International Symposium on Statistical Theory and Methodology for Large Complex Data

November 16-18, 2018

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)
Aki Ishii	(Tokyo University of Science)

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 Research (Exploratory) 17K19956 (Project Period: 2017-2018)
"New developments for big data by non-sparse modeling"
(Principal Investigator: Makoto Aoshima)

Grant-in-Aid for Scientific Research (S) 18H05290 (Project Period: 2018-2022)
"Introduction of general causality to various observations and the innovation for its optimal statistical inference"
(Principal Investigator: Masanobu Taniguchi)

Program

November 16 (Friday)

14:00∼14:05 Opening

14:05∼14:45 Aki Ishii*,[a], Kazuyoshi Yata[b] and Makoto Aoshima[b]
[a](Department of Information Sciences, Tokyo University of Science)
[b](Institute of Mathematics, University of Tsukuba)
Tests of high-dimensional mean vectors under the SSE model

14:55∼15:35 Hiroumi Misaki (Faculty of Engineering, Information and Systems, University of Tsukuba)
Comparison of financial volatility estimators: RK, TS, PA and SIML

15:50∼16:30 Shota Katayama
(Department of Industrial Engineering and Economics, Tokyo Institute of Technology)
Robust and sparse Gaussian graphical modelling under cell-wise contamination

16:40∼17:20 Mariko Yamamura (Graduate School of Education, Hiroshima University)

Estimation of spatiotemporal effects by the fused lasso for densely sampled spatial data using body condition data set from common minke whales

November 17 (Saturday)

9:25∼10:05 Kazuyoshi Yata* and Makoto Aoshima (Institute of Mathematics, University of Tsukuba)

A high-dimensional quadratic classifier after feature selection

10:15∼10:55 Junichi Hirukawa[a,*] and Sangyeol Lee[b]

[a](Faculty of Science, Niigata University)

[b](Department of Statistics, Seoul National University)

Asymptotic properties of mildly explosive processes with locally stationary disturbance

11:05∼11:45 Hirokazu Yanagihara

(Department of Mathematics, Graduate School of Science, Hiroshima University)

High-dimensionality adjusted asymptotically loss efficient GCp criterion in normal multivariate linear regression models

11:45∼13:15 Lunch

13:15∼17:30 **Special Invited and Keynote Sessions**

18:30∼ Dinner

November 18 (Sunday)

9:40∼10:20 Kei Hirose[a,*] and Hiroki Masuda[b]

[a](Institute of Mathematics for Industry, Kyushu University)

[b](Faculty of Mathematics, Kyushu University)

Robust relative error estimation

10:30∼11:10 Tomonari Sei (Graduate School of Information Science and Technology, University of Tokyo)

Inconsistency of diagonal scaling under high-dimensional limit

11:20∼12:00 Takafumi Kanamori

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

Statistical inference with unnormalized models

12:00∼ 12:10 Closing

(∗ Speaker)

Special Invited Session

13:15～14:05 **Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage**

Speaker: Johan Lim
(Department of Statistics, Seoul National University)

14:20～15:10 **Greedy active learning algorithm for logistic regression models**

Speaker: Ray-Bing Chen
(Department of Statistics, National Cheng Kung University)

15:25～16:15 **Inference for LSHD time series models and applications to sensor monitoring and financial engineering**

Speaker: Ansgar Steland
(Institut fuer Statistik und Wirtschaftsmathematik, RWTH Aachen University)

Keynote Session

16:30～17:30 **Do we necessarily have more cost-effective information with big data?**

Speaker: Nitis Mukhopadhyay
(Department of Statistics, University of Connecticut-Storrs)

Discussion Leader: Chikara Uno
(Faculty of Education and Human Studies, Akita University)

Tests of high-dimensional mean vectors under the SSE model

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

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

1 Introduction

We considered statistical inference on mean vectors in the high-dimension, low-sample-size (HDLSS) context. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be a random sample of size $n(\geq 4)$ from a p-variate distribution with an unknown mean vector $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. In the HDLSS context, the data dimension p is very high and n is much smaller than p. We define the eigen-decomposition of $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ is a diagonal matrix of eigenvalues, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, and $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ is an orthogonal matrix of the corresponding eigenvectors. We write the sample mean vector and the sample covariance matrix as $\overline{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j / n$ and $\boldsymbol{S} = \sum_{j=1}^n (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^T/(n-1)$.

In this talk, we discussed the one-sample test:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{1.1}$$

where $\boldsymbol{\mu}_0$ is a candidate mean vector. We assume $\boldsymbol{\mu}_0 = \boldsymbol{0}$ without loss of generality. One should note that Hotelling's T^2-statistic is not available because $\boldsymbol{S}^{-1}$ does not exist in the HDLSS context. [6, 7] considered the test when $\boldsymbol{X}$ is Gaussian. When $\boldsymbol{X}$ is non-Gaussian, [5] considered the test. Let us consider the following eigenvalue condition:

$$\frac{\lambda_1^2}{\text{tr}(\boldsymbol{\Sigma}^2)} \to 0 \ \text{ as } p \to \infty. \tag{1.2}$$

Under (1.2), H_0 and some regularity conditions, [5] and [1, 2] showed the asymptotic normality for their test statistics.

[3] called the eigenvalue condition (1.2) the "non-strongly spiked eigenvalue (NSSE) model" and drew attention that high-dimensional data do not fit the NSSE model on several occasions. In order to overcome this inconvenience, [3] proposed the "strongly spiked eigenvalue (SSE) model" defined by

$$\liminf_{p\to\infty} \left\{ \frac{\lambda_1^2}{\text{tr}(\boldsymbol{\Sigma}^2)} \right\} > 0 \tag{1.3}$$

and gave a data transformation technique from the SSE model to the NSSE model.

2 A new test procedure under the SSE model

Let $\Psi_r = \text{tr}(\boldsymbol{\Sigma}^2) - \sum_{s=1}^{r-1} \lambda_s^2 = \sum_{s=r}^{p} \lambda_s^2$ for $r = 1, ..., p$. We assumed the following model:

(A-ii) There exists a fixed integer k (≥ 1) such that

(i) When $k \geq 2$, $\lambda_1, ..., \lambda_k$ are distinct in the sense that

$$\liminf_{p\to\infty} (\lambda_r/\lambda_s - 1) > 0 \ \text{ for } 1 \leq r < s \leq k;$$

(ii) λ_k and λ_{k+1} satisfy

$$\liminf_{p\to\infty} \frac{\lambda_k^2}{\Psi_k} > 0 \ \text{ and } \ \frac{\lambda_{k+1}^2}{\Psi_{k+1}} \to 0 \ \text{ as } p \to \infty.$$

Note that (A-ii) is one of the SSE models. According to [3, 4], we considered transforming the data from the SSE model to the NSSE model by using the projection matrix

$$\boldsymbol{A} = \boldsymbol{I}_p - \sum_{j=1}^{k} \boldsymbol{h}_j \boldsymbol{h}_j^T = \sum_{j=k+1}^{p} \boldsymbol{h}_j \boldsymbol{h}_j^T.$$

We have that $E(\boldsymbol{A}\boldsymbol{x}_j) = \boldsymbol{A}\boldsymbol{\mu}$ $(= \boldsymbol{\mu}_*,$ say) and

$$\text{Var}(\boldsymbol{A}\boldsymbol{x}_j) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A} = \sum_{j=k+1}^{p} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T \ (= \boldsymbol{\Sigma}_*, \text{ say}).$$

Note that $\text{tr}(\boldsymbol{\Sigma}_*^2) = \Psi_{k+1}$ and $\lambda_{\max}(\boldsymbol{\Sigma}_*) = \lambda_{k+1}$, where $\lambda_{\max}(\boldsymbol{\Sigma}_*)$ denotes the largest eigenvalue of $\boldsymbol{\Sigma}_*$. Then, it holds that

$$\lambda_{\max}^2(\boldsymbol{\Sigma}_*)/\text{tr}(\boldsymbol{\Sigma}_*^2) \to 0 \ \text{ as } p \to \infty \text{ under (A-ii).}$$

Thus, the transformed data has the NSSE model.

By using the transformed data, we consider the following quantity:

$$T_{\text{DT}} = \|\boldsymbol{A}\overline{\boldsymbol{x}}\|^2 - \frac{\text{tr}(\boldsymbol{A}\boldsymbol{S})}{n} = 2\frac{\sum_{l<l'}^{n} \boldsymbol{x}_l^T \boldsymbol{A} \boldsymbol{x}_{l'}}{n(n-1)} = 2\frac{\sum_{l<l'}^{n} \left(\boldsymbol{x}_l^T \boldsymbol{x}_{l'} - \sum_{j=1}^{k} x_{jl} x_{jl'}\right)}{n(n-1)},$$

where

$$x_{jl} = \boldsymbol{h}_j^T \boldsymbol{x}_l \ \text{ for all } j, l.$$

We discussed the asymptotic null distribution and the power of T_{DT}. We applied the findings to the construction of confidence regions on the mean vector under the SSE model. We further discussed multi-sample problems under the SSE models. Finally, we demonstrated the new test procedure by using actual microarray data sets.

References

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

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

[4] Aoshima, M., Yata, K. (2018b). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics,* in press (doi:10.1007/s10463-018-0655-z).

[5] Bai, Z., Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica, 6,* 311–329.

[6] Dempster, A.P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics, 29,* 995–1010.

[7] Dempster, A.P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics, 16,* 41–50.

Comparison of financial volatility estimators: RK, TS, PA and SIML

Hiroumi Misaki [1]

In the recent decade, several estimation methods of volatility of financial asset prices with high-frequency data have been developed to deal with the problem of market microstructure noise (MMN). Bandorff-Nielsen *et al.* (2008) proposed the *realized kernel* (RK) method using autocovariances and kernels, which is widely used for empirical analysis in academic literatures. The *two-scale* (TS) estimator by Zhang *et al.* (2005) combines sparsely and densely (sub)sampled realized volatility to eliminate the bias. Jacod *et al.* (2009) developed *pre-averaging* (PA) approach for the estimation problem with MMN.

Kunitomo and Sato (2011, 2013) have proposed the *separating information maximum likelihood* (SIML) method. Misaki and Kunitomo (2015) and Kunitomo, Misaki and Sato (2015) have further investigated the properties of the SIML estimation with the data that is randomly sampled and includes MMN. The analytical investigations and the batch of simulations have shown that the SIML estimator has reasonable asymptotic properties as well as finite sample properties.

Although some alternative methods have been proposed as described above, the relative properties of these methods in actual market structures are yet to be assessed. Therefore, in this presentation, we make a comparative study on RK, TS, PA and SIML.

We assume that the underlying continuous process $X(t)$ $(0 \leq t \leq 1)$ is that

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

where $B(s)$ represents standard Brownian motion and $\sigma_x(s)$ is the instantaneous volatility function. The main statistical goal 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 $y(t_i^n)$ that are generated by $y(t_i^n) = h\left(X(t_i^n), y(t_{i-1}^n), u(t_i^n)\right)$, where $u(t_i^n)$ is market MMN.

We investigate the finite sample properties of the five estimators for the integrated volatility based on a set of simulations. Generally, we can assume deterministic volatility function $\sigma_x^2(s) = \sigma(0)^2 \left[a_0 + a_1 s + a_2 s^2\right]$ where a_i $(i = 0, 1, 2)$ are constants and $\sigma_x(s)^2 > 0$ for $s \in [0, 1]$. In this case the integrated volatility is given by

$$\sigma_x^2 = \int_0^1 \sigma_x(s)^2 ds = \sigma_x(0)^2 \left[a_0 + \frac{a_1}{2} + \frac{a_2}{3}\right] .$$

[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

We use several non-linear transformation models for the form of MMN in addition to the standard 'addtive plus noise' model: $h(x,y,u) = x + u$. Each model corresponds to

$$
\begin{aligned}
&\text{Model 1} \qquad && h(x,y,u) = g_\eta(x+u), \\
&\text{Model 2} \qquad && h(x,y,u) = y + b(x-y), \\
&\text{Model 3} \qquad && h(x,y,u) = y + g_\eta(x-y+u), \\
&\text{Model 4} \qquad && h(x,y,u) = y + g_\eta(x-y) + u, \\
&\text{Model 5} \qquad && h(x,y,u) = y + u + \begin{cases} b_1(x-y) & \text{if } x \geq y \\ b_2(x-y) & \text{if } y < x \end{cases} ,
\end{aligned}
$$

where $g_\eta(x) = \eta\,[x/\eta]$, and b, b_1, b_2 are constants, respectively.

In the Tables we show the some of simulation results except for PA. Figures of MSE for each estimator are attached. As we can see from the Figures, the RK is severely biased to the round-off errors when the noise is small, whereas the TS is sensitive to the linear price adjustment when the noise is relatively large. The SIML and LSIML are not the best in most of the cases including the standard one, but they seem to have robustness to the form of noise, compared to the other estimators. We have shown the more findings at the session.

References

[1] Barndorff-Nielsen, O., P. Hansen, A. Lunde and N. Shephard (2008), "Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise," *Econometrica*, Vol.76-6, 1481-1536.

[2] Jacod, J., Y. Li, P. A. Mykland, M. Podolskij and M. Vetter (2009), "Microstructure noise in the continuous case: The pre-averagin approach," *Stochastic Processes and their Applications*, Vol. 119, 2249-2276.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Zhang, L. , P. Mykland and Ait-Sahalia, Y., (2005), "A tale for two time scale : determining integrated volatility with noisy-high frequency data," *Journal of the American Statistical Association*, Vol. 100(472), 1394-1411.

Robust and sparse Gaussian graphical modelling under cell-wise contamination

Shota Katayama[1], Hironori Fujisawa[2] and Mathias Drton[3]

[1]Tokyo Institute of Technology, Japan
[2]The Institute of Statistical Mathematics, Japan
[3]University of Washington, USA

Let $\boldsymbol{Y} = (Y_1, \dots, Y_p)^T$ be a p-dimensional random vector representing a multivariate observation. The conditional independence graph of $\boldsymbol{Y}$ is the undirected graph $G = (V, E)$ whose vertex set $V = \{1, \dots, p\}$ indexes the individual variables and whose edge set E indicates conditional dependences among them. More precisely, $(i, j) \notin E$ if and only if Y_i and Y_j are conditionally independent given $Y_{V\setminus\{i,j\}} = \{Y_k : k \neq i, j\}$. For a Gaussian vector, the edge set E corresponds to the support of the precision matrix. Indeed, it is well known that if $\boldsymbol{Y}$ follows a multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then $(i, j) \notin E$ if and only if $\boldsymbol{\Omega}_{ij} = 0$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

Inference of the conditional independence graph sheds light on direct as opposed to indirect interactions and has received much recent attention (Drton and Maathuis, 2017). In particular, for high-dimensional Gaussian problems, several techniques have been developed that exploit available sparsity in inference of the support of the precision matrix $\boldsymbol{\Omega}$. Meinshausen and Bühlmann (2006) suggested fitting node-wise linear regression models with ℓ_1 penalty to recover the support of each row. Yuan and Lin (2007), Benerjee et al. (2008) and Friedman et al. (2008) considered the graphical lasso (Glasso) that involves the ℓ_1 penalized log-likelihood function. Cai et al. (2011) proposed the constrained ℓ_1 minimization for inverse matrix estimation (CLIME), which may be formulated as a linear program.

In fields such as bioinformatics and economics, data are often not only high-dimensional but also subject to contamination. While suitable for high dimensionality, the above mentioned techniques are sensitive to contamination. Moreover, traditional robust methods may not be appropriate when the number of variables is large. Indeed, they are based on the model in which an observation vector is either without contamination or fully contaminated. Hence, an observation vector is treated as an outlier even if only one of many variables is contaminated. As a result these methods down-weight the entire vector regardless of whether it contains 'clean' values for some variables. Such information loss can become fatal as the dimension increases. As a more realistic model in high dimensional data, Alqallaf et al. (2002) considered cell-wise contamination: the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with p variables are generated by

$$\boldsymbol{X}_i = (\boldsymbol{I}_p - \boldsymbol{E}_i)\boldsymbol{Y}_i + \boldsymbol{E}_i \boldsymbol{Z}_i, \quad i = 1, \dots, n. \tag{1}$$

Here, $\boldsymbol{I}_p$ is the $p \times p$ identity matrix and each $\boldsymbol{E}_i = \mathrm{diag}(E_{i1}, \dots, E_{ip})$ is a diagonal random matrix with the E_{ij}'s independent and Bernoulli distributed with $P(E_{ij} = 1) = \varepsilon_j$. The random vectors $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ are independent, and $\boldsymbol{Y}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ corresponds to a clean sample while $\boldsymbol{Z}_i$ makes contaminations in some elements of $\boldsymbol{X}_i$.

Our goal is to develop a robust estimation method for the conditional independence graph G of $\boldsymbol{Y}_i$ from the cell-wise contaminated observations $\boldsymbol{X}_i$. Techniques such as node-wise regression, Glasso and CLIME process an estimate of the covariance matrix. Our strategy is thus simply to apply these procedures using a covariance matrix estimator that is robust against cell-wise contamination. However, while many researchers have considered the traditional 'whole-vector' contamination framework (see, e.g., Maronna et al., 2006), there are fewer existing methods for cell-wise contamination. Specifically, we are aware of three approaches, namely, use of alternative t-distributions Finegold and Drton (2011), use of rank correlations (Loh and Tan, 2015; Öllerer and Croux, 2015), and a pairwise covariance estimation method by Tarr et al. (2016) who adopt an idea of Gnanadesikan and Kettenring (1972). In contrast, in this talk, we provide a robust covariance matrix estimator via γ-divergence as proposed by Fujisawa and Eguchi (2008). The γ-divergence can automatically reduce the impact of contaminations, and it is known to be robust even when the number of contaminations is large.

Estimation of spatiotemporal effects by the fused lasso for densely sampled spatial data using body condition data set from common minke whales

Mariko Yamamura[1], Hirokazu Yanagihara[2], Keisuke Fukui[3],

Hiroko Solvang[4], Nils Øien[4], Tore Haug[4]

[1]Graduate School of Education, Hiroshima University, [2]Graduate School of Science, Hiroshima University, [3]Research & Development Center, Osaka Medical College,

[4]Institute of Marine Research, Norway,

Samples evenly distributed all over the population are not always available for real data analysis. As an example of a spatial data, a data from common mink whales in Norwegian water provides values showing their body conditions with whaling locations such as longitudes and latitudes. Though whales are distributed all over the Norwegian water, whaling locations are almost the same every year, therefore samples are dense at particular locations. The space to be analyzed is subdivided into several, and we estimate the spatial effect by using fused lasso with combining spatial effects from subdivided space.

Let $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ be n-dimensional vectors obtained by stacking the vectors of response variables and error variables of the j-th space, respectively, i.e., $\boldsymbol{y} = (\boldsymbol{y}_1', \dots, \boldsymbol{y}_m')'$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \dots, \boldsymbol{\varepsilon}_m')'$, and let $\boldsymbol{X}$ and $\boldsymbol{B}$ be $n \times k$ and $n \times b$ matrices obtained by stacking the matrices of explanatory variables and basis functions of the j-th space, respectively, i.e., $\boldsymbol{X} = (\boldsymbol{X}_1', \dots, \boldsymbol{X}_m')'$ and $\boldsymbol{B} = (\boldsymbol{B}_1', \dots, \boldsymbol{B}_m')'$. As a whole of space, the additive model is written such as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{B\alpha} + \boldsymbol{R\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)'$ and $\boldsymbol{R}$ is an $n \times m$ matrix defined by

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{1}_{n_1} \otimes \boldsymbol{e}_1' \\ \vdots \\ \boldsymbol{1}_{n_m} \otimes \boldsymbol{e}_m' \end{pmatrix}.$$

Here, $\boldsymbol{e}_j$ is the m-dimensional vector of which the j-th element is 1 while all the other elements are 0, and $\otimes$ indicates the Kronecker product of the two matrices.

Yanagihara (2012) shows that choosing the smoothing parameters in the penalized smoothing spline is equivalent to choosing the ridge parameters in the generalized ridge regression using the matrix of transformed basis function values as the matrix of explanatory variables. And then Yanagihara (2018) considers optimization of the ridge parameters in generalized ridge regression by minimizing a model selection criterion, i.e., generalized cross-validation (GCV). From Yanagihara (2012, 2018), estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ after optimizing smoothing parameters by GCV are given by

$$\hat{\boldsymbol{\alpha}}_\gamma = \boldsymbol{Q}_\gamma \boldsymbol{V}_\gamma \boldsymbol{D}_\gamma^{-1/2} \boldsymbol{z}_\gamma, \quad \hat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{R}\hat{\boldsymbol{\mu}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}}_\gamma),$$

respectively, where $\boldsymbol{z}_\gamma = (z_1, \ldots, z_\gamma)'$, and $\boldsymbol{V}_\gamma$ is a $\gamma \times \gamma$ diagonal matrix as

$$\boldsymbol{V}_\gamma = \mathrm{diag}(\nu_{\gamma,1}, \ldots, \nu_{\gamma,\gamma}), \quad \nu_{\gamma,j} = I\left(z_j^2 > s^2_{\gamma,a^*_\gamma}\right)\left(1 - \frac{s^2_{\gamma,a^*_\gamma}}{z_j^2}\right) \ (j = 1, \ldots, \gamma),$$

where the γ is optimized by the GCV.

The penalized residual sum of squares (PRSS_λ) for the adaptive fused lasso is given by

$$\mathrm{PRSS}_\lambda(\boldsymbol{\mu}|\hat{f}) = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{B}\hat{\boldsymbol{\alpha}} - \boldsymbol{R}\boldsymbol{\mu}\|^2 + \lambda \sum_{j=1}^{m} \sum_{\ell \in \mathcal{D}_j} w_{j\ell}|\mu_j - \mu_\ell|,$$

where λ is the non-negative regularization parameter. The spatial effect $\boldsymbol{\mu}$ is estimated by minimizing PRSS_λ as

$$\hat{\boldsymbol{\mu}}_\lambda = \arg\min_{\boldsymbol{\mu} \in \mathbb{R}^m} \mathrm{PRSS}_\lambda(\boldsymbol{\mu}|\hat{f}).$$

The above minimization problem can be solved by the coordinate descent algorithm in Friedman *et al.* (2007).

There is the geographic distribution of the five International Whaling Commission (IWC) management areas. We subdivide each areas to have about 300 samples, and estimate $\hat{\boldsymbol{\mu}}$ of subdivided areas. If the $\hat{\boldsymbol{\mu}}$ of a subdivided area is equal to the one of its neighbor area, the subdivided area and the neighbor are united. From the fused lasso estimation result, the subdivided areas are narrowed down the 11 spaces, and whales have the thickest blubber in the northernmost space.

References

[1] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1**, 302–332.

[2] Yanagihara, H. (2012). A non-iterative optimization method for smoothness in penalized spline regression. *Statistics and Computing*, **22**, 527–544.

[3] Yanagihara, H. (2018). Explicit solution to the minimization problem of generalized cross-validation criterion for selecting ridge parameters in generalized ridge regression. *Hiroshima Mathematical Journal*, **48**, 203–222.

A high-dimensional quadratic classifier after feature selection

Kazuyoshi Yata and Makoto Aoshima
Institute of Mathematics, University of Tsukuba, Ibaraki, Japan

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 situation, here p is the data dimension and n is the sample size. In this paper, we mainly focus on the case when "$n/p \to 0$". Suppose we have independent and p-variate two populations, $\pi_i,\ i=1,2$, having an unknown mean vector $\boldsymbol{\mu}_i=(\mu_{i1},...,\mu_{ip})^T$ and unknown positive-definite covariance matrix $\boldsymbol{\Sigma}_i$ for each i. Let

$$\boldsymbol{\mu}_{12}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=(\mu_{121},...,\mu_{12p})^T.$$

We assume $\limsup_{p\to\infty}|\mu_{12j}|<\infty$ for all j. Note that $\limsup_{p\to\infty}||\boldsymbol{\mu}_{12}||^2/p<\infty$, where $||\cdot||$ denotes the Euclidean norm. Let $\sigma_{i(j)}$ be the j-th diagonal element of $\boldsymbol{\Sigma}_i$ for $j=1,...,p$ $(i=1,2)$. We assume that $\sigma_{i(j)}\in(0,\infty)$ as $p\to\infty$ for all i,j. For a function, $f(\cdot)$, "$f(p)\in(0,\infty)$ as $p\to\infty$" implies that $\liminf_{p\to\infty}f(p)>0$ and $\limsup_{p\to\infty}f(p)<\infty$. Here, "$\liminf_{p\to\infty}f(p)$" and "$\limsup_{p\to\infty}f(p)$" are the limit inferior and the limit superior of $f(p)$, respectively. Then, it holds that $\text{tr}(\boldsymbol{\Sigma}_i)/p\in(0,\infty)$ as $p\to\infty$ for $i=1,2$. We do not assume $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i=\boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T$, where $\boldsymbol{\Lambda}_i=\text{diag}(\lambda_{i1},...,\lambda_{ip})$ is a diagonal matrix of eigenvalues, $\lambda_{i1}\ge\cdots\ge\lambda_{ip}>0$, and $\boldsymbol{H}_i=[\boldsymbol{h}_{i1},...,\boldsymbol{h}_{ip}]$ is an orthogonal matrix of the corresponding eigenvectors. We have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}$, from each π_i, where $\boldsymbol{x}_{ik}=(x_{i1k},...,x_{ipk})^T,\ k=1,...,n_i$. We assume $n_i\ge 2,\ i=1,2$. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by

$$\overline{\boldsymbol{x}}_{in_i}=(\overline{x}_{i1n_i},...,\overline{x}_{ipn_i})^T=\sum_{k=1}^{n_i}\boldsymbol{x}_{ik}/n_i$$

and $\boldsymbol{S}_{in_i}=\sum_{k=1}^{n_i}(\boldsymbol{x}_{ik}-\overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ik}-\overline{\boldsymbol{x}}_{in_i})^T/(n_i-1)$. Let $s_{in_i(j)}$ be the j-th diagonal element of $\boldsymbol{S}_{in_i}$ for $j=1,...,p$ $(i=1,2)$. Let $\boldsymbol{x}_0=(x_{01},...,x_{0p})^T$ be an observation vector of an individual belonging to one of the two populations. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. Let

$$n_{\min}=\min\{n_1,n_2\}\ \text{ and }\ m=\min\{p,n_{\min}\}.$$

Note that the divergence condition "$p\to\infty$, $n_1\to\infty$ and $n_2\to\infty$" is equivalent to "$m\to\infty$.

Let $|\boldsymbol{M}|$ be the determinant of a square matrix $\boldsymbol{M}$. When π_is are Gaussian, the Bayes optimal rule (the minimum-error rate discriminant function) is given as follows: One classifies the individual into π_1 if

$$(\boldsymbol{x}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_1)-\log|\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}|<(\boldsymbol{x}_0-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}_0-\boldsymbol{\mu}_2) \tag{1.1}$$

and into π_2 otherwise. Since $\boldsymbol{\mu}_i$s and $\boldsymbol{\Sigma}_i$s are unknown, one usually considers the following typical classifier:

$$(\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1})^T \boldsymbol{S}_{1n_1}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{1n_1}) - \log |\boldsymbol{S}_{2n_2} \boldsymbol{S}_{1n_1}^{-1}| < (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2})^T \boldsymbol{S}_{2n_2}^{-1} (\boldsymbol{x}_0 - \overline{\boldsymbol{x}}_{2n_2}).$$

The classifier usually converges to the Bayes optimal classifier when $n_{\min} \to \infty$ while p is fixed or $n_{\min}/p \to \infty$. However, in the HDLSS context, the inverse matrix of $\boldsymbol{S}_{in_i}$ does not exist. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Bickel and Levina (2004) considered an inverse matrix defined by only diagonal elements of the pooled sample covariance matrix. When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Dudoit et al. (2002) considered an inverse matrix defined by only diagonal elements of $\boldsymbol{S}_{in_i}$. Aoshima and Yata (2011) considered using $\{\mathrm{tr}(\boldsymbol{S}_{in_i})/p\}^{-1}\boldsymbol{I}_p$ instead of $\boldsymbol{S}_{in_i}^{-1}$ from a geometrical background of HDLSS data and proposed geometric classifiers. Here, $\boldsymbol{I}_p$ denotes the identity matrix of dimension p. Chan and Hall (2009) and Aoshima and Yata (2014, 2018) considered distance-based classifiers and Aoshima and Yata (2014) gave the misclassification rate adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds.

In this talk, we considered classifiers by the diagonal elements of $\boldsymbol{S}_{in_i}$. We provided a DQDA type classifier by feature selection and show that it has the consistency property even when $n_{\min}/p \to 0$.

References

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

[2] Aoshima M, Yata K (2014) A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. Ann I Stat Math 66:983–1010

[3] Aoshima M, Yata K (2018) Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. Ann I Stat Math, in press (doi:10.1007/s10463-018-0655-z)

[4] Bickel PJ, Levina E (2004) Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10:989–1010

[5] Chan YB, Hall P (2009) Scale adjustments for classifiers in high-dimensional, low sample size settings. Biometrika 96:469–478

[6] Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97:77–87

Asymptotic Properties of Mildly Explosive Processes with Locally Stationary Disturbance

Junichi Hirukawa and Sangyeol Lee
Niigata University and Seoul National University

1. Introduction Autoregressive processes of the form $y_t = \rho y_{t-1} + \varepsilon_t$ with an explosive root $|\rho| > 1$, where ε_t are iid standard normal random variables, were first studied in White (1958) and Anderson (1959). By assuming a zero initial value for y_t, a Cauchy limit theory is derived for the least squares estimate $\widehat{\rho}_n = (\sum_{t=1}^n y_{t-1}y_t)(\sum_{t=1}^n y_{t-1}^2)^{-1}$:

$$\frac{\rho^n}{\rho^2 - 1}(\widehat{\rho}_n - \rho) \Rightarrow C, \tag{1}$$

where C denotes a Cauchy random variable and $\Rightarrow$ denotes convergence in distribution when n tends to infinity. The Gaussian assumption imposed on the innovation sequence $\{\varepsilon_t\}$ plays an important role and cannot be relaxed to obtain the same asymptotic distribution as in (1): see Anderson (1959) who demonstrate that the limit distribution of the least squares estimate depends upon the distributional assumptions imposed on the error sequence.

However, this difficulty can be avoided when the explosive root approaches unity as the sample size tends to infinity. Phillips and Magdalinos (2007a) consider autoregressive processes with root $\rho_n = 1 + c/k_n$, where k_n is a positive real sequence with $k_n = o(n)$. When $c > 0$, such roots are explosive in finite samples and approach unity at the rate slower than $O(n^{-1})$. It is well known that the asymptotic behavior of such mildly explosive autoregressions is more uniform than their purely explosive counterparts. Under the second moment condition on the iid innovations, Phillips and Magdalinos (2007a) establish limit theorems for sample moments generated by mildly explosive processes and obtain the following Cauchy limit result:

$$\frac{1}{2c}k_n\rho_n^n(\widehat{\rho}_n - \rho_n) \Rightarrow C. \tag{2}$$

This limit result is unaffected by both the distribution of the initial condition y_0 as far as $y_0 = o(k_n)$. The result was extended by Phillips and Magdalinos (2007b) to a class of weakly dependent innovations. Aue and Horv*á*th (2007) relax the moment conditions on the innovations by considering an iid innovation sequence that belongs to the domain of attraction of a stable law. Magdalinos and Phillips (2008) give multivariate extensions and Magdalinos (2012) considers mildly explosive autoregressions generated by a linear process that may exhibit long-range dependence. Oh et al. (2017) recently study mildly explosive autoregressions with strong mixing innovation sequence, showing that the least squares estimate has the same limit distribution as the iid innovation case.

However, the time homogenous assumption on the residuals seems to be restrictive. The analysis of relatively long stretches of time series data that may contain either slow or rapid changes in the spectrum is of interest in a number of areas. Although the idea of having locally approximately a stationary process was also the starting point of Priestley ’ s theory of processes with evolutionary spectra (Priestley (1965)), recently one of the most important classes of non-stationary processes has been formulated in a rigorous asymptotic framework by Dahlhaus (1996a, 1996b, 1996c, 1997), called locally stationary processes. Locally stationary processes have time varying spectral densities whose spectral structures smoothly change in time. Dahlhaus (2012) also gave the extensive review about locally stationary processes. In this study, we investigate the asymptotic distribution of the LSE for the autoregression with locally stationary error process. The limit behavior of the LSE from the unit root and near stationary autoregressions with locally stationary disturbance was considered by Hirukawa and Skdakata (2012). For the iid unit root process case, see Chan and Wei (1980) and Lee and Wei (1999).

Although we mainly focus on the limit behavior of the LSE, we develop a method for identifying the onset and the end of a bubble period of an econometric time series as an application, originally considered in Phillips and Yu (2009) and Phillips et al. (2011). For this task, we investigate the limiting distribution of the Dickey-Fuller tests when the underlying process is either a unit-root process or explosively mild process and finally demonstrate that they are consistent.

2. Limit distribution of LSE In this section, we consider the following mildly explosive process with locally stationary disturbance:

$$\begin{aligned} y_{t,n} &= \rho_{t,n} y_{t-1,n} + u_{t,n}, \ t = 1, \dots, n, \\ &= \left(\prod_{k=1}^{t} \rho_{k,n} \right) y_{0,n} + \sum_{j=1}^{t} \left(\prod_{k=j+1}^{t} \rho_{k,n} \right) u_{j,n}, \end{aligned}$$

where $\left\{ u_{j,n} \right\}$ is generated from the time varying MA (∞) model:

$$u_{t,n} = \sum_{l=0}^{\infty} \alpha_l \left(\frac{t}{n} \right) \varepsilon_{t-l} = \sum_{l=0}^{\infty} \alpha_l \left(\frac{t}{n} \right) L^l \varepsilon_t := \alpha \left(\frac{t}{n}, L \right) \varepsilon_t,$$

where $\varepsilon_t \sim i.i.d. \left(0, \sigma^2 \right)$, $\rho_{t,n} = 1 + \frac{1}{k_n} \beta \left(\frac{t}{n} \right)$ with $\beta \in C[0,1]$, the class of continuous real-valued functions on [0,1], satisfying $0 < \beta(u) < \infty$, the MA coefficients $\alpha \left(\frac{t}{n}, L \right) := \sum_{l=0}^{\infty} \alpha_l \left(\frac{t}{n} \right) L^l$ with lag operator L, satisfying

$$\sum_{l=0}^{\infty} l \sup_{0 \le u \le 1} |\alpha_l(u)| < \infty \quad \text{and} \quad \sum_{l=0}^{\infty} l \sup_{0 \le u \le 1} \left| \frac{\partial}{\partial u} \alpha_l(u) \right| < \infty,$$

and $y_{0,n} = o_P \left(\sqrt{k_n} \right)$ with $k_n = o(n)$ being a sequence of positive real numbers.

We then consider the normalized serial correlation coefficient:

$$S_n := \frac{\frac{1}{k_n \left(\prod_{k=1}^{n} \rho_{k,n} \right)} \sum_{t=1}^{n} y_{t-1,n} u_{t,n}}{\frac{1}{k_n^2 \left(\prod_{k=1}^{n} \rho_{k,n} \right)^2} \sum_{t=1}^{n} y_{t-1,n}^2} := \frac{U_n}{V_n}. \tag{3}$$

Note that if $\rho_{t,n} \equiv \rho_n = 1 + \frac{\beta}{k_n}$, $t = 1, \dots, n$, we have

$$S_n = k_n \left(\prod_{k=1}^{n} \rho_{k,n} \right) \left(\widehat{\rho}_n - \rho_n \right), \tag{4}$$

where $\widehat{\rho}_n := \frac{\sum_{t=1}^{n} y_{t-1,n} y_{t,n}}{\sum_{t=1}^{n} y_{t-1,n}^2}$ is the least squares estimator (LSE) of the AR(1) process with constant coefficient:

$$y_{t,n} = \rho_n y_{t-1,n} + u_{t,n}, \quad t = 1, \dots, n, \quad \rho_n = 1 + \frac{\beta}{k_n}.$$

Therefore, the $\widehat{\rho}_n$ has bias for the estimation of $\rho_{t,n}$. We then obtain the following theorem.

Theorem 1. *Let* $\kappa = 2\sqrt{\beta(0)\beta(1)} \left| \frac{\alpha(1,1)}{\alpha(0,1)} \right|$. *Then,*

$$S_n \Rightarrow \kappa C.$$

Therefore, if $\rho_{t,n} \equiv \rho_n = 1 + \frac{\beta}{k_n}$, $t = 1, \dots, n$,

$$k_n \left(\prod_{k=1}^{n} \rho_{k,n} \right) \left(\widehat{\rho}_n - \rho_n \right) \Rightarrow \kappa C$$

from (4).

High-dimensionality Adjusted Asymptotically Loss Efficient GC_p Criterion in Normal 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

The multivariate linear regression model is one of basic models of multivariate analysis. This model is introduced in many multivariate statistical textbooks (see e.g., Srivastava, 2002, chap. 9; Timm, 2002, chap. 4), and even now is widely used in chemometrics, engineering, econometrics, psychometrics, and many other fields, for the predication of multiple responses to a set of explanatory variables. Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots . \boldsymbol{y}_n)'$ be an $n \times p$ matrix of p response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of non-stochastic k explanatory variables, where n is the sample size. In order to ensure the possibility of estimating the model and the existence of a variable selection criterion, we assume that $\mathrm{rank}(\boldsymbol{X}) = k\ (< n)$ and $n - p - k - 1 > 0$. Suppose that j denotes a subset of $\omega = \{1, \dots, k\}$ containing k_j elements, and $\boldsymbol{X}_j$ denotes the $n \times k_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j, where k_A denotes the number of elements of a set A, i.e., $k_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth columns of $\boldsymbol{X}$. We then consider the following multivariate linear regression model with k_j explanatory variables as the candidate model:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_j \boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients, and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown covariance matrix. We identify the candidate model by the the set j and call the candidate model in (1) the model j. Especially, the model with $\boldsymbol{X}_\omega$ (namely $\boldsymbol{X}$) is called the full model. We will assume that the data are generated from the following true model:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_{j_*} \boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n), \tag{2}$$

where j_* is a set of integers indicating the subset of explanatory variables in the true model, $\boldsymbol{\Theta}_*$ is a $k_{j_*} \times p$ matrix of true regression coefficients, and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true covariance matrix with $\mathrm{rank}(\boldsymbol{\Sigma}_*) = p$. We call the model in (2) the true model j_*. Henceforth, for simplicity, we represent $\boldsymbol{X}_{j_*}$ and k_{j_*} as $\boldsymbol{X}_*$ and k_*, respectively.

In this paper, we focus on a variable selection method by minimizing a generalized C_p (GC_p) criterion, which is called a GC_p-minimization method, when p may be large but still smaller than n. The GC_p criterion is defined by adding a positive constant value α times the number of parameters in the mean structure to the minimum value of the residual sum of squares (RSS). The GC_p criterion in an univariate linear regression model was proposed by Atkinson (1980) and that in a multivariate linear regression model was proposed by Nagai *et al.* (2012). The family of GC_p criteria contains many widely known variable selection criteria, e.g., Mallows C_p criterion proposed by Sparks *et al.* (1983) (the original C_p was proposed by Mallows, 1973, under the univariate linear regression model), the modified C_p (MC_p) criterion proposed by Fujikoshi and Satoh (1997), which is the completely bias-corrected version of C_p criterion. Since we deal with the multivariate linear regression model of which the dimension p may be large, the following asymptotic framework is used for assessing an asymptotic property of a variable-selection method:

- Large-sample and high-dimensional (LSHD) asymptotic framework: n and p/n approach ∞ and $c_0 \in [0, 1)$, respectively. For simplicity, we will write it "$n \to \infty, p/n \to c_0$".

It should be emphasized that we do not care about whether p goes to ∞ or not in the above asymptotic framework. Suppose that regression coefficients of the ath explanatory variable ($a \in j_*$) and diagonal elements of the true covariance matrix are corresponding to sequences $\{\beta^*_{a,i}\}_{i=1,2,\dots}$ and $\{\psi^*_i\}_{i=1,2,\dots}$, respectively. These mean that new elements are added to current regression coefficients and diagonal elements of the true covariance matrix when p increases.

There are two important properties of a variable selection method by minimizing a variable selection criterion. One is a consistency property that a selection probability of the true model by a variable selection criterion goes to 1 asymptotically, the other is an efficiency property that a ratio of a loss function of a selected model by a variable selection criterion and the minimum loss function goes to 1 asymptotically (see e.g., Shibata, 1980, 1981; Shao, 1997). Recently, Yanagihara (2016) clarified a sufficient condition of α of GC_p criterion in (1) to satisfy a consistency property under the LSHD asymptotic framework. An aim of this paper is to derive a sufficient condition of α of GC_p criterion in (1) to satisfy an efficiency property under the LSHD asymptotic framework. Then, we can propose asymptotically loss efficient GC_p criterion even under high-dimensionality of p. The asymptotically loss efficient GC_p criterion is defined by the following α:

$$\alpha(\beta) = \frac{2n}{n-p} + \beta, \quad \beta > 0 \quad s.t. \lim_{n\to\infty, p/n\to c_0} \sqrt{p}\beta = \infty \quad \text{and} \quad \lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\beta = 0. \tag{3}$$

Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage

Young-Geun Choi

Data R&D Center, SK Telecom, Seoul, Korea

Johan Lim

Department of Statistics, Seoul National University, Seoul, Korea

Anindya Roy and Junyong Park

Department of Mathematics and Statistics, University of Maryland,
Baltimore County, MD, USA

Abstract

In this work, we study the positive definiteness (PDness) problem in covariance matrix estimation. For high dimensional data, many regularized estimators are proposed under structural assumptions on the true covariance matrix including sparsity. They are shown to be asymptotically consistent and rate-optimal in estimating the true covariance matrix and its structure. However, many of them do not take into account the PDness of the estimator and produce a non-PD estimate. To achieve the PDness, researchers consider additional regularizations (or constraints) on eigenvalues, which make both the asymptotic analysis and computation much harder. In this paper, we propose a simple modification of the regularized covariance matrix estimator to make it PD while preserving the support. We revisit the idea of linear shrinkage and propose to take a convex combination between the first-stage estimator (the regularized covariance matrix without PDness) and a given form of diagonal matrix. The proposed modification, which we denote as FSPD (Fixed Support and Positive Definiteness) estimator, is shown to preserve the asymptotic properties of the first-stage estimator, if the shrinkage parameters are carefully selected. It has a closed form expression and its computation is optimization-free, unlike existing PD sparse estimators. In addition, the FSPD is generic in the sense that it can be applied to any non-PD matrix including the precision matrix. The FSPD estimator is numerically compared with other sparse PD estimators to understand its finite sample properties as well as its computational gain. It is also applied to two multivariate procedures relying on the covariance matrix estimator

— the linear minimax classification problem and the Markowitz portfolio optimization problem — and is shown to substantially improve the performance of both procedures.

Key words: Covariance matrix; fixed support; high dimensional estimation; linear minimax classification problem; linear shrinkage; mean-variance portfolio optimization; precision matrix; positive definiteness.

Greedy Active Learning Algorithm for Logistic Regression Models

Ray-Bing Chen

Department of Statistics,

National Cheng Kung University

We study a logistic model-based active learning procedure for binary classification problems, in which we adopt a batch subject selection strategy with a modified sequential experimental design method. Moreover, accompanying the proposed subject selection scheme, we simultaneously conduct a greedy variable selection procedure such that we can update the classification model with all labeled training subjects. The proposed algorithm repeatedly performs both subject and variable selection steps until a prefixed stopping criterion is reached. Our numerical results show that the proposed procedure has competitive performance, with smaller training size and a more compact model compared with that of the classifier trained with all variables and a full data set. We also apply the proposed procedure to a well-known wave data set (Breiman et al., 1984) and a MAGIC gamma telescope data set to confirm the performance of our method.

Inference for LSHD Time Series Models and Applications to Sensor Monitoring and Financial Engineering

Ansgar Steland

Institut fuer Statistik und Wirtschaftsmathematik, RWTH Aachen University

In data science applications such as sensor monitoring or financial portfolio optimization the available data for modeling, computations and analysis of the covariance matrix has often be done in a low-sample-size-high-dimensional (LSHD) regime. Especially, if the dimension is larger than the sample size, classic methods fail and need to be replaced by procedures which are designed for high-dimensional data.

Methods for testing and change detection in the covariance matrix can be based on recent results on LSHD asymptotics of bilinear forms of the sample covariance matrix. This approach allows us to detect and infer changes in averages of covariances of variables or in the variance of projections. The theoretical results hold without the need to constraint the dimension relative to the sample size. For the statistical estimation of unknowns one often needs a (large) learning sample. To circumvent this, we propose in-sample estimators not requiring a learning sample.

Simulations show that the proposed methods work reliable for realistic mathematical models. As a real world application the method is applied to analyze monitoring data from ozone sensors. The sensor data is compressed by projecting it onto sparse principal directions obtained by a sparse principal component analysis (SPCA). It turns out that the SPCA automatically learns the spatial locations of the sensors and leads to a spatial segmentation. Analyzing the projections for a change-point provides a mean to detect changes in the spatial dependence structure of the sensor network measuring ozone.

Do We Necessarily Have More Cost-Effective Information with Big Data?*

Nitis Mukhopadhyay
Department of Statistics
University of Connecticut-Storrs

Abstract

It is a normally held belief that more data provide more information without mentioning what kind of information one may be looking for. Such a feeling is widespread especially in the face of big data movement. In order to hold a reasonable discourse, pros and cons both, I begin working under a certain stochastic models of one kind or another and then try to grasp what additional useful information may entail as more data come in.

I will illustrate situations where a sentiment that more data may amount to more information is actually nearly valid. But, I will also show that the same belief may not be entirely justified in other situations where the relative gain in information is not very substantial as one utilizes more available data. Indeed, the relative gain in information may go down with more data accrued.

In this presentation, I will share preliminary ideas and show some analysis to validate my thoughts. It is my earnest belief that there must be more to it than just using more and more available data simply because they are out there!

Robust relative error estimation

Kei Hirose [1] and Hiroki Masuda [2]

[1] Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

[2] Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

E-mail: hirose@imi.kyushu-u.ac.jp, hiroki@math.kyushu-u.ac.jp

In regression analysis, many analysts use the (penalized) least squares estimation, which aims at minimizing the mean squared prediction error [5]. On the other hand, the relative (percentage) error is often more useful and/or adequate than the mean squared error. For example, in econometrics, the comparison of prediction performance between different stock prices with different units should be made by relative error; we refer to [9] and [11] among others. Additionally, the prediction error of photovoltaic power production or electricity consumption is evaluated by not only mean squared error but also relative error (see, e.g., [10]).

In relative error estimation, we minimize a loss function based on the relative error. An advantage of using such a loss function is that it is scale free or unit free. Recently, several researchers have proposed various loss functions based on relative error [9, 11, 1, 8, 2, 3]. In practice, a response variable $y (> 0)$ can turn out to be extremely large or close to zero. For example, the electricity consumption of a company may be low during holidays and high on exceptionally hot days. These responses may often be considered to be outliers, to which the relative error estimator is sensitive because the loss function diverges when $y \to \infty$ or $y \to 0$. Therefore, a relative error estimation that is robust against outliers must be considered. Recently, Chen et al. [2] discussed the robustness of various relative error estimation procedures by investigating the corresponding distributions, and concluded that the distribution of least product relative error estimation (LPRE) proposed by [2] has heavier tails than others, implying that the LPRE might be more robust than others in practical applications. However, our numerical experiments show that the LPRE is not as robust as expected, so that the robustification of the LPRE is yet to be investigated from the both theoretical and practical viewpoints.

To achieve a relative error estimation that is robust against outliers, this paper employs the γ-likelihood function for regression analysis by Kawashima and Fujisawa [7], which is constructed by the γ-cross entropy [4]. An analysis of electricity consumption data is presented to illustrate the usefulness of our procedure. For detail of the theoretical properties, algorithms, and Monte Carlo simulations, please refer to Hirose and Masuda [6].

References

[1] K. Chen, S. Guo, Y. Lin, and Z. Ying. Least Absolute Relative Error Estimation. *Journal of the American Statistical Association*, 105(491):1104–1112, Sept. 2010.

[2] K. Chen, Y. Lin, Z. Wang, and Z. Ying. Least product relative error estimation. *Journal of Multivariate Analysis*, 144:91–98, Feb. 2016.

[3] H. Ding, Z. Wang, and Y. Wu. A relative error-based estimation with an increasing number of parameters. *Communications in Statistics—Theory and Methods*, 47(1):196–209, Nov. 2017.

[4] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, Oct. 2008.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New York, New York, NY, second edition, 2009.

[6] K. Hirose and H. Masuda. Robust relative error estimation. *Entropy*, 20(9):632, 2018.

[7] T. Kawashima and H. Fujisawa. Robust and Sparse Regression via γ-Divergence. *Entropy*, 19(12):608–21, Dec. 2017.

[8] Z. Li, Y. Lin, G. Zhou, and W. Zhou. Empirical likelihood for least absolute relative error regression. *TEST*, 23(1):86–99, Sept. 2013.

[9] H. Park and L. A. Stefanski. Relative-error prediction. *Statistics & Probability Letters*, 40 (3):227–236, 1998.

[10] D. W. van der Meer, J. Widén, and J. Munkhammar. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. *Renewable and Sustainable Energy Reviews*, 81(Part 1):1484–1512, Jan. 2018.

[11] J. Ye. Price Models and the Value Relevance of Accounting Information. *SSRN Electronic Journal*, 2007.

Inconsistency of diagonal scaling under high-dimensional limit[1]

Tomonari Sei
Graduate School of Information Science and Technology,
The University of Tokyo.

1 Summary

We claim that diagonal scaling of a sample covariance matrix is asymptotically inconsistent if the ratio of the dimension to the sample size converges to a positive constant, where the population is assumed to be Gaussian with a spike covariance model. Our non-rigorous proof relies on the replica method developed in statistical physics. In contrast to similar results known in the literature on principal component analysis, strong inconsistency is not observed. Numerical experiments support the derived formulas.

2 Main results

Let $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$ be independent and identically distributed according to the p-dimensional Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Denote the (uncentered) sample covariance matrix by $\boldsymbol{S} = (1/n)\sum_{t=1}^{n} \boldsymbol{x}_{(t)}\boldsymbol{x}_{(t)}^{\top}$. We assume $n \geq p$, which implies that $\boldsymbol{S}$ is positive definite with probability one, unless otherwise stated.

Let $\mathbb{R}_+$ be the set of positive numbers. By a diagonal scaling theorem, there exists a unique vector $\hat{\boldsymbol{w}} \in \mathbb{R}_+^p$ such that

$$\hat{\boldsymbol{W}}\boldsymbol{S}\hat{\boldsymbol{W}}\mathbf{1} = \mathbf{1}, \tag{1}$$

where $\hat{\boldsymbol{W}} = \mathrm{diag}(\hat{\boldsymbol{w}})$ and $\mathbf{1} = \mathbf{1}_p = (1, \ldots, 1)^{\top}$. In other words, all row sums of the scaled matrix $\hat{\boldsymbol{W}}\boldsymbol{S}\hat{\boldsymbol{W}}$ are unity.

[1] This study was supported by Kakenhi Grant Numbers JP17K00044 and JP26108003.

Let $\boldsymbol{w}_0$ be the population counterpart of $\hat{\boldsymbol{w}}$, which means $\boldsymbol{W}_0\boldsymbol{\Sigma}\boldsymbol{W}_0\boldsymbol{1} = \boldsymbol{1}$, $\boldsymbol{W}_0 = \mathrm{diag}(\boldsymbol{w}_0)$. If p is fixed and $n \to \infty$, a standard argument of asymptotic statistics shows that $\hat{\boldsymbol{w}}$ converges almost surely to the true parameter $\boldsymbol{w}_0$ because $\boldsymbol{S}$ converges to $\boldsymbol{\Sigma}$. However, if p is getting large as well as n, then the limiting behavior of $\hat{\boldsymbol{w}}$ is not obvious. We are interested in the behavior of $\hat{\boldsymbol{w}}$ if $\alpha_p := n/p$ converges to some $\alpha \in [1, \infty)$ as $p \to \infty$.

In principal component analysis, this type of high-dimensional asymptotics has been deeply investigated. In particular, the angle between the first eigenvectors of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ converges to a non-zero value. Furthermore, the limit becomes $\pi/2$ if α is less than a threshold. We call these phenomena inconsistency and strong inconsistency, respectively.

We obtained similar conclusions for the diagonal scaling problem, at least numerically. The following formula is derived with the help of the replica method from statistical physics. See [1] for details.

Claim 1. *Let $\boldsymbol{\Sigma} = \boldsymbol{I}$. Suppose that $\alpha_p = n/p$ converges to some $\alpha \in [1, \infty)$ as $p \to \infty$. Then we have*

$$\lim_{p\to\infty} \frac{\hat{\boldsymbol{w}}^\top \boldsymbol{w}_0}{\|\hat{\boldsymbol{w}}\|\|\boldsymbol{w}_0\|} = \frac{1 - \frac{3}{8\alpha}}{\sqrt{1 - \frac{1}{2\alpha}}}. \tag{2}$$

The right-hand side falls within $(5\sqrt{2}/8, 1)$.

We also established formulas for a class of covariance matrices

$$\boldsymbol{\Sigma} = \Omega\frac{\boldsymbol{1}\boldsymbol{1}^\top}{p} + \boldsymbol{I}, \tag{3}$$

where Ω is a positive constant expressing the signal-to-noise ratio.

References

[1] Sei, T. (2018). Inconsistency of diagonal scaling under high-dimensional limit: a replica approach, Preprint, arXiv:1808.05781.

Statistical Inference with Unnormalized Models

Takafumi Kanamori[1]

[1]Tokyo Institute of Technology/RIKEN AIP

Abstract

In this talk, we propose a class of estimators for unnormalized models, and show that our estimators have the property of the efficiency in the parametric inference. The key concept of our method is the density-ratio matching under Bregman divergences. This is the joint work with T. Takenouchi of Future University Hakodate/RIKEN AIP, and M. Uehara and X.-L. Meng of Harvard University.

Statistical inference with parametric statistical models is an important issue in the fields of machine learning and statistics. The maximum likelihood estimation (MLE) or its variants is often used for the parameter estimation. However, the direct application of the MLE to, say, the Boltzmann machine[6] is not tractable. A difficulty comes from the calculation of the normalization constant. Its computational cost is of the exponential order in the variable dimension. The problem of computational cost is common to many probabilistic models, and several solutions for the estimation of unnormalized models have been suggested; see Markov Networks[3], the Boltzmann machine (with hidden variables)[6, 1, 2], models in independent component analysis[7], truncated distribution[8], exponential-polynomial distribution[4], and references therein.

When the normalization constant in a parametric model is not computationally tractable, we need to deal with unnormalized statistical model, $p(x;\theta)$, $\theta \in \Theta$. Here, the integral of the probability density $\int p(x;\theta)dx$ is not necessarily equal to one, i.e., it may depend on the parameter θ. An approach for the statistical inference with unnormalized models is to approximate the unnormalized model by a tractable model by the mean-field approximation, which considers a model assuming independence of variables [9]. Another approach such as the contrastive divergence [5] avoids the exponential time calculation by the Markov Chain Monte Carlo (MCMC) sampling.

In this talk, we propose a class of estimators for unnormalized models. Our method does not require calculation of the normalization constant. The proposed estimator is defined by minimization of Bregman divergence in which density-ratio matching is used. Our approach works for statistical models with both discrete and continuous random variables. We show that in many case our estimator achieves the Fisher efficiency and also possesses the convex property in the parameter for the unnormalized exponential family. So far, estimators for unnormalized model rarely achieves the fisher efficiency except the estimator based on the pseudo-spherical divergence on discrete sample space[10]. Our estimator is regarded as an extension of [10] to the model on the continuous sample space. We show that not only pseudo-spherical divergence but other Bregman divergences yield efficient estimators for unnormalized models. Some theoretical background and numerical experimenters will be shown in the presentation. The paper including the content of the presentation is now in preparation[11].

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[2] S. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. In *IEEE Transactions on Neural Networks*, volume 3, pages 260–271, 1992.

[3] 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.

[4] J. Hayakawa and A. Takemura. Estimation of exponential-polynomial distribution by holonomic gradient descent. *Communications in Statistics -Theory and Methods*, 45(23):6860–6882, 2016.

[5] G.E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[6] Geoffrey E Hinton and Terrance J Sejnowski. Learning and relearning in boltzmann machines. *MIT Press, Cambridge, Mass*, 1:282–317, 1986.

[7] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2005.

[8] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Wiley, New York, 1995.

[9] M. Opper and D. Saad, editors. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA, 2001.

[10] Takashi Takenouchi and Takafumi Kanamori. Empirical localization of homogeneous divergences on discrete sample spaces. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 820–828. Curran Associates, Inc., 2015.

[11] M. Uehara, T. Kanamori, T. Takenouchi, and X. L. Meng. Unified efficient estimation framework for unnormalized models. in preparation.