International Symposium on New Developments of Theories and Methodologies for Large Complex Data

November 5-6, 2021

Venue:

Conference Room 102, Tsukuba International Congress Center
2-20-3 Takezono, Tsukuba, Ibaraki 305-0032, Japan (Hybrid Symposium with Zoom)

Organizers:

Makoto Aoshima	(University of Tsukuba)
Mika Sato-Ilic	(University of Tsukuba)
Kazuyoshi Yata	(University of Tsukuba)
Aki Ishii	(Tokyo University of Science)
Yugo Nakayama	(Kyoto University)

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) 19K22837 (Project Period: 2019-2021)
"Tackling individualized modeling with ultra-high dimensional data"
(Principal Investigator: Makoto Aoshima)

Program (UTC+9)

November 5 (Friday)

14:00∼14:10 Opening

14:10∼15:00 Koji Tsukuda[*,a] and Shun Matsuura[b]
[a](Faculty of Mathematics, Kyushu University)
[b](Faculty of Science and Technology, Keio University)

On high-dimensional testing for common principal components model

15:00∼15:50 Takahiro Nishiyama[*,a], Masashi Hyodo[b] and Tatjana Pavlenko[c]
(Zoom) [a](Department of Business Administration, Senshu University)
[b](Department of Economics, Kanagawa University)
[c](Department of Mathematics, KTH Royal Institute of Technology)

A two sample Behrens-Fisher problem for factor models in high dimensions

15:50∼16:40 Shao-Hsuan Wang
(Zoom) (Graduate Institute of Statistics, National Central University)

On the asymptotic result of Kronecker envelope principal component analysis in high dimension low sample size data

16:45∼19:00 **Special Invited and Keynote Sessions 1**
(Zoom)

November 6 (Saturday)

10:00∼10:50 Yugo Nakayama[*,a], Kazuyoshi Yata[b] and Makoto Aoshima[b]
[a](Graduate School of Informatics, Kyoto University)
[b](Institute of Mathematics, University of Tsukuba)

Asymptotic properties of high-dimensional kernel PCA and its applications

10:50∼11:40 Thong Pham[*,a], Paul Sheridan[b] and Hidetoshi Shimodaira[c]
(Zoom) [a](RIKEN AIP)
[b](Tupac Bio)
[c](Graduate School of Informatics, Kyoto University, and RIKEN AIP)

Estimating preferential attachment in growing networks

11:40∼13:00 Lunch

13:00∼13:50 Tsubasa Ito[a] and Shonosuke Sugasawa[*,b]
(Zoom) [a](M&D Data Science Center, Tokyo Medical and Dental University)
[b](Center for Spatial Information Science, The University of Tokyo)

Grouped generalized estimating equations for heterogeneous longitudinal data

13:50∼14:40 Shota Katayama
(Zoom) (Faculty of Economics, Keio University)

Hypothesis testing on high dimensional parameter under confounding

14:40∼15:30 Tsutomu T. Takeuchi
(Division of Particle and Astrophysical Science, Nagoya University)

New quantification of galaxy evolution via manifold learning

15:35∼17:50 **Special Invited and Keynote Sessions 2**
(Zoom)

17:50∼18:00 Closing

(∗ Speaker)

Special Invited and Keynote Sessions 1

November 5 (Friday)

Special Invited Talk

16:45∼17:45 **Double data piling and negatively ridged classifiers in high dimensions**

Speaker: Sungkyu Jung
(Department of Statistics, Seoul National University)

Chair: Kazuyoshi Yata (Institute of Mathematics, University of Tsukuba)

Keynote Talk

17:50∼19:00 **Using the classical Growth Curve model in non-standard situations**

Speaker: Dietrich von Rosen
(Department of Energy and Technology, Swedish University of Agricultural Sciences)

Discussion Leader: Shinpei Imori
(Graduate School of Advanced Science and Engineering, Hiroshima University)

Special Invited and Keynote Sessions 2

November 6 (Saturday)

Special Invited Talk

15:35∼16:35 **Normal-reference tests for high-dimensional hypothesis testing**

Speaker: Jin-Ting Zhang
(Department of Statistics and Data Science, National University of Singapore)

Chair: Shota Katayama (Faculty of Economics, Keio University)

Keynote Talk

16:40∼17:50 **Generalized information criterion for high-dimensional PCA rank selection**

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

Discussion Leader: Yuan-Tsung Chang
(Department of Social Information, Mejiro University)

On high-dimensional testing for common principal components model *

Koji Tsukuda† and Shun Matsuura‡

In this presentation, a high-dimensional statistical test procedure for the common principal components (CPC) hypothesis on two population covariance matrices was discussed. The CPC hypothesis means that the population covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ can be simultaneously diagonalized; that is to say, there exists an orthogonal matrix $\boldsymbol{U}$ such that both $\boldsymbol{U}^\top \boldsymbol{\Sigma}_x \boldsymbol{U}$ and $\boldsymbol{U}^\top \boldsymbol{\Sigma}_y \boldsymbol{U}$ are diagonal. Testing the CPC hypothesis was firstly considered by Flury [2], and has been studied in several works; see, for example, Boente, Pires, and Rodrigues [1] and Hallin, Paindaveine and Verdebout [3, 4]. Unlike these studies, we considered the case where sample sizes are smaller than the number of observed variables.

Our approach is based on the asymptotic normality of

$$M = \frac{1}{p^2\sqrt{n_a n_b n_c n_d}} \mathrm{tr}(\boldsymbol{T}_a(n_a)\boldsymbol{T}_b(n_b)\boldsymbol{T}_c(n_c)\boldsymbol{T}_d(n_d)),$$

that is the trace of the product of four $p \times p$ independent Wishart matrices $\boldsymbol{T}_a(n_a), \boldsymbol{T}_b(n_b), \boldsymbol{T}_c(n_c), \boldsymbol{T}_d(n_d)$ with respective degrees-of-freedom n_a, n_b, n_c, n_d and respective scale matrices $\boldsymbol{\Sigma}_a, \boldsymbol{\Sigma}_b, \boldsymbol{\Sigma}_c, \boldsymbol{\Sigma}_d$, under a high-dimensional asymptotic regime

$$n_a, n_b, n_c, n_d \asymp p^\delta,\ 0 < \delta < 1.$$

The test statistic is shown to be asymptotically distributed as the standard normal distribution under the null hypothesis. It is also shown that the asymptotic power of the test goes to 1 under the alternative hypothesis. The performances of the test for finite samples are numerically examined.

This presentation was based on the article [5].

*This work is supported by JSPS KAKENHI (grant number: 18K13454,20K11713,21K13836)

†Faculty of Mathematics, Kyushu University

‡Faculty of Science and Technology, Keio University

References

[1] Boente, G., Pires, A.M., and Rodrigues, I.M. (2009). Robust tests for the common principal components model. *J. Statist. Plann. Inference* **139**, no.4, 1332–1347.

[2] Flury, B.N. (1984). Common principal components in k groups. *J. Amer. Statist. Assoc.* **79**, no.388, 892–898.

[3] Hallin, M., Paindaveine, D., Verdebout, T. (2010). Testing for common principal components under heterokurticity. *J. Nonparametr. Stat.* **22**, no.7, 879–895.

[4] Hallin, M., Paindaveine, D., Verdebout, T. (2013). Optimal rank-based tests for common principal components. *Bernoulli* **19**, no.5B, 2524–2556.

[5] Tsukuda, K., Matsuura, S. (2021). Limit theorem associated with Wishart matrices with application to hypothesis testing for common principal components. *J. Multivariate Anal.* **186**, 104822.

A two sample Behrens-Fisher problem for factor models in high dimensions

Takahiro Nishiyama[a], Masashi Hyodo[b] and Tatjana Pavlenko[c]

[a] Department of Business Administration, Senshu University
[b] Faculty of Economics, Kanagawa University
[c] Department of Mathematics, KTH Royal Institute of Technology

Let $\mathbf{x}_{gi} = (x_{gi1}, \ldots, x_{gip})^\top \sim \mathcal{F}_g$ be iid p-dimensional random vectors collected from the ith subject in the gth population, where $\mathcal{F}_g$ denotes the distribution function for gth population, $i \in \{1, \ldots, n_g\}$, $g \in \{1, 2\}$. A factor model assumes that for each $g \in \{1, 2\}$, the observable vector $\mathbf{x}_{gi}$ is decomposable into a latent factor and an idiosyncratic component as follows:

$$\mathbf{x}_{gi} = \boldsymbol{\mu}_g + \mathbf{B}_g \mathbf{z}_{gi} + \boldsymbol{\Psi}_g^{1/2} \boldsymbol{\epsilon}_{gi}, \tag{1}$$

where $\boldsymbol{\mu}_g \in \mathbb{R}^p$ is a deterministic intercept vector, $\mathbf{z}_{gi} = (\mathrm{z}_{gi1}, \ldots, \mathrm{z}_{gid_g})^\top$ is the d_g-dimensional latent factor vector, and $\boldsymbol{\epsilon}_{gi} = (\epsilon_{gi1}, \ldots, \epsilon_{gip})^\top$ is the p-dimensional error vector which is uncorrelated with the latent factor. In what follows, we assume that $d_g \in \mathbb{N}$ is a fixed number. Further, $\mathbf{B}_g = (\mathbf{b}_{g1}, \ldots, \mathbf{b}_{gp})^\top$ denotes the loading matrix where for each $j \in \{1, \ldots, p\}$, $\mathbf{b}_{gj} = (b_{gj1}, \ldots, b_{gjd_g})^\top \in \mathbb{R}^{d_g}$ is a non-random vector, and $\boldsymbol{\Psi}_g = \mathrm{diag}(\psi_{g1}, \ldots, \psi_{gp})$ is the non-random $p \times p$ diagonal matrix whose elements are $\psi_{g1} > 0, \ldots, \psi_{gp} > 0$. For the latent vector $\mathbf{z}_{gi}$ and error vector $\boldsymbol{\epsilon}_{gi}$, we further assume that $\mathrm{z}_{gi\ell}$ are iid with $\mathrm{E}(\mathrm{z}_{gi\ell}) = 0$, $\mathrm{E}(\mathrm{z}_{gi\ell}^2) = 1$ and $\mathrm{E}(\mathrm{z}_{gi\ell}^4) = \kappa_{\mathrm{z}_g} < \infty$, and ϵ_{gij} are iid with $\mathrm{E}(\epsilon_{gij}) = 0$, $\mathrm{E}(\epsilon_{gij}^2) = 1$ and $\mathrm{E}(\epsilon_{gij}^4) = \kappa_{\epsilon_g} < \infty$ for $g \in \{1, 2\}$, $i \in \{1, \ldots, n_g\}$, $j \in \{1, \ldots, p\}$ and $\ell \in \{1, \ldots, d_g\}$. Structural assumptions of the model (1) imply that

$$\mathrm{E}(\mathbf{x}_{gi}) = \boldsymbol{\mu}_g, \ \mathrm{cov}(\mathbf{x}_{gi}) = \mathbf{B}_g \mathbf{B}_g^\top + \boldsymbol{\Psi}_g := \boldsymbol{\Sigma}_g,$$

where $\boldsymbol{\Sigma}_g \in \mathbb{R}_{>0}^{p \times p}$ and $\mathbb{R}_{>0}^{p \times p}$ denotes the space of real, symmetric, positive definite, $p \times p$ matrices.

By using the data generated by (1), we designed a high-dimensinal test procedure for testing

$$\mathcal{H} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \quad \mathcal{A} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{2}$$

The proposed test will be valid for high-dimensional, non-normal, unbalanced data under two-sample Behrens-Fisher setting for the low-dimensional factor model.

Recently, a number of useful two-sample tests have been proposed for high-dimensional settings. Nevertheless, their applicability relies on a critical assumption, that is, requiring elements of $\mathbf{x}_{gi}$ to be weakly dependent. Being expressed in terms of the latent factor model (1) and assuming the common covariance matrix $\boldsymbol{\Sigma}$, the weak dependence is equivalent to an assumption stated as $\mathrm{tr}(\boldsymbol{\Sigma}^4)/\mathrm{tr}^2(\boldsymbol{\Sigma}^2) \to 0$, when $p \to \infty$; see e.g. assumption

(3.6) in Chen and Qin (2010). This assumption is crucial for establishing the asymptotic normality of the test statistic proposed by Chen and Qin (2010); see Theorem 1 of their paper. However, besides of being difficult to verify in practice, the regularity assumptions imposed on the covariance structure in e.g. Bai and Saranadasa (1996) and Chen and Qin (2010) can be easily violated in covariance models where the eigenvalues of $\mathbf{\Sigma}_g$ are dominated by few top ones. This is precisely the type of covariance structure underlying $\mathbf{x}_{gi}$: under the low-dimensional latent factor model (1), the first d_g eigenvalues of $\mathbf{\Sigma}_g$ are considerably larger than than the rest. So, Ma et al (2015) proposed the mean difference test for the model (1) under homoscedastisity. However, an assumption of common covariance matrix is a very strong assumption which is hard to be practically verified in $p \gg n_g$ settings, this in turn limits the applicability of the procedure developed by these authors. We do not assume that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$; out test statistics, along with their limit properties are studied under heteroscedasticity, i.e. solves a general, two-sample Behrens-Fisher problem for the latent factor model (1).

For this problem, we defined the data-driven test statistic as

$$T_{FA} = \frac{n}{p}\left\{\|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2 - \frac{1}{n_1}\widehat{\mathrm{tr}(\mathbf{\Psi}_1)} - \frac{1}{n_2}\widehat{\mathrm{tr}(\mathbf{\Psi}_2)}\right\},$$

where, for $g \in \{1,2\}$, $\overline{\mathbf{x}}_g = (1/n_g)\sum_{i=1}^{n_g}\mathbf{x}_{gi}$ and $\widehat{\mathrm{tr}(\mathbf{\Psi}_g)} = \mathrm{tr}(\mathbf{S}_g) - \sum_{\ell=1}^{\widehat{d}_g}\lambda_\ell(\mathbf{S}_g)$. Here, $\lambda_\ell(\mathbf{S}_g)$ is the ℓth largest eigenvalue of matrix $\mathbf{S}_g = \{1/(n_g-1)\}\sum_{i=1}^{n_g}(\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)(\mathbf{x}_{gi} - \overline{\mathbf{x}}_g)^\top$ and $\widehat{d}_g$ is a consistent estimator of d_g based on the ER method proposed by Ahn and Horenstein (2013). Besides, we derived the limiting null distribution of T_{FA} under some assumptions and constructed test procedure for testing (2). Also, we compared, through simulations, the performance of the proposed test and existing procedures suitable for a two-sample, Behrens-Fisher problem in high-dimensional data in terms of size control and power.

References

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

[2] Bai, Z.D., Saranadasa, H., 1996. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, **6**, 311–329.

[3] Chen, S.X., Qin, Y.L., 2010. A two-sample test for high dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808–835.

[4] Ma, Y., Lan, W., Wang, H., 2015. A high dimensional two-sample test under a low dimensional factor structure. *J. Multivariate Anal.*, **140**, 162–170.

On the asymptotic result of Kronecker envelope principal component analysis in high dimension low sample size data

Shao-Hsuan Wang
(Graduate Institute of Statistics, National Central University)

Abstract

In the analysis of high dimensional data, the Kronecker envelope principal component analysis (KEPCA) serves as an efficient alternative to the classical PCA at pre-processing steps. We derive the consistency and the asymptotic normality of Kronecker envelope PCA in High dimension low sample size (HDLSS) setting and compare it with classical PCA. Face database example is used to demonstrate our method.

Double data piling and negatively ridged classifiers in high dimensions

Sungkyu Jung

(Department of Statistics, Seoul National University)

Abstract

Data piling refers to the phenomenon that training data vectors from each class project to a single point for classification. While this interesting phenomenon has been a key to understanding many distinctive properties of high-dimensional discrimination, the theoretical underpinning of data piling is far from properly established. In this work, high-dimensional asymptotics of data piling is investigated under a spiked covariance model, which reveals its close connection to the well-known ridged linear classifier. In particular, by projecting the ridge discriminant vector onto the subspace spanned by the leading principal component directions and the maximal data piling vector, we show that a negatively ridged discriminant vector can asymptotically achieve data piling of independent test data, essentially yielding a perfect classification. The second data piling direction is obtained purely from training data and shown to have a maximal property. Furthermore, asymptotic perfect classification occurs only along the second data piling direction. This interesting phenomenon is shown to also occur in multi-category classification problems, in which the second data piling subspaces are estimated by negatively ridged discriminant subspaces. We demonstrate that negative ridge parameters can be optimal in classification of well-known image and microarray datasets.

Using the classical Growth Curve model in non-standard situations

Dietrich von Rosen
(Department of Energy and Technology, Swedish University of Agricultural Sciences)

Abstract

More than fifty years ago the Growth Curve model (GMANOVA) was introduced. It is a multivariate bilinear regression model which under a normality assumption belongs to the curved exponential family. The model is of great interest since explicit maximum likelihood estimators exist. In the presentation we will focus on estimation based on the likelihood. Four different types of models will be discussed via the Growth Curve model and all types are connected to high-dimensional problems.

The first case is a discussion of the Growth Curve model when the number of independent observations are less than the number of repeated measurements. It turns out that the Moore-Penrose generalized inverse of a singular Wishart matrix can be used. Unfortunately, moments for the Moore-Penrose inverse are not available if the dispersion matrix is unstructured and we discuss what to do in this situation.

The second case which will be considered is about estimation in Partial Least Squares (PLS). PLS is often connected with different algorithms. In this talk, when estimating parameters, we will set up a model which gives the same result as algorithms. In PLS a Krylov space plays a central role.

The third case which will be discussed is when in the Growth Curve model there are rank restrictions on the mean parameters of the model. When there are a huge number of background variables this type of model can be useful. If one can imagine that there are a few number of latent processes which govern the background variables the rank restrictions make sense.

In the fourth case we extend the third one by also allowing for rank restrictions on the dispersion matrix. Rank restrictions on an unknown dispersion matrix has not been studied much and we will show how the mathematics for finding estimators can be carried out. However, there are still many open problems connected to the interpretability of the results and connected to the decision of the rank on both the mean parameters and the dispersion of the model.

Asymptotic properties of high-dimensional kernel PCA and its applications

Yugo Nakayama[a], **Kazuyoshi Yata**[b] **and Makoto Aoshima**[b]
[a]Graduate School of Informatics, Kyoto University
[b]Institute of Mathematics, University of Tsukuba

1 Introduction

In this talk, we considered the kernel principal component analysis (KPCA) for high-dimension, low-sample-size (HDLSS) data. We proposed clustering by using asymptotic properties of the KPCA and applied it to outlier detection.

2 Asymptotic properties of high-dimensional kernel PCA

Suppose there are independent and d-variate populations, $\Pi_i,\ i = 1, 2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ for each i. Suppose we have a $d \times n$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, where $\boldsymbol{x}_j$s are independently taken from Π_1 or Π_2. Let

$$n_i = \#\big\{j | \boldsymbol{x}_j \in \Pi_i \ \text{ for } j = 1, ..., n\big\},$$

where $\#A$ denotes the number of elements in a set A. Note that $n = n_1 + n_2$. We assume that n and n_is are independent of d, and $n_i \ge 1$ for $i = 1, 2$. For the sake of simplicity, we assume that $\mathrm{tr}(\boldsymbol{\Sigma}_1) \le \mathrm{tr}(\boldsymbol{\Sigma}_2)$ and

$$\boldsymbol{x}_j \in \Pi_1,\ j = 1, ..., n_1, \quad \boldsymbol{x}_j \in \Pi_2,\ j = n_1 + 1, ..., n. \tag{1}$$

In this section, we consider asymptotic properties of the KPCA with the linear kernel $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \boldsymbol{x}_j^\top \boldsymbol{x}_{j'}$ in the HDLSS context that $d \to \infty$ while n is fixed. Let $\boldsymbol{K}$ be an $n \times n$ gram matrix with the (j, j') element $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$. Let $\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^\top$, where $\boldsymbol{I}_n$ denotes the n-square identity matrix and $\boldsymbol{1}_n = (1, ..., 1)^\top$. We define the (centroid) gram matrix by

$$\boldsymbol{K}_0 = \boldsymbol{P}_n\boldsymbol{K}\boldsymbol{P}_n.$$

Note that $\mathrm{rank}(\boldsymbol{K}_0) \le n - 1$. Let $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_{n-1}$ be the eigenvalues of $\boldsymbol{K}_0$. Then, we define the eigen-decomposition of $\boldsymbol{K}_0$ by

$$\boldsymbol{K}_0 = \sum_{i=1}^{n-1} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^\top,$$

where $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, ..., \hat{u}_{in})^\top$ denotes a unit eigenvector corresponding to the $\hat{\lambda}_i$. The ith (normalized) PC score of $\boldsymbol{x}_j$ is given by $s_{ij} = \sqrt{n}\hat{u}_{ij}$. Since the sign of an eigenvector is arbitrary, we assume that $(\mathbf{1}_{n_1}^\top, -\mathbf{1}_{n_2}^\top)\hat{\boldsymbol{u}}_1 \geq 0$ without loss of generality.

We assume the following condition:

(A-i) $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2|\boldsymbol{x} \in \Pi_i) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$ as $d \to \infty$ for $i = 1, 2$.

Note that $\mathrm{E}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2|\boldsymbol{x} \in \Pi_i) = \mathrm{tr}(\boldsymbol{\Sigma}_i)$ for $i = 1, 2$. If Π_is are Gaussian, it holds that $\mathrm{Var}(\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2|\boldsymbol{x} \in \Pi_i) = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ for $i = 1, 2$, so that (A-i) naturally holds. Let $\Delta_\mu = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ and $\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|$, where $\|\cdot\|$ denotes the Euclidean norm. Here, we assume the following conditions:

(A-ii) $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_\mu^2 = o(1)$ as $d \to \infty$ for $i = 1, 2$.

(A-iii) $\limsup_{d\to\infty} \dfrac{\Delta_\Sigma}{n_1\Delta_\mu} < 1$ when $n_2 \geq 2$.

Nakayama et al. (2021) gave the following result.

Theorem 2.1. *Assume (A-i) to (A-iii). Then, it holds that as $d \to \infty$*

$$s_{1j} = \begin{cases} \sqrt{n_2/n_1} + o_P(1), & j = 1, ..., n_1, \\ -\sqrt{n_1/n_2} + o_P(1), & j = n_1 + 1, ..., n. \end{cases} \tag{2}$$

Thus from Theorem 2.1, one can classify $\boldsymbol{x}_j$s into two groups by the sign of the first PC scores.

Remark 1. If Δ_μ/Δ_Σ is small, we do not recommend to use the linear kernel. In such case, you can use the Gaussian kernel $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) = \exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_{j'}\|^2/\gamma)$, where $\gamma > 0$. Note that the Gaussian kernel can draw information about heteroscedasticity via the difference of $\boldsymbol{\Sigma}_i$s.

We considered applying the proposed clustering method to outlier detection. We set $n_1 = 1$ and $\boldsymbol{x}_1$ is regarded as an outlier. From (2), under some regularity conditions, it holds as $d \to \infty$

$$s_{1j} = \begin{cases} \sqrt{n-1} + o_P(1), & j = 1, \\ -1/\sqrt{n-1} + o_P(1), & j = 2, ..., n. \end{cases}$$

Thus by using the first PC scores, one can detect the outlier.

References

[1] Y. Nakayama, K. Yata, M. Aoshima, Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings, J. Multivariate Anal. 185 (2021) in print.

Estimating preferential attachment in growing networks

Thong Pham*[1], Paul Sheridan[2], and Hidetoshi Shimodaira[3,1]

[1]RIKEN Center for AIP
[2]Tupac Bio, Inc.
[3]Kyoto University

We consider the problem of estimating the preferential attachment (PA) function A_k from empirical data that does not contain any information about the growth process of the network. Consider a growing network that grows from time-step $t = 1$ to T and denote its snapshot at time-step t by G_t. Traditionally, the estimation of A_k is often considered when the growth process of the network can be observed at at least two time-steps [6, 5, 7]. However, what if we cannot observe anything about the growth process and have to content ourselves with only the one snapshot G_T? When it comes to estimating PA in general growing networks without time-resolved data, no satisfactory methods exist. All existing methods assume either unrealistic network types or unnecessarily restrictive functional forms for A_k [1, 4, 2, 3].

We propose a method called PAFit-oneshot to nonparamterically estimate the PA function of a growing network from its final snapshot, G_T, alone [9]. Our method does not assume any functional form for the PA function, and can be applied to any real-world network snapshot.

Our method can be summarized as follows. From Theorem 1 in [9], one gets the basic equation:

$$A_k \approx \frac{\mathbb{E}\sum_{j>k} n_j(T)}{\mathbb{E}n_k(T)}, \tag{1}$$

where $n_k(t)$ is the number of degree k nodes at time-step t. If one estimates $\mathbb{E}\sum_{j>k} n_j(T)$ and $\mathbb{E}n_k(T)$ by $\sum_{j>k} n_j(T)$ and $n_k(T)$, one arrives at the baseline estimator:

$$\hat{A}_k^{\text{baseline}} = \frac{\sum_{j>k} n_j(T)}{n_k(T)}, \tag{2}$$

which is the estimator of Gao et al. [4]. However, there is a severe underestimation of A_k in the region of large k by this method (see Fig. 1 of [9]). We call this underestimation the waterfall artefact.

The root of the waterfall artefact is that $n_k(T)$ is a poor estimator of $\mathbb{E}n_k(T)$ when k is large. Given G_T, we only estimate the A_k values for the degrees k observed in G_T, which means $n_k(T)$ is positive *a priori*. Therefore, $n_k(T)$ is actually an estimator for the conditional expectation $\mathbb{E}\left[n_k(T) \mid n_k(T) > 0\right]$, which is equal to $\mathbb{E}n_k(T)/\mathbb{P}(n_k(T) > 0)$. Surprisingly, the presence of this bias, let alone a proposed correction for it, has never been discussed in the literature.

[*] Corresponding author. Email: thong.pham@riken.jp

To remove this bias, we need to estimate $p_k = \mathbb{P}(n_k(T) > 0)$, the probability that k exists in G_T. This resembles the problem of selective inference. Selective inference typically considers model selection in a regression setting and adjusts the bias of regression coefficients for the selected predictors; one must correct for the effect of choosing the predictor. In our problem, model selection is equivalent to choosing over which values of k to use to estimate A_k, which is where k exists in G_T. Correcting for p_k leads us to the following equation:

$$A_k \approx \frac{\sum_{j>k} n_j(T)}{n_k(T) p_k}. \tag{3}$$

Starting from initial rough estimations of p_k and A_k, our method iteratively improves these estimations using Monte Carlo simulations (see Fig. 6 of [9]). The proposed method is implemented in the R package `PAFit` [8].

References

[1] Ivona Bezáková, Adam Kalai, and Rahul Santhanam. Graph model selection using maximum likelihood. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 105–112, New York, NY, USA, 2006. ACM.

[2] George T. Cantwell, Guillaume St-Onge, and Jean-Gabriel Young. Inference, model selection, and the combinatorics of growing trees. *Phys. Rev. Lett.*, 126:038301, Jan 2021.

[3] Harry Crane and Min Xu. Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):639–668, 2021.

[4] Fengnan Gao, Aad van der Vaart, Rui Castro, and Remco van der Hofstad. Consistent estimation in general sublinear preferential attachment trees. *Electron. J. Statist.*, 11(2):3979–3999, 2017.

[5] H Jeong, Z Néda, and AL Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(61):567–572, 2003.

[6] MEJ Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[7] Thong Pham, Paul Sheridan, and Hidetoshi Shimodaira. PAFit: A statistical method for measuring preferential attachment in temporal complex networks. *PLOS ONE*, (9):e0137796, 9 2015.

[8] Thong Pham, Paul Sheridan, and Hidetoshi Shimodaira. PAFit: An R package for the non-parametric estimation of preferential attachment and node fitness in temporal complex networks. *Journal of Statistical Software, Articles*, 92(3):1–30, 2020.

[9] Thong Pham, Paul Sheridan, and Hidetoshi Shimodaira. Non-parametric estimation of the preferential attachment function from one network snapshot. *Journal of Complex Networks*, 9(5), 09 2021. cnab024.

Grouped generalized estimating equations for heterogeneous longitudinal data

Tsubasa Ito[1] and Shonosuke Sugasawa[2]

[1]M&D Data Science Center, Tokyo Medical and Dental University

[2]Center for Spatial Information Science, The University of Tokyo

Longitudinal data in which responses (repeated measurements) within the same subject are correlated is appeared in many scientific applications such as biomedical statistics and social science. For analyzing longitudinal data, it is typically difficult to correctly specify the underlying correlation structures among response variables within the same subject, and one of the standard approaches is the generalized estimating equations (GEE) developed by Liang and Zeger (1986), which uses "working" correlation structures specified by users. The advantage of the GEE approach is that the estimator is still consistent even when the working correlation is misspecified. However, the existing GEE methods assumes homogeneous regression coefficients that are common to all the subject, which could be restrictive in practical applications since there might be potential heterogeneity among subjects or clusters as confirmed in several applications.

In this work, we extend the standard GEE analysis to take account of potential heterogeneity in longitudinal data. Specifically, we develop grouped GEE analysis by adopting the grouping approach that is widely adopted in literatures for panel data analysis. We assume that subjects in longitudinal data can be classified to a finite number of groups, and subjects within the same group share the same regression coefficients, that is, the regression coefficients are homogeneous over subjects in the same groups. Since the grouping assignment of subjects are unknown, we treat it as unknown parameters and estimate them and group-wise regression coefficient simultaneously. Given the grouping parameters, the standard GEE can be performed to obtain group-wise estimators of regression coefficients. On the other hand, given the group-wise regression coefficients, we consider estimating the grouping parameters using a kind of Mahalanobis distance between response variables and predictors with taking account of potential correlations via working correlation matrix. In other words, we employ the working correlation not only in performing GEE analysis in each group but also estimating the grouping assignment. We will show that the grouped GEE method can be easily carried out by a simple iterative algorithm similar to *k-means* algorithm that combines the existing algorithm for the standard GEE and simple optimization steps for grouping assignment. Moreover, we adopt the cross validation to carry out data-dependent selection of the number of groups.

We derive the statistical properties of the grouped GEE estimator in an asymptotic framework where both n (the number of subjects) and T (the number of repeated measurements) tend to infinity, but we here allow T to grow considerably slower than n, namely, $n/T^{\nu} \to 0$ for some large ν. Hence, our method can be applicable when T is much smaller n as observed in many applications using longitudinal data. As theoretical difficulties of the grouped estimation in longitudinal data analysis, the true correlations within the same subject can be considerably high, so the existing theoretical argument assuming negligibly small correlations imposed typically by mixing conditions for the underlying true correlations is no more applicable. To overcome the limitation of the existing theoretical argument, we consider grouping assignment using a kind of Mahalanobis distance with working correlation, and we will show that such grouping strategy leads to consistent estimation of the grouping parameters as long as the working correlation is relatively close to the true one. Therefore, even when the underlying correlations within the same subject is not weak, we can successfully estimate the grouping parameters using a reasonable working correlation matrix. Then, we will also establish consistency and asymptotic normality of the grouped GEE estimator of the regression coefficients, and also provide a consistent estimator of asymptotic variances.

Hypothesis testing on high dimensional parameter under confounding

Shota Katayama

Faculty of Economics, Keio University, Japan

1 Introduction

In high dimensional data analysis, especially in differential gene expression analysis, detecting difference of a huge number of features between two groups is essential problem. Each individual is assigned to one of two groups according to difference in conditions, e.g., whether they have been treated or not. In situations where one can assume a random assignment, the two groups are different due to the condition alone, and thus a pure comparison is possible. As in Potter (2003) and Heller et al. (2009), however, the random assignment is not feasible in gene data analysis where the condition may be cancer or non-cancer. In this case, the difference between the two groups may stem from variables other than the condition of interest, and hence such *confounding* variables need to be adjusted appropriately. Moreover, gene data are not necessary continuous values, but can be binary or count.

This talk provides an unified inference on high dimensional parameter for the comparison of two groups. Suppose we have p dimensional response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$, q dimensional confounding vector $\boldsymbol{X} = (X_1, \ldots, X_q)^T$ with $X_1 = 1$ for intercept, and an indicator variable $D \in \{0, 1\}$ introducing the difference in conditions. Assume that

$$\mathbb{E}(Y_j \,|\, D, \boldsymbol{X}) = h\big(D\tau_j^* + \boldsymbol{X}^T\boldsymbol{\theta}_j^*\big), \quad j = 1, 2, \ldots, p, \tag{1}$$

$$\mathbb{P}(D = 1 \,|\, \boldsymbol{X}) = 1/(1 + \exp(-\boldsymbol{X}^T\boldsymbol{\gamma}^*)), \tag{2}$$

where $\tau_j^* \in \mathbb{R}, \boldsymbol{\theta}_j^* \in \mathbb{R}^q$, $\boldsymbol{\gamma}^* \in \mathbb{R}^q$ and $h(\cdot)$ is a known differentiable inverse link function. The first equation (1) models the dependence between Y_j and $(D, \boldsymbol{X})$ under a general framework to deal with various type of responses. For instance, $h(\cdot)$ would be identity function when Y_j is continuous and would be $\exp(\cdot)$ when Y_j is count. The second (2) assumes the logistic model for D. Finally, suppose we have n i.i.d. observations $\{(\boldsymbol{Y}_i, D_i, \boldsymbol{X}_i)\}_{i=1}^n$ following (1) and (2) and consider the high dimensional setting where p is greater than n, but q is smaller.

In (1) the parameter τ_j^* represents the effect of D on Y_j, and thus the high dimensional parameter vector $\boldsymbol{\tau}^* = (\tau_1^*, \dots, \tau_p^*)^T$ is of interest. Under the strong ignorability, $\boldsymbol{\tau}^*$ can also be regarded as a causal parameter. This talk provides an inference on $\boldsymbol{\tau}^*$ under the high dimensional asymptotic framework $(n, p) \to \infty$.

2 Proposed methodology

Applying the maximum likelihood estimation to each Y_j as the generalized linear regression, we can obtain an estimator of τ_j^*. In the case where $\boldsymbol{X}$ is confounded, however, the MLE of τ_j^* is not efficient since it ignores the model (2). Chernozhukov et al. (2018) and Vansteelandt and Dukes (2020) provide an efficient score function of τ_j^* given by

$$S_i(\tau_j\,; \boldsymbol{\theta}_j^*, \boldsymbol{\gamma}^*) = \frac{Y_{ij} - h(D_i\tau_j + \boldsymbol{X}_i^T\boldsymbol{\theta}_j^*)}{h'(D_i\tau_j + \boldsymbol{X}_i^T\boldsymbol{\theta}_j^*)}(D_i - \pi_i(\boldsymbol{\gamma}^*)), \tag{3}$$

where $\pi_i(\boldsymbol{\gamma}^*) = 1/(1 + \exp(-\boldsymbol{X}_i^T\boldsymbol{\gamma}^*))$. The nuisance parameters $(\boldsymbol{\theta}_j^*, \boldsymbol{\gamma}^*)$ can be estimated by MLE or other methods and plugged into the score function above, but we need the cross-fitting approach (Chernozhukov et al. (2018)) to control the asymptotic behavior of (3) with estimated nuisance parameters.

Based on the estimate $\hat{\tau}_j$ and its variance estimate $\hat{\sigma}_j^2$ by the cross-fitting, this talk gives a maximum type test statistics for the global hypothesis $H_0 : \boldsymbol{\tau}^* = \mathbf{0}$. A multiple testing procedure with controlling the false discovery rate is also provided.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal*, **21**(1), C1–C68.

Heller, R., Manduchi, E. and Small, D. S. (2009). Matching methods for observational microarray studies. *Bioinformatics*, **25**(7), 904–909.

Potter, J. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. *Trends in Genetics*, **19**, 690–695.

Vansteelandt, S. and Dukes, O. (2020). Assumption-lean inference for generalised linear model parameters. http://arxiv.org/abs/2006.08402

New Quantification of Galaxy Evolution via Manifold Learning

Tsutomu T. TAKEUCHI[1,2], Suchetha COORAY[1,†], Kai T. KONO[1]

1. Division of Particle and Astrophysical Science, Nagoya University, Nagoya 464-8602, Japan

2. The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

† JSPS Research Fellow (DC1)

1. Galaxy Formation and Evolution

Matter in the early Universe was almost uniform, and a slightly dense region grew by gravity, finally into a galaxy. It was attempted to develop a theory to deal with the star formation and associated history of heavy element synthesis, under an assumption that a galaxy has formed from a single, huge gas cloud. While the research in this direction was once completed in the first half of 1980s, this was not the end of the studies of galaxy evolution. Cosmological research that has progressed in parallel has revealed that galaxies merge and grow. This indicates that the galaxy evolution is a very complicated process that strongly depends on the density of the surrounding galaxies and the gas density. In order to formulate the galaxy evolution, it is necessary to determine such a huge system of equations. Though astrophysicists have constructed the governing equations from the physical laws from the first principle before, such a method is not realistic anymore when the quantity space exceeds 10 dimensions. Galaxy surveys as of the 21st century provides hundreds of physical quantities for hundreds of millions of galaxies, typical big data in both quality and quantity indeed. The feature space of the galaxy to be analyzed exceeds 100 dimensions. Therefore, the characterization of the galaxy evolution is no longer possible by the traditional method relying on physical intuition.

2. Galaxy Manifold

2.1 Rise, fall, and revival of the galaxy manifold

From 1970s to the mid-1980s, classical multivariate analysis methods such as the principal component analysis (PCA) were used to combine physical quantities of galaxies in a high-dimensional space. Various (logarithmic) linear relations, so-called galactic scaling relations, have been discovered. Research to unify the scaling relations and find the fundamental relationships has led to the concept of galaxy manifolds. However, the galaxy manifold has once been almost forgotten because the classical PCA could treat only linear relations, and it remained a limited concept, though they are still useful for exploring (log)linear relations of galaxies.

Recently, we discovered a galaxy manifold that expresses the basics of galactic evolution by the Fisher EM algorithm. Because of its strongly nonlinear spatial structure, it could have never been found in previous studies based on the classical PCA. To understand the manifold, a more sophisticated method beyond a mere classification is needed. We focused on a method known as the manifold learning, one of the latest methods of data science that is completely different from conventional methodologies

2.2 Galaxy manifold constructed by manifold learning

We adopt the algorithm Isomap and UMAP (Uniform Manifold Approximation and Projection). Isomap defines the neighboring points by using input-space distance and the distant points as a sequence of “short hops” between neighboring points. Isomap tries to find shortest paths in a graph with edges connecting neighboring data points. By construction, Isomap preserves the “surface density” of data points in the feature space. UMAP is based on differential geometry and algebraic topology. The algorithm is founded on three assumptions: 1) the data are uniformly distributed on a Riemannian manifold, 2) the Riemannian metric is locally constant (or can be approximated as such), and 3) the manifold is locally connected. From these assumptions it is possible to model the manifold with a fuzzy topological structure. Sine it defines the manifold so that the data points distribute as homogeneously as possible, it does not preserve the surface density of data points. UMAP also preserves some important structural properties, and it is more robust against noise than Isomap. Manifold learning algorithm can “unfold” a curved and/or rolled manifold in the feature space, and provide a local coordinate system on it. The resulting manifolds with local coordinates from Isomap and UMAP are presented in Fig. 1. From Figure 1, we clearly see that the galaxy manifold is two-dimensional. We also stress that two different algorithms, Isomap and UMAP yield similar two-dimensional manifolds. The difference of the two estimated manifolds is clearly seen in Fig. 1. Since Isomap preserves the density of data point cloud, we observe that the manifold has a density structure, i.e., dense and sparse regions on the manifold.

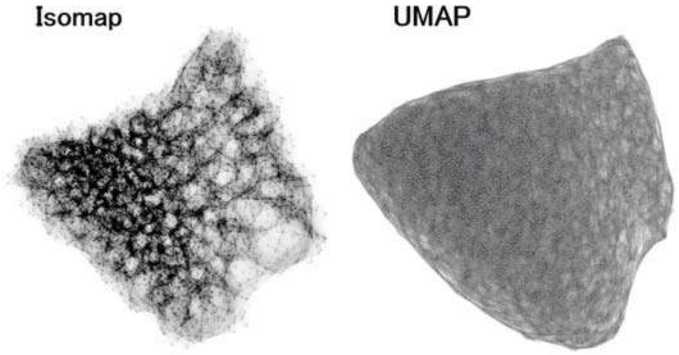

Figure 1: The “unfolded” galaxy manifold by a manifold learning algorithm Isomap and UMAP. Left and right panels show the manifolds from Isomap and UMAP, respectively. Though the global shape is slightly different from each other, they share common features on the manifold.

2.3 Result

The galaxy manifold obtained with Isomap preserve this information and reveal the speed of galaxy evolution at various stages along the manifold. e.g., galaxies passes the green valley very fast. In contrast, the galaxy manifold obtained with UMAP is imposed uniformity on the galaxy data, leading to a more robust and representative description of the observed galaxy properties e.g., galaxies evolve continuously in the feature space, without a discontinuity or “jump” on their evolutionary tracks. Thus, the galaxy manifold provides a clue to the evolutionary path of galaxies on the manifold. The SFR and stellar mass fields do not show the same evolutionary path. This supports that the galaxy merger without star formation plays a significant role in the growth of stellar mass. Next step is to fully parametrize the evolution equation of galaxies.

Normal-Reference Tests for High-Dimensional Hypothesis Testing

Jin-Ting Zhang

National University of Singapore

Abstract

In the past two decades, much attention has been paid for high-dimensional hypothesis testing. Several centralized or non-centralized L2-norm based test statistics have been proposed. Most of them imposed strong assumptions on the underlying covariance structure of the high-dimensional data so that the associated test statistics are asymptotically normally distributed. In real data analysis, however, these assumptions are hardly checked so that the resulting tests have a size control problem when the required assumptions are not satisfied. To overcome this difficulty, in this talk, we investigate a so-called normal-reference test which can control the size well. In the normal-reference test, the null distribution of a test statistic is approximated with that of a chi-square-type mixture which is obtained from the test statistic when the null hypothesis holds and when the samples are normally distributed. The distribution of the chi-square-type mixture can be well approximated by a three-cumulant matched x^2-approximation with the approximation parameters consistently estimated from the data. Two simulation studies demonstrate that in terms of size control, the proposed normal- reference test performs well regardless of whether the data are nearly uncorrelated, moderately correlated, or highly correlated and it performs much better than two existing competitors. A real data example illustrates the proposed normal-reference test.

KEY WORDS: x^2-type mixtures; high-dimensional data; three-cumulant matched x^2-approximation; two-sample Behrens–Fisher problem.

Generalized information criterion for high-dimensional PCA rank selection

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

Abstract

Principal component analysis (PCA) is a commonly used statistical tool for dimension reduction. An important issue in PCA is to determine the rank, which is the number of dominant eigenvalues of the covariance matrix. Among information-based criteria, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are the two most common ones. Both use the number of free parameters for assessing model complexity, which requires the validity of the simple spiked covariance model. As a result, AIC and BIC may suffer from the problem of model misspecification when the tail eigenvalues do not follow the simple spiked model assumption. To alleviate this difficulty, we adopt the idea of the generalized information criterion (GIC) to propose a model complexity measure for PCA rank selection. The proposed model complexity takes into account the sizes of eigenvalues and, hence, is more robust to model misspecification. Asymptotic properties of our GIC are established under the high-dimensional setting, where n goes to infinity and p/n goes to a constant $c > 0$. Our asymptotic results show that GIC is better than AIC in excluding noise eigenvalues, and is more sensitive than BIC in detecting signal eigenvalues. Numerical examples will be presented.