International Symposium on Statistical Analysis for Large Complex Data

November 21-23, 2016

Venue:

D509 Institute of Natural Sciences
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan

Organizers:

Makoto Aoshima	(University of Tsukuba)
Mika Sato-Ilic	(University of Tsukuba)
Kazuyoshi Yata	(University of Tsukuba)

Supported by

Grant-in-Aid for Scientific Research (A) 15H01678 (Project Period: 2015-2019)
“Theories and Methodologies for Large Complex Data”
(Principal Investigator: Makoto Aoshima)

Grant-in-Aid for Challenging Exploratory Research 26540010 (Project Period: 2014-2016)
“Statistics for Big Data: Development of Theories and Tackling the 3Vs”
(Principal Investigator: Makoto Aoshima)

Program

November 21 (Monday)

14:00∼14:05 Opening

14:05∼14:35 Aki Ishii (Graduate School of Pure and Applied Sciences, University of Tsukuba)
Two-sample tests for high-dimension, low-sample-size data under a strongly spiked eigenvalue model

14:45∼15:25 Hiroumi Misaki (Faculty of Engineering, Information and Systems, University of Tsukuba)
Recent developments in the SIML estimation of integrated volatility with high frequency financial data

15:40∼16:20 Hidetoshi Matsui (The Center for Data Science Education and Research, Shiga University)
Sparse regularization for functional logistic regression models

16:30∼17:10 Hirokazu Yanagihara
(Department of Mathematics, Graduate School of Science, Hiroshima University)
A high-dimensionality-adjusted consistent Cp-type criterion in multivariate linear regression models

November 22 (Tuesday)

9:30∼10:10 Mariko Yamamura (Graduate School of Education, Hiroshima University)

Canonical correlation analysis for geographical and chronological responses

10:20∼11:00 Takafumi Kanamori[a,*] and Takashi Takenouchi[b]

[a](Department of Computer Science and Mathematical Informatics, Nagoya University)

[b](Future University Hakodate)

Statistical inference using graph-based divergences on discrete sample spaces

11:10∼11:50 Masahiro Mizuta

(Advanced Data Science Laboratory, Information Initiative Center, Hokkaido University)

Mini data approach to big data

11:50∼13:25 Lunch

13:25∼17:30 **Special Invited Session**

18:30∼ Dinner

November 23 (Wednesday)

9:30∼10:10 Kazuyoshi Yata* and Makoto Aoshima

(Institute of Mathematics, University of Tsukuba)

Statistical inference in strongly spiked eigenvalue models

10:20∼11:00 Shinpei Imori (Graduate School of Engineering Science, Osaka University)

Regression with auxiliary variable

11:10∼11:50 Taiji Suzuki

(Department of Mathematical and Computing Sciences, Tokyo Institute of Technology)

Some convergence results of nonparametric tensor estimators

11:50∼ 12:00 Closing

(∗ Speaker)

Special Invited Session

November 22 (Tuesday)

13:25~14:15 **Business analytics and big data**

Speaker: Haipeng Shen
(Faculty of Business and Economics, University of Hong Kong, Hong Kong)

14:30~15:20 **Order selection for predictions in high-dimensional AR model: the case of $I(d)$ processes**

Speaker: Shu-Hui Yu
(Institute of Statistics, National University of Kaohsiung, Taiwan)

15:35~16:25 **On high-dimensional cross-validation**

Speaker: Ching-Kang Ing
(Institute of Statistical Science, Academia Sinica, Taiwan)

16:40~17:30 **Classification and variable selection for high-dimensional data with applications to proteomics**

Speaker: Inge Koch
(School of Mathematical Sciences, University of Adelaide, Australia)

Two-sample tests for high-dimension, low-sample-size data

Aki Ishii

Graduate School of Pure and Applied Sciences, University of Tsukuba, Ibaraki, Japan

1 Introduction

One of the features of modern data is that the data dimension is extremely high, however, the sample size is relatively low. We call such data "HDLSS" or "large p, small n" data, where p is the data dimension and n is the sample size. Suppose we have two classes $\pi_i,\ i=1,2$, and define independent $p\times n_i$ data matrices, $\boldsymbol{X}_i=[\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}],\ i=1,2$, from $\pi_i,\ i=1,2$, where $\boldsymbol{x}_{ij},\ j=1,...,n_i$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i=\boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i=\mathrm{diag}(\lambda_{1(i)},...,\lambda_{p(i)})$ having $\lambda_{1(i)}\geq\cdots\geq\lambda_{p(i)}(\geq 0)$ and $\boldsymbol{H}_i=[\boldsymbol{h}_{1(i)},...,\boldsymbol{h}_{p(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. We considered the two-sample test:

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1:\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2.$$

We defined $\overline{\boldsymbol{x}}_{in_i}=\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i}=\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij}-\overline{\boldsymbol{x}}_{in_i})^T/(n_i-1)$ for $i=1,2$. Note that $\boldsymbol{S}_{in_i}^{-1}$ does not exist in the HDLSS context such as $n_i/p\to 0$. Under the assumption that π_1 and π_2 are Gaussian, there are a lot of literatures about the two-sample problem in the HDLSS context. When π_1 and π_2 are non-Gaussian, Chen and Qin [4] and Aoshima and Yata [1, 2] considered the two-sample test under heteroscedasticity, $\boldsymbol{\Sigma}_1\neq\boldsymbol{\Sigma}_2$. We note that the above literatures considered constructing two-sample test procedures under the eigenvalue condition as follows:

$$\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}\to 0 \ \text{ as } p\to\infty \text{ for } i=1,2. \tag{1.1}$$

However, (1.1) sometimes fails in actual high-dimensional analyses. Aoshima and Yata [3] proposed to develop high-dimensional inference not only for the "non-strongly spiked eigenvalue (NSSE) model" defined by (1.1) but also for the "strongly spiked eigenvalue (SSE) model" defined by

$$\liminf_{p\to\infty}\left\{\frac{\lambda_{1(i)}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}\right\}>0 \ \text{ for } i=1 \text{ or } 2. \tag{1.2}$$

They discussed the two-sample test by using the high-dimensional asymptotic theory where $p\to\infty$ and $n_i\to\infty$. In this talk, we focused on the SSE model and constructed two-sample test procedures when $p\to\infty$ while n_is are fixed.

2 A new two-sample test under the SSE model

Ishii [7] constructed a new two-sample test procedure by using the noise-reduction (NR) methodology given by Yata and Aoshima [8]. Let $\hat{\lambda}_{1(i)}\geq\cdots\geq\hat{\lambda}_{p(i)}\geq 0$ be the eigenvalues of $\boldsymbol{S}_{in_i}$. By using the NR method, $\lambda_{j(i)}$s are estimated by

$$\tilde{\lambda}_{j(i)}=\hat{\lambda}_{j(i)}-\frac{\mathrm{tr}(\boldsymbol{S}_{in_i})-\sum_{s=1}^{j}\hat{\lambda}_s}{n_i-1-j} \quad (i=1,2;\ j=1,...,n_i-2).$$

Note that $\tilde{\lambda}_{j(i)} \geq 0$ w.p.1 for $j = 1, ..., n_i - 2$. Yata and Aoshima [8, 9] and Ishii et al. [5, 6] showed that $\tilde{\lambda}_{j(i)}$ has several consistency properties in high-dimensional context.

We started with the following test statistic:

$$T_n = ||\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}||^2 - \sum_{i=1}^{2} \text{tr}(\boldsymbol{S}_{in_i})/n_i.$$

Note that T_n was discussed by Chen and Qin [4] and Aoshima and Yata [1, 2] under the NSSE model. We evaluated T_n under the SSE model and gave a new test statistic as follows.

$$F = u_n \frac{T_n + \sum_{i=1}^{2} \tilde{\lambda}_{1(i)}/n_i}{\tilde{\lambda}_{1n}},$$

where $u_n = (1/n_1 + 1/n_2)^{-1}$ and $\tilde{\lambda}_{1n} = (n_1 + n_2 - 2)^{-1} \sum_{i=1}^{2}(n_i - 1)\tilde{\lambda}_{1(i)}$. We discussed the asymptotic null distribution and the power of F. We also gave another test statistic and compared test procedures by using computer simulations.

References

[1] Aoshima, M. and Yata, K. (2011). Two-Stage Procedures for High-Dimensional Data. *Sequantial Analysis* (*Editor's special invited paper*) 30: 356-399.

[2] Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodology and Computing in Applied Probability* 17: 419-439.

[3] Aoshima, M. and Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, to appear.

[4] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* 38: 808-835.

[5] Ishii, A., Yata, K., and Aoshima, M. (2014). Asymptotic distribution of the largest eigenvalue via geometric representations of high-dimension, low-sample-size data. *Sri Lankan Journal of Applied Statistics*, Special Issue: Modern Statistical Methodologies in the Cutting Edge of Science (ed. Mukhopadhyay, N.), 81-94.

[6] Ishii, A., Yata, K., and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference*, 170: 186-199.

[7] Ishii, A. (2016). A two-sample test for high-dimension, low-sample-size data under the strongly spiked eigenvalue model, submitted.

[8] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* 105, 193-215.

[9] Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis* 122: 334-354.

Recent developments in the SIML estimation of integrated volatility with high frequency financial data

Hiroumi Misaki [1]

Estimating the volatility and covariance of asset prices has been a key issue in finance, and considerable interest has been paid on the estimation problem by using high-frequency data in financial econometrics. We assume that the underlying continuous process $X(t)$ $(0 \leq t \leq 1)$ is not necessarily the same as the observed (log-) price at $t_i^n (i = 1, \cdots, n^*)$ and

$$X(t) = X(0) + \int_0^t \sigma_x(s) dB(s) \;\; (0 \leq t \leq 1),$$

where $B(s)$ is the standard Brownian motion, $\sigma_x(s)$ is the instantaneous volatility function. The main statistical objective is to estimate the integrated volatility

$$\sigma_x^2 = \int_0^1 \sigma_x^2(s) ds$$

of the underlying continuous process $X(t)$ from the set of discretely observed prices $y(t_i^n)$ which are generated by $y(t_i^n) = h\left(X(t_i^n), y(t_{i-1}^n), u(t_i^n)\right)$.

It has been well known that the realized volatility works poorly when there exist micro-market noise. Kunitomo and Sato (2011, 2013) have proposed a new statistical method called the Separating Information Maximum Likelihood (SIML) method under the presence of micro-market noises. They have shown that the SIML estimator has reasonable asymptotic properties as well as finite sample properties.

Misaki and Kunitomo (2015) and Kunitomo, Misaki and Sato (2015) have further investigated the properties of the SIML estimation when we have the micro-market noises and randomly sampled data at the same time. We have shown the asymptotic robustness in the sense that it is consistent and it has the asymptotic normality under a set of fairly general conditions.

We have investigated the finite sample properties of the SIML estimator for the integrated volatility based on a set of simulations. In all cases, the estimates obtained by realized volatility are badly-biased, which have been well known in the analysis of high frequency financial data. The SIML estimate, on the other

[1] Faculty of Engineering, Information and Systems, University of Tsukuba, Tennodai 1-1-1, Tsukuba City, Ibaraki 305-8577, JAPAN, hmisaki@risk.tsukuba.ac.jp

hand, gives reasonable estimate and the variance of the SIML estimator is within a reasonable range for practical purposes.

In empricial studies, we have analysed high frequency financial data in the Japanese stock market. Our main purpose is to estimate daily volatility, covariance and other related quantities by using SIML estimator and to compare them to some alternative estimators. We have found that the SIML estimation provides reasonable results in any case whereas most of the examined alternatives are severely biased. In addition, we have found that the SIML estimates are similar to the realized volatilities and covariances based on relatively long intervals in respect of the summary statistics across the sample period. Our detailed analysis, however, have indicated they are not always coincide to each other. In conclusion, our investigation suggest that the SIML estimation is useful to estimate the daily integrated volatility, covariance, and other related quantities in actual markets.

We have also discussed the finite sample estimation of the variance or standard deviation of estimators. One of the possible approach to obtain the variance and its functionals of an estimator in complex models is to utilize the bootstrap methods, introduced by Efron (1979). In our formulation there would be two possible approaches to exploit bootstrap methods for the SIML estimator. One is to resample the transformed variables z_k which is mutually independent (but not identical), and the other is to apply the MBB to the difference of log price y_t. We have given the preliminary report on estimating the finite sample variance of the SIML estimator.

References

[1] Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, Vol. 7, 1, 1-26.

[2] Kunitomo, N., H. Misaki and S. Sato (2015), "The SIML Estimation of Integrated Covariance and Hedging Coefficients with Micro-market noises and Random Sampling," *Asia-Pacific Financial Markets*, Vol. 22, 3, 333-368.

[3] Kunitomo, N. and S. Sato (2011), "The SIML Estimation the Integrated Volatility of Nikkei-225 Futures and Hedging Coefficients with Micro-Market Noise," *Mathematics and Computers in Simulations*, Elsevier, 81, 1272-1289.

[4] Kunitomo, N. and S. Sato (2013), "Separating Information Maximum Likelihood Estimation of the Integrated Volatility and Covariance with Micro-Market Noise," *North American Journal of Economics and Finance*, Vol. 26, 282-309.

[5] Misaki, H. and N. Kunitomo (2015), "On robust properties of the SIML estimation of volatility under micro-market noise and random sampling," *International Review of Economics & Finance*, 40, 265-281.

Sparse regularization for functional logistic regression models

Hidetoshi Matsui

The Center for Data Science Education and Research, Shiga University
1-1-1 Banba, Hikone, Shiga, 522-8522, Japan.

hmatsui@biwako.shiga-u.ac.jp

1 Introduction

Sparse regularization have attracted attentions as they provide a unified approach to problems of estimating and selecting variables, and for this reason they are broadly applied in several fields (Hastie et al., 2015). In this work we consider applying the sparse regularization to the analysis of longitudinal data and selecting genes that have effect on classification.

When the data to be classified have been measured repeatedly over time, they can be represented by a functional form. Ramsay and Silverman (2005) established this type of analysis and called it functional data analysis (FDA). FDA is one of the most useful methods for effectively analyzing discretely observed data, and it has received considerable attention in various fields.

In this work we consider the problem of using L_1-type regularization to select the variables for classifying functional data by using the multiclass logistic regression model. In particular, we apply two types of L_1-type penalties and then describe the effect of them. Then we report results of the analysis of multiple sclerosis data and yeast cell cycle gene expression data.

2 Multiclass logistic regression model for functional data

Suppose we have n sets of functional data and a class label $\{(x_i(t), g_i);\ i = 1, \ldots, n\}$, where $x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^T$ are predictors given as functions and $g_i \in \{1, \ldots, L\}$ are the classes to which x_i belongs. In the classification setting, we apply the Bayes rule, which assigns x_i to class $g_i = l$ with the maximum posterior probability given x_i, denoted by $\Pr(g_i = l|x_i) = \pi_l(x_i; b)$ with a parameter vector b incluced in the model. Then the logistic regression model is given by the log-odds of the posterior probabilities:

$$\log\left\{\frac{\pi_l(x_i; b)}{\pi_L(x_i; b)}\right\} = \beta_{0l} + \sum_{j=1}^{p} \int x_{ij}(t)\beta_{jl}(t)dt, \tag{1}$$

where β_{0l} is an intercept and $\beta_{jl}(t)$ are coefficient functions. We assume that both $x_{ij}(t)$ and β_{jl} are expressed by basis expansions. Furthermore, we define the vectors of the response variables y_i, which indicate the class labels. Then the functional logistic regression model (1) has the probability function

$$f(y_i|x_i;b) = \prod_{l=1}^{L-1} \pi_l(x_i;b)^{y_{il}} \pi_L(x_i;b)^{1-\sum_{h=1}^{L-1} y_{ih}}.$$

3 Estimation by sparse regularization

We consider estimating the parameter b by maximizing the penalized log-likelihood function

$$\ell_{\lambda,\alpha}(b) = \sum_{i=1}^{n} \log f(y_i|x_i;b) - P_{\lambda,\alpha}(b),$$

where $P_{\lambda,\alpha}(b)$ is a penalty function controlled by tuning parameters $\lambda > 0$ and $\alpha \in [0,1]$. Here we apply following two types of penalties for $P_{\lambda,\alpha}(b)$:

$$P_{\lambda,\alpha}(b) = \frac{1}{2}(1-\alpha)\sum_{j=1}^{p} \lambda_j \sum_{l=1}^{L-1} \|b_{jl}\|_2^2 + \alpha \sum_{j=1}^{p} \lambda_j \left\{ \sum_{l=1}^{L-1} \|b_{jl}\|_2^2 \right\}^{\frac{1}{2}}, \tag{2}$$

$$P_{\lambda,\alpha}(b) = n(1-\alpha)\sum_{j=1}^{p} \lambda_j \left\{ \sum_{l=1}^{L-1} \|b_{jl}\|_2^2 \right\}^{1/2} + n\alpha \sum_{j=1}^{p} \lambda_j \sum_{l=1}^{L-1} \|b_{jl}\|_2, \tag{3}$$

where λ_j are tuning parameters controlled by λ. Penalty (2) is the elastic net-type penalty (Zou and Hastie, 2005) and has the property that it select variables in functional logistic regression model (Kayano et al., 2016). On the other hand, (3) is the sparse group lasso-type penalty (Friedman et al., 2010) and it can select both variables and decision boundaries in the functional logistic regression model.

References

Friedman, J., Hastie, T., and Tibshirani, R. (2010), A note on the group lasso and a sparse group lasso, *arXiv preprint arXiv:1001.0376*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Kayano, M., Matsui, H., Yamaguchi, R., Imoto, S., and Miyano, S. (2016), Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to timedependent biomarker detection, *Biostatistics*, 17, 235–248.

Ramsay, J. and Silverman, B. (2005), *Functional data analysis 2nd ed.*, New York: Springer.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. Ser. B*, 67, 301–320.

A High-dimensionality-adjusted Consistent C_p-type Criterion in Multivariate Linear Regression Models

Hirokazu Yanagihara

Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

Suppose that k-variate explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ and p-variate mutually correlated response variables $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})'$ $(i = 1, \ldots, n)$ are observed, where n is the sample size. The set of n-vectors of the response variables $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, and the set of the n-vectors of the k explanatory variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are written in matrix notation as an $n \times p$ matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and an $n \times k$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, respectively. A multivariate linear regression model in which the normality of $\boldsymbol{y}_i$ is assumed, called a normal multivariate linear regression (NMLR) model, is defined as follows:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n),$$

where $\boldsymbol{\Theta}$ is a $k \times p$ matrix of the unknown regression coefficients, and $\boldsymbol{\Sigma}$ is a $p \times p$ unknown variance-covariance matrix that is positive definite. In order to ensure that it is possible to estimate the NMLR model, we assume that $\mathrm{rank}(\boldsymbol{X}) = k$ $(< n)$ and $n - k > p$.

Let us express the indices of the relevant explanatory variables as the elements of a set j defined as $j \subseteq \omega = \{1, \ldots, k\}$ and $k_j = \#(j)$. We denote the matrix of explanatory variables indexed by the elements of j as $\boldsymbol{X}_j$, i.e., $\boldsymbol{X}_j$ is the $n \times k_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth columns of $\boldsymbol{X}$, and $k_j = 3$. We then consider the candidate model to be the following NMLR model, which uses the k_j relevant explanatory variables that were extracted from the set of all possible explanatory variables:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n),$$

where $\boldsymbol{\Theta}_j$ is the $k_j \times p$ matrix of unknown regression coefficients, and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown variance-covariance matrix that is assumed to be positive definite.

A commonly used method for variable selection is to choose the best subset of explanatory variables by minimizing a variable selection criterion. The best subset of explanatory variables is the subset in the candidate model that results in the smallest value for the variable selection criterion. One of the most widely known variable selection criteria is the C_p criterion proposed by Mallows (1973; 1995). The criterion is defined by adding twice the number of parameters in the mean structure to the minimum value of the residual sum of squares (RSS), which is the sum of the squared differences between the data and the fitted value. The C_p criterion was proposed for selecting variables in a linear regression model in which there is only one response variable. A multiple-response version of the C_p criterion was proposed by Sparks *et al.* (1983) and Fujikoshi and Satoh (1997) proposed a modified C_p (MC_p) criterion that corrected the bias of that proposed by Sparks *et al.* For a linear regression with a single response, Atkinson (1980) proposed a generalized C_p (GC_p) criterion, which is obtained by adding a penalty term that is α times the number of parameters in the mean structure to the RSS, where α is some positive constant. A multiple-response version of the GC_p criterion was proposed by Nagai *et al.* (2012). By varying α, the GC_p can express a wide variety of selection criteria, e.g., the GC_p with $\alpha = 2$ coincides with the C_p, and the GC_p with $\alpha = (n-k)/(n-k-p)$ is essentially equivalent to the MC_p.

An important property of a variable selection criterion is its consistency, which means that the true subset of explanatory variables can be selected with probability tending asymptotically to unity. We express the subset of true explanatory variables as $j_* \subseteq \omega$. A large sample (LS) asymptotic framework in which the sample size n goes to ∞ is commonly used for evaluating consistency. Although consistency is an asymptotic property, we can expect that there is a high probability that a consistent variable selection criterion can choose the true subset j_* as the best one for a moderate sample size. On the other hand, we often see a high-dimensional dataset in which the dimension of the response variables vector is large, but it is still smaller than the sample size; this situation is called moderately high-dimensional data (see e.g., Yao *et al.*, 2015). With moderately high-dimensional data, a consistent variable selection criterion developed for the LS asymptotic framework will often choose a subset other than j_* even when the sample size is large. This occurs because the asymptotic distribution evaluated from the LS asymptotic framework has poor accuracy with moderately high-dimensional data. We can improve the accuracy of the asymptotic distribution by using not the LS asymptotic framework but a high-dimensional (HD) asymptotic framework, in which the sample size n and the dimension of the response variables vector p go to ∞ simultaneously under the condition that $p/n \to c_0 \in [0, 1)$. We can expect that a variable selection criterion that is judged to be consistent by an evaluation within the HD asymptotic framework can choose the true subset j_* as the best one with high probability under a moderate sample size even when p is large. However, when p is very small, there remains a possibility that such a criterion will choose a subset other than j_* even under a large sample size, because then, the accuracy of the asymptotic distribution evaluated within the HD asymptotic framework becomes low.

The aim of this paper is to propose a C_p-type criterion that will meet the conditions for consistency regardless of the asymptotic framework that is used to evaluate it; that is, with high probability, this criterion is expected to select the true subset j_* for a moderate sample size regardless of the size of p. To achieve our aim, the following asymptotic framework is used for evaluating consistency: $n \to \infty$ and $p/n \to c_0 \in [0, 1)$.

A sufficient condition to ensure the consistency of GC_p when $n \to \infty$ and $p/n \to c_0 \in [0, 1)$ is that the following conditions are satisfied simultaneously:

$$\lim_{n\to\infty, p/n\to c_0} \sqrt{p}\left(-\frac{n}{n-p} + \alpha\right) = \infty, \quad \lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\left(-\frac{n}{n-p} + \alpha\right) = 0.$$

By using the result, we propose new GC_p, which is consistent even under high dimensionality; we call it the high-dimensionality-adjusted consistent GC_p ($HCGC_p$). The $HCGC_p$ criterion is defined by the following α:

$$\alpha = \frac{n}{n-p} + \beta, \quad \beta > 0 \ \ s.t. \lim_{n\to\infty, p/n\to c_0} \sqrt{p}\beta = \infty, \quad \lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\beta = 0.$$

Canonical correlation analysis for geographical and chronological responses

Mariko Yamamura

Graduate School of Education, Hiroshima University, 1-1-1 Kagamiyama,
Higashi-Hiroshima, 739-8524, Japan

Data containing information about observed location and time are called geographical and chronological data. Yamamura *et al.* (2016) extended the application potency of the varying coefficient model in Tonda *et al.* (2010) by applying the model not only for geographical, but also for chronological data. As is often the case with real data sets, we sometimes need to analyze with multiple response variables. Tonda *et al.* (2010) has only proposed the varying coefficient model for a single response variable. One method of treating multiple response variables is that we create a synthesis variable from them and apply a regression model which procedure corresponds to canonical correlation analysis (CCA). The purpose of this paper is to propose how we can analyze geographical and chronological data with multiple response variables by innovating the varying coefficient model in CCA.

As numerical background, we propose to apply an approach where we use a body condition data set from common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the Barents Sea (Solvang *et al.* (2016)). Over the period 1993-2013, the body condition data were obtained from a total of 10,556 common minke whales taken in Norwegian scientific and commercial whaling operations in the Northeast Atlantic during the months April to September. Blubber thickness (BT1, BT2, and BT3) measurements were made perpendicular from the skin surface to the muscle-connective tissue interface. Length and girth measurements were made to the nearest centimeter, while blubber measurements were to the nearest millimeter. For all whales, the year, month, day, and latitude / longitude were recorded.

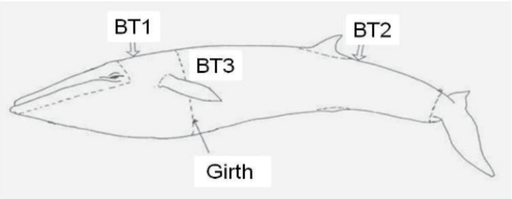

Figure 1: Measurement sites.

The synthesis variable $u = \boldsymbol{\alpha}'\boldsymbol{y}$ is assumed to have a liner structure, where $\boldsymbol{\alpha}$ is a parameter vector and columns of $\boldsymbol{y}$ are $(y_1, y_2, y_3)' = (\text{“length”}, \text{“BT1”}, \text{“BT3”})'$. Explanatory variables $\boldsymbol{a}$ take values 1. We fit the linear model to estimate the varying coefficient cubic plane or plane curve, i.e. $\hat{\boldsymbol{\beta}}(z_1, z_2, z_3, z_4) = \hat{\boldsymbol{\theta}}'\boldsymbol{w}(z_1, z_2, z_3, z_4)$, where $(z_1, z_2, z_3, z_4)' = (\text{“latitude”}, \text{“longitude”}, \text{“year”}, \text{“calendar day”})'$. The $\boldsymbol{w}(z_1, z_2, z_3, z_4) = (\boldsymbol{w}_1(z_1, z_2)', \boldsymbol{w}_2(z_3)', \boldsymbol{w}_3(z_4))'$ is assumed to have one of either linear, quadratic or cubic expression with their interaction each.

All variables used in the estimation are standardized, since estimated coefficient values should be compared on the same basis, which is common practice in a real data analysis with CCA.

The best variables for $\boldsymbol{y}$ and best expression form of $\boldsymbol{w}(z_1, z_2, z_3, z_4)$ were selected by the Bayesian information criterion (BIC) in Schwarz (1978). For clear interpretation of results, estimated varying coefficients are graphically expressed by geography in Fig.2.

Fig.2 shows estimated varying coefficients cubic plane curves by sex. White markers are actual catching points, and coefficient values are showed by contour plots which become higher in warmer colored areas.

Contours take values between -1.5 and 1 and are almost flat in male, meaning that body condition considered by length, BT1, and BT3 are not much different in any geographical areas in males. In females, the low contour in dark blue at the bottom left in the map, Fig.2 and the high in yellow at the top, might signify habits of whales that they migrate from south with hunger and move northward to take enough nourishment in Barents Sea, or might take extreme large or small values since $\boldsymbol{w}_1(z_1, z_2)$ have a cubic expression.

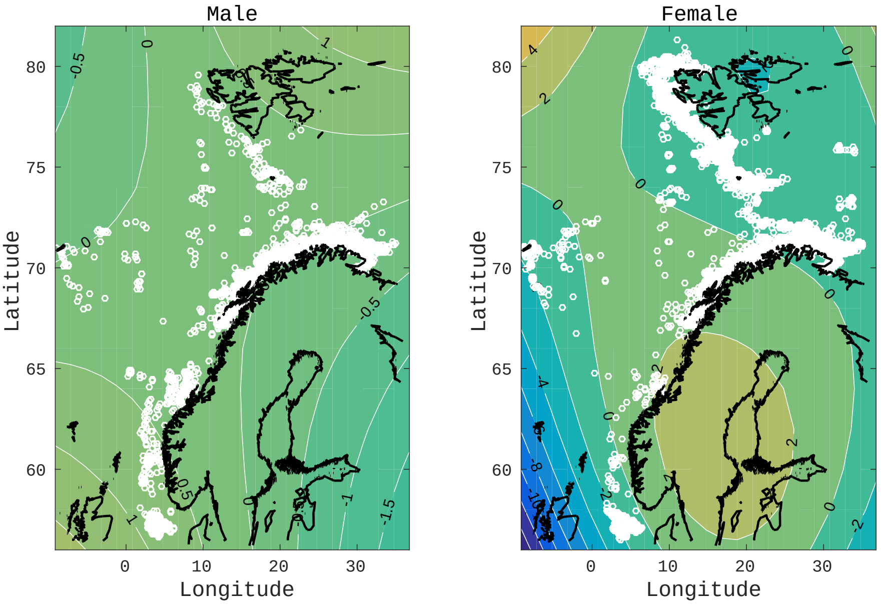

Figure 2: Varying coefficient cubic plane curves (length, BT1 & BT3).

From the estimation results of minke whales body condition data, male and female whales gain their body condition as fall approaches, which is the well known as their general habits in the Barents Sea; the nourishment during summer result in fat deposition and leads to fatter body conditions in the fall. Windsland *et al.* (2008) suggested the possibility of food reduction for whales caused by ecological change in Barents Sea.

References

[1] Yamamura M, Fukui K, Yanagihara H. Illustration of the varying coefficient model for a tree growth analysis from the age and space perspectives. *FORMATH* 2016;**15**:1-9.

[2] Tonda T, Satoh K, Yanagihara H. Statistical inference on a varying coefficient surface using interaction model for spatial data. *Japanese J Appl Statist* 2010;**39**:59-70 (in Japanese).

[3] Solvang HK, Yanagihara H, Øien N, Haug T. Temporal and geographical variation in body condition of common minke whales (*Balaenoptera acutorostrata acutorostrata*) in the northeast Atlantic. *TR No 16-05, Statistical Research Group, Hiroshima University*, Hiroshima; 2016.

[4] Schwarz G. Estimating the dimension of a model. *Ann Statist* 1978;**6**:461-464.

[5] Windsland K, Lindstrøm U, Nilssen K. T., Haug T. Relative abundance and size composition of prey in the common minke whale diet in selected areas of the northeast Atlantic during 2000-04. *J Cetacean Res Manag* 2008;**9**:167-178.

Statistical Inference using Graph-based Divergences on Discrete Sample Spaces

Takafumi Kanamori[1] and Takashi Takenouchi[2]

[1]Nagoya University
[2]Future University Hakodate

This paper proposes a general framework of statistical inference using unnormalized statistical models on discrete sample spaces. One of the most common methods in statistical inference is the maximum likelihood estimator (MLE), which is obtained by maximizing the empirical mean for the log-likelihood of the statistical model. The MLE has some nice properties such as the statistical consistency and efficiency. The computation of the normalization constant in the statistical model is, however, often intractable in high-dimensional sample domains.

Several approaches have been proposed to deal with the normalization constant. Monte Carlo method is a popular method to approximate integrals and total sums using random sampling [5, 9, 19]. Alternatively, the log-likelihood can be replaced with another scoring rule that measures the goodness of fit of the model to observed samples. Scoring rules such as pseudo-likelihood, composite likelihood, and ratio matching [1, 2, 7, 10, 11, 12, 13, 14, 15, 18], which do not require the normlization constant, are thought to be computationally efficient. As a general rule, the computational cost can be reduced by localizing the scoring rule over the sample space. Dawid, et al. [4] argued the theoretical properties of scoring rules that can be expressed as the sum of localized scoring rules.

In this paper, we study the statistical consistency of local scoring rules. In general, the scoring rule is closely related to the Bregman divergence $D(p, q)$, which measures the discrepancy between two probability distributions, p and q [6, 8, 16]. The Bregman divergence takes non-negative real numbers and $D(p, p) = 0$ holds for any probability distribution. An important axiom of the Bregman divergence is the coincidence axiom, which states that $p = q$ holds when $D(p, q) = 0$ [3, 17]. For the Bregman divergence with the coincidence axiom, the associated scoring rule has the property of the asymptotic consistency under a mild assumption. The asymptotic consistency of the estimator using the scoring rule is guaranteed by the sufficient condition of asymptotic consistency for M-estimator [20].

We also demonstrate the relation between the local scoring rules and localized Bregman divergences, and use it to investigate the statistical consistency of local scoring rules. We show that the consistency depends on the structure of the neighborhood system defined on discrete sample spaces.

References

[1] A. U. Asuncion, Q. Liu, A. T. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In Yee W. Teh and D. M. Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, pages 33–40, 2010.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

[3] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex prog ramming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[4] A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *Annals of Statistics*, 40:593–608, 2012.

[5] C. J. Geyer. Markov chain monte carlo maximum likelihood. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[6] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

[7] M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI2011)*, pages 283–290, 2011.

[8] A. D. Hendrickson and R. J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42:19161921, 1971.

[9] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[10] A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007.

[11] A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51:2499–2512, 2007.

[12] P. Liang and M. I. Jordan . An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 584–591, New York, NY, USA, 2008. ACM.

[13] B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80(1), 1988.

[14] S. Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 359–366, Arlington, Virginia, United States, 2009. AUAI Press.

[15] B. Marlin and N. de Freitas. Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. In *Uncertainty in Artificial Intelligence (UAI)*, Corvallis, Oregon, 2011. AUAI Press.

[16] J. McCarthy. Measures of the value of information. *Proc Natl Acad Sci U S A.*, 42(9):654–655, 1956.

[17] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

[18] M. Pihlaja, M. Gutmann, and A. Hyvärinen. A Family of Computationally Efficient and Simple Estimators for Unnormalized Statistical Models. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 442–449, Corvallis, Oregon, 2010. AUAI Press.

[19] R. Salakhutdinov. Learning and evaluating boltzmann machines. Technical report, UTML TR 2008-002, Dept. of Computer Science, University of Toronto, 2008.

[20] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

Mini Data Approach to Big Data

Masahiro Mizuta
Advanced Data Science Laboratory, Information Initiative Center,
Hokkaido University
N.11, W.5, Kita-ku, Sapporo 060-0811, Japan

Abstract: We deal with a question whether the visualization is always effective for big data? "Big data" is an important key word for business, academic, medical and so on. Many persons insist "In a Big Data World we need visualization" or "What we need to deal with big data is a visualization!!" We discuss a mechanism and a limitation of visualization and generalize the visualization approach to the mini data approach[1].

Key words: data analysis, data reduction, symbolic data analysis

１. Introduction

Nowadays, Big Data is an important key word. But I am afraid that we, statisticians, do not have clear strategies to deal with Big data. We try to discuss this problem.

２. Is Visualization ALWAYS effective for BIG DATA?

The question is "Is Visualization ALWAYS effective for BIG DATA?" Valuable clues are the size of data and the size of the plot. The size of IRIS DATA is 4 kilobytes at most. The size of the plot is more than 20 kilobytes. In principle, all information is contained in the plot, in other words, we can recover the Iris data from the plot.

But, in the case of Big data, the situation is completely changed. The size of Big data is beyond several gigabytes or terabytes. Even if ultra-high resolution display is used, the size of the plot is several gigabytes. This means it is impossible to visualize big data completely. The visualization of big data is the task of data reduction. The task is a field of statistics.

３. Mini data approach

We have proposed a concept of mini data approach for big data (Mizuta & Minami, 2015). *Mini data* of big data are defined as data set which contains an important information about the big data, but its size and/or structure are realistic to deal with. Visualization is a kind of mini data. Contingency table is also a kind of mini data. There are two steps in mini data approach; How to build Mini Data, and how to analyze Mini Data.

Here are methods or tools to build mini data from big data; Sampling, Variable selection, Dimension reduction, Feature extraction, and Symbolization.

[1] A part of this manuscript is a translated version of Mizuta (2016a).

Another important problem is how to utilize mini data or how to get fruitful results from mini data. We have many methods that can be used for the problem, including conventional multidimensional data analysis, symbolic data analysis.

４． Concluding remarks

We begin with the question “Is Visualization ALWAYS effective for analysis of data?” The answer is definitely NO. Of course, when the size of data is small, visualization is completely effective, e.g. we can find an excellent plot of iris data. But the size of the data set is big, for example more than 100Mbytes, Visualization is a kind of data reduction or data compression!! In general, it is a HARD WORK. We must focus on the process of the visualization and generalize it. It is a mini data approach in my presentation.

References

Billard L., Diday E. (2012) Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, 2012.

Kelly K. (2011) Web 2.0 Expo and Conference, March 29, 2011. Video available at: www.web2expo.com/webexsf2011/public/schedule/proceedings.

McKinsey (2011) Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Mizuta M., and Minami H. (2015) Methods for Analyzing Joint Distribution Valued Data and Actual Data Sets -Mini Data Approach to Big Data-, Symbolic Data Analysis Workshop (SDA 2015), 2015.

Mizuta (2016a) 水田 正弘　ビッグデータに対するミニデータアプローチ －可視化は常に有効か？－ 医用画像情報学会雑誌 Vol. 33(2016) No. 1 pp.1-3

Mizuta (2016b) 水田正弘:ビッグデータに対する統計学の役割、日本経営情報学会誌　日本経営情報学会誌 2016 年 6 月号、Vol.36, No.4, pp.12-17.

Ramsay J., Silverman B. (2015) Functional Data Analysis (2nd Edition). Springer, 2005.

Business Analytics and Big Data

Haipeng Shen

Faculty of Business and Economics, University of Hong Kong

Big data are becoming increasingly common in our modern digital business world. More and more data are being collected with ever - increasing volume, dimensionality, and complexity. We are blessed with the flood of data, as business analytics techniques can be called upon to "mine" important features from the data, i.e. finding the needles. At the same time, high - dimensionality and complexity can be a curse, as the needles are often hidden in a haystack, or multiple haystacks, and classical methods sometimes fail to work for big data. This presentation uses real examples, including customer service call center workforce management and healthcare delivery systems, to provide a statistician's perspective on how innovative data - analytical techniques can assist business decision making by asking the right questions, having the right data, and collaborating with the right people.

Order selection for predictions in high-dimensional AR models: the cases of *I*(*d*) processes

Shu-Hui Yu
Institute of Statistics, NUK

Abstract

Most order selection methods in high-dimensional autoregressive models are devised for processes of integrated of order 0 ($I(d)$ processes, $d = 0$). We consider in this paper an $I(d)$ autoregressive (AR) process, $d \geq 0$ is an unknown integer and the lag order may be finite or infinite. The number of lags considered, P_n, goes to infinity, when the sample size, n, does. While Sin and Yu (2016) show that Akaike's information criterion (AIC) is asymptotically inefficient (in terms of prediction) when the lag order is *finite*; this paper shows that when the lag order is *infinite* with algebraically decaying AR coefficients, neither Bayesian information criterion (BIC) nor Hanan Quninn information criterion (HQIC) is asymptotically inefficient. These results motivate us to combine the strengths of AIC and BIC/HQIC, yielding a so-called two-stage information criterion (TSIC) for a general $I(d)$ AR process. We show that TSIC is asymptotically efficient in the aforementioned two scenarios, as well as the scenario of exponentially decaying AR coefficients. This paper concludes with a simulation study which compares various information criteria with the least absolute shrinkage and selection operator (Lasso) and the adaptive Lasso. Although the (modified) Lasso-type methods perform comparably with, if not marginally outperform, the TSIC for some processes, the TSIC performs substantially better for some other processes.

On high-dimensional cross-validation

Ching-Kang Ing
Institute of Statistical Science, Academia Sinica

Cross validation (CV) has been one of the most popular methods for model selection. By splitting n data points into a training sample of size n_{c} and a validation sample of size n_{v} in which n_{v}/n approaches 1 and n_{c} tends to infinity, Shao (1993) showed that subset selection based on CV is consistent in a regression model of p candidate variables with p<< n. However, in the case of p >> n, not only does CV's consistency remain undeveloped, but subset selection is also practically infeasible. Instead of subset selection, in this talk, we suggest using CV as a backward elimination tool for excluding redundant variables that enter regression models through high-dimensional variable screening methods such as LASSO, LARS, ISIS and OGA. By choosing an n_{v} such that n_{v}/n converges to 1 at a rate faster than the one suggested by Shao (1993), we establish the desired consistency property. We further illustrate the finite sample performance of the proposed procedure via Monte Carlo simulations. Moreover, applications of our method to the analysis of wafer yields are also provided.

Classification and variable selection for high-dimensional data with applications to proteomics

Inge Koch

School of Mathematical Sciences, University of Adelaide

Australian Mathematical Sciences Institute, University of Melbourne

For two-class classification problems Fisher's discriminant rule performs well provided the dimension is smaller than the sample size. As the dimension increases, Fisher's rule may no longer be adequate, and can perform as poorly as random guessing. For high-dimension low sample size (HDLSS) data, dimension reduction and feature selection have become essential prior to applying any classification rule.

In this talk we look at different ways of incorporating feature selection into Fisher's classification rule and the nave Bayes rule including the 'Features Annealed Independence Rule' (FAIR) of Fan and Fan (2008), and the 'Nave Canonical Correlation' approach of Tamatani, Koch and Naito (2012). We examine the behavior of such rules and look at asymptotic properties as the dimension and sample size grow.

Proteomics is a rapidly growing research area within bioinformatics which focuses on identification of proteins, peptides and biomarkers from peptide concentrations. We consider proteomics imaging mass spectrometry data − HDLSS data with an underlying spatial distribution − here from patients with endometrial cancer. For these data we examine the performance of feature selection and classification rules in predicting which patients' cancer will metastasise.

Statistical inference in strongly spiked eigenvalue models

Kazuyoshi Yata and Makoto Aoshima

Institute of Mathematics, University of Tsukuba

1. Introduction

A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. This is the so-called "HDLSS" or "large p, small n" data, where p is the data dimension, n is the sample size and $p/n \to \infty$. Statistical inference on this type of data is becoming increasingly relevant, especially in the areas of medical diagnostics, engineering and other big data. Suppose we have independent samples of p-variate random variables from two populations, $\pi_i,\ i=1,2$, having an unknown mean vector $\boldsymbol{\mu}_i$ and unknown positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each π_i. We do not assume the normality of the population distributions. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ $(i=1,2)$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T = \sum_{j=1}^{p}\lambda_{ij}\boldsymbol{h}_{ij}\boldsymbol{h}_{ij}^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1},...,\lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$, and $\boldsymbol{H}_i = [\boldsymbol{h}_{i1},...,\boldsymbol{h}_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. Note that λ_{i1} is the largest eigenvalue of $\boldsymbol{\Sigma}_i$ for $i=1,2$. For the eigenvalues, Aoshima and Yata (2016) proposed the two disjoint models: the strongly spiked eigenvalue (SSE) model, which is defined by

$$\liminf_{p\to\infty}\left\{\frac{\lambda_{i1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}\right\} > 0 \quad \text{for } i=1 \text{ or } 2, \tag{1.1}$$

and the non-SSE (NSSE) model, which is defined by

$$\frac{\lambda_{i1}^2}{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \to 0 \quad \text{as } p\to\infty \text{ for } i=1,2. \tag{1.2}$$

In this talk, we considered statistical inference for high-dimensional data under the SSE model.

2. Two-sample test for SSE model

We consider the two-sample test:

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Chen and Qin (2010) and Aoshima and Yata (2011, 2015) considered the test under heteroscedasticity, $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Particularly, Aoshima and Yata (2011) proposed a test procedure to ensure prespecified accuracies regarding the size and power. We note that those two-sample tests were constructed under (1.2). However, if (1.1) is met, one cannot use those two-sample tests.

In this talk, we investigated the test statistic under the SSE model by considering strongly spiked eigenstructures. We created a new test procedure by estimating the eigenstructures for the SSE model.

3. Classifier for SSE model

We consider the distance-based classifier given by Aoshima and Yata (2014).

Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to one of the two populations. Aoshima and Yata (2014) considered the distance-based classifier as follows: Let

$$W(\boldsymbol{x}_0) = \Big(\boldsymbol{x}_0 - \frac{\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2}{2}\Big)^T (\overline{\boldsymbol{x}}_2 - \overline{\boldsymbol{x}}_1) - \frac{\mathrm{tr}(\boldsymbol{S}_{1n_1})}{2n_1} + \frac{\mathrm{tr}(\boldsymbol{S}_{2n_2})}{2n_2}.$$

Then, one classifies $\boldsymbol{x}_0$ into π_1 if $W(\boldsymbol{x}_0) < 0$ and into π_2 otherwise. Aoshima and Yata (2014) showed the asymptotic normality of $W(\boldsymbol{x}_0)$ under the NSSE model. However, the asymptotic normality does not hold under the SSE model.

In this talk, we proposed a new classifier by estimating the eigenstructures for the SSE model. We verified that the proposed classifier is asymptotically distributed as a normal distribution under the SSE model. We also showed that it gives preferable performances for the SSE model.

References

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

Aoshima, M. and Yata, K. (2014). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Ann. Inst. Stat. Math.* **66**, 983-1010.

Aoshima, M. and Yata, K. (2015). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodol. Comput. Appl. Probab.* **17**, 419-439.

Aoshima, M. and Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, to appear (arXiv:1602.02491).

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

Regression with auxiliary variables

Shinpei Imori

Graduate School of Engineering Science, Osaka University

1 Introduction

In this paper, we attempt to construct a parametric model for random variables of interest, which are called objective variables. Along with the objective variables, we sometime observe random variables related to the objective variables but of no interest, like as metadata. Such variables are referred to as auxiliary (or secondary) variables. For example, in Flickr (https://www.flickr.com), which is an image sharing web service, we can observe many images as well as their auxiliary variables including tags, geographical information and user information among others (see McAuley and Leskovec, 2012).

Because we can expect that the auxiliary variables have a lot of information about the objective variables, we try to improve the parametric model of the objective variables to fit a future observation by using the auxiliary variables. In fact, Mercatanti, Li & Mealli (2015) had already reported that the auxiliary variables are able to improve the precision of the parameter estimation in the Gaussian mixture model.

In a regression framework, a simple way to use the auxiliary variables is to incorporate them into the regression model as explanatory variables. Unless the true regression structure is completely explained by the objective variables, we will be able to expect that using auxiliary variables as the explanatory variables improves the accuracy of the regression model. However, if the future observation does not include the auxiliary variables, the regression model with the auxiliary variables cannot be applied to the future observation because this indicates that part of explanatory variables are missing. Indeed, it may be difficult to collect the metadata as mentioned before rather than the image data.

In this paper, we construct a regression model that can be applied to the future observation without the auxiliary variables although this regression model includes the information of the auxiliary variables on the objective variables. Concretely speaking, at first, we consider a joint model of the objective and the auxiliary variables, and then, we derive a marginal model of the objective variables from the joint model i.e., the auxiliary variables are deleted. We would like to note that the auxiliary variables are sometimes used in missing data analysis, however, this is different from our purpose.

2 Regression with auxiliary variable

In this section, we explain a regression framework with the auxiliary variables. Let Y be a response variable and X be a p-dimensional explanatory variable. The true conditional density function of Y given X is denoted by $q(y|x)$. Our aim is to construct a good regression model to estimate $q(y|x)$, and then a candidate model $p_y(y|x;\alpha)$ is considered where $\alpha \in \mathcal{A}$ is unknown parameters.

Here, we have the auxiliary variables A that are a q-dimensional random variable. A joint model of (Y, A) given X is assumed to be $p(y, a|x;\theta)$ where $\theta \in \Theta$ is unknown parameters. More directly, we consider the regression model of Y given (A, X) as $p_y(y|a, x;\theta)$ when we regard A as covariates. These models $p(y, a|x;\theta)$ and $p_y(y|a, x;\theta)$ may be able to explain the event of interest and to predict the future behavior of Y more precisely than $p_y(y|x;\alpha)$. However, since we allow the auxiliary variables not to be collected in the future observation, we do not apply the regression model $p_y(y|a, x; \theta)$ to the future observation. Thus, we consider an alternative model as follows:

$$p(y|x; \theta) \equiv \int p(y, a|x; \theta)da = \int p_y(y|x, a; \theta)p(a|x; \theta)da.$$

Hence, if we specify the regression model of A given X, $p(a|x; \theta)$, then we can define the marginal model $p(y|x; \theta)$. The unknown parameters θ are estimated from the joint model $p(y, a|x; \theta)$ in order to utilize the information of the auxiliary variables A.

Note that when the regression model $p_y(y|x; \alpha)$ is correctly specified, i.e., there exists $\alpha_0 \in \mathcal{A}$ such that $p_y(y|x; \alpha_0) = q(y|x)$, it may not need to consider the joint model $p(y, a|x; \theta)$ (or its marginal model $p(y|x; \theta)$) because the maximum likelihood estimator (MLE) of α will converge to the true value α_0.

3 Auxiliary variable selection

It follows form the results of the previous section that a goodness of fit of the regression model with the auxiliary variables depends strongly on the regression model for the auxiliary variables given the explanatory variables, $p(a|x; \theta)$. Hence, it needs to select the best regression model among candidate models.

In this study, the goodness of fit of the candidate model is measured by KL-divergence (Kullback and Leibler, 1951):

$$\mathcal{L}(\theta) = -\int q(y|x)q(x) \log \frac{p(y|x; \theta)}{q(y|x)} dydx,$$

whereas the unknown parameters θ will be estimated from $p(y, a|x; \theta)$. This implies that the adequacy of usual selection methods by information criteria such as AIC proposed by Akaike (1974) and BIC proposed by Schwarz (1978) is not guaranteed. Hence, we consider the new variable selection procedure, and we attempt to select the valid auxiliary variables and model.

References

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716–723.

2. Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79–86.

3. McAuley, J. & Leskovec, J. (2012, October). Image labeling on a network: using social-network metadata for image classification. In European Conference on Computer Vision, 828–841. Springer Berlin Heidelberg.

4. Mercatanti, A., Li, F. & Mealli, F. (2015). Improving inference of Gaussian mixtures using auxiliary variables. Statistical Analysis and Data Mining, 8(1), 34–48.

5. Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461–464.

Some convergence results of nonparametric tensor estimators

Taiji Suzuki
Tokyo Institute of Technology; JST, PRESTO; RIKEN, AIPC

Abstract

We investigate the statistical efficiency and computational complexity of some nonparametric estimators for a nonlinear tensor estimation problem. Low-rank tensor estimation has been used as a method to learn higher order relations among several data sources in a wide range of applications, such as multi-task learning, recommendation systems, and spatiotemporal analysis. We consider a general setting where a common linear tensor learning is extended to a nonlinear learning problem in reproducing kernel Hilbert space and propose two nonparametric estimators such as a Bayes estimator [5] and an alternating minimization procedure [10]. It is shown that the Bayes estimator achieves a near minimax optimal convergence rate without any strong convexity assumption, such as restricted strong convexity. We also show that the alternating minimization method achieves linear convergence as an optimization algorithm and that the generalization error of the resultant estimator yields the minimax optimality.

1 Problem formulation

Suppose that we are given n input-output samples $\{(x_i, y_i)\}_{i=1}^n$. The input x_i is a concatenation of K variables, i.e., $x_i = (x_i^{(1)}, \cdots, x_i^{(K)}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_K = \mathcal{X}$, where each $x_i^{(k)}$ is an element of a set $\mathcal{X}_k$. We consider the regression problem where these samples are generated according to the non-parametric model [8]:

$$y_i = \sum_{r=1}^{d} \prod_{k=1}^{K} f^*_{(r,k)}(x_i^{(k)}) + \epsilon_i, \tag{1}$$

where $\{\epsilon_i\}_{i=1}^n$ represents an i.i.d. zero-mean noise. In this regression problem, our objective is to estimate the true function $f^*(x^{(1)}, \ldots, x^{(K)}) = \sum_{r=1}^d \prod_{k=1}^K f^*_{(r,k)}(x^{(k)})$.

This model captures the effect of non-linear higher order interactions among the input components $\{x^{(k)}\}_{k=1}^K$ to the output y, and thus, is useful for a regression problem where the output is determined by complex relations between the input components. This type of regression problem appears in several applications, such as multi-task learning, recommendation systems and spatiotemporal data analysis [6, 7, 1].

To understand the model in Eq. (1), it is helpful to consider a linear case as a special case [2, 11]. In general, the linear tensor model is formulated as

$$Y_i = \langle A^*, X_i \rangle + \epsilon_i. \tag{2}$$

Here, X_i, A^* are tensors in $\mathbb{R}^{M_1 \times \cdots \times M_K}$ and the inner product $\langle \cdot, \cdot \rangle$ is defined by $\langle A, X \rangle = \sum_{i_1,\ldots,i_K=1}^{M_1,\ldots,M_K} A_{i_1 \ldots i_K} X_{i_1 \ldots i_K}$. A^* is assumed to be low rank in the sense of CP-rank [3, 4], i.e., A^* is decomposed as $\sum_{r=1}^d u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}$, where the vector $u_r^{*(k)} \in \mathbb{R}^{M_k}$ and the symbol $\circ$ represents the vector outer product. If we also assume X_i is rank-1, i.e., $X_i = x_i^{(1)} \circ \cdots \circ x_i^{(K)}$, then the inner product in Eq.(2) is written as: $\langle A^*, X_i \rangle = \left\langle \sum_{r=1}^d u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}, x_i^{(1)} \circ \cdots \circ x_i^{(K)} \right\rangle = \sum_{r=1}^d \prod_{k=1}^K \left\langle u_r^{*(k)}, x_i^{(k)} \right\rangle$. This is equivalent to the case where we limit $f^*_{(r,k)}$ in Eq. (1) to the linear function $\left\langle u_r^{*(k)}, x^{(k)} \right\rangle$. Hence, the linear model based on CP-decomposition can be understood as a special case of our proposed model.

[5] proposed a nonparametric Bayesian estimator and [10] proposed an alternating minimization estimator for the nonlinear model (1).

2 Convergence analysis

[5, 10] have shown that the convergence rates of the predictive risks of the Bayes estimator and the alternating minimization achieve the minimax optimal rate.

Let $\|\widehat{f} - f^*\|_n^2 := \frac{1}{n} \sum_{i=1}^n (\widehat{f}(x_i) - f^*(x_i))^2$ and $s_{(r,k)}$ be the covering number exponent of the RKHS in which the true function $f^*_{(r,k)}$ is supposed to be included. Now let $\widehat{f}$ be the Bayes estimator proposed in [5]. Then the following risk bound is obtained.

Theorem 1 Under some assumptions, there exists a constant $C > 0$ such that

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\widehat{f} - f^*\|_n^2\right] \le C\left\{\sum_{r=1}^{d}\sum_{k=1}^{K} n^{-\frac{1}{1+s_{(r,k)}}} + \frac{d}{n}\log\left(\frac{1}{\kappa}\right)\right\},$$

where $\mathrm{E}_{Y_{1:n}|x_{1:n}}$ indicates the expectation with respect to the outputs $Y_1, \dots, Y_n$ conditioned by the inputs $x_1, \dots, x_n$, and κ is a constant that is determined by the prior distribution of the rank.

It was shown that this bound actually achieves the minimax optimal rate [5].

On the other hand, the convergence rate of the alternating minimization method proposed [10] have been also analyzed. Let $(\widehat{f}^{(t)}, \hat{v}^{(t)})$ be the estimator at the tth iteration of the alternating minimization method. Then the following risk bound of the alternating minimization method is obtained [10].

Theorem 2 Suppose that $(\widehat{f}^{(1)}, \hat{v}^{(1)})$ is sufficiently close to the true function f^*, then we have

$$\|\check{f}^{(t)} - f^*\|_{L_2}^2 = O\left(\tau dKn^{-\frac{1}{1+s}}\log(dK) + \tau dK\left(\frac{3}{4}\right)^t\right)$$

with probability $1 - 3\exp(-\tau)$.

This theorem indicates that after $T = O(\log(n))$ iterations, we obtain the estimation accuracy of $O(dKn^{-\frac{1}{1+s}}\log(dK))$. The estimation accuracy bound $O(dKn^{-\frac{1}{1+s}}\log(dK))$ is intuitively natural because we are estimating $d \times K$ functions $\{f^*_{(r,k)}\}_{r,k}$ and the optimal sample complexity to estimate one function $f^*_{(r,k)}$ is known as $n^{-\frac{1}{1+s}}$ [9]. Indeed, this accuracy bound is minimax optimal [5].

References

[1] M. T. Bahadori, Q. R. Yu, and Y. Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in Neural Information Processing Systems 27*.

[2] W. Chu and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR Workshop and Conference Proceedings*, 2009.

[3] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6:164–189, 1927.

[4] F. L. Hitchcock. Multilple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7:39–79, 1927.

[5] H. Kanagawa, T. Suzuki, H. Kobayashi, N. Shimizu, and Y. Tagami. Gaussian process nonparametric tensor estimator and its minimax optimality. In *International Conference on Machine Learning (ICML2016)*, pages 1632–1641, 2016.

[6] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems 2010*, pages 79–86, 2010.

[7] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML2013)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1444–1452, 2013.

[8] M. Signoretto, L. D. Lathauwer, and J. A. K. Suykens. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. *CoRR*, abs/1310.4977, 2013.

[9] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.

[10] T. Suzuki, H. Kanagawa, H. Kobayashi, N. Shimizu, and Y. Tagami. Minimax optimal alternating minimization for kernel nonparametric tensor learning. In *Annual Conference on Neural Information Processing Systems (NIPS2016)*, page to appear, 2016.

[11] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems 24*, pages 972–980, 2011. NIPS2011.