International Symposium on Recent Advances in Theories and Methodologies for Large Complex Data

December 7-9, 2023

Venue:

Conference Room 101, Tsukuba International Congress Center
2-20-3 Takezono, Tsukuba, Ibaraki 305-0032, Japan (Hybrid Symposium with Zoom)

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 7 (Thursday)

13:50∼14:00 Opening

14:00∼14:40 Kento Egashira[*,a], Kazuyoshi Yata[b] and Makoto Aoshima[b]
[a](Department of Information Sciences, Tokyo University of Science)
[b](Institute of Mathematics, University of Tsukuba)
Asymptotic properties of kernel k-means for high dimensional data

14:50∼15:30 Yohji Akama
(Mathematical Institute, Tohoku University)
Broken-stick components retention rule for equi-correlated normal population

15:40∼16:20 Takayuki Morimoto
(School of Science, Kwansei Gakuin University)
Forecasting high-dimensional covariance matrices using high-dimensional principal component analysis

16:30∼17:10 Shao-Hsuan Wang

(Graduate Institute of Statistics, National Central University)

A geometric algorithm for contrastive principal component analysis in high dimension

17:20∼18:00 Taiji Suzuki[*,a], Denny Wu[b], Atsushi Nitanda[c] and Kazusato Oko[a]

(Zoom) [a](Department of Mathematical Informatics, The University of Tokyo / RIKEN AIP)

[b](Center for Data Science, New York University)

[c](Department of Artificial Intelligence, Kyushu Institute of Technology / RIKEN AIP)

Feature learning via mean field neural networks and anisotropic features

December 8 (Friday)

9:00∼9:40 Akifumi Okuno

(The Institute of Statistical Mathematics / RIKEN AIP)

Statistical estimation with integral-based loss functions

9:50∼10:30 Masaaki Imaizumi

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

Non-sparse high-dimensional statistics and its applications

10:40∼11:20 Tsutomu T. Takeuchi

(Division of Particle and Astrophysical Science, Nagoya University)

Statistical challenges to dimensionality in astronomical big data

11:30∼12:10 Yuan-Tsung Chang[*,a], Nobuo Shinozaki[b] and William, E. Strawderman[c]

[a](Department of Social Information, Mejiro University)

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

[c](Department of Statistics, Rutgers University)

Predictive density estimation for two ordered normal means under α-divergence loss

12:10∼13:40 Lunch

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

19:00∼21:00 Dinner

December 9 (Saturday)

9:00∼9:40 Shogo Nakakita

(Komaba Institute for Science, The University of Tokyo)

On approximate sampling from non-log-concave non-smooth distributions via a Langevin-type Monte Carlo algorithm

9:50∼10:30 Kou Fujimori[*,a] and Koji Tsukuda[b]

[a](Department of Economics, Shinshu University)

[b](Faculty of Mathematics, Kyushu University)

Two step estimations via the Dantzig selector for ergodic time series models

10:40∼11:20 Junichi Hirukawa[*,a] and Kou Fujimori[b]

[a](Faculty of Science, Niigata University)

[b](Department of Economics, Shinshu University)

Innovation algorithm of fractionally integrated ($I(d)$) process and applications on the estimation of parameters

11:30∼12:10 Kengo Kamatani

(The Institute of Statistical Mathematics)

Scaling limits of Markov chains/processes in Monte Carlo methods

12:20∼13:00 Takahiro Nishiyama[*,a] and Masashi Hyodo[b]

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

[b](Department of Economics, Kanagawa University)

On a general linear hypothesis testing problem for latent factor models in high dimensions

13:00∼13:10 Closing

(∗ Speaker)

Special Invited and Keynote Sessions

December 8 (Friday)

Special Invited Session

13:40∼14:30 **On the efficiency-loss free ordering-robustness of product-PCA**

Speaker: Hung Hung
(Institute of Health Data Analytics and Statistics, National Taiwan University)

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

14:40∼15:30 **Learning ordinality in high-dimensional data**

Speaker: Jeongyoun Ahn
(Department of Industrial and Systems Engineering, KAIST)

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

Keynote Session

15:50∼16:50 **Normal-reference test for high-dimensional covariance matrices**

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

Discussion Leader: Aki Ishii (Department of Information Sciences, Tokyo University of Science)

17:00∼18:00 **Testing high-dimensional general linear hypotheses through spectral shrinkage**

Speaker: Debashis Paul
(Zoom) (Department of Statistics, University of California, Davis / Indian Statistical Institute, Kolkata)

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

Asymptotic properties of kernel k-means for high dimensional data

Kento Egashira[a], **Kazuyoshi Yata**[b], **Makoto Aoshima**[b]

[a]**Department of Information Sciences, Tokyo University of Science**
[b]**Institute of Mathematics, University of Tsukuba**

Cluster analysis can be divided into two types: hierarchical and partitional. Hierarchical clustering groups data into dendrograms based on their cluster similarities determined by a preset linkage function. A dendrogram enables the observation of the process of merging or dividing clusters. For discussions on hierarchical cluster analyses, see the works of Everitt et al. [4] and Hastie et al. [6], among others. Partitional clustering, as its name suggests, divides data into a pre-determined number of clusters. K-means can be given on behalf of partitional clustering. Notably, k-means has been approved as a useful tool for analyzing microarray gene expression data. A characteristic of such data is that the number of variables was considerably larger than the sample size, giving high-dimensional, low-sample-size (HDLSS) scenarios. Substantial work on HDLSS asymptotic clustering has been performed in recent years. For example, Liu et al. [8] proposed a two-way split statistical-significance-of-clustering (SigClust) method for HDLSS data. Ahn et al. [1] proposed hierarchical divisive clustering for high-dimensional asymptotics. Huang et al. [5] modified SigClust using a soft thresholding approach. Kimes et al. [7] proposed a method for sequentially testing the statistical significance of hierarchical clustering by controlling the family-wise error rate in HDLSS settings. Yata and Aoshima [10] presented the consistency properties of sample principal component scores and applied them to clustering in high-dimensional settings. Nakayama et al. [9] investigated HDLSS clustering using kernel principal component analysis. Borysov et al. [2] studied the behaviors of hierarchical clustering under several asymptotic settings from a moderate dimension for HDLSS; however, the theoretical assumptions were considered to be strict for HDLSS data owing to several simultaneous asymptotic settings. Egashira et al. [3] explores practical assumptions to indicate the behavior of hierarchical clustering and obtained theoretical results in multiclass settings. Given this background, asymptotic properties of k-means in the HDLSS settings seems to have not been studied sufficiently.

In this talk, we investigated k-means when both the dimension and sample size approach infinity at first. Then, we explored kernel k-means in the HDLSS context theoretically. Especially, we mensioned kernel k-means with gaussian kernel function and compared performance of it to conventional k-means in the multiclass HDLSS context.

References

[1] Ahn, J., Lee, M.H., Yoon, Y.J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, 22, 443–464.

[2] Borysov, P., Hannig, J., Marron, J.S. (2014). Asymptotics of hierarchical clustering for growing dimension. *Journal of Multivariate Analysis*, 124, 465–479.

[3] Egashira, K., Yata, K., Aoshima, M. (2023). Asymptotic properties of hierarchical clustering in high-dimensional settings. *Journal of Multivariate Analysis*, 199, 105251.

[4] Everitt, B.S., Landau, S., Leese, M. (2001). Cluster Analysis. Arnold, New York.

[5] Huang, H., Liu, Y., Yuan, M., Marron, J.S. (2015). Statistical Significance of Clustering using Soft Thresholding. *Journal of Computational and Graphical Statistics*, 24, 975–993.

[6] Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, New York.

[7] Kimes, P.K., Liu, Y., Neil, H.D., Marron, J.S. (2017) Statistical significance for hierarchical clustering. *Biometrics*, 73, 811–821.

[8] Liu, Y., Hayes, D.N., Nobel, A., Marron, J.S.(2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103, 1281–1293.

[9] Nakayama, Y., Yata, K., Aoshima, M. (2021). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis*, 185, 104779.

[10] Yata, K., Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47, 899–921.

Broken-stick components retention rule for equi-correlated normal population (報告書)

赤間陽二

2023 年 12 月 9 日

株価の低頻度サンプリング時系列や, 大規模心理学的調査などでは, 変数の間の相関が顕著になる. そのような状況で因子モデルの因子の個数を標本相関行列 $\mathbf{C}$ から推定することを考える. 代表的な推定方法としては, 閾値法 (thresholding) の他に,「ランダム」データから因子などをシミュレーションして, $\mathbf{C}$ の対応物と比較する方法がある. 閾値法としては, Guttman-Kaiser rule や adjusted correlation thresholding [6] があり, 後者のシミュレーションによる因子数推定法としては, 伝統的な broken-stick rule や, 頻繁に使用される parallel analysis [3] がある. 閾値法もシミュレーションによる因子推定法も, 実データを用いて振る舞いが研究されてきた. 閾値法の理論的背景は [8], [6], [2] などが考察しているが, broken-stick rule や parallel analysis の理論的背景は十分理解されてない [5].

そこで broken-stick rule の理論研究のために次のような設定を考えた. どの異なる変数の組みも共通の定数 $\rho \in [0, 1)$ を相関係数として持つ p 次元正規母集団 (等相関正規母集団と呼ぶことにする) に着目し, サイズ n のサンプルのサンプル相関行列 $\mathbf{C}$ に対する broken-stick rule の結果の, 比例的漸近枠組み $n, p \to \infty,\ p/n \to c > 0$ での極限を研究することにした. このような $\mathbf{C}$ の絶対最大非対角要素は, さまざまな漸近枠組みで ρ に関する相転移を現象を持つ [7]. 我々は, ρ の 0 または正で, Guttman-Kaiser rule の極限挙動に相転移 [2] を見たように, broken-stick rule の極限挙動に相転移を見ることを目標とした.

本研究では, 比例的漸近枠組みにおいて, 等相関正規母集団の標本相関行列 $\mathbf{C}$ の第二固有値が $(1 - \rho)(1 + \sqrt{c})^2$ に高確率で収束することを証明した. その証明には, 比例的漸近枠組みでの非有界スペクトルを持つ標本分散行列 $\mathbf{S}$ の固有値分布 [4], $\mathbf{S}$ と $\mathbf{C}$ の漸近的関係 [1], および, Weyl の不等式を用いた. この第二固有値に関する命題と, $\mathbf{C}$ の最大固有値の発散 [1] より, broken-stick rule が $\mathbf{C}$ から抽出する因子の個数が, 性質 $\rho > 0$ の指示関数に高確率で収束することが証明できた. この定理により, broken-stick rule は, 母相関係数が正である equi-correlation block の個数を数えていることが説明できた. 対応しそうな現象が, binary multiple sequence alignment のデータ [9] や, Fama-French 100 portfolios のデータ [6] の平均相関係数の低頻度サンプリング時系列に確認できた.

しかし, シンポジウムで高次元ファイナンスの専門家が次のような指摘をした: 高頻度サンプリング時系列分析では, 正規分布より裾が重い分布, 例えば t-分布などを用い, 因子モデルの因子の個数を推定するのに [10] の Shrinkage Principal Orthogonal complEment Thresholding (S-POET) という閾値法が有用である.

さらに, シンポジウムで矢田・青嶋のグループは, 一般的なスパイク固有値モデルの標本分散行列の固有値に関する分布自由な結果や, 比例的漸近枠組み以外の漸近枠組みに関する結果を示唆し, S-POET は noise reduction PCA [11, 12] に関係していると指摘した.

以上から Yata & Aoshima や Fan の結果が, 本研究の理論展開に重要であることを認識した. またシンポ

ジウムでは本研究に有用なさまざまな知見を得た.

参考文献

[1] Y. Akama. Correlation matrix of equi-correlated normal population: fluctuation of the largest eigenvalue, scaling of the bulk eigenvalues, and stock market. *Int. J. Theor. Appl. Finance*, 26:2350006, 2023.

[2] Y. Akama and A. Husnaqilati. A dichotomous behavior of Guttman-Kaiser criterion from equi-correlated normal population. *J. Indones. Math. Soc.*, 28(3):272–303, 2022.

[3] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multiv. Behav. Res.*, 27:509–540, 1992.

[4] T. Cai, X. Han, and G. Pan. Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Ann. Stat.*, 48(3):1255–1280, 2020.

[5] E. Dobriban and A. B. Owen. Deterministic Parallel Analysis: An Improved Method for Selecting Factors and Principal Components. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 81(1):163–183, 11 2018.

[6] J. Fan, J. Guo, and S. Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *J. Am. Stat. Assoc.*, 117(538):852–861, 2022.

[7] J. Fan and T. Jiang. Largest entries of sample correlation matrices from equi-correlated normal populations. *Ann. Probab.*, 47(5):3321–3374, 2019.

[8] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory*, 60(8):5040–5053, 2014.

[9] A. A. Quadeer, R. H. Louie, K. Shekhar, A. K. Chakraborty, I. Hsing, and M. R. McKay. Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus nonstructural protein 3 exposes targets for immunogen design. *J. Virol.*, 88(13):7628–7644, 2014.

[10] Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342 – 1374, 2017.

[11] Kazuyoshi Yata and Makoto Aoshima. Effective pca for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, 105(1):193–215, 2012.

[12] Kazuyoshi Yata and Makoto Aoshima. Pca consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis*, 122:334–354, 2013.

Forecasting High-Dimensional Covariance Matrices Using High-Dimensional Principal Component Analysis

Hideto Shigemoto [*], Takayuki Morimoto [†]

November 17, 2023

Abstract

We modify the recently proposed forecasting model of high-dimensional covariance matrices (HDCM) of asset returns using high-dimensional principal component analysis (PCA). It is well-known that when the sample size is smaller than the dimension, eigenvalues estimated by classical PCA have a bias. In particular, a very small number of eigenvalues are extremely large and they are called spiked eigenvalues. High-dimensional PCA gives eigenvalues which correct the biases of the spiked eigenvalues. This situation also happens in the financial field, especially in situations where high-frequency and high-dimensional data are handled. The research aims to estimate the HDCM of asset returns using high-dimensional PCA for the realized covariance matrix using the Nikkei 225 data, it estimates 5- and 10-minute intraday asset-returns intervals. We construct time-series models for eigenvalues which are estimated by each PCA, and forecast HDCM. Our simulation analysis shows that the high-dimensional PCA has better estimation performance than classical PCA for the estimating integrated covariance matrix. In our empirical analysis, we show that we will be able to improve the forecasting performance using the high-dimensional PCA and make a portfolio with smaller variance.

[*] Group Risk Management Department, Nomura Holdings, Inc., 2-2-2 Otemachi, Chiyoda-ku, Tokyo 100-8130, Japan.

[†] Corresponding author: Department of Mathematical Sciences, Kwansei Gakuin University, 1, Gakuen Uegahara, Sanda, Hyogo 669-1330, Japan.
morimot@kwansei.ac.jp

Keywords: covariance forecasting; high-dimensional covariance; principal component analysis; high-frequency data; time series

Summary

In this study, we constructed the HDCM forecasting models using high-dimensional PCA. In particular, the previous studies show that to estimate the latent factors, POET is used. However, it is known that when the dimension is greater than the sample size, the eigenvalues estimated by classical PCA have biases. Therefore, in order to estimate the eigenvalues more accurately, we adopted SPOET which corrects biases of empirical eigenvalues. In addition, we combined eigenvalues and time-series models to forecast eigenvalues and covariance matrix.

In the simulation study, we generated the asset returns based on the estimated HDCM as the integrated covariance matrix and it shows that SPOET is also effective for the price process. Especially, the empirical eigenvalues of SPOET were closer to the true values than those of POET.

In the empirical analysis, we constructed some forecasting models of HDCM using a number of individual stocks traded on Nikkei 225. Almost all our proposed models which use SPOET show better performance than the other models which use POET. In addition, in terms of economic performance, our models can generate a smaller variance than benchmarks in most cases. This study applied SPOET discussed under the i.i.d. setting to the continuous Itô semi-martingale setting for simulation study and empirical analysis. Thus, theoretical results are needed in the future.

This study is partly supported by the Institute of Statistical Mathematics (ISM) cooperative research program (2022-ISMCRP-2024), JSPS KAKENHI Grant Number 21K01433, and Grant-in-Aid for JSPS Fellows Grant Number 22J10285. The views expressed in this study are those of the authors and do not necessarily reflect the official views of Nomura Holdings, Inc. or Kwansei Gakuin University.

A Geometric Algorithm for Contrastive Principal Component Analysis in High Dimension

Shao-Hsuan Wang

Graduate Institute of Statistics, National Central University

Abstract: Principal component analysis (PCA) has been widely used in exploratory data analysis. Contrastive PCA (Abid et al., 2018), a generalized method of PCA, is a new tool used to capture features of a target dataset relative to a background dataset while preserving the maximum amount of information contained in the data. With high dimensional data, contrastive PCA becomes impractical due to its high computational requirement of forming the contrastive covariance matrix and associated eigenvalue decomposition for extracting leading components. In this work, we propose a geometric curvilinear-search method to solve this problem and provide a convergence analysis. Our approach offers significant computational efficiencies. Specifically, it reduces the time complexity from O((n∨m)p^2) to a more manageable O((n∨m)pr), where n, m are the sample sizes of the target data and background data, respectively, p is the data dimension and r is the number of leading components. Additionally, we streamline the space complexity from O(p^2), necessary for storing the contrastive covariance matrix, to a more economical O((n∨m)p), sufficient for storing the data alone. Numerical examples are presented to show the merits of the proposed algorithm.

Feature learning via mean-field neural networks and anisotropic features

Taiji Suzuki[1,2]
Joint work with Denny Wu[3], Atsushi Nitanda[2,4], Kazusato Oko[1,2]
[1]Graduate School of Information Science and Technology, the University of Tokyo,
[2]RIKEN Center for Advanced Intelligence Project,
[3]Center for Data Science, New York University,
[4]Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology.

In this presentation, I have presented the feature learning perspective of mean-field neural networks and compared it with kernel methods. Especially, I presented the generalization ability of mean field neural network trained by mean field Langevin dynamics for learning k-sparse parity functions. I summarize the technical details given in the presentation as follows.

1 Introduction

In this work, we bridge the aforementioned gap by presenting a simple and general framework to establish sample complexity of MFLD in learning binary classification problems. We then apply this framework to the sparse k-parity problem, and obtain improved rate of convergence for the fully time- and space-discretized algorithm. More specifically, our contributions can be summarized as follows.

- We present a general framework to analyze MFLD in the learning of binary classification tasks. Our framework has two main ingredients: (i) an annealing procedure that applies to common classification losses that removes the exponential dependence on regularization parameters in the *logarithmic Sobolev inequality*, and (ii) a novel local Rademacher complexity analysis for the distribution of parameters optimized by MFLD. As a result, we can obtain learning guarantees for the mean-field neural network in discrete-time and finite-width settings.
- We apply our general framework to the k-sparse parity problem, and derived learning guarantees with improved rate of convergence and dimension dependence. Specially, in the $n \asymp d^2$ regime we obtain exponentially converging classification error, whereas in the $n \asymp d$ regime we achieve linear dimension dependence. Note that this improves upon the NTK analysis (which gives a sample complexity of $n = \Omega(d^k)$) in that it "decouples" the degree k from the exponent in the dimension dependence. Our theoretical results are supported by empirical findings.

2 Problem Setting

We consider a classification problem given by the following model:

$$Y = \mathbf{1}_A(Z) - \mathbf{1}_{A^c}(Z) \in \{\pm 1\}$$

where $Z = (Z_1, \ldots, Z_d)$ is the input random variable on $\mathbb{R}^d$ and $\mathbf{1}_A$ is the indicator function corresponding to a measurable set $A \in \mathbb{B}(\mathbb{R}^d)$, i.e., $\mathbf{1}_A(Z) = 1$ if $Z \in A$ and $\mathbf{1}_A(Z) = 0$ if $Z \notin A$. Let P_Z be the distribution of Z. We are given input-output pairs $D_n = (z_i, y_i)_{i=1}^n$ independently identically distributed from this model as training data. Then, we construct a binary classifier that predicts the label for the test input data as accurate as possible. To achieve this, we learn a two-layer neural network model in the mean-field regime via the *mean-field Langevin dynamics*.

One important problem setting for our analysis is the k-sparse parity problem defined as follows.

Example 1 (k-sparse parity problem)**.** *P_Z is the uniform distribution on the grid $\{\pm 1/\sqrt{d}\}^d$ and $A = \{\zeta = (\zeta_1, \ldots, \zeta_d) \in \{\pm 1/\sqrt{d}\}^d \mid \zeta_1 \cdots \zeta_k > 0\}$*[1].

Mean-field two-layer network. Given input z, let $h_x(z)$ be one neuron in a two-layer neural network with parameter $x = (x_1, x_2, x_3) \in \mathbb{R}^{d+1+1}$ defined as

$$h_x(z) = \bar{R}[\tanh(z^\top x_1 + x_2) + 2\tanh(x_3)]/3,$$

where $\bar{R} \in \mathbb{R}$ is a hyper-parameter determining the scale of the network. We place an extra tanh activation for the bias term $x_3 \in \mathbb{R}$ because the boundedness of h_x is required in the convergence analysis. Let $\mathcal{P}$ be the set of probability measures on $(\mathbb{R}^{\bar{d}}, \mathcal{B}(\mathbb{R}^{\bar{d}}))$ where $\bar{d} = d + 2$ and $\mathcal{B}(\mathbb{R}^{\bar{d}})$ is the Borel σ-algebra on $\mathbb{R}^{\bar{d}}$ and $\mathcal{P}_p$ be the subset of $\mathcal{P}$ such that its p-th moment is bounded: $\mathbb{E}_\mu[\|X\|^p] < \infty$ ($\mu \in \mathcal{P}$). The mean-field neural network is defined as an integral over neurons h_x,

$$f_\mu(\cdot) = \int h_x(\cdot)\mu(\mathrm{d}x),$$

for $\mu \in \mathcal{P}$. To evaluate the performance of f_μ, we define the empirical risk and the population risk as

$$L(\mu) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(y_i f_\mu(z_i)), \quad \bar{L}(\mu) := \mathbb{E}[\ell(Y f_\mu(Z))],$$

respectively, where $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a convex loss function. In particular, we consider the logistic loss $\ell(f, y) = \log(1 + \exp(-yf))$ for $y \in \{\pm 1\}$ and $f \in \mathbb{R}$. To avoid overfitting, we consider a regularized empirical risk $F(\mu) := L(\mu) + \lambda \mathbb{E}_{X\sim\mu}[\lambda_1 \|X\|^2]$, where $\lambda, \lambda_1 \geq 0$ are regularization parameters. One advantage of this mean-field definition is that f_μ is a linear with respect to μ, and hence the functional $L(\mu)$ becomes a convex functional. Let $\hat{\mu}$ be the minimizer of $F(\mu)$. We put the following assumption.

Assumption 1. *There exists $c_0 > 0$ and $R > 0$ such that the following conditions are satisfied:*

- *For some $\bar{R}$, there exists $\mu^* \in \mathcal{P}$ such that $\mathrm{KL}(\nu, \mu^*) \leq R$ and $L(\mu^*) \leq \ell(0) - c_0$.*
- *For any $\lambda < c_0/R$, the regularized expected risk minimizer $\mu_{[\lambda]} := \operatorname{argmin} \bar{L}(\mu) + \lambda \mathrm{KL}(\nu, \mu)$ satisfies $Y f_{\mu_{[\lambda]}}(X) \geq c_0$ almost surely.*

Type I: Perfect Classification with Exponentially Decaying Error Under the margin assumption of f_{μ^*} (Assumption 1), we have that $f_{\hat{\mu}}$ also yields a Bayes optimal classifier. More precisely, we have the following theorem.

Theorem 1. *Suppose Assumption 1 holds. Let $M_0 = (\epsilon^* + 2(\bar{R} + 1))/\lambda$. Moreover, suppose that $\lambda < c_0/R$ and*

$$Q := c_0^2 - \frac{4\bar{R}^2}{n\lambda^2}\left[\lambda\left(4\bar{R} + \frac{\lambda}{32\bar{R}^2 n}\right) + 8\bar{R}^2(4 + \log\log_2(8n^2 M_0 \bar{R})) + n\lambda\epsilon^*\right] > 0,$$

then $f_{\hat{\mu}}$ yields perfect classification, i.e., $P\left(Y f_{\hat{\mu}}(Z) > 0\right) = 1$, with probability $1 - \exp(-\frac{n\lambda^2}{32\bar{R}^4}Q)$.

Type II: Polynomial Order Classification Error We can also obtain a result that has a milder dependency on λ and hence a better sample complexity.

Theorem 2. *Suppose Assumption 1 holds. Let $\lambda < c_0/R$ and $M_0 = (\epsilon^* + 2(\bar{R} + 1))/\lambda$. Then, with probability $1 - \exp(-t)$, the classification error of $f_{\hat{\mu}}$ is bounded as*

$$P(Y f_{\hat{\mu}}(Z) \leq 0) \leq 2\psi(c_0)\left[\frac{8\bar{R}^2}{n\lambda}\left(4 + t + \log\log_2(8n^2 M_0 \bar{R})\right) + \frac{1}{n}\left(4\bar{R} + \frac{\lambda}{32\bar{R}^2 n}\right) + \epsilon^*\right].$$

We notice that the right hand side scales with $O(1/(n\lambda))$, which is better than $O(1/(n\lambda^2))$ in Theorem 1; this implies that a sample size linear in the dimensionality is sufficient to achieve small classification error. The reason for such improvement in the λ-dependence is that the stronger L^∞-norm convergence is not used in the proof; instead, only the convergence of the loss is utilized. On the other hand, this analysis does not guarantee a perfect classification.

[1] We present the axis-aligned setting for conciseness, but the same result holds under orthogonal transforms.

Statistical estimation with integral-based loss functions

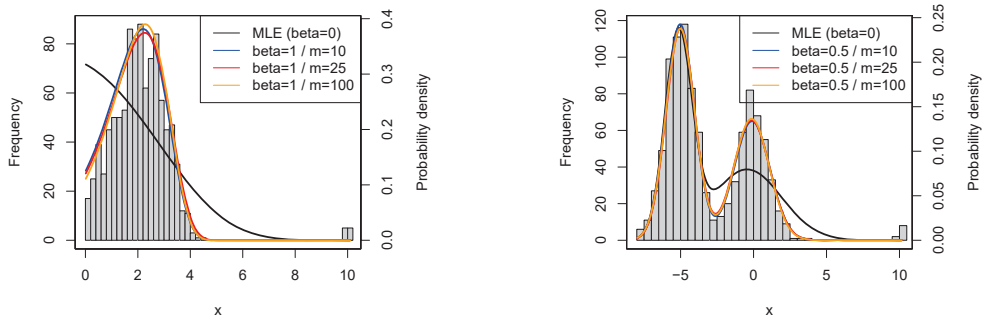
Akifumi Okuno[1,2]

[1]Inst. Stat. Math., [2]RIKEN AIP

This study considers statistical estimation problems using integral-based loss functions, which encompass the following two specific applications.

(1) **Robust estimation using density-power divergence** discussed in Okuno (2023a). Therein, they minimize the following robust loss function equipped with a parametric density p_θ arbitrarily specified by users (e.g., Gaussian mixture model, Gompertz model):

$$-\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} p_\theta(x_i)^\beta + \frac{1}{1+\beta}\int p_\theta(z)^{1+\beta}\mathrm{d}z.$$

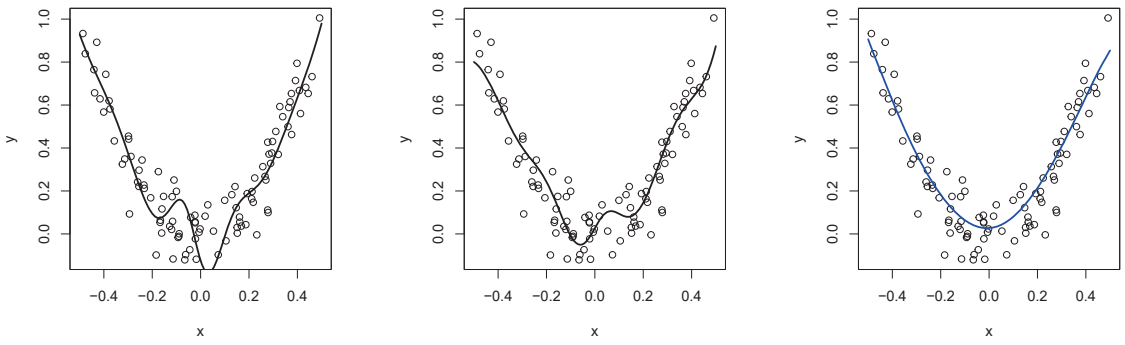


(a) Gompertz density (b) Gaussian mixture

Figure 1: Computational intractable models (Gompertz and Gaussian mixture) are estimated by the proposed approach.

(2) **Higher-order variation regularization** discussed in Okuno (2023b). Therein, they minimize the following penalized loss function equipped with a function f_θ arbitrarily specified by users (e.g., neural network, generalized additive model):

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f_\theta(x_i)\}^2 + \sum_{k=0}^{K}\eta_k\int\left|\frac{\partial^k}{\partial z^k}f_\theta(z)\right|^q\mathrm{d}z.$$



(a) No regularization (b) Ridge regularization (c) Proposal

Figure 2: Single-layer perceptron (with $L = 200$ hidden units) trained with the proposed approach.

Due to the difficulty in evaluating the integral terms, much of the existing research has concentrated on (i) straightforward models (e.g., normal density estimation, spline estimation) where the explicit form of these integral terms can be obtained, or (ii) numerical integration, which is computationally intensive. Notably, statisticians have long focused on optimization with full-batch methods (e.g., Newton-Raphson method), and the obsession with full-batch methods has made computations challenging.

However, if we rewind the long history of research all the way back to the beginning, such integral-based loss functions have been known to be simply minimized by stochastic gradient descent (Robbins and Monro, 1951). To bridge the substantial, unseen gap between the practical applications in statistics and the profound advancements in stochastic optimization theory, this talk intentionally sheds light on the potential utility of the stochastic optimization techniques.

References

Okuno, A. (2023a). A stochastic optimization approach to minimize robust density power-based divergences for general parametric density models. *arXiv preprint arXiv:2307.05251*.

Okuno, A. (2023b). A stochastic optimization approach to train non-linear neural networks with a higher-order variation regularization. *arXiv preprint arXiv:2308.02293*.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

NON-SPARSE HIGH-DIMENSIONAL STATISTICS AND ITS APPLICATIONS

MASAAKI IMAIZUMI[1,2]

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

Abstract. In this talk, we present several results in non-sparse high-dimensional statistics. Specifically, the generalization and Bayesian estimation of high-dimensional linear regression models, statistical inference for high-dimensional generalized linear models, and the regret analysis on contextual bandit problems applying high-dimensional linear models. The analysis in these studies uses the theory of benign overfitting using spectrum, the risk analysis using the convex Gaussian minimax theorem, and the statistical inference using approximate message propagation methods.

1. Outline

1.1. **Linear Regression.** We consider a linear regression problem 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 $\theta_0 \in \mathbb{R}^p$:

$$Y_i = \langle X_i, \theta_0 \rangle + \xi_i,\ i = 1, ..., n,$$

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

The goal of this problem is to estimate the parameter θ_0 from the observations. Here, we consider the high-dimensional setting, specifically, we consider $p \gg n$ or where $p = \infty$ regardless of n. Also, we do not impose the assumption of sparsity on the true parameter θ_0. In this setting, the notion of benign overfitting is actively studied.

We investigate whether benign overfitting-like phenomena occur in situations in which we relax key assumptions in this foundational model. [NI22] considers the case where the covariates are dependent in the sample direction and shows that similar benign overfitting occurs depending on the strength of the dependence. [WI23] develops an informative prior distribution that allows for a Bayesian estimator and its distribution approximation that is valid even in non-sparse high dimensions. [TI23] considers that the noise ξ_i is dependent on the covariance and gives conditions for the estimation error convergence, by utilizing the Gaussian minimax inequality.

1.2. **Generalized Linear Regression.** We next consider the generalized linear model (GLM): for a pair (X, Y) of p-dimensional random features X and random responses Y, we consider the following model

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

where $g : \mathbb{R} \to \mathbb{R}$ is an inverse link function that monotonically increases, and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ 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 $Y_i \in \mathcal{Y}$ that follow the GLM (1), where $\mathcal{Y}$ is a response space, such as $\mathbb{R}, \mathbb{R}_+, \{0, 1\}, \{0, 1, 2, \ldots\}$, and so on.

We consider the proportional high-dimensional regime: we are particularly interested in the proportional limit of the coefficient dimension p and sample size n:

$$n, p \to \infty \text{ and } p/n \to \exists \kappa \in (0, \infty).$$

In this regime, for the logistic regression as the special case, statistical inference on β has been actively studied without the sparsity of β.

[SUI23] develops a methodology for statistical inference for a broad class of GLMs, by deriving the asymptotic normality of an estimator. This method is based on the analysis of a state evolution equation by a vector approximate massage passiong (VAMP) and its application to statistical models.

1.3. **Bandit with Linear Context.** We study a bandit problem with K arms associated with a linear model for its rewards with p-dimensional context vectors. For each round $t \in [T] := \{1, 2, ..., T\}$ and arm $i \in [K]$, we define a context $X_t^{(i)}$ which is a p-dimensional zero-mean sub-Gaussian vector, which is independent among rounds t. An agent chooses an arm $I(t) \in [K]$ based on $X_t^{(i)}$ of all the arms $k \in [K]$, and then observes a reward that follows a linear model as shown in

$$Y^{(I(t))} = \langle X^{(I(t))}, \theta^{I(t)} \rangle + \xi(t).$$

The unknown true parameters $\theta^{(i)}$ for each arm $i \in [K]$ lie in a parameter space $\mathbb{R}^p$, and the independent sub-Gaussian noise $\xi(t)$ with zero mean and variance $\sigma^2 > 0$. We define $i^*(t) := \mathrm{argmax}_{i \in [K]} \langle X_t^{(i)}, \theta^{(i)} \rangle$ as the (ex ante) optimal arm at round t.

Our goal is to design an algorithm that maximizes the total reward, which is equivalent to minimizing the following expected regret. [KI23] considers the high-dimensional setting $p \gg n$ or $p = \infty$ without the sparsity, then derive a novel explore-then-commit strategy to achieve minimize the expected regret.

References

[KI23] Junpei Komiyama and Masaaki Imaizumi. High-dimensional contextual bandit problem without sparsity. *Advances in Neural Information and Processing Systems*, 2023.

[NI22] Shogo Nakakita and Masaaki Imaizumi. Benign overfitting in time series linear model with over-parameterization. *arXiv preprint arXiv:2204.08369*, 2022.

[SUI23] Kazuma Sawaya, Yoshimasa Uematsu, and Masaaki Imaizumi. Feasible adjustments of statistical inference in high-dimensional generalized linear models. 2023.

[TI23] Toshiki Tsuda and Masaaki Imaizumi. Benign overfitting of non-sparse high-dimensional linear regression with correlated noise. *arXiv preprint arXiv:2304.04037*, 2023.

[WI23] Tomoya Wakayama and Masaaki Imaizumi. Bayesian analysis for over-parameterized linear model without sparsity. *arXiv preprint arXiv:2305.15754*, 2023.

Statistical Challenges to Dimensionality in Astronomical Big Data

Tsutomu T. TAKEUCHI[1,2], Kazuyoshi YATA[3], Makoto AOSHIM[3], Kohji YOSHIKAWA[3], Kento EGASHIRA[4], Aki ISHII[4], Suchetha COORAY[5], Kouichiro NAKANISHI[5], Kotaro KOHNO[6], Ryusei R. KANO[1], Yoh-Ichi MOTOTAKE[7], Taisei YAMAGATA[1], Aina May SO[1,8], Hai-Xia MA[1], Shunya UCHIDA[1], Wen E. SHI[1], Sena A. MATSUI[1], and Kai T. KONO[1]

1. Nagoya University, Japan, 2. Institute of Statistical Mathematics, Japan, 3. University of Tsukuba, Japan, 4. Tokyo University of Science, Japan, 5. National Astronomical Observatory of Japan, 6. The University of Tokyo, Japan, 7. Hitotsubashi University, Japan, 8. Gakushuin University, Japan

1. Introduction: Galaxy Formation and Evolution with Big Data

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. It is a high time to revolutionize the methodology for galaxy evolution studies.

2. Application of High-Dimensional Statistical Analysis to Astrophysical data

In astronomy, if we denote the dimension of data as d and the number of samples as n, we often meet a case with $n << d$. Traditionally, such a situation is regarded as ill-posed, and there was no choice but to throw away most of the information in data dimension to let $d < n$. The data with $n << d$ 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. We first introduce the high-dimensional statistical analysis to the astronomical community.

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 R-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 data with small sample size and large dimension. We are also trying to develop a method to analyze absorption line systems in the spectra of

distant radio quasars.

3. 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.

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. 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.

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.

Predictive Density Estimation for Two Ordered Normal Means Under α-Divergence Loss

Yuan-Tsung Chang (Mejiro University) Nobuo Shinozaki (Keio University)
William, E. Strawderman (Rutgers University)

When the underlying loss metric is α-divergence, $D(\alpha)$, loss introduced by siszàr (1967), we consider stochastic and Pitman closeness domination in predictive density estimation problems when there are restrictions given on two means. The underlying distributions considered are normal location-scale models, including the distribution of the observables, the distribution of the variable whose density is to be predicted, and the estimated predictive density which will be taken to be of the plug-in type. The scales may be known or unknown. The main contents are as follows:

1. First, we introduce a general expression which derived by Chang and Strawderman (2014) for the α-divergence loss as following:

   If the true density function of Y is $N(\mu, \sigma^2)$ and the estimated predictive density of Y, is $N(\hat{\mu}, \hat{\sigma}^2)$ then
   a) for $-1 < \alpha < 1$,

$$D_\alpha(N(\tilde{y}|\hat{\mu}, \hat{\sigma}^2), N(\tilde{y}|\mu, \sigma^2)) = \frac{4}{1-\alpha^2}\left(1 - d(\sigma^2, \hat{\sigma}^2)e^{-A(\sigma^2,\hat{\sigma}^2)\frac{(\hat{\mu}-\mu)^2}{2}}\right),$$

   where

$$d(\sigma^2, \hat{\sigma}^2) = \frac{\sigma^{(\alpha-1)/2}\tau}{\hat{\sigma}^{(\alpha+1)/2}}, \quad A(\sigma^2, \hat{\sigma}^2) = \left(\frac{1-\alpha}{2\sigma^2}\right)\left(1 - \frac{(1-\alpha)\tau^2}{2\sigma^2}\right) > 0, \quad \frac{1}{\tau^2} = \left(\frac{1+\alpha}{2\hat{\sigma}^2} + \frac{1-\alpha}{2\sigma^2}\right).$$

   Further, $d(\sigma^2, \hat{\sigma}^2) < 1$ and $A(\sigma^2, \hat{\sigma}^2) > 0$.

   b) (Reverse KL)

$$D_{+1}(N(\tilde{y}|\hat{\mu}, \hat{\sigma}^2), N(\tilde{y}|\mu, \sigma^2)) = \frac{1}{2}\left[\left(\frac{\hat{\sigma}^2}{\sigma^2} - \log\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) + \frac{(\hat{\mu}-\mu)^2}{\sigma^2}\right].$$

   c) (KL)

$$D_{-1}(N(\tilde{y}|\hat{\mu}, \hat{\sigma}^2), N(\tilde{y}|\mu, \sigma^2)) = \frac{1}{2}\left[\left(\frac{\sigma^2}{\hat{\sigma}^2} - \log\frac{\sigma^2}{\hat{\sigma}^2} - 1\right) + \frac{(\hat{\mu}-\mu)^2}{\hat{\sigma}^2}\right].$$

   Also note that in each case, the $\{D(\alpha)\}$ loss is a concave monotone function of squared error loss $|\hat{\mu} - \mu|^2$ and is also a function of the variances. In this set-up and show that it is a concave monotone function of quadratic loss, and also of the variances (predicand, and plug-in).

2. Next, we demonstrate $D(\alpha)$ stochastic domination and Pitman closeness of certain plug-in predictive densities over others for the entire class of metrics simultaneously when "usual" stochastic domination and Pitman closeness holds in the related problem of estimating two ordered means with respect to quadratic loss(Oono, Shinozaki (2005), Chang, Oono and Shinozaki(2012), Chang, Fukuda and Shinozaki(2017)).

3. We also discuss improving the generalized Bayesian predictive densities suggested by Corcuera and Giummole (1999) under $D(\alpha)$ loss.

   Based on the data $X_{ij} \sim N(\mu_i, \sigma_i^2), i = 1, 2,\ j = 1, \cdots, n_i$, we predict the density $\tilde{Y} \sim N(\mu_i, \sigma_i^2), i = 1, 2$. We denote its density function by $p(\tilde{y}; \mu_i, \sigma_i)$, where μ_i and σ_i^2 are unknown.

   When $-1 \leq \alpha < 1$, Corcuera and Giummole (1999) have established that the best invariant predictive density of $p(\tilde{y}; \mu_i, \sigma_i)$ based solely on $x_{i1}, \cdots x_{in_i}$ is

$$\hat{p}_\alpha(\tilde{y}; \bar{x}_i, \tilde{\sigma}_i) \propto \left[1 + \frac{1-\alpha}{2n_i + 1 - \alpha}\left(\frac{y - \bar{x}_i}{\tilde{\sigma}_i}\right)^2\right]^{-(2n_i - 1 - \alpha)/2(1-\alpha)},$$

   where $\bar{x}_i$ is the sample mean and $\tilde{\sigma}_i^2 = ((n_i - 1)/n_i)s_i^2$ is the sample variance. Corcuera and Giummole (1999) have also shown that $\hat{p}_\alpha(\tilde{y}; \bar{x}_i, \tilde{\sigma}_i)$ is the generalized Bayesian predictive density for the prior density $f(\mu_i, \sigma_i) \propto 1/\sigma_i, 0 < \sigma_i < \infty$. It is to be noted that $\hat{p}_\alpha(\tilde{y}; \bar{x}_i, \tilde{\sigma}_i)$ is not a normal distribution, although the plug-in density $N(\bar{x}_i, s_i^2)$ is the generalized Bayes rule when $\alpha = 1$.

   We consider the following two cases separately where order restrictions on μ_i and/or σ_i^2 are present,

   i) Case when $\mu_1 \leq \mu_2$.
   ii) Case when $\mu_1 \leq \mu_2$ and $\sigma_1^2 \leq \sigma_2^2$.

   Examples of $D(\alpha)$ stochastic (Pitman closeness) domination presented relate to the problem of estimating the predictive density of the variable with the restrictions on two normal means.

keywords: Predictive density, α-divergence, stochastic dominance, ordered normal means, Pitman closeness criterion

References

[1] Aitchison, J., Goodness of prediction fit, Biometrika, 62 (1975) 545-554.

[2] Chang, Y.-T., Oono, Y., Shinozaki, N., Improved estimators for the common mean and ordered means of two normal distributions with ordered variances. *Journal of Statistical Planning and Inference*, 142 (2012) 2619-2628.

[3] Chang, Y.-T., Strawderman, W. E., Stochastic domination in predictive density estimation for ordered normal means under α-divergence loss. *Journal of Multivariate Analysis*, 128 (2014) 1-9.

[4] Chang, Y.-T., Fukuda, K., Shinozaki, N., Estimation of two ordered normal means when a covariance matrix is known. *Statistics*, 5 (2017) 1095-1104.

[5] Chang, Y.-T., Shinozaki, N, Strawderman, W. E., Pitman closeness domination in predictive density estimation for two-ordered normal means under α-divergence loss. *Japanese Journal of Statistics and Data Science*, (2020) 3:1-21.

[6] Csiszàr, I., Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.*, 2 (1967) 299-318.

[7] Fourdrinier, D., Marchand, E., Righi, A., Strawderman, W. E., On Improved predictive density with parametric constraints. *Electronic Journal of Statistics*, 5 (2011) 172-191.

[8] Oono, Y., Shinozaki, N., Estimation of two order restricted normal means with unknown and possibly unequal variances. *Journal of Statistical Planning and Inference*, 131 (2005) (2), 349-363.

On the efficiency-loss free ordering-robustness of product-PCA

Hung Hung

[1]Institute of Health Data Analytics and Statistics, National Taiwan University, Taiwan

Abstract

This article studies the robustness of the eigenvalue ordering, an important issue when estimating the leading eigen-subspace by principal component analysis (PCA). In Yata and Aoshima (2010), cross-data-matrix PCA (CDM-PCA) was proposed and shown to have smaller bias than PCA in estimating eigenvalues. While CDM-PCA has the potential to achieve better estimation of the leading eigen-subspace than the usual PCA, its robustness is not well recognized. In this article, we first develop a more stable variant of CDM-PCA, which we call product-PCA (PPCA), that provides a more convenient formulation for theoretical investigation. Secondly, we prove that, in the presence of outliers, PPCA is more robust than PCA in maintaining the correct ordering of leading eigenvalues. The robustness gain in PPCA comes from the random data partition, and it does not rely on a data down-weighting scheme as most robust statistical methods do. This enables us to establish the surprising finding that, when there are no outliers, PPCA and PCA share the same asymptotic distribution. That is, the robustness gain of PPCA in estimating the leading eigen-subspace has no efficiency loss in comparison with PCA. Simulation studies and a face data example are presented to show the merits of PPCA. In conclusion, PPCA has a good potential to replace the role of the usual PCA in real applications whether outliers are present or not.

Key words: cross-data-matrix PCA; dimension reduction; efficiency loss; ordering of eigenvalues; random partition; robustness.

Learning Ordinality in High-Dimensional Data

Jeongyoun Ahn
Department of Industrial and Systems Engineering, KAIST

Numerous real-world applications involve naturally ordinal outcomes, such as cancer stages or tumor grades. Despite the recent surge in high-dimensional statistical methodologies, high-dimensional learning with ordinal outcomes has been largely overlooked in the HDLSS literature. In this talk, I will introduce recent projects on ordinality in high-dimensional data. The first three topics concern supervised learning aimed at predicting ordinal labels. All three ordinal methods assume sparsity and equal covariance population structure, leading us to term them 'ordinal sparse high-dimensional LDA'. They operate on the principle that a classification rule primarily reliant on variables that are monotonically associated with the outcome should be preferable. They all result in a low-dimensional discriminant subspace where classes are sequentially aligned. The first FWOC method weights features based on their rank correlations with class labels, integrating these weights into the LDA framework. The second SOBL method combines sparsity and ordinality regularizations in a high-dimensional generalized eigenvalue problem. The third SODA approach applied regularization to optimal scores within the sparse LDA framework. In addition, in scenarios where an ordinal outcome is unobserved, one may search for an ordinal signal in the data. This leads us to develop 'monotone clustering' that is designed to identify subgroups interpretable in an ordinal manner.

Normal-reference test for high-dimensional covariance matrices

Jin-Ting Zhang

National University of Singapore

Abstract

In the past decade, much attention has been paid for testing the equality of high-dimensional covariance matrices. Several test statistics have been proposed for this purpose. Some of them imposed strong assumptions, aiming to yield the asymptotic normality of the associated test statistics. In practice, however, these assumptions are often challenging to verify, resulting in size control issues when the required assumptions are not met. To address this challenge, in this talk, we investigate a normal-reference test which can effectively control the size. 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 under the null hypothesis, assuming normality of the data samples. To accurately approximate the distribution of the chi-square-type mixture, we employ a three-cumulant matched x^2-approximation with the approximation parameters being consistently estimated from the data. Two simulation studies demonstrate that in terms of size control, the proposed normal-reference test performs well across a range of scenarios and it outperforms several 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.

International Symposium on Recent Advances in Theories and Methodologies for Large Complex Data

December 7-9, 2023

Venue: Tsukuba International Congress Center

Speaker: Debashis Paul (University of California, Davis & Indian Statistical Institute, Kolkata)

Title: Testing high-dimensional general linear hypotheses through spectral shrinkage

Abstract: We consider the problem of testing linear hypotheses associated with a high-dimensional multivariate linear regression model under the setting where the dimensionality of the response is comparable to the sample size, and the dimensionality of the predictors is finite. Classical solutions involving likelihood ratio tests for such problems suffer from significant loss of power within this asymptotic framework. We propose regularization schemes that modify the likelihood ratio statistics by applying nonlinear shrinkage to the eigenvalues of the empirical covariance matrix of the regression residuals. We propose two different classes of regularized tests to deal with different types of structural assumptions on the covariance matrix of the noise in the linear regression model: (a) the spectral measure of the noise covariance converges to a nontrivial limit; and (b) the noise covariance has a spiked covariance structure. We show that in each case, the proposed tests significantly improve on the performance of the likelihood ratio test. We also address the problem of finding the optimal regularization parameter within a decision-theoretic framework by adopting a probabilistic formulation of the alternatives. As an application, we consider the problem of detecting possible associations among human behavioral measurements and volumetric measurements for various brain regions.

(This is a joint work with Haoran Li, Alexander Aue and Jie Peng).

ON APPROXIMATE SAMPLING FROM NON-LOG-CONCAVE NON-SMOOTH DISTRIBUTIONS VIA A LANGEVIN-TYPE MONTE CARLO ALGORITHM

SHOGO NAKAKITA

We consider the problem of sampling from a Gibbs distribution $\pi(\mathrm{d}x) \propto \exp(-U(x))\mathrm{d}x$ on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$, where $U : \mathbf{R}^d \to [0, \infty)$ is a non-negative potential function. One of the extensively used types of algorithms for the sampling is the Langevin type motivated by the Langevin dynamics, the solution of the following d-dimensional stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla U\left(X_t\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ X_0 = \xi, \tag{0.1}$$

where $\{B_t\}_{t\geq 0}$ is a d-dimensional Brownian motion and ξ is a d-dimensional random vector with $|\xi| < \infty$ almost surely. Since the 2-Wasserstein or total variation distance between π and the law of X_t is convergent under mild conditions, we expect that the laws of Langevin-type algorithms inspired by X_t should converge to π. However, most of the theoretical guarantees for such algorithms are based on the convexity of U, the twice continuous differentiability of U, or the Lipschitz continuity of the gradient ∇U, which do not hold in some modelling in statistics and machine learning. The main interest of this study is proposal of a Langevin-type algorithm whose convergence can be given under minimal assumptions.

To see what difficulties we need to deal with, we review a typical analysis [6] based on the smoothness of U, that is, the twice continuous differentiability of U and the Lipschitz continuity of ∇U. Firstly, the twice continuous differentiability simplifies discussions or plays significant roles in studies of functional inequalities such as Poincaré inequalities and logarithmic Sobolev inequalities [e.g., 1, 2]. Since the functional inequalities for π are essential in analysis of Langevin algorithms, the assumption that U is of class $\mathcal{C}^2$ frequently appears in previous studies. In the second place, the Lipschitz continuity combined with weak conditions ensures the representation of the likelihood ratio between $\{X_t\}$ and $\{Y_t\}$, which is critical when we bound the Kullback–Leibler divergence. Liptser and Shiryaev [4] exhibit much weaker conditions than Novikov's or Kazamaki's condition for the explicit representation if (0.1) has the unique strong solution. Since the Lipschitz continuity of ∇U is sufficient for the existence and the uniqueness of the strong solution of (0.1), the framework of Liptser and Shiryaev [4] is applicable.

Our approaches to overcome the non-smoothness of U are mollification, a classical approach to dealing with non-smoothness in differential equations, and the 'misuse' of moduli of continuity for possibly discontinuous functions. We consider the convolution $\bar{U}_r := U * \rho_r$ on U with a weak gradient, and some sufficiently smooth non-negative function ρ_r with compact support in a ball of centre $\mathbf{0}$ and radius $r \in (0, 1]$. We can let $\bar{U}_r$ be of class $\mathcal{C}^2$ and obtain bounds for the constant of Poincaré inequalities for $\bar{\pi}^r(\mathrm{d}x) \propto \exp(-\bar{U}_r(x))\mathrm{d}x$, which suffice to show the convergence of the law of the mollified

Komaba Institute for Science, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
E-mail address: nakakita@g.ecc.u-tokyo.ac.jp.
The author was supported by JSPS KAKENHI Grant Number JP21K20318 and JST CREST Grant Numbers JPMJCR21D2 and JPMJCR2115.

dynamics $\{\bar{X}_t^r\}$ defined by the SDE

$$\mathrm{d}\bar{X}_t^r = -\nabla\bar{U}_r\left(\bar{X}_t^r\right)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,\ \bar{X}_0^r = \xi$$

to the corresponding Gibbs distribution $\bar{\pi}^r$ in 2-Wasserstein distance owing to Bakry et al. [1], Liu [5], and Lehec [3]. Since the convolution $\nabla\bar{U}_r$ is Lipschitz continuous if the modulus of continuity of a representative ∇U is finite (the convergence to zero is unnecessary), a concise representation of the likelihood ratios between the mollified dynamics $\{\bar{X}_t^r\}$ and $\{Y_t\}$ is available, and we can evaluate the Kullback–Leibler divergence under weak assumptions.

As our analysis relies on mollification, the bias–variance decomposition in estimation of $\nabla\bar{U}_r$ rather than ∇U is crucial. This decomposition enables us to propose new algorithms for U without continuous differentiability. Concretely speaking, we propose a new algorithm named the spherically smoothed Langevin Monte Carlo (SS-LMC) algorithm, whose errors can be arbitrarily small under the dissipativity of U and the boundedness of the modulus of continuity of weak gradients. In addition, we argue zeroth-order versions of these algorithms which are naturally obtained via integration by parts.

References

[1] Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. (2008). A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66.

[2] Cattiaux, P., Guillin, A., and Wu, L.-M. (2010). A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148:285–304.

[3] Lehec, J. (2021). The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *To appear in the Annals of Applied Probability*.

[4] Liptser, R. S. and Shiryaev, A. N. (2001). *Statistics of Random Processes: I. General theory*. Springer, 2nd edition.

[5] Liu, Y. (2020). The poincaré inequality and quadratic transportation-variance inequalities. *Electronic Journal of Probability*, 25(1):1–16.

[6] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1674–1703.

Two step estimations via the Dantzig selector for ergodic time series models

Kou Fujimori[1] and Koji Tsukuda[2]

[1]Faculty of Economics and Law, Shinshu University.
[2]Faculty of Mathematics, Kyushu University.

Let us consider the following time series model.

$$X_t = S(\alpha^\top \phi(\boldsymbol{X}_{t-1}), \beta^\top Z_{t-1}) + u_t, \quad \mathrm{E}[u_t^2|\mathcal{F}_{t-1}] = \sigma^2(\boldsymbol{X}_{t-1}; h),$$

where $\boldsymbol{X}_{t-1} = (X_{t-1}, \ldots, X_{t-d})$, without loss of generality. Let $\theta_0 = (\alpha_0^\top, \beta_0^\top)^\top$ be the true value of $\theta = (\alpha^\top, \beta^\top)^\top$, $\Theta = \Theta_\alpha \times \Theta_\beta \subset \mathbb{R}^{p+q}$ a parameter space for θ and H a metric space equipped with a metric d_H. Put $T_{10} := \{j : \alpha_{0j} \neq 0\}$, $T_{20} := \{j + p_1 : \beta_{0j} \neq 0\}$ and $T_0 = T_{10} \cup T_{20}$. We observe $(X_1, Z_1), \ldots, (X_n, Z_n)$. Our aim is to estimate $\theta = (\alpha^\top, \beta^\top)^\top$. Hereafter, we fix an initial value $(X_0, \ldots, X_{1-d}) = (x_0, \ldots, x_{1-d})$ and put $p = p_1 + p_2$, $s = s_1 + s_2$, where s_1 and s_2 are the numbers of elements in T_{10} and T_{20}, respectively.

We first construct the estimator $\hat{\theta}_n^{(1)}$ for θ by the following Dantzig selector type estimator:

$$\hat{\theta}_n^{(1)} := \arg\min_{\theta \in \mathcal{C}_n} \|\theta\|_1, \quad \mathcal{C}_n := \{\theta \in \mathbb{R}^p : \|\psi_n^{(1)}(\theta)\|_\infty \leq \lambda_n\},$$

where

$$\psi_n^{(1)}(\theta) = \frac{1}{n}\sum_{t=1}^n \frac{\partial}{\partial \theta} S(\alpha^\top \phi(\boldsymbol{X}_{t-1}), \beta^\top Z_{t-1})\{X_t - S(\alpha^\top \phi(\boldsymbol{X}_{t-1}), \beta^\top Z_{t-1})\}$$

and λ_n is a tuning parameter. Moreover, we define the following estimator $\hat{T}_n$ for T_0:

$$\hat{T}_n := \{j : |\hat{\theta}_{nj}| > \tau_n\},$$

where τ_n is a threshold. For the second step, we construct a consistent estimator for h, by using $\hat{\theta}_n^{(1)}$. Finally, using $\hat{T}_n$ and $\hat{h}_n$, we consider the estimator $\hat{\theta}_n^{(2)}$ for θ as a solution to the following equation:

$$\Psi_{n\hat{T}_n}(\theta_{\hat{T}_n}, \hat{h}_n) = 0, \quad \tilde{\theta}_{n\hat{T}_n^c} = 0,$$

where

$$\begin{aligned}\Psi_{nT}(\theta_T, \hat{h}_n) &= \frac{1}{n}\sum_{t=1}^{n} \frac{\frac{\partial}{\partial \theta_T} S(\alpha_{T_1}^\top \phi(\boldsymbol{X}_{t-1})_{T_1}, \beta_{T_2}^\top Z_{t-1T_2})}{\sigma^2(\boldsymbol{X}_{t-1}; \hat{h}_n)} \\ &\quad \cdot\{X_t - S(\alpha_{T_1}^\top \phi(\boldsymbol{X}_{t-1})_{T_1}, \beta^\top Z_{t-1T_2})\}\end{aligned}$$

for every $T = T_1 \cup T_2$.

In this talk, we establish the rate of convergence of $\hat{\theta}_n^{(1)}$, and the asymptotic normality of $\tilde{\theta}_{n\hat{T}_n}$ under some regularity conditions under high-dimensional and sparse settings. Moreover, we discuss the integer-valued autoregressive models as an example.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 21K13271(K.F.), 21K13836 (K.T.).

Innovation algorithm of fractionally integrated ($I(d)$) process and applications on the estimation of parameters

Junichi Hirukawa and Kou Fujimori
Niigata University and Shinshu University

ABSTRACT

The long memory phenomena frequently occur in the empirical studies of various fields. The fractionally integrated process is the one of the suitable candidate which appropriately represents the long memory property. There are two recursive algorithms for determining the one-step predictors of time series, that is, the Durbin-Levinson algorithm and the innovation algorithm. The Durbin-Levinson algorithm for the fractionally integrated process is well-known and widely used, which naturally derives the Cholesky factorization of the inverse matrix of the covariance matrix of the process. In this paper, we derive the innovation algorithm for the fractionally integrated process. The result is also applied to the derivations of the Cholesky factorization of the covariance matrix and the Gaussian likelihood of the process in the explicit forms. Moreover, the asymptotic theory of Gaussian maximum likelihood estimator (GMLE) is derived in terms of the innovation algorithm.

1 Introduction

An ARMA (p,q) process $\{x_t\}$ is often called a short memory process since the covariance between x_t and x_{t+j} decreases rapidly as $j \to \infty$. However, the long memory phenomena frequently occur in the empirical studies of various fields (see e.g., Hurst (1951)). In this paper, we consider one of the long memory process so-called the fractionally integrated ($I(d)$) process defined by

$$(1-L)^d z_t = \varepsilon_t, \quad (t = 1, \ldots, n), \tag{1}$$

where $d \in (-1/2, 1/2)$, $(d \neq 0)$, L is the lag operator and $\{\varepsilon_t\} \overset{i.i.d.}{\sim} \left(0, \sigma^2\right)$. Using the expansion of the lag operator

$$\Delta(L) = (1-L)^d = \frac{1}{\Gamma(-d)} \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(j+1)} L^j = \sum_{j=0}^{\infty} \varphi_j L^j, \tag{2}$$

this can be rewritten as

$$\varepsilon_t = \sum_{j=0}^{\infty} \varphi_j L^j z_t = \sum_{j=0}^{\infty} \varphi_j z_{t-j}. \tag{3}$$

Then, $\{z_t\}$ is a stationary long memory process generated by

$$\begin{aligned} z_t &= (1-L)^{-d} \varepsilon_t \\ &= \frac{1}{\Gamma(d)} \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)} \varepsilon_{t-j} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \end{aligned}$$

where the coefficients satisfy $\psi_j = O\left(j^{d-1}\right)$, so that the degree of decreasing is quite slow as $j \to \infty$.

2 Main result

In this section, we provide the main results.

2.1 The Gaussian MLE for $I(d)$ process

In this section, we impose the Gaussian assumption on $I(d)$ process. Then, we have the Gaussian log-likelihood of $I(d)$ process for $\boldsymbol{\theta} = \left(d, \sigma^2\right)'$

$$l(\boldsymbol{\theta}) = l\left(d, \sigma^2\right) = -\frac{n}{2}\log\{2\pi\} - \frac{1}{2}\sum_{j=1}^{n}\log v_{j-1}(\boldsymbol{\theta}) - \frac{1}{2}\sum_{j=1}^{n}\frac{u_{j-1}(d)^2}{v_{j-1}(\boldsymbol{\theta})}. \tag{4}$$

Now, we have the following main results. First, we describe the consistency of GMLE.

Theorem 1. *Let $\{z_t\}$ is the Gaussian $I(d)$ process defined in (1) with $d \in (-1/2, 1/2)$, $(d \neq 0)$. And let $\widehat{\boldsymbol{\theta}} = \left(\widehat{d}, \widehat{\sigma}^2\right)'$ is the Gaussian MLE (GMLE) of $\boldsymbol{\theta} = \left(d, \sigma^2\right)'$ which maximizes the Gaussian log-likelihood (4). Then, the GMLE $\widehat{\boldsymbol{\theta}}$ has consistency, that is,*

$$\widehat{\boldsymbol{\theta}} \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0,$$

where $\boldsymbol{\theta}_0 = \left(d_0, \sigma_0^2\right)'$ is the true value of $\boldsymbol{\theta}$.

Next, we have the following asymptotic normality.

Theorem 2. *Let $\{z_t\}$ is the Gaussian $I(d)$ process defined in (1) with $d \in (-1/2, 1/2)$, $(d \neq 0)$. And let $\widehat{\boldsymbol{\theta}} = \left(\widehat{d}, \widehat{\sigma}^2\right)'$ is the Gaussian MLE (GMLE) of $\boldsymbol{\theta} = \left(d, \sigma^2\right)'$ which maximizes the Gaussian log-likelihood (4). Then, the GMLE $\widehat{\boldsymbol{\theta}}$ satisfies the asymptotic normality, that is,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{A}),$$

where $\boldsymbol{\theta}_0 = \left(d_0, \sigma_0^2\right)'$ is the true value of $\boldsymbol{\theta}$.

References

Akaike, H. (1969). Power Spectrum Estimation through Autoregressive Model Fitting. *Ann. Inst. Stat. Math.* **21**, 407–419.

Berk, K. N. (1974). Consistent Autoregressive Spectral Estimates. *Ann. Statist.* **2**, 489–502.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer.

Hosking, J. R. M. (1981). Fractional Differencing. *Biometrika* **1**, 165–176.

Hurst, H. E. (1951). Long-Term Storage Capacity of Reservoirs. *Trans. Amer. Soc. Civil Eng.* **116**, 770–799.

Li, W. K. and McLeod, A. I. (1986). Fractional Time Series Modelling. *Biometrika* **73**, 217-221.

Slater, L J. (1966). *Generalized Hypergeometric Functions*. New York: Cambridge University Press.

Yajima, Y. (1985). On Estimation of Long - Memory Time Series Models. *Australian Journal of Statistics* **27**, 303–320.

Scaling Limits of Markov Chains/Processes in Monte Carlo Methods

Kengo Kamatani (Institute of Statistical Mathematics, JST CREST)

In this presentation, we will explore the recent results of scaling limit of piecewise deterministic Markov processes for anisotropic targets. Suppose we wish to sample from

$$\Pi(\mathrm{d}x) = \exp(-\mathcal{H}(x))\mathrm{d}x$$

where $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function. For the Bayesian context, this probability distribution is the posterior distribution of interest. If we have an i.i.d. sample from Π, we can approximate Π-integral of any function $f(x)$ by the law of large numbers. In most of the cases, direct i.i.d. sampling is impossible or computationally very expensive. For these cases, the Markov chain Monte Carlo method is useful which originated with the classic paper by Metropolis et al. (1953) almost 70 years ago. The Markov chain Monte Carlo method is designed to construct an ergodic Markov kernel P which is Π-invariant. If a Markov chain $X_1, X_2, \dots$ is generaetd from the Markov kernel P then the law of large numbers is satisfied. The Markov chain Monte Carlo is now a gold standard for Bayesian inference.

Recently, its continuous process version, the Markov **process** Monte Carlo method is of substantial interest for Monte Carlo analysis. Known Markov process Monte Carlo methods rely on an auxiliary variable trick which uses an auxiliary variable v with a probability density ν on Ξ and considers the joint probability distribution $\mu := \Pi(\mathrm{d}x) \otimes \nu(\mathrm{d}v)$ as an extended target distribution on $\mathcal{Z} = \mathbb{R}^d \times \Xi$. The original target distribution is a marginal distribution of the extended target distribution. Since Brownian motion does not have an absolutely continuous path, we can not simulate processes driven by Brownian motion exactly. For our Monte Carlo analysis, exact sampling is necessary. Therefore, the Markov processes of interest should not have a Brownian part. Known processes consist of a deterministic part and a pure jump part. These processes are known as the **piecewise deterministic Markov processes**.

Here we follow Azaïs et al. (2014) for the expression of the piecewise deterministic Markov processes. The processes are constructed by characteristics $(\phi, \lambda_k, Q_k : l = 1, \dots, K)$. The flow $\phi : \mathcal{Z} \times \mathbb{R} \to \mathcal{Z}$ is continuous, $\phi(\cdot, t)$ is a homeomorphism for each $t \in \mathbb{R}$ and $\phi(\phi(\cdot, s), t) = \phi(\cdot, s + t)$. For each $k = 1, \dots, K$, the jump rate $\lambda_k : \mathcal{Z} \to \mathbb{R}_+$ determines the jump time of pure jump processes, and Q_k is a Markov kernel on $\mathcal{Z}$. Let $\Lambda_k(z, t) = \int_0^t \lambda_k(\phi(z, s))\mathrm{d}s$.

The Markov process is defined by the following way. Suppose $z(0) = (x(0), t(0)) \in \mathcal{Z}$. Let $T_1, \dots, T_K$ be independent processes with $\mathbb{P}(T_k \geqslant t) = \exp(-\Lambda_k(z, t))$. Let $T_* = \min_{k=1,\dots,K} T_k$. If $T_k = T_*$, then Z is generated from $Q_k(\phi(z, T_*), \cdot)$ and set

$$X(t) = \begin{cases} \phi(z(0), t) & \text{for } t < T_* \\ Z & \text{for } t = T_*. \end{cases}$$

After T_*, the process evolves in the same way with starting value Z. There are several choices of characteristics. Two popular piecewise deterministic Markov processes use the same flow ϕ defined by $x'(t) = v(t)$ and $v'(t) = 0$. The **Zig-Zag sampler** proposed by Bierkens et al. (2019)

uses d Markov kernels $Q_1, \ldots, Q_d$ with d jump rates $\lambda_1, \ldots, \lambda_d$. For each $i = 1, \ldots, d$, the Markov kernel is a deterministic kernel Q_i defined by a map $(x, v) \mapsto (x, F_i(v))$ where F_i is an operator that flips the i-th coordinate of x. The jump rate is defined by $\lambda_i((x, v)) = \max\{0, \partial_i \mathcal{H}(x) v_i\}$.

The **bouncy particle sampler** proposed by Peters and de With (2012), Bouchard-Côté et al. (2018) uses two Markov kernels Q_{bounce} and Q_{ref} with corresponding jump rates λ_{bounce} and λ_{ref}. The kernel Q_{bounce} is a deterministic kernel defined by a map $(x, v) \mapsto (x, \kappa(x, v))$:

$$\kappa(x, v) = v - 2\frac{\langle \nabla \mathcal{H}(x), v \rangle}{\|\nabla \mathcal{H}(x)\|^2} \nabla \mathcal{H}(x)$$

and $\lambda_{\text{bounce}}(x, v) = \max\{0, \langle \nabla \mathcal{H}(x), v \rangle\}$. The jump rate λ_{ref} is a positive constant, and Q_{ref} is a μ-invariant Markov kernel. For our analysis, for simplicity, we assume $Q_{\text{ref}}((x, v), \mathrm{d}(y, w)) = \nu(\mathrm{d}w)$.

We have several critical findings. For the Zig-Zag algorithm, its performance is intricately linked to the orientation of the target's anisotropy; specific alignments with the algorithm's operational axes lead to enhanced efficiency, while others can hinder its effectiveness. The BPS algorithm, on the other hand, exhibits a deterministic dynamical behaviour in its limiting form with a better rate of convergence.

This is joint work with Joris Bierkens (TU Delft) and Gareth O. Roberts (Warwick). See our paper on arxiv `https://arxiv.org/abs/2305.00694` for the detail.

References

Christophe Andrieu, Alain Durmus, Nikolas Nüsken, and Julien Roussel. Hypocoercivity of piecewise deterministic markov process-monte carlo. *The Annals of Applied Probability*, 31, 10 2021. ISSN 1050-5164. doi: 10.1214/20-AAP1653.

Romain Azaïs, Jean-Baptiste Bardet, Alexandre Génadot, Nathalie Krell, and Pierre-André Zitt. Piecewise deterministic Markov process—recent results. In *Journées MAS 2012*, volume 44 of *ESAIM Proc.*, pages 276–290. EDP Sci., Les Ulis, 2014. doi: 10.1051/proc/201444017.

Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Ann. Statist.*, 47(3):1288–1320, 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1715.

Joris Bierkens, Kengo Kamatani, and Gareth O. Roberts. High-dimensional scaling limits of piecewise deterministic sampling algorithms. *The Annals of Applied Probability*, 32, 10 2022. ISSN 1050-5164. doi: 10.1214/21-AAP1762.

Alexandre Bouchard-Côté, Sebastian J. Vollmer, and Arnaud Doucet. The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method. *Journal of the American Statistical Association*, 113(522):855–867, 2018. doi: 10.1080/01621459.2017.1294075.

George Deligiannidis, Daniel Paulin, Alexandre Bouchard-Côté, and Arnaud Doucet. Randomized hamiltonian monte carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates. *The Annals of Applied Probability*, 31, 12 2021. ISSN 1050-5164. doi: 10.1214/20-AAP1659.

N. Metropolis, W. A. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Elias AJF Peters and G. de With. Rejection-free monte carlo sampling for general potentials. *Physical Review E*, 85(2):026703, 2012.

On a general linear hypothesis testing problem for latent factor models in high dimensions

Takahiro Nishiyama[a] and Masashi Hyodo[b]

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

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, $i \in \{1, \ldots, n_g\}$, $g \in \{1, \ldots, k\}$. A factor model assumes that for each $g \in \{1, \ldots, k\}$, the observable vector $\mathbf{x}_{gi}$ is decomposable into a latent factor and an idiosyncratic component as follows:

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

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 a d_g-dimensional latent factor vector, and $\boldsymbol{\epsilon}_{gi} = (\epsilon_{gi1}, \ldots, \epsilon_{gip})^\top$ is a 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, $\mathbf{F}_g = (\mathbf{f}_{g1}, \ldots, \mathbf{f}_{gp})^\top$ denotes a loading matrix where for each $j \in \{1, \ldots, p\}$, $\mathbf{f}_{gj} = (f_{gj1}, \ldots, f_{gjd_g})^\top \in \mathbb{R}^{d_g}$ is a non-random vector, and $\boldsymbol{\Psi}_g = \mathrm{diag}(\psi_{g1}, \ldots, \psi_{gp})$ is a 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 $\mathrm{E}(\mathrm{z}_{gi\ell}^4) = \kappa_{z_g} < \infty$, 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 $g \in \{1, \ldots, k\}$, $i \in \{1, \ldots, n_g\}$, $j \in \{1, \ldots, p\}$ and $\ell \in \{1, \ldots, d_g\}$. Structural assumptions of the model (1) imply that

$$\mathrm{E}(\mathbf{x}_{gi}) = \boldsymbol{\mu}_g, \ \mathrm{cov}(\mathbf{x}_{gi}) = \mathbf{F}_g \mathbf{F}_g^\top + \boldsymbol{\Psi}_g := \boldsymbol{\Sigma}_g, \tag{2}$$

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 design a high-dimensinal test procedure for a general linear hypothesis testing (GLHT) problem:

$$\mathcal{H} : \widetilde{\mathbf{G}}\mathbf{M} = \mathbf{O}, \quad \mathcal{A} : \widetilde{\mathbf{G}}\mathbf{M} \neq \mathbf{O}, \tag{3}$$

where $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^\top$ is a $k \times p$ matrix and $\widetilde{\mathbf{G}}$ is a $q \times k$ known coefficient matrix with full row rank $q < k$. By setting $\widetilde{\mathbf{G}}$ to be any $(k-1) \times k$ contrast matrix, i.e., any $(k-1) \times k$ matrix with linearly independent rows and zero row sums, the GLHT problem (3) reduces to the one-way MANOVA problem:

$$\mathcal{H} : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k, \quad \mathcal{A} :\neq \mathcal{H}. \tag{4}$$

Also, various post hoc and contrast tests can be written in the form of (3).

From Zhang et al. (2017) and Zhang et al. (2022), we re-write (3) into the following equivalent form:

$$\mathcal{H} : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}, \quad \mathcal{A} : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}, \tag{5}$$

where $\mathbf{C} = \mathbf{G} \otimes \mathbf{I}_p$ ($qp \times kp$ matrix), $\mathbf{G} = (\widetilde{\mathbf{G}}\mathbf{D}\widetilde{\mathbf{G}}^\top)^{-1/2}\widetilde{\mathbf{G}}$ with $\mathbf{D} = \text{diag}(1/n_1, \cdots, 1/n_k)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_k^\top)^\top$.

Let $\mathbf{H} = \mathbf{G}^\top\mathbf{G}$ and $\widehat{\boldsymbol{\mu}} = (\overline{\mathbf{x}}_1^\top, \ldots, \overline{\mathbf{x}}_k^\top)^\top$ where $\overline{\mathbf{x}}_g = (1/n_g)\sum_{i=1}^{n_g}\mathbf{x}_{gi}$ for $g \in \{1, \ldots, k\}$. Then, tor testing (5), we defined the test statistic as

$$\begin{aligned} T_{nh} &= \frac{1}{p}\left\{\|\mathbf{C}\widehat{\boldsymbol{\mu}}\|^2 - \sum_{g=1}^{k} a_{gg}\widehat{\text{tr}(\boldsymbol{\Psi}_g)}\right\} \\ &= \frac{1}{p}\left\{\widehat{\boldsymbol{\mu}}^\top(\mathbf{H}\otimes\mathbf{I}_p)\widehat{\boldsymbol{\mu}} - \sum_{g=1}^{k} a_{gg}\widehat{\text{tr}(\boldsymbol{\Psi}_g)}\right\} \end{aligned}$$

where, for $g \in \{1, \ldots, k\}$, a_{gg} is the diagonal element of the matrix $\mathbf{A} = \mathbf{D}^{1/2}\mathbf{H}\mathbf{D}^{1/2}$ and $\widehat{\text{tr}(\boldsymbol{\Psi}_g)} = \text{tr}(\mathbf{S}_g) - \sum_{\ell=1}^{\widehat{d}_g}\lambda_\ell(\mathbf{S}_g)$. 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_{nh} under some assumptions and constructed test procedure for testing (5). Also, we compared, through simulations, the performance of the proposed test and existing procedures suitable for one-way MANOVA 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] Zhang, J.-T., Guo, J., Zhou, B., 2017. Linear hypothesis testing in high-dimensional one-way MANOVA. *J. Multivar. Anal.*, **155**, 200–216.

[3] Zhang, J.-T., Zhou, B., Guo, J., 2022. Linear hypothesis testing in high-dimensional heteroscedastic one-way MANOVA: A normal reference L^2-norm based test. *J. Multivar. Anal.*, **187**, 104816.