

provided that $\mathrm{E}[w_1^3] = 1$.

This result shows that under the stated assumptions on the covariance matrix, the third-moment match wild bootstrap is second-order accurate in the high-dimensional setting such that $d \gg n$ even when applied to a non-studentized statistic.

The full version of the paper is available at arXiv: `https://arxiv.org/abs/2404.05006`.

On a test for assessing vector correlation for latent factor models in high-dimensional settings

Takahiro Nishiyama[a], Masashi Hyodo[b] and Shoichi Narita[c]

[a] Department of Business Administration, Senshu University
[b] Faculty of Economics, Kanagawa University
[c] Graduate School of Economics, Kanagawa University

We let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-dimensional random sample with a population mean vector $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. We further partition $\mathbf{x}_i$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ into 2 components:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{x}_{gi}$ and $\boldsymbol{\mu}_g$ are $p_g \times 1$ vectors, and $\boldsymbol{\Sigma}_{gh}$ is a $p_g \times p_h$ matrix, $g, h \in \{1, 2\}$. Note that $p = p_1 + p_2$. The test for assessing the vector correlation can be fomulated as

$$\mathcal{H} : \boldsymbol{\Sigma}_{12} = \mathbf{O} \ \text{ vs. } \ \mathcal{A} : \boldsymbol{\Sigma}_{12} \neq \mathbf{O}. \tag{1}$$

Also, the data generation model is assumed to be a latent factor model expressed as

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{B}\mathbf{f} + \boldsymbol{\epsilon}. \tag{2}$$

Here, $\boldsymbol{\mu} \in \mathbb{R}^p$ is the population mean vector, $\mathbf{B}$ is the $p \times d$ non-random matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)^\top$ that satisfies $\mathrm{rank}(\mathbf{B}) = d$, and elements $\mathbf{b}_1, \ldots, \mathbf{b}_p$ are referred to as factor loadings. $\mathbf{f} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ are random vectors for common and specific factors, respectively. We assume that $\mathbf{f}$ and $\boldsymbol{\epsilon}$ are independent. We let $\mathbf{f} = (\mathrm{f}_1, \ldots, \mathrm{f}_d)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^\top$. Furthermore, we assume that f_i is iid with $\mathrm{E}(\mathrm{f}_i) = 0$, $\mathrm{E}(\mathrm{f}_i^2) = 1$, and $\mathrm{E}(\mathrm{f}_i^4) = \kappa + 3 < \infty$. and ϵ_j are iid with $\mathrm{E}(\epsilon_j) = 0$, $0 < \mathrm{E}(\epsilon_j^2) = \psi_j < \infty$, $\mathrm{E}(\epsilon_j^4) = \psi_j^2(\kappa + 3) < \infty$ for $i \in \{1, \ldots, d\}$, and $j \in \{1, \ldots, p\}$. Under these assumptions, $\mathrm{E}(\mathbf{f}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{f}) = \mathbf{I}_d$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$.

We further partition $\mathbf{B}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\epsilon}$ into 2 components:

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \ \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_1 & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Psi}_2 \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{pmatrix},$$

where $\mathbf{B}_g$ is $p_g \times d$ nonrandom matrix that satisfies $\mathrm{rank}(\mathbf{B}_g) = d_g > 0$, $\boldsymbol{\Psi}_g$ is $p_g \times p_g$ diagonal matrix, and $\boldsymbol{\epsilon}_g$ is p_g-dimensional random vector.

To construct test (1), we introduced the following ρV coefficient of $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ given by [2]:

$$\rho V_{12} = \frac{\|\boldsymbol{\Sigma}_{12}\|_F^2}{\|\boldsymbol{\Sigma}_{11}\|_F \|\boldsymbol{\Sigma}_{22}\|_F},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The ρV-coefficient measures the correlation between two probability vectors. Because $\boldsymbol{\Sigma}_{12} = \mathbf{O}$ and $\rho V_{12} = 0$ are equivalent, the

estimator of ρV_{12} can be used to hypothesize testing (1). The RV coefficient introduced by [4] can be interpreted as a naive estimator of ρV-coefficient. However, [3] states that the RV coefficient takes high values when the sample size n is small, and when both p_1 and p_2 are large. Therefore, we defined the estimator of ρV_{12} with a high-dimensionality adjustment as

$$MRV_{12} = \frac{\widehat{\|\boldsymbol{\Sigma}_{12}\|_F^2}}{\widehat{\|\boldsymbol{\Sigma}_{11}\|_F}\widehat{\|\boldsymbol{\Sigma}_{22}\|_F}}.$$

Here, for $g \in \{1, 2\}$, $\widehat{\|\boldsymbol{\Sigma}_{gh}\|_F^2}$ is an unbiased estimator of $\|\boldsymbol{\Sigma}_{gh}\|_F^2$ derived by [5].

Then, to construct a hypothesis test (1), we defined the test statistic as

$$T = nMRV_{12} + \frac{\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_1)}\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_2)}}{\sqrt{\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_1^2)}\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_2^2)}}},$$

where, for $\in \{1, 2\}$, $\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g)} = \sum_{i=1}^{\widehat{d}_g} \widehat{\lambda}_{gi}$, $\widehat{\mathrm{tr}(\boldsymbol{\Lambda}_g^2)} = \sum_{i=1}^{\widehat{d}_g} \widehat{\lambda}_{gi}^2$ and $\widehat{\lambda}_{gi} = \lambda_i(\mathbf{S}_{gg})/p_g$ for $i \in \{1, 2, \ldots, \widehat{d}_g\}$. Here, $\lambda_i(\mathbf{S}_{gg})$ is the i-th largest eigenvalue of matrix $\mathbf{S}_{gg} = \{1/(n_g - 1)\} \sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)(\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)^\top$, $\mathbf{x}_{gi} = (1/n_g) \sum_{i=1}^{n_g} \mathbf{x}_{gi}$ and $\widehat{d}_g$ is a consistent estimator of d_g based on the ER method proposed by [1]. Besides, we derived the limiting null distribution of T under some assumptions, and constructed test procedure for testing (1). Also, we compared, through simulations, the performance of the proposed test and existing procedures suitable for test for assessing vector correlation in terms of size control and power.

References

[1] Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.

[2] Escoufier, Y. (1973). Le Traitement des variables vectorielles. *Biometrics* **29**, 751–760.

[3] Hyodo, M., Nishiyama, T., and Pavlenko, T. (2020). Testing for independence of high-dimensional variables:ρV-coefficient based approach. *J. Multivariate Anal.* **178**, 104627.

[4] Josse, J. and Holmes, S. (2016). Measuring multivariate association and beyond. *Statistics Surveys* **10**, 132–167.

[5] Yamada, Y., Hyodo, M., and Nishiyama, T. (2017). Testing block-diagonal covariance structure for high-dimensional data under non-normality. *J. Multivariate Anal.* **155**, 305–316.

Asymptotic locations of bounded and unbounded eigenvalues of sample correlation matrices of certain factor models – application to a components retention rule

Yohji Akama
The Mathematical Institute, Tohoku University,
Aramaki, Aoba, Sendai, 980-8578, Japan

Peng Tian
Laboratoire Jean Alexandre Dieudonné, Université Côte d'Azur,
28, Avenue Valrose, 06108 Nice Cedex 2, France

December 9, 2024

Let the dimension N of data and the sample size T tend to ∞ with $N/T \to c > 0$. The spectral properties of a sample correlation matrix $\mathbf{C}$ and a sample covariance matrix $\mathbf{S}$ are asymptotically equal whenever the population correlation matrix $\mathbf{R}$ is bounded [1]. We demonstrate this also for general linear models for *unbounded* $\mathbf{R}$, by examining the behavior of the singular values of multiplicatively perturbed matrices. By this, we establish: Given a factor model of an idiosyncratic noise variance σ^2 and a rank-r factor loading matrix $\mathbf{L}$ which rows all have common Euclidean norm L. Then, the kth largest eigenvalues λ_k $(1 \le k \le N)$ of $\mathbf{C}$ satisfy almost surely: (1) λ_r diverges, (2) $\lambda_k/s_k^2 \to 1/(L^2+\sigma^2)$ $(1 \le k \le r)$ for the kth largest singular value s_k of $\mathbf{L}$, and (3) $\lambda_{r+1} \to (1-\rho)(1+\sqrt{c})^2$ for $\rho := L^2/(L^2+\sigma^2)$. Whenever s_r is much larger than $\sqrt{\log N}$, then broken-stick rule [2, 3], which estimates $\operatorname{rank}\mathbf{L}$ by a random partition (Holst, 1980) of $[0, 1]$, tends to r (a.s.). We also provide a natural factor model where the rule tends to "essential rank" of $\mathbf{L}$ (a.s.) which is smaller than $\operatorname{rank}\mathbf{L}$.

References

[1] 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.

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

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