**科研費シンポジウム**

「多様な分野における統計科学に関する理論と方法論の革新的展開」
（Innovative development of theory and methodology on statistical science in various fields）

日時：2021 年 9 月 3 日（金）～9 月 5 日（日）
(Date: Friday, 3, – Sunday, 5, September, 2021)
場所：新潟大学駅南キャンパスときめいと 講義室 A,B
(Place: Niigata University, Station South Campus Tokimate, Lecture Room A, B)
（TEL: 025-248-8141）

科学研究費・基盤研究（A）（課題番号：20H00576）
「大規模複雑データの理論と方法論の革新的展開」
研究代表者：青嶋 誠（筑波大学）
Makoto Aoshima (University of Tsukuba)
開催責任者：蛭川 潤一（新潟大学）
Junichi Hirukawa (Niigata Universiry)

## Program

**Friday, 3, September**

**Reception** 13:00-13:15

**Opening** 13:15-13:20 **Junichi HIRUKAWA** (Niigata University)

**Afternoon Session I (in Japanese)** 13:20-14:45
Chair **宮田 庸一** (高崎経済大学)
1. 13:20-14:00 **太刀岡 勇気**
デンソーアイティーラボラトリ
深層学習モデルの内部状態クラスタリングへのノイズ掃き出し法とクロスデータ行列法の利用
2. 14:05-14:45 **劉慶豊（Qingfeng Liu）**
小樽商科大学（Otaru University of Commerce）
Machine Collaboration

**Coffee Break** 14:45-15:00

**Afternoon Session II (in English)** 15:00-17:10
Chair **Hiroshi Shiraishi** (Keio University)
3. 15:00-15:40 **張元宗（Chang Yuan-Tsung）**
目白大学（Mejiro University）
Simultaneous estimation of multiplicative Poisson means in two-way contingency tables
4. 15:45-16:25 **栗栖 大輔（Daisuke Kurisu）**
東京工業大学（Tokyo Institute of Technology）
非定常な関数時系列データの統計分析（Statistical analysis of nonstationary functional time series）
5. 16:30-17:10 **明石 郁哉 (Fumiya Akashi)**
Faculty of Economics, University of Tokyo
Statistical inference for nonstationary heavy-tailed time series models by L1 approach

**Saturday, 4, September**

**Morning Session (in Japanese)** 9:20-11:30
Chair **Muneya Matsui** (Nanzan University)
6. 9:20-10:00 **桃﨑 智隆** [1]・長 光司 [1]・中川 智之 [2]・富澤 貞男 [2]
[1] 東京理科大学大学院 理工学研究科
[2] 東京理科大学 理工学部
ベイズ法を用いた分割表における尺度の推定
7. 10:05-10:45 **永井 勇**
中京大学 教養教育研究院
高次元小標本データにおける多変量線形回帰モデルでの罰則を用いない推定法
8. 10:50-11:30 **阿部俊弘**
法政大学 経済学部
Cauchy 型分布の単純な EM アルゴリズムについて

**Lunch** 11:30-13:20

**Afternoon Session I (in Japanese)** 13:20-14:00
Chair **阿部俊弘** (法政大学)
9. 13:20-14:00 **松井 宗也 (Muneya Matsui)**
南山大学 (Nanzan University)
ランダムフィールドにおける距離の共分散 (Distance covariance for random fields)

**Coffee Break** 14:00-15:00

**Afternoon Session II (in English)** 15:00-16:25
Chair **Fumiya Akashi** (University of Tokyo)
10. 15:00-15:40 澁木 涼太郎（Shibuki Ryotaro）, 中村 知繁（Nakamura Tomoshige）, **白石 博（Shiraishi Hiroshi）**
慶應義塾大学理工学研究科 (Graduate School of Science and Technology, Keio University),
慶應義塾大学理工学部 (Faculty of Science and Technology, Keio University),
慶應義塾大学理工学部 (Faculty of Science and Technology, Keio University)
ランダムフォレストを用いた時系列分位点回帰 (Time Series Quantile Regressions by using Random Forest)
11. 15:45-16:25 **竹内 努（Tsutomu T. TAKEUCHI）**
名古屋大学素粒子宇宙物理学専攻
（Division of Particle and Astrophysical Science, Nagoya University, Japan）
高次元統計学の方法による銀河の分光マップの解析
（Analysis of Spatially Resolved Spectral Map of a Galaxy by the High-Dimensional Statistical Method）
**Coffee Break** 16:25-16:40

**Afternoon Session III (Guest speakers session)** 16:40-18:10
Chair **Junichi Hirukawa** (Niigata University)
12. 16:40-17:20 **Ngai Hang Chan**
The Chinese University of Hong Kong
Statistical Inference for Glaucoma Detection
13. 17:30-18:10 **Konstantinos Fokianos**
University of Cyprus
Mixtures of Nonlinear Poisson Autoregressions

**Sunday, 5, September**

**Morning Session (in Japanese)** 10:00-11:25
Chair **阿部俊弘** (法政大学)
14. 10:00-10:40 **宮田 庸一**
高崎経済大学 経済学部
ブリッジ推定量におけるチューニングパラメーターの選択について
15. 10:45-11:25 **種市 信裕**・関谷 祐里・外山 淳
北海道教育大学（札幌校）・北海道教育大学（釧路校）・数学利用研究所
3 次元分割表における条件付き独立性検定の変換統計量について

**Closing** 11:25-11:30 **Junichi HIRUKAWA** (Niigata University)

# 深層学習モデルの内部状態クラスタリングへの<br>ノイズ掃き出し法とクロスデータ行列法の利用

太刀岡 勇気 (デンソーアイティーラボラトリ)

## 1　はじめに

複数の話者やスタイルを制御できる end-to-end 音声合成システムで、タグづけにより話者・スタイルを制御する方法を提案した [1]。ところで統計学の分野では、データの次元数 $d$ と標本数 $n$ の間に $d \gg n$ の関係がある場合、高次元小標本 (high-dimension, low-sample-size; HDSS) データと呼ばれ、特殊な取り扱いが必要である [2]。埋め込み表現を主成分分析 (PCA) や t-SNE により次元圧縮して可視化することが、埋め込み表現は $2^9$ 程度の次元、話者・スタイル数は $2^2 \sim 2^4$ 程度であるため、この問題は HDSS のクラスタリング問題であると考えられる。ここでは、音声合成器内部の埋め込み表現に対して、PCA と、HDSS データを扱う代表的な 2 手法であるノイズ掃き出し法 (NRM)[3] とクロスデータ行列法 (CDM)[4] での分析を比較した。

## 2　実験

学習話者 ftak,fhar の 2 人の女性話者について、$4(=n)$ スタイルでの内部表現を PCA した結果を、図 1 に示す。話者によって分離できていることと、h,t,z,k の順に上から並んでいることから、スタイルが同じものは近くに配置されていることが推測される。NRM による結果を図 2 に示す。ラベルによる並び順は異なるがほぼ同傾向といえる。CDM による結果を図 3 に示す。話者毎に集合分割すると、スタイル別にクラスタリングできることがわかった。

スパイク指数の推定結果を図 4 に示す。点推定の結果、その平均値、中央値を示す。外れ値もあるがそれなりに収束している。下段にはべき乗近似により得られた $\alpha_s$ を示している。これより $\alpha_s > 0.8$ であり、$\gamma > 0.2$ であれば標本固有値が一致性を持つことがわかり、ここでのは指数 $\gamma = 2/9 \simeq 0.22$ であるから問題ない範囲であるといえる。ただし一般化スパイクモデルが仮定している $\alpha_s$ の $s$ に関する単調減少性は満たされていない。

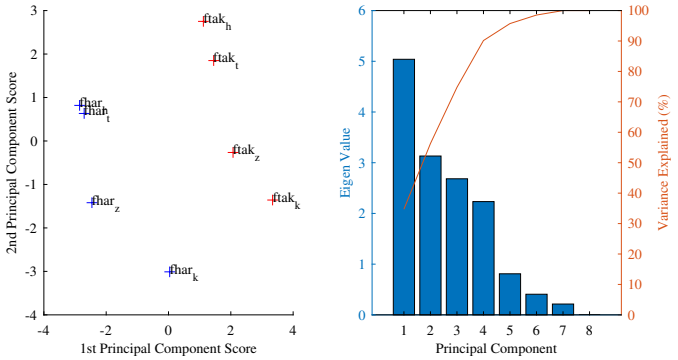


図 1　Principal component analysis (PCA) of embedded vectors.

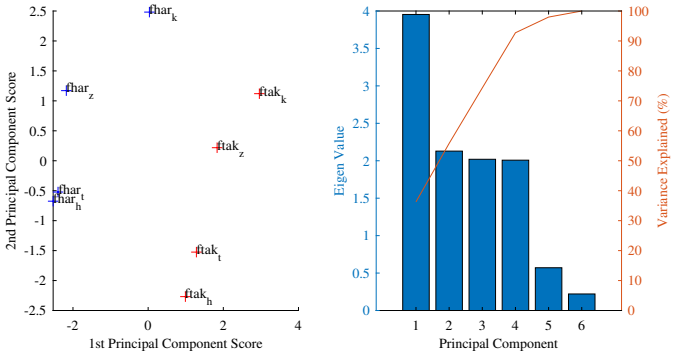


図 2　Noise reduction methodology.

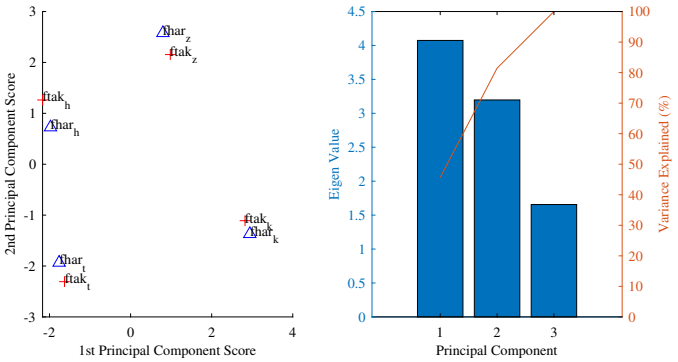


図 3　Cross-data-matrix methodology.

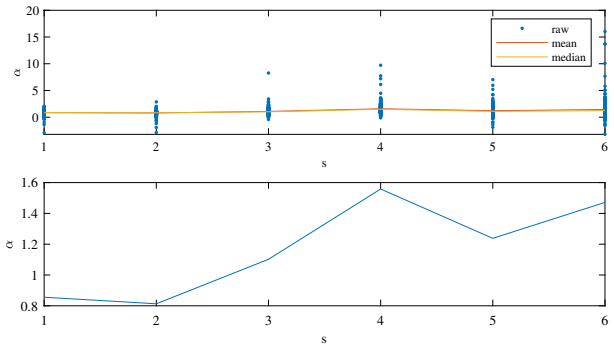


図 4　Estimated $\alpha$ (embedded vector, style).

# 3　まとめ

合成音声器の内部状態を分析する際の問題が HDSS の設定であることに着目し、NRM と CDM を適用した。NRM は通常の PCA よりも若干固有値が割り引かれるものの結果は大差がなかった。CDM は初期の集合の分割によって結果が大きく変わることがわかった。

# 参考文献

[1] 太刀岡勇気，“話者タグによる tacotron2 の話者性制御，” 日本音響学会研究発表会講演論文集 (秋季)，pp.857–858 (2020).
[2] 青嶋誠，矢田和善，高次元の統計学，共立出版 (2019).
[3] K. Yata and M. Aoshima, “PCA consistency for the power spiked model in high-dimensional settings,” Journal of Multivariate Analysis, **122**, 334–354 (2013).
[4] K. Yata and M. Aoshima, “High-dimensional inference on covariance structures via the extended cross-data-matrix methodology,” Journal of Multivariate Analysis, **151**, 151–166 (2016).

# Machine Collaboration*

Qingfeng Liu† and Yang Feng‡

August 2, 2021

Ensemble learning has emerged and been extensively studied by many in the past few decades. In general, the idea of ensemble learning is to combine the predictions obtained from different learning methods (hereafter, base machines), or predictions based on different subsamples to improve prediction performance. Bagging, stacking, and boosting are three prominent examples. In bagging (Breiman, 1996)/stacking, base machines first run in parallel and independently, and then the final prediction is constructed as a simple/weighted average of the predictions from these base machines. In boosting (Schapire et al., 1998), the base machines work jointly in a top-down manner. In the aforementioned algorithms, the output from each base machine is fixed after being calculated. Like human collaboration, an idea that may yield potential improvement is to let the base machines communicate with each other and update their outputs after observing the predictions of the other base machines. Based on this idea, we propose the *Machine Collaboration* or MaC learning framework. Compared with bagging, stacking, and boosting, MaC has the following desirable features. Figure 1 provides the schematic for bagging, stacking, boosting, and MaC. As illustrated, bagging and stacking are parallel & independent, boosting is sequential & top-down, while MaC is *circular* & *interactive*. For MaC, pieces of information are passed repeatedly between base machines around a "round table," but not one-way or top-down. In this type of scheme, the base machines update their structures and/or parameters according to the information received from the other machines. We demonstrate that MaC can deliver competitive performance when compared with the base machines or the other ensemble methods.

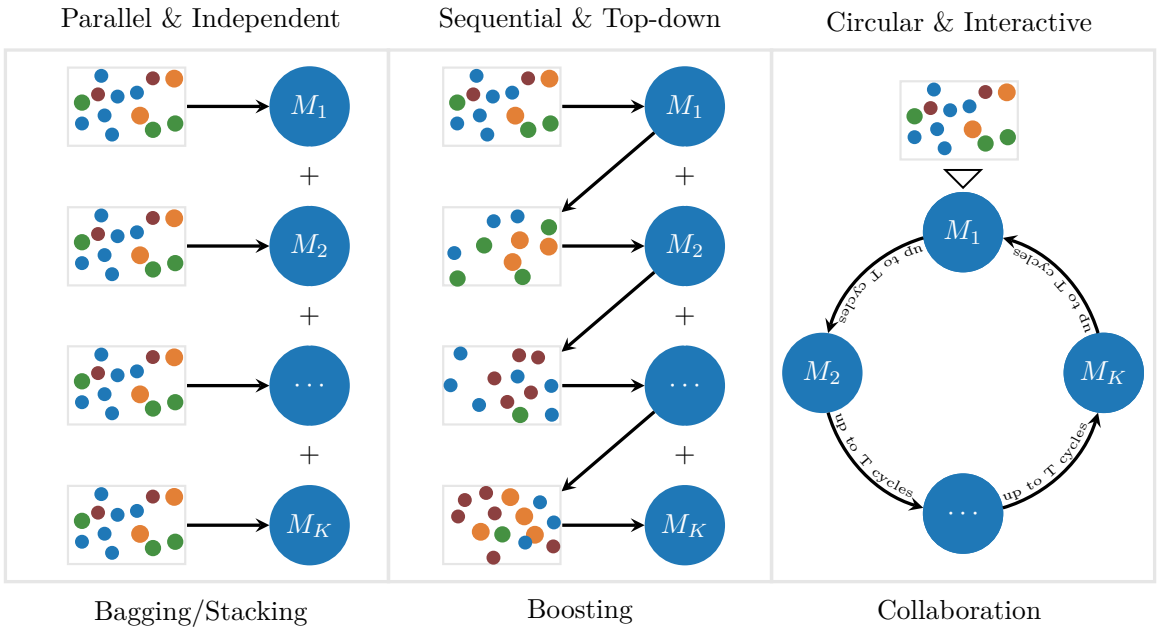


Figure 1: Bagging, boosting, and machine collaboration

The main contributions of this work are fourfold. First, we propose a new type of ensemble

*The authors gratefully acknowledge the support of the Japan Society for the Promotion of Science through KAKENHI Grant No. JP19K01582 (Liu), the Nomura Foundation for Social Science Grant No. N21-3-E30-010 (Liu) and a National Science Foundation CA-REER Grant No. DMS-1554804 (Feng).

†Corresponding author. Department of Economics, Otaru University of Commerce, Otaru City, Hokkaido, Japan. Email: qliu@res.otaru-uc.ac.jp.

‡School of Global Public Health at New York University, NY, NY, USA. Email: yang.feng@nyu.edu.

learning framework, MaC, which is *circular & interactive*. The *circular & interactive* aspect could be a potential direction for exploring new methods of ensemble learning. Second, we present some desirable finite statistical properties of MaC. Third, we demonstrate via extensive simulations that MaC performs better than all individual base machines and the ensemble methods SL and LS-Boost. Lastly, in the analysis of real data, we compare MaC with the competing methods on 119 benchmark datasets in the Penn Machine Learning Benchmarks (PMLB) (Olson et al., 2017) for evaluating and comparing machine learning algorithms. The results of this analysis demonstrate the notable advantages of MaC for most datasets.

# References


Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Breiman, L., Friedman, J., Stone, C., Olshen, R., 1984. Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Dasarathy, B.V., Sheela, B.V., 1979. A composite classifier system design: Concepts and methodology. Proceedings of the IEEE 67, 708–713.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Frontiers of Computer Science 14, 241–258.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189 – 1232.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Ho, T.K., 1995. Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, IEEE. pp. 278–282.

van der Laan, M.J., Polley, E.C., Hubbard, A.E., 2007. Super learner. Statistical Applications in Genetics and Molecular Biology 6, 1544–6115.

Lu, X., Van Roy, B., 2017. Ensemble sampling, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 3259–3267.

Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D., 2012. Ensemble approaches for regression: A survey. Acm computing surveys (csur) 45, 1–40.

Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J., Moore, J.H., 2017. Pmlb: a large benchmark suite for machine learning evaluation and comparison. BioData Mining 10, 36.

Qi, Y., Liu, B., Wang, Y., Pan, G., 2019. Dynamic ensemble modeling approach to nonstationary neural decoding in brain-computer interfaces, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 6087–6096.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, e1249.

Schapire, R.E., 1990. The strength of weak learnability. Machine Learning 5, 197–227.

Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. The annals of statistics 26, 1651–1686.

Yu, Z., Luo, P., Liu, J., Wong, H.S., You, J., Han, G., Zhang, J., 2018. Semi-supervised ensemble clustering based on selected constraint projection. IEEE Transactions on Knowledge and Data Engineering 30, 2394–2407.

**Simultaneous estimation of multiplicative Poisson means in two-way contingency tables**

Yuan-Tsung Chang (Mejiro University), Shinozaki Nobuo (Keio University)


# Abstract

Shrinkage estimation of Poisson means is considered when observations are given in the form of a two-way contingency table. Assuming a multiplicative Poisson model, estimators which shrink to the specified values or an order statistic in one dimension and in two dimensions are considered and are shown to dominate the maximum likelihood estimator (MLE) under normalized squared error loss.


# 1 Introduction

We consider two-way multiplicative model where $x_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J$, are independent random Poisson random variables with means

$$\lambda_{ij} = \lambda \alpha_i \beta_j, \quad i = 1, \ldots, I,\ \ j = 1, \ldots, J,$$

where $\alpha_i \geq 0$ and $\beta_j \geq 0$ satisfy $\sum_{i=1}^{I} \alpha_i = 1$ and $\sum_{j=1}^{J} \beta_j = 1$, respectively. We denote the one-dimensional frequencies and the total frequency by

$$x_{i+} = \sum_{j=1}^{J} x_{ij},\ i = 1, \ldots, I,\ x_{+j} = \sum_{i=1}^{I} x_{ij},\ j = 1, \ldots, J,\ x_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}.$$

As discussed in Hara and Takemura (2006) complete sufficient statistics are $\boldsymbol{x}_1 = (x_{1+}, \ldots, x_{I+})$ and $\boldsymbol{x}_2 = (x_{+1}, \ldots, x_{+J})$. The MLE of $\lambda_{ij}$ is

$$\hat{\lambda}_{ij}^{ML} = \begin{cases} \dfrac{x_{i+} x_{+j}}{x_{++}} & \text{if} \quad x_{++} \neq 0 \\ 0 & \text{if} \quad x_{++} = 0. \end{cases}$$

They have given a class of improved estimators which shrink the MLE toward the origin under the normalized squared error loss. The simple one is

$$\delta_{ij}^{HT} = \frac{x_{i+} x_{+j}}{x_{++}} \left\{ 1 - \frac{d}{x_{++} + d} \right\}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

Next section we consider one-dimensional shrinkage to an order statistic or a specified point.

# 2 One-dimensional shrinkage to an order statistic or a specified point

First, we consider one-dimensional shrinkage to an order statistic.

Let $x_{(\ell)+}$ be the $\ell$-th smallest observation among $x_{1+}, \ldots, x_{I+}$. We assume that $I \geq \ell + 2$ and consider the following estimator which shrinks $x_{i+}$ toward $x_{(\ell)+}$ when $x_{i+} \geq x_{(\ell)+}$:

$$\delta_{ij}^{(1)} = \frac{x_{+j}}{x_{++}} \left\{ x_{i+} - \varphi(W) \frac{(x_{i+} - x_{(\ell)+})^+}{W + d} \right\}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

where $W = \sum_{i=1}^{I} (x_{i+} - x_{(\ell)+})^+$, $a^+ = \max(0, a)$ and $d$ is a positive constant. Then we have the following.

**Theorem 2.1.** Suppose that $\varphi(W)$ is a non-decreasing function satisfying $0 \leq \varphi(W) \leq 2(I - \ell - 1)$ and that $d \geq \sup \varphi(W)/2$. Then $\delta_{ij}^{(1)}, i = 1, \ldots, I$ improves upon the MLE $\lambda_{ij}^{ML}, i = 1, \ldots, I$ under the loss function $\sum_{i=1}^{I} (\hat{\lambda}_{ij} - \lambda_{ij})^2 / \lambda_{ij}$ for any $j = 1, \ldots, J$.

Next we consider the estimators shrink $\hat{\lambda}_{ij}^{ML}$ to a specified non-negative values, $b_i$.

Let $b_i \geq 0, i = 1, \ldots, I$ be given numbers and we propose the following shrinkage estimator which shrinks $x_{i+}$ to $b_i$ when $x_{i+} \geq b_i$:

$$\delta_{ij}^{(2)} = \frac{x_{+j}}{x_{++}} \left\{ x_{i+} - \varphi(N, W) \frac{(x_{i+} - b_i)^+}{W + d(N)} \right\}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J,$$

where $W = \sum_{i=1}^{I} (x_{i+} - b_i)^+$ and $N = \#\{i | x_{i+} \geq b_i\}$. Then we have the following.

**Theorem 2.2.** Suppose that $\varphi(N, W)$ is a non-decreasing function of $W$ and satisfies $0 \leq \varphi(N, W) \leq 2(N-1)^+$ for any $0 \leq N \leq I$. Suppose that $d(N) \geq \sup_W \varphi(N, W)/2$. Then $\delta_{ij}^{(2)}, i = 1, \ldots, I$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}, i = 1, \ldots, I$ under the loss function $\sum_{i=1}^{I} (\hat{\lambda}_{ij} - \lambda_{ij})^2 / \lambda_{ij}$ for any $j = 1, \ldots, J$.
It may be noticed that the shrinkage is made only when $N \geq 2$.

**Remark** Theorems 2.1 and 2.2 can be generalized directly to the case of Poisson multiplicative model for a multi-way contingency tables.

Next section we also consider two-dimensional shrinkage to order statistics or to a specified point.

# 3 Two-dimensional shrinkage to order statistics.

Let $x_{(\ell)+}$ and $x_{+(m)}$ be the $\ell$-th and $m$-th smallest observation among $x_{1+}, \ldots, x_{I+}$ and $x_{+1}, \ldots, x_{+J}$, respectively. We assume that $I \geq \ell + 2$ and $J \geq m + 2$ and consider the estimator which shrinks $x_{i+}$ toward $x_{(\ell)+}$ when $x_{i+} \geq x_{(\ell)+}$ in the first dimension and shrinks $x_{+j}$ toward $x_{+(m)}$ when $x_{+j} \geq x_{+(m)}$ in the second dimension simultaneously. To improve upon the MLE $\hat{\lambda}_{ij}^{ML}$, we propose the following estimator :

$$\delta_{ij}^{(3)} = \frac{1}{x_{++}} \left\{ x_{i+} - \varphi_1(W_1) \frac{(x_{i+} - x_{(\ell)+})^+}{W_1 + d_1} \right\} \left\{ x_{+j} - \varphi_2(W_2) \frac{(x_{+j} - x_{+(m)})^+}{W_2 + d_2} \right\}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J, \qquad (2.4)$$

where $W_1 = \sum_{i=1}^{I} (x_{i+} - x_{(\ell)+})^+$ and $W_2 = \sum_{j=1}^{J} (x_{+j} - x_{+(m)})^+$ and $d_1$ and $d_2$ are positive constants. Then we have the following.

**Theorem 3.1.** Suppose that $\varphi_1(W_1)$ and $\varphi_2(W_2)$ are non-decreasing functions satisfying $0 \leq \varphi_1(W_1) \leq I - \ell - 1$ and $0 \leq \varphi_2(W_2) \leq J - m - 1$, respectively. If $d_1 \geq (I-\ell-1)/(I-\ell) \sup \varphi_1(W_1)$ and $d_2 \geq (J-m-1)/(J-m) \sup \varphi_2(W_2)$. Then $\delta_{ij}^{(3)}, i = 1, \ldots, I, j = 1, \ldots, J$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}$ under the loss function $\sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{\lambda}_{ij} - \lambda_{ij})^2 / \lambda_{ij}$.

Next, we consider Two-dimensional shrinkage to a specified point.

Let $b_i \geq 0, i = 1, \ldots, I$ and $c_j \geq 0, j = 1, \ldots, J$ be given numbers. Assuming that $I, J \geq 2$, we shrink $x_{i+}$ to $b_i$ when $x_{i+} \geq b_i$ and $x_{+j}$ to $c_j$ when $x_{+j} \geq c_j$. To improve upon the MLE $\hat{\lambda}_{ij}^{ML}$, we propose the following estimator

$$\delta_{ij}^{(4)} = \frac{1}{x_{++}} \left\{ x_{i+} - \varphi_1(N_1, W_1) \frac{(x_{i+} - b_i)^+}{W_1 + d_1(N_1)} \right\} \left\{ x_{+j} - \varphi_2(N_2, W_2) \frac{(x_{+j} - c_j)^+}{W_2 + d_2(N_2)} \right\}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J, \qquad (2.5)$$

where $W_1 = \sum_{i=1}^{I} (x_{i+} - b_i)^+, W_2 = \sum_{j=1}^{J} (x_{+j} - c_j)^+$, $N_1 = \#\{i | x_{i+} \geq b_i, i = 1, \ldots, I\}$ and $N_2 = \#\{j | x_{+j} \geq c_i, j = 1, \ldots, J\}$. Although it may be natural to put the condition $\sum_{i=1}^{I} b_i = \sum_{j=1}^{J} c_j$, we do not need it in the following.

**Theorem 3.2.** Suppose that $\varphi_i(N_i, W_i)$ is a non-decreasing function of $W_i$ and satisfies $0 \leq \varphi_i(N_i, W_i) \leq (N_i - 1)^+$ for any $N_i \geq 0$, and that $d_i(N_i) \geq (N_i - 1)^+ / N_i \sup_{W_i} \varphi_i(N_i, W_i)$, for any $N_i \geq 0$, $i = 1, 2$. Then $\delta_{ij}^{(4)}$ improves upon the MLE $\hat{\lambda}_{ij}^{ML}$ under the loss function $\sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{\lambda}_{ij} - \lambda_{ij})^2 / \lambda_{ij}$.

It may be noticed that the shrinkage in the $i-$th dimension is made only when $N_i \geq 2$.

# Statistical analysis of nonstationary functional time series

Daisuke Kurisu
Tokyo Institute of Technology

## 1 Introduction

In this study, we develop an asymptotic theory for estimating the time-varying characteristics of a locally stationary functional time series. The notion of a locally stationary functional time series is an extension of that of a locally stationary process introduced by Dahlhaus (1997). For some probability space $(\Omega, \mathcal{A}, P)$ and for $p \geq 1$, let $L^p(\Omega, \mathcal{A}, P)$ denote the space of real-valued random variables such that $\|X\|_p = (E[|X|^p])^{1/p} < \infty$. Let $D \subset \mathbb{R}^d$ be a compact set and let $L^2 = L^2(D)$ denote a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle x, y \rangle = \int_D x(t)y(t)dt, \ x, y \in H = L^2.$$

Further, let $L_H^p = L_H^p(\Omega, \mathcal{A}, P)$ denote the space of $H$-valued random functions $X$ such that

$$(E[\|X\|^p])^{1/p} = \left( E\left[ \left( \int X^2(t)dt \right)^{p/2} \right] \right)^{1/p} < \infty.$$

## 2 Settings

### 2.1 local stationarity

The definition of an $H$-valued locally stationary process is as follows: The $H$-valued stochastic process $\{X_{t,T}\}$ in $L_H^p$ is locally stationary if for each rescaled time point $u \in [0, 1]$, there exists an associated $H$-valued process $\{X_t^{(u)}\}$ in $L_H^p$ with the following properties:

(i) $\{X_t^{(u)}\}_{t \in \mathbb{Z}}$ is strictly stationary.

(ii) It holds that

$$\|X_{t,T} - X_t^{(u)}\| \leq \left( \left| \frac{t}{T} - u \right| + \frac{1}{T} \right) U_{t,T}^{(u)} \ a.s.,$$

for all $1 \leq t \leq T$ where $\{U_{t,T}^{(u)}\}$ is a process of positive variables satisfying $E[(U_{t,T}^{(u)})^\rho] < C$ for some $\rho > 0$, $C < \infty$ that is independent of $u, t$, and $T$.

This definition is a natural extension of the notion of local stationarity for real-valued time series introduced in Dahlhaus (1997).

### 2.2 dependence structure

For each $u \in [0, 1]$, we assume that $\{X_t^{(u)}\}$ is $L^p$-$m$-approximable, that is, $X_t^{(u)} \in L_H^p$ is of the form

$$X_t^{(u)} = f_u(\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots),$$

where $\varepsilon_j$ are i.i.d. elements taking values in a measurable space $S$, and $f_u$ is a measurable function $f_u : S^\infty \to H$. Note that $X_t^{(u)}$ is strictly stationary. We also assume that if $\{\varepsilon_{j,k}^{(m)}\}$ is an independent

copy of $\{\varepsilon_j\}$ defined on the same probability space, then letting

$$X_{m,t}^{(u)} = f_u(\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-m+1}, \varepsilon_{t,t-m}^{(m)}, \varepsilon_{t,t-m-1}^{(m)}, \ldots),$$

we have

$$\sum_{m\geq 1} v_p(X_t^{(u)} - X_{m,t}^{(u)}) < \infty.$$

We can show that a wide class of functional time series (e.g., functional AR(1) process and functional ARCH(1) process) is $L^p$-$m$-approximable under some regularity conditions. See Hörmann and Kokoszka (2010) and Horváth and Kokoszka (2012) for details on the properties of $L^p$-$m$-approximable random functions.

## 3 Results

We introduce a kernel-based method to estimate the time-varying covariance operator and the time-varying mean function of a locally stationary functional time series. Subsequently, we derive the convergence rate of the kernel estimator of the covariance operator and associated eigenvalue and eigenfunctions. We also establish a central limit theorem for the kernel-based locally weighted sample mean. As applications of our results, we discuss the prediction of locally stationary functional time series and methods for testing the equality of time-varying mean functions in two functional samples. We also discuss a problem to estimate the number of principal components to be used. To the best of our knowledge, this is the first paper that develops an asymptotic theory of estimating time-varying characteristics of locally stationary functional time series based on kernel methods.

## References


Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1-37.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845-1884.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.

Kurisu, D. (2021). On the estimation of locally stationary functional time series. arXiv2105.11873.

# Statistical inference for nonstationary heavy-tailed time series models by L1 approach

Fumiya Akashi (University of Tokyo)*

**Abstract**

This talk proposes a robust $L_1$-estimation methods for a time-varying autoregressive (AR) model. First, we construct a robust local linear estimator based on the self-weighting approach (Ling (2005, Journal of the Royal Statistical Society, Series B)), and show the asymptotic normality of the proposed estimator. Second, the generalized empirical likelihood statistic is proposed to test the hypothesis of the coefficients of the model. The simulation experiments are given in the talk.

## 1 The model and self-weighted statistics

Suppose that an observed stretch $\{Y_{1,T}, ..., Y_{t,T}\}$ is generated by the following time-varying autoregressive (AR) model

$$Y_{t,T} = \beta(t/T)^\top X_{t-j,T} + \epsilon_t, \tag{1}$$

where $\beta(u) = (\beta_1(u), ..., \beta_p(u))^\top$ is a function $[0,1] \to \mathbb{R}^p$, $X_{t-1,T}$ is a $p$-dimensional vector defined as $X_{t-1,T} := (Y_{t-1,T}, ..., Y_{t-p,T})^\top$, and $\{\epsilon_t : t \in \mathbb{Z}\}$ is a sequence of i.i.d. zero-median random variables. We assume some conditions for $\{\epsilon_t : t \in \mathbb{Z}\}$. In particular, the tail distribution of $\epsilon_t$ is assumed to satisfy $x^\alpha P(\epsilon_t > x) \to qC$ and $x^\alpha P(\epsilon_t < -x) \to (1-q)C$ as $x \to \infty$ with some $\alpha > 0$, $q \in [0,1]$ and $C \geq 0$. In particular, $\epsilon_t$ do not have the infinite variance when $\alpha < 2$.

We define the self-weighed local linear $L_1$ estimator for $\beta(u_0)$ ($u_0 \in [0,1]$) as

$$(\hat{\beta}(u_0), \hat{\gamma}) := \arg\min_{\beta,\gamma} \sum_{t=p+1}^{T} K\left(\frac{t-t_0}{Th}\right) w_{t-1,T} |e_{t-1,T}(\beta, \gamma)|,$$

*This talk is partially based on a joint work with Junichi Hirukawa (Niigata University) and Konstantinos Fokianos (University of Cyprus).

where $K$ is a kernel function, $h$ is a bandwidth parameter which goes to zero as $T \to \infty$, $t_0$ is an integer satisfying $|t_0/T - u_0| < 1/T$, $\gamma$ is a nuisance parameter and

$$e_{t-1,T}(\beta, \gamma) := Y_{t,T} - \left[\beta + \left(\frac{t}{T} - u_0\right)\gamma\right]^\top X_{t-1,T} \quad (\beta, \gamma \in \mathbb{R}^p).$$

In particular, $w_{t-1,T}$ $(t = p+1, ..., T)$ is called the self-weight, which is a measurable function of $X_{t-1,T}$. The self-weighting approach for time series model was originally proposed by Ling (2005) for a stationary AR-model.

The limit distribution of the proposed estimator is given in the following theorem.

**Theorem 1.** *Under some regularity conditions, we have*

$$\sqrt{Th}\left[\hat{\beta}(u_0) - \beta(u_0)\right] \xrightarrow{d} N\left(0_p, \frac{\kappa_2}{4f(0)^2}\Sigma(u_0)^{-1}\Omega(u_0)\Sigma(u_0)^{-1}\right) \quad (T \to \infty),$$

*where $\Sigma(u_0)$ and $\Omega(u_0)$ are nonsingular matrices.*

Another important issue in statistical inference for infinite variance process is hypothesis testing for the coefficients of the model (1). Let us consider the hypothesis $H : \beta(u) = \beta_0(u)$ $(\forall u \in [0,1])$, and we apply the generalized empirical likelihood (GEL) test statistic. Let us define the self-weighted moment function

$$g_{t,T}(b) := w_{t-1,T}\text{sign}\left(Y_{t,T} - b^\top X_{t-1,T}\right) X_{t-1,T} \quad (b \in \mathbb{R}^p)$$

and the self-weighted GEL test statistic as

$$r_T(\beta(\cdot)) := 2 \sup_{\lambda \in \hat{\Lambda}_T(\beta(\cdot))} \sum_{t=p+1}^{T} \rho\left\{\lambda^\top g_{t,T}(\beta(t/T))\right\},$$

where $\rho$ is a score function whose domain is $\mathcal{V}_\rho$ $(\subset \mathbb{R})$, and

$$\hat{\Lambda}_T(\beta(\cdot)) := \left\{\lambda : \lambda \in \mathbb{R}^p \text{ and } \lambda^\top g_{t,T}(\beta(t/T)) \in \mathcal{V}_\rho \text{ for all } t \in \{1, ..., T\}\right\}.$$

Then, we get the following theorem.

**Theorem 2.** *Under some regularity conditions, we have $r_T(\beta_0(\cdot)) \xrightarrow{p} \chi^2_p$ as $T \to \infty$.*

Remarkably, the proposed statistics always converge to pivotal limit distributions regardless of whether the model has finite or infinite variance. Finite sample performance of the proposed statistics is also illustrated by some simulation experiments.

# References

Ling, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 67*(3), 381–393.

# ベイズ法を用いた分割表における尺度の推定

桃﨑 智隆 [1]　　長 光司 [1]　　中川 智之 [2]　　富澤 貞男 [2]

[1] 東京理科大学大学院 理工学研究科
[2] 東京理科大学 理工学部

2 元分割表の解析では, 行変数と列変数の間の連関尺度を用いて独立性からの隔たりの程度を測る. 連関尺度には, 例えば, Cramér 係数 (Cramér, 1946) などがある. 一方で, 行と列が同じ分類からなる正方分割表の解析では, 行と列の変数間の独立性に代わり対称性からの隔たりの程度を測る. Tomizawa *et al.* (1998) は power divergence (または, diversity index) を用いて, 対称性からの隔たりの程度を測る様々な尺度を提案した.

分割表におけるこれらの尺度は分割表のセル確率の関数として表されるので, 尺度の値は未知の値を取る. そのため尺度の推定には, 標本比率を用いた尺度のプラグイン推定量が用いられる. サンプル数が十分にあるとき, 標本比率のプラグイン推定量は近似不偏推定量となる. しかし, サンプル数が十分にない場合は, 推定量のバイアスや平均二乗誤差 (MSE) が大きくなってしまう.

Tomizawa *et al.* (2007) は, バイアスの高次のオーダーを導出し, バイアス補正を行うことで尺度の推定量を改良した. これらの研究では数値実験を用いて, 改良型推定量は標本比率のプラグイン推定量よりも速く尺度の真値に近づくことが示された. しかしながら, 改良型推定量にはいくつかの問題点がある. 確かに, サンプル数が小さい場合でも改良型推定量はバイアスを小さくできているが, MSE も小さくできるとは限らない. さらに, 改良型推定量の値域と尺度の値域が異なる可能性もある. 例えば, Tomizawa *et al.* (1998) の尺度の値域は 0 以上 1 以下であるが, 改良型推定量の値域は 0 以上 1 以下を超えている. また, このような場合, 改良型推定量を用いた尺度の推定量の信頼区間の解釈は容易ではない.

本講演では, ベイズ法を用いて, 十分なサンプルサイズがなくてもバイアスや MSE を小さくできる尺度の推定量を提案した.

2 元 $r \times c$ 分割表を考える. 確率変数ベクトル $\boldsymbol{n} = (n_{11}, n_{12}, \ldots, n_{1c}, n_{21}, \ldots, n_{rc})^\top$ は多項分布 $M(n, \boldsymbol{p})$ に従うとする. ここで, $n = \sum_{i,j} n_{ij}$, $p_{ij}$ は $(i,j)$ セル確率, $\boldsymbol{p} = (p_{11}, p_{12}, \ldots, p_{1c}, p_{21}, \ldots, p_{rc})^\top$, “$\top$” は転置を表す.

セル確率ベクトル $\boldsymbol{p}$ がディリクレ事前分布

$$p(\boldsymbol{p}|\alpha) = \frac{\Gamma(rc\alpha)}{(\Gamma(\alpha))^{rc}} \prod_{i,j} p_{ij}^{\alpha-1}$$

に従うとすると, $\boldsymbol{p}$ の事後平均は以下のようになる.

$$\hat{\boldsymbol{p}}^{(\alpha)} = (\hat{p}_{11}^{(\alpha)}, \hat{p}_{12}^{(\alpha)}, \ldots, \hat{p}_{1c}^{(\alpha)}, \hat{p}_{21}^{(\alpha)}, \ldots, \hat{p}_{rc}^{(\alpha)})^\top$$

ただし

$$\hat{p}_{ij}^{(\alpha)} = \frac{n_{ij} + \alpha}{n + rc\alpha}$$

であり, $\Gamma(\cdot)$ はガンマ関数である. ここで, $\alpha = 0$ のときは, 事後平均 $\hat{\boldsymbol{p}}^{(\alpha)}$ は標本比率 $\hat{\boldsymbol{p}} = n^{-1}\boldsymbol{n}$ に対応している.

関数 $f(\cdot)$ を分割表の尺度とする．ここで, $f(\cdot)$ は $\boldsymbol{p}$ において少なくとも 4 回微分可能であるとする．尺度 $f(\boldsymbol{p})$ の値は, $f(\boldsymbol{p})$ において $\boldsymbol{p}$ を $\hat{\boldsymbol{p}}$ に置き換えた推定量 $f(\hat{\boldsymbol{p}})$ を用いて推定できる．また, $f(\boldsymbol{p})$ において $\boldsymbol{p}$ を $\hat{\boldsymbol{p}}^{(\alpha)}$ に置き換えた推定量 $f(\hat{\boldsymbol{p}}^{(\alpha)})$ で推定することも可能である．本講演では, ディリクレパラメータ $\alpha$ を選択する方法の 1 つとして, $f(\hat{\boldsymbol{p}}^{(\alpha)})$ の MSE を最小化する $\alpha$ を考えた．そのため, $f(\hat{\boldsymbol{p}}^{(\alpha)})$ の MSE を漸近的に評価することで, $f(\hat{\boldsymbol{p}}^{(\alpha)})$ の MSE を最小化する $\alpha$ を導出した．すなわち, 以下のようなディリクパラメータの導出を行った.

$$\tilde{\alpha} = \underset{\alpha}{\operatorname{argmin}} \lim_{n\to\infty} n^2 \mathrm{MSE}[f(\hat{\boldsymbol{p}}^{(\alpha)})]$$

推定量 $f(\hat{\boldsymbol{p}}^{(\alpha)})$ の MSE は

$$\mathrm{MSE}[f(\hat{\boldsymbol{p}}^{(\alpha)})] = \frac{1}{n^2}\left(A_1\alpha^2 - 2A_2\alpha\right) + (\text{the terms independent of } \alpha) + o(n^{-2})$$

となる．ここで

$$\begin{aligned}
A_1 =& \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}^\top}\right)(rc\boldsymbol{p} - \mathbf{1}_{rc})(rc\boldsymbol{p} - \mathbf{1}_{rc})^\top\right], \\
A_2 =& \frac{1}{2}(rc\boldsymbol{p} - \mathbf{1}_{rc})^\top\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\mathrm{tr}\left[\left(\frac{\partial^2 f(\boldsymbol{p})}{\partial \boldsymbol{p}\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right] \\
&+ rc\ \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right] \\
&+ \mathrm{tr}\left[\left(\frac{\partial f(\boldsymbol{p})}{\partial \boldsymbol{p}}\right)(rc\boldsymbol{p} - \mathbf{1}_{rc})^\top\left(\frac{\partial^2 f(\boldsymbol{p})}{\partial \boldsymbol{p}\partial \boldsymbol{p}^\top}\right)(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\right],
\end{aligned}$$

$\mathbf{1}_{rc}$ は全ての要素が 1 の $rc \times 1$ ベクトル, $\mathrm{diag}(\boldsymbol{p})$ は主対角線の要素が $\boldsymbol{p}$ の対角行列である．よって, $f(\hat{\boldsymbol{p}}^{(\alpha)})$ の MSE を最小化するディリクレパラメータ $\alpha$ は

$$\tilde{\alpha} = \underset{\alpha}{\operatorname{argmin}} \lim_{n\to\infty} n^2 \mathrm{MSE}[f(\hat{\boldsymbol{p}}^{(\alpha)})] = \frac{A_2}{A_1} \tag{1}$$

となる. 以上より, 尺度 $f(\cdot)$ の推定量として $f(\hat{\boldsymbol{p}}^{(\tilde{\alpha})})$ を提案した.

本講演では, 数値実験を用いて提案推定量が多くの場面において標本比率のプラグイン推定量や改良型推定量よりもバイアスや MSE を小さくできることを示した．さらに, ディリクレパラメータに $\alpha = 1$(一様事前分布) や $\alpha = 1/2$(Jeffreys 事前分布) を用いたり, Fienberg and Holland (1973) の方法で選択したディリクレパラメータを用いたりした場合における事後平均のプラグイン推定量よりも, 提案手法によって選択されるディリクレパラメータを用いた場合における事後平均のプラグイン推定量の方が, バイアスや MSE を小さくできることも明らかになった．また, 導出したディリクレパラメータを用いることで, 尺度の信用区間をモンテカルロシミュレーションにより構成可能であることも示した.

## 参考文献


Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, N.J., Princeton Univ. Press.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, **68**(343), 683–691.

Tomizawa, S., Miyamato, N., and Ohba, N. (2007). Improved approximate unbiased estimators of measure of asymmetry for square contingency tables. *Advances and Applications in Statistics*, **7**(1), 47–63.

Tomizawa, S., Seo, T., and Yamamoto, H. (1998). Power-divergence-type measure of departure from symmetry for square contingency tables that have nominal categories. *Journal of Applied Statistics*, **25**(3), 387–398.

# 高次元小標本データにおける多変量線形回帰モデルでの罰則を付けない推定法

永井 勇[†]

[†] 中京大学 教養教育研究院

2021 年 9 月 4 日

多様な分野における統計科学に関する理論と方法論の革新的展開



## 概要

- 多変量線形回帰モデルでよく置かれる前提条件

**前提条件**

サンプル数 $\geq$ 説明変数の次元数

- この前提を満たさない場合など

つまり 説明変数からなる行列が列フルランクではない場合

→ この前提条件を**回避**するために様々な手法を用いて推定
  - 列フルランクでない場合でも推定できる手法

⇒ よりシンプルな手法での新しい推定法の提案

- **陽に求まる**ように一工夫も
  - 提案する推定量の問題点など

⇒ 改良した推定量の構築







## モデルとリスク関数

- 多変量線形回帰モデル; $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}}$
  $\boldsymbol{Y}$; $n \times p$ 目的変数行列
  $\mathbf{1}_n$; 1 が $n$ 個並んだ定数ベクトル
  $\boldsymbol{\mu}$; $p$ 次元未知ベクトル
  $\boldsymbol{A}$; $n \times k$ 説明変数行列
  $\boldsymbol{\Xi}$; $k \times p$ 未知行列
  $\boldsymbol{\mathcal{E}}$; $n \times p$ 誤差行列
  - $p = 1$; 重回帰モデル

仮定 $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$, $\mathbf{0}_k$ は $k$ 次元の 0 ベクトル, $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$, $\mathrm{rank}(\boldsymbol{\Sigma}) = p$ ($\boldsymbol{\Sigma}$ は未知)

リスク $R(\boldsymbol{\mu}, \boldsymbol{\Xi}) \overset{\text{def.}}{=} \mathrm{tr}\{(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{A}\boldsymbol{\Xi})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{A}\boldsymbol{\Xi})'\}$

推定 $\arg\min_{\boldsymbol{\mu}, \boldsymbol{\Xi}} R(\boldsymbol{\mu}, \boldsymbol{\Xi})$



## 従来の推定での問題点と本研究の目的

推定量 $\hat{\boldsymbol{\mu}} = \boldsymbol{Y}'\mathbf{1}_n/n$, $\boldsymbol{A}'\boldsymbol{A}\hat{\boldsymbol{\Xi}} = \boldsymbol{A}'\boldsymbol{Y}$ の解

- $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Xi}}) = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Xi}} R(\boldsymbol{\mu}, \boldsymbol{\Xi})$

→ $\mathrm{rank}(\boldsymbol{A}) = k$ ならば $\hat{\boldsymbol{\Xi}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}$

**$\mathrm{rank}(\boldsymbol{A}) = k$ の前提条件**

$n \geq k$ ($\because$ $\boldsymbol{A}$ が $n \times k$ 行列)

- 大体の場合は $\mathrm{rank}(\boldsymbol{A}) = k \leq n$ が仮定されている

問題点 $n < k$ の状況で, どうやって $\boldsymbol{\Xi}$ を推定するのか？

⇒ $\mathrm{rank}(\boldsymbol{A}) < k$ のときの $\boldsymbol{\Xi}$ の推定法
- 高次元小標本データ (例; 遺伝子データ) に対応

**本研究の大まかな目的**

$\mathrm{rank}(\boldsymbol{A}) < k$ のときの $\boldsymbol{\Xi}$ の推定法の構築



## $\mathrm{rank}(\boldsymbol{A}) < k$ の状況での従来の推定法

- $\mathrm{rank}(\boldsymbol{A}) < k$ での $\boldsymbol{\Xi}$ の従来の推定法
  - $n < k$ での推定法

(i) 罰則付推定法
  例 Lasso 型 (Katayama & Imori, 2014), Ridge 型 (Yanagihara & Satoh, 2010), elastic net 型 (Lasso + Ridge) など

(ii) 変数選択 (Oda & Yanagihara, 2021) してから推定
  - $\boldsymbol{A}$ の列数 ($k$) を減らす

(iii) 主成分分析を用いてから推定
  - $\boldsymbol{A}$ を主成分分析後に推定
  名前 主成分回帰 (Fujiwara, Minamidani, Nagai & Wakaki, 2013)

(iv) 一般逆行列を用いて推定
  別名 疑似逆行列 (e.g., Srivastava, 2002, A.6)

本研究 **もっとシンプルな**推定法の提案



## 提案するモデルとリスク関数

元 $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}} \to \boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)$ と分割
- $\boldsymbol{A}$ の分割に対応して $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1', \ldots, \boldsymbol{\Xi}_r')'$ と分割
- $\boldsymbol{A}_i$; $n \times k_i$ 行列, $1 \leq r \leq k$
  $\boldsymbol{\Xi}_i$; $k_i \times p$ 行列 (未知)

仮定 $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ $(i = 1, \ldots, r)$

→ $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}' + \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i + \boldsymbol{\mathcal{E}}$

リスク $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r) \overset{\text{def.}}{=}$

$$\mathrm{tr}\left\{\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \sum_{i=1}^{r} \boldsymbol{A}_i\boldsymbol{\Xi}_i\right)'\right\}$$

Note $R(\boldsymbol{\mu}, \boldsymbol{\Xi}) = R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$



## 提案するモデルのリスク関数の問題点

- $\hat{\boldsymbol{\Xi}}_i \overset{\text{def.}}{=} \arg\min_{\boldsymbol{\Xi}_i} R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ $(i = 1, \ldots, r)$

問題点 $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ を直接展開した場合の問題点

**直接計算した際の問題点**

$\hat{\boldsymbol{\Xi}}_i$ のために $\hat{\boldsymbol{\Xi}}_j$ $(i \neq j)$ が**全部必要**

- 初期値設定が必要
- $\hat{\boldsymbol{\Xi}}_1, \ldots, \hat{\boldsymbol{\Xi}}_r$ が収束するまで繰り返し計算が必要

- 問題点の原因; $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ を展開すると, $\boldsymbol{\Xi}_i$ に関連する項 $= \boldsymbol{\Xi}_j$ $(i \neq j)$ が含まれる形

→ $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ の展開を一工夫すると, $\boldsymbol{\Xi}_i$ に関連する項 $= \boldsymbol{\Xi}_j$ $(i < j)$ **だけ**が含まれる形



## 提案モデルでの推定量

- $R'(\boldsymbol{\mu}, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_r)$ の展開に一工夫すると

**工夫して計算した結果**

$\hat{\boldsymbol{\Xi}}_i$ のために $\hat{\boldsymbol{\Xi}}_j$ $(i < j)$ **だけ**が必要

**$\hat{\boldsymbol{\Xi}}_i$ $(i = 1, \ldots, r)$ を求めた結果**

$$\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right),$$

ただし, $\sum_{j>r}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j = \mathbf{0}_n\mathbf{0}_p'$.

- $\hat{\boldsymbol{\Xi}}_r = (\boldsymbol{A}_r'\boldsymbol{A}_r)^{-1}\boldsymbol{A}_r'\boldsymbol{Y}$

結局 $r$ 番目から順に**陽に**求まる
- $\hat{\boldsymbol{\Xi}}_1$ から順に求まるようにもできる



## 提案モデルでの推定の利点

- $\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right)$ の利点

利点 $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ $(i = 1, \ldots, r)$ でよい

Note $\mathrm{rank}(\boldsymbol{A}) = k \Rightarrow \mathrm{rank}(\boldsymbol{A}_i) = k_i$
$\mathrm{rank}(\boldsymbol{A}_i) = k_i \not\Rightarrow \mathrm{rank}(\boldsymbol{A}) = k$
- $r = k$ とすると, $\boldsymbol{A}_i'\boldsymbol{A}_i \neq 0$ で推定可能
  - $r = k$ のときは $\boldsymbol{A}_i$ は $n \times 1$ ベクトル

- この前提条件; $n \geq \max(k_1, \ldots, k_r)$

- $\hat{\boldsymbol{\Xi}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}$ の前提条件; $n \geq k = \sum_{i=1}^{r} k_i$



## 提案モデルでの推定の問題点

- $\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right)$ の問題点

(I) $r$ の選択
- $\boldsymbol{A}$ や $\boldsymbol{\Xi}$ をいくつに分割するか？
- $r$ を大きくする → 前提条件や仮定を満たしやすい
  $r$ を小さくする → 前提条件や仮定を満たしにくい

(II) $\boldsymbol{A}_i$ $(i = 1, \ldots, r)$ の選択
- $\boldsymbol{A}$ の分割の仕方はどうするか？
- $\mathrm{rank}(\boldsymbol{A}_i) = k_i$ の条件下での選択

(III) $\hat{\boldsymbol{\Xi}}_i$ は**不偏推定量ではない**
- $E[\hat{\boldsymbol{\Xi}}_i] \neq \boldsymbol{\Xi}_i$ $(i = 1, \ldots, r)$
  後述 $(\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'$ の項を工夫すると, 不偏推定量にもなる



## もっと改良する

- 提案した推定量;

$$\hat{\boldsymbol{\Xi}}_i = (\boldsymbol{A}_i'\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\hat{\boldsymbol{\Xi}}_j\right)$$

注意点 $E[\hat{\boldsymbol{\Xi}}_i] \neq \boldsymbol{\Xi}_i$ $(i = 1, \ldots, r)$

→ $E[\boldsymbol{Y}] = \mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi}$ の不偏推定量**ではない**

⇔ 一般逆行列を用いた場合; $E[\boldsymbol{Y}]$ の不偏推定量

⇒ $\hat{\boldsymbol{\Xi}}_i$ を不偏推定量になるようにすれば...
より予測精度を上げられるのでは？

## 改良結果

- 求め方を変えると，
  - 改めて $\tilde{\Xi}_i = \arg\min_{\Xi} R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ $(i = 1, \ldots, r)$

**$\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ を求めた結果**

$$\tilde{\Xi}_i = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\tilde{\Xi}_j\right),$$

ただし $\sum_{j>r} \boldsymbol{A}_j\tilde{\Xi}_j = \boldsymbol{0}_n\boldsymbol{0}_p'$, ここで $\boldsymbol{Q}_1 = \boldsymbol{A}_1$,

$$\boldsymbol{Q}_i = \left(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{Q_j}\right)\boldsymbol{A}_i \ (i = 2, \ldots, r) \text{ とし，}$$

$\mathrm{rank}(\boldsymbol{Q}_i) = k_i$ を仮定.

**$\tilde{\Xi}_i$ の性質; 不偏推定量**



## 一工夫の実際

- $R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ を展開結果 ($\Xi_1, \ldots, \Xi_r$ に関する項);

$$-2\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_i') + \sum_{i=1}^{r}\sum_{j=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

Step 1 この項を分解すると，

$$\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i') + \sum_{i=1}^{r}\sum_{j\neq i} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

Step 2 この項を $j > i$ と $j < i$ で分割

→ $\mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j') = \mathrm{tr}(\boldsymbol{A}_j\Xi_j\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i')$ を使うと，

この項 $= 2\sum_{i=1}^{r}\sum_{j>i} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$

Step 3 微分して 0 を解いて，$\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ を求めた



## 一工夫の途中までで止めてみる

- Step 2 前で止めると，
  $R'(\boldsymbol{\mu}, \Xi_1, \ldots, \Xi_r)$ を展開結果 ($\Xi_1, \ldots, \Xi_r$ に関する項);

$$-2\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_i') + \sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_i'\boldsymbol{A}_i')$$

$$+\sum_{i=1}^{r}\sum_{j\neq i} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_j'\boldsymbol{A}_j')$$

- この項この項で $i = \ell \neq j$ と $i \neq \ell = j$ の場合で分割

→ $\Xi_\ell$ に関する項だけピックアップすると，

$$-2\mathrm{tr}(\boldsymbol{A}_\ell'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_\ell') + \mathrm{tr}(\boldsymbol{A}_\ell\Xi_\ell\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$
$$+2\sum_{i\neq\ell} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$

→ これを最小にする $\Xi_\ell$ を求める



## $\tilde{\Xi}_\ell$ とその変形 (1/3)

- $\Xi_\ell$ について，最小化する関数;

$$-2\mathrm{tr}(\boldsymbol{A}_\ell'\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\Xi_\ell') + \mathrm{tr}(\boldsymbol{A}_\ell\Xi_\ell\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$
$$+2\sum_{i\neq\ell} \mathrm{tr}(\boldsymbol{A}_i\Xi_i\boldsymbol{\Sigma}^{-1}\Xi_\ell'\boldsymbol{A}_\ell')$$

→ $\boldsymbol{A}_\ell'\boldsymbol{A}_\ell\tilde{\Xi}_\ell = \boldsymbol{A}_\ell'\left(\boldsymbol{Y} - \sum_{i\neq\ell} \boldsymbol{A}_i\tilde{\Xi}_i\right)$ $(\ell = 1, \ldots, r)$

→ $\tilde{\Xi}_1 = (\boldsymbol{A}_1'\boldsymbol{A}_1)^{-1}\boldsymbol{A}_1'\left(\boldsymbol{Y} - \sum_{i>1} \boldsymbol{A}_i\tilde{\Xi}_i\right)$,

$\boldsymbol{A}_2'\boldsymbol{A}_2\tilde{\Xi}_2 = \boldsymbol{A}_2'\left(\boldsymbol{Y} - \boldsymbol{A}_1\tilde{\Xi}_1 - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right)$.



## $\tilde{\Xi}_\ell$ の変形 (2/3) と仮定

- $\tilde{\Xi}_1$ を $\tilde{\Xi}_2$ に代入して変形すると，

$$\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\boldsymbol{A}_2\tilde{\Xi}_2 = \boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\left(\boldsymbol{Y} - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right),$$

ここで，$\boldsymbol{P}_M = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'$.

仮定 $\mathrm{rank}((\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\boldsymbol{A}_2) = k_2$ を仮定すると，

$$\tilde{\Xi}_2 = \{\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\boldsymbol{A}_2\}^{-1}\boldsymbol{A}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})\left(\boldsymbol{Y} - \sum_{i>2} \boldsymbol{A}_i\tilde{\Xi}_i\right)$$

⇒ $\mathrm{rank}((\boldsymbol{I}_n - \boldsymbol{P}_{A_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})A_2})\boldsymbol{A}_3) = k_3$ も仮定すると，

$$\tilde{\Xi}_3 = \{\boldsymbol{A}_3'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})A_2})\boldsymbol{A}_3\}^{-1} \times$$
$$\boldsymbol{A}_3'(\boldsymbol{I}_n - \boldsymbol{P}_{A_1} - \boldsymbol{P}_{(\boldsymbol{I}_n - \boldsymbol{P}_{A_1})A_2})\left(\boldsymbol{Y} - \sum_{i>3} \boldsymbol{A}_i\tilde{\Xi}_i\right)$$



## $\tilde{\Xi}_\ell$ の変形 (3/3) の結果

- 繰り返していくと，

**$\tilde{\Xi}_i$ $(i = 1, \ldots, r)$ を求めた結果**

$$\tilde{\Xi}_i = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\left(\boldsymbol{Y} - \sum_{j>i}^{r} \boldsymbol{A}_j\tilde{\Xi}_j\right),$$

ただし $\sum_{j>r} \boldsymbol{A}_j\tilde{\Xi}_j = \boldsymbol{0}_n\boldsymbol{0}_p'$, ここで $\boldsymbol{Q}_1 = \boldsymbol{A}_1$,

$$\boldsymbol{Q}_i = \left(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{Q_j}\right)\boldsymbol{A}_i \ (i = 2, \ldots, r) \text{ とし，}$$

$\mathrm{rank}(\boldsymbol{Q}_i) = k_i$ を仮定.

Note $(\boldsymbol{I}_n - \sum_{j<i} \boldsymbol{P}_{Q_j})$ は射影行列

予想 $\mathrm{rank}(\boldsymbol{Q}_r) = k_r$ のために $n \geq k$ が必要



## 変形結果の性質

- こんな変形して何が嬉しい？

**$\tilde{\Xi}_i$ の性質; 不偏性**

$E[\tilde{\Xi}_i] = \Xi_i$ $(i = 1, \ldots, r)$

Note $E[\tilde{\Xi}_i] \neq \Xi_i$ $(i = 1, \ldots, r)$

- $E[(\tilde{\Xi}_1', \ldots, \tilde{\Xi}_r')'] = (\Xi_1', \ldots, \Xi_r')' = \Xi$

⇒ $\mathrm{rank}(\boldsymbol{A}) < k$ でも
$\Xi$ の不偏推定量が得られる

Note $E[\boldsymbol{Y}] = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\Xi$ の不偏推定量でもある

Note 一般逆行列を用いた場合は，
$E[\boldsymbol{Y}]$ の不偏推定量だが，$E[\Xi]$ **の不偏推定量でない**



## $E[\tilde{\Xi}_i] = \Xi_i$ の証明の概略

- $E[\tilde{\Xi}_i] = \Xi_i$ $(i = 1, \ldots, r)$ の証明の概略

1. $\boldsymbol{Q}_i'\boldsymbol{1}_n = \boldsymbol{0}_{k_i}$ $(i = 1, \ldots, r)$ を示す
   ∵ $\boldsymbol{A}_i'\boldsymbol{1}_n = \boldsymbol{0}_{k_i}$
2. $\boldsymbol{Q}_i'\boldsymbol{A}_j = \boldsymbol{0}_{k_i}\boldsymbol{0}_{k_j}'$ $(j < i \leq r)$ を示す
3. $(\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'\boldsymbol{A}_i = \boldsymbol{I}_{k_i}$ $(i = 1, \ldots, r)$ を示す
4. $E\left[\boldsymbol{Y} - \sum_{j>i} \boldsymbol{A}_j\tilde{\Xi}_j\right] = \boldsymbol{1}_n\boldsymbol{\mu}' + \sum_{j<i} \boldsymbol{A}_j(\Xi_j - E[\tilde{\Xi}_j]) + \boldsymbol{A}_i\Xi_i$
   を示す
5. $E[\tilde{\Xi}_i] = (\boldsymbol{Q}_i'\boldsymbol{Q}_i)^{-1}\boldsymbol{Q}_i'E\left[\boldsymbol{Y} - \sum_{j>i} \boldsymbol{A}_j\tilde{\Xi}_j\right]$ と
   これまで示したことを使う



## まとめ

- 多変量線形回帰モデル $\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\Xi + \boldsymbol{\mathcal{E}}$ での従来の推定の問題点
  $n \geq k = \mathrm{rank}(\boldsymbol{A})$ を前提にしている
  → $\mathrm{rank}(\boldsymbol{A}) < k$ のときの推定法を構築
- 従来 罰則付推定，$\boldsymbol{A}$ を変数選択して推定，
  $\boldsymbol{A}$ を主成分分析して回帰，一般逆行列を使う
  ⇒ 目的；もっと**シンプルな**推定法の構築
- $\boldsymbol{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_r)$, $\Xi = (\Xi_1', \ldots, \Xi_r')'$ として，
  $\Xi$ の推定 ⇔ $\Xi_1, \ldots, \Xi_r$ の推定
- 結果 前提条件が**かなり**緩く，反復計算が**不要**な推定法
  - 分割数 $(r)$ を大きくすれば，前提条件をかなり緩和可能
- 仮定を追加して，$\Xi$ の不偏推定量の構築



## 今後の課題

1. $r$ ($\boldsymbol{A}$ の分割数) や分割の中身の選択
2. $\mathrm{rank}(\boldsymbol{A}) < k$ での $\Xi$ の不偏推定量の数値実験
3. (予想が正しい場合) $n \geq \sum_{i=1}^{j} k_i$ で $n < \sum_{i=1}^{j+1} k_i$ のような
   $n$ で $\Xi$ の一部の不偏推定量の構築
   予想! 不偏推定量のために $n \geq k = \sum_{i=1}^{r} k_i$ が必要？
   - ある程度の標本数で，途中までの不偏性を持つ推定量ができないか？
4. 罰則付推定量の構築
   - $\mathrm{rank}(\boldsymbol{A}) < k$ で PMSE をより小さくする推定量が作れないか？
5. 他のモデル (GMANOVA モデル) などへの拡張



## 参考文献


1. Fujiwara, M., Minamidani, T., Nagai, I. and Wakaki, H. (2013). Principal components regression by using generalized principal components analysis. *J. Jpn. Stat. Soc.*, **43**, 57–78.
2. Katayama, S. & Imori, S. (2014) Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *J. Multivariate Anal.*, **132**, 138–150.
3. Oda, R. & Yanagihara, H. (2021) A consistent likelihood-based variable selection method in normal multivariate linear regression. *Smart Systems and IoT: Innovations in Computing*, (in press).
4. Srivastava, M. S. (2002) *Methods of Multivariate Statistics*, Wiley-Interscience.
5. Yanagihara, H. and Satoh, K. (2010). A unbiased $C_p$ criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.




終

# Cauchy 型分布の単純な EM アルゴリズムについて

法政大学 経済学部
阿部 俊弘

対称分布の非対称化として, Azzalini (1985) に始まる skew-symmetric 分布が良く知られている. これの発展については, Azzalini (2005) や Azzalini & Capitanio (1999) にまとめられている. Lin et al. (2007) では確率表現を活用して歪正規分布の EM アルゴリズムを与えている. この歪正規分布の多変量化はいくつのものが提案されているが, Sahu et al. (2003) の多変量歪正規分布についての EM アルゴリズムは Lin (2009) により与えられている.

Lin et al. (2007) や Lin (2009) による EM アルゴリズムでは, skew パラメータに対する推定式が陽に解けないため, 数値的手法が必要となる. これに対して, Abe et al. (2021) では, 歪正規分布の確率表現に overparameter を導入することにより, EM アルゴリズムの陽的表現を導出し, さらに, Chen et al. (2014) と比較し, 推定手法として優れていることを示した. 歪正規分布は様々な良い性質がある一方で, 歪正規分布の Fisher information matrix は skew パラメータが 0 の周りで, 特異になる分布であることが指摘されている (Azzalini, 2013; Hallin & Ley, 2014).

この講演では, まず初めに対称な Cauchy 型分布の EM アルゴリズムを統一的な視点で与え, 関連したモデルについても EM アルゴリズムを与える: 対数 Cauchy 分布は裾の重い分布の一例であり, その名前から明らかなように, $Z$ が Cauchy 分布に従うときに $X = \exp(Z)$ が対数 Cauchy 分布に従う. 対数 Cauchy 分布は, Cauchy 分布と同様に一切の (非自明) モーメントが無限大になり, しばしば非常に裾の重い分布 (super-heavy tailed distribution) とされる. 対数 Cauchy 分布のパラメータ推定に関して, 標本の自然対数をとったものの中央値は, $\mu$ のロバストな推定量になり, 標本の自然対数をとったものの中央絶対偏差は $\sigma$ のロバストな推定量になる. 対数 Cauchy 分布は, 有意な外れ値または極値が発生するようなある種の生存過程や生物種の豊富度パターンのモデル化に用いることができる. 例としては, ヒト免疫不全ウイルスの感染から発症までの時間が挙げられる.

上記の Cauchy 分布と同様にして, 密度関数の形から $t$ 分布の EM アルゴリズムも与えていく. Liu & Rubin (1995) でも $t$ 分布の確率表現から EM アルゴリズムの陽的表現を導出しており, 彼らのものとは別の観点のものであるが, M-step での解は本質的に彼らのものと同じものを導出する.

上で述べた歪正規分布にはいくつかの別形があるが, $t$ 分布の歪分布化についても様々なものが提案されている. skew-symmetric 分布の pdf については有名なもので次のような表現がある (see Azzalini, 2005): まず, $G$ を $x = 0$ の周りで対称な分布の cdf とし, $f_0$ を $x = 0$ の周りで対称な分布の pdf とする. このとき, skew-symmetric 分布の pdf

$$2G(\lambda x) f_0(x)$$

において $x \mapsto \dfrac{x-\mu}{\sigma}$ とすると,

$$\frac{2}{\sigma} G\left(\lambda \frac{x-\mu}{\sigma}\right) f_0\left(\frac{x-\mu}{\sigma}\right)$$

となる. Azzalini & Capitanio (2003) では, 歪正規分布に scale mixture をすることにより, pdf が

$$f_{ACST}(x;\nu) = \frac{2}{\sigma} F_T\left(\lambda \frac{x-\mu}{\sigma} \sqrt{\frac{\nu+1}{(x-\mu)^2/\sigma^2+\nu}}; 1, \nu+1\right) f_T\left(\frac{x-\mu}{\sigma}; \nu\right) \tag{1}$$

である skew-$t$ 分布を導出している．ここで, $f_T$, $F_T$ はそれぞれ, 自由度 $\nu > 0$ の $t$ 分布の pdf と cdf

$$f_T(x;\nu) = \frac{1}{\sqrt{\nu}B\left(\frac{\nu}{2},\frac{1}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad F_T(x;\sigma,\nu) := \frac{1}{\sigma}\int_{-\infty}^{x} f_T\left(\frac{t}{\sigma};\nu\right)dt \tag{2}$$

であり, $B(\cdot,\cdot)$ は Beta 関数である．

## References


Abe, T., Fujisawa, H., Kawashima, T. & Ley, C. (2021). EM algorithm using overparameterization for multivariate skew-normal distribution. Econometrics and Statistics, **19**, 151–168.

Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics **12(2)**, 171–178.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. Scandinavian Journal of Statistics **32** (2), 159–200.

Azzalini, A. (2013). The Skew-Normal and Related Families. vol. 3. Cambridge University Press.

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61(3)**, 579–602.

Azzalini, A. & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew $t$-distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **65(2)**, 367–389.

Chen, L., Pourahmadi, M. & Maadooliat, M. (2014). Regularized multivariate regression models with skew-$t$ error distributions. Journal of Statistical Planning and Inference **149**, 125–139.

Hallin, M. & Ley, C. (2014). Skew-symmetric distributions and Fisher information: the double sin of the skew-normal. Bernoulli **20**, 1432–1453.

Lin, T.I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. Journal of Multivariate Analysis **100** (2), 257–265.

Lin, T.I., Lee, J.C. & Yen, S.Y. (2007). Finite mixture modelling using the skew normal distribution. Statistica Sinica **17**, 909–927.

Liu, C. & Rubin, D. B. (1995). ML estimation of the $t$ distribution using EM and its extensions, ECM and ECME. Statistica Sinica, 19–39.

Ma, Y. & Genton, M. G. (2004). Flexible class of skew-symmetric distributions. Scandinavian Journal of Statistics, **31(3)**, 459–468.

Sahu, S.K., Dey, D.K. & Branco, M.D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. Canadian Journal of Statistics **31** (2), 129–150.

# Distance covariance for random fields

Nanzan University Muneya Matsui

## Abstract

We study an independence test based on distance correlation for random fields $(X, Y)$. We consider the situations when $(X, Y)$ is observed on a lattice with equidistant grid sizes and when $(X, Y)$ is observed at random locations. We provide asymptotic theory for the sample distance correlation in both situations and show bootstrap consistency. The latter fact allows one to build a test for independence of $X$ and $Y$ based on the considered discretizations of these fields. We illustrate the performance of the bootstrap test in a simulation study involving fractional Brownian and infinite variance stable fields. The independence test is applied to Japanese meteorological data, which are observed over the entire area of Japan.

## Introduction

It is well known that two $q$- and $r$-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively, are independent if and only if their joint characteristic function factorizes, i.e.,

$$\varphi_{\mathbf{X},\mathbf{Y}}(\mathbf{s},\mathbf{t}) = \mathbb{E}[\exp(i\,\mathbf{s}^\top\mathbf{X} + i\mathbf{t}^\top\mathbf{Y})] = \mathbb{E}[\exp(i\,\mathbf{s}^\top\mathbf{X})]\,\mathbb{E}[\exp(i\,\mathbf{t}^\top\mathbf{Y})] = \varphi_{\mathbf{X}}(\mathbf{s})\,\varphi_{\mathbf{Y}}(\mathbf{t})\,,\ \mathbf{s}\in\mathbb{R}^q\,,\mathbf{t}\in\mathbb{R}^r.$$

Since this identity should hold with any $(\mathbf{s},\mathbf{t}) \in \mathbb{R}^q \times \mathbb{R}^r$, in general it is recommended to use a weighted $L^2$-distance between $\varphi_{\mathbf{X},\mathbf{Y}}$ and $\varphi_{\mathbf{X}}\,\varphi_{\mathbf{Y}}$ for the test of independence.

We specifically call the weighted $L^2$-distance as *distance covariance between* $\mathbf{X}$ *and* $\mathbf{Y}$ if the weight function is given as

$$T_\beta(\mathbf{X},\mathbf{Y}) = c_q c_r \int_{\mathbb{R}^{q+r}} \left|\varphi_{\mathbf{X},\mathbf{Y}}(\mathbf{s},\mathbf{t}) - \varphi_{\mathbf{X}}(\mathbf{s})\varphi_{\mathbf{Y}}(\mathbf{t})\right|^2 |\mathbf{s}|^{-(q+\beta)}|\mathbf{t}|^{-(r+\beta)}\,d\mathbf{s}\,d\mathbf{t}\,,\ \beta\in(0,2),$$

where the constants $c_d$ for $d \geq 1$ are chosen such that $c_d \int_{\mathbb{R}^d}(1-\cos(\mathbf{s}'\mathbf{x}))\,|\mathbf{x}|^{-(d+\beta)}d\mathbf{x} = |\mathbf{s}|^\beta$. The quantity $T_\beta(\mathbf{X},\mathbf{Y})$ is finite under suitable moment conditions on $\mathbf{X},\mathbf{Y}$. The corresponding *distance correlation* is given by

$$R_\beta(\mathbf{X},\mathbf{Y}) = \frac{T_\beta(\mathbf{X},\mathbf{Y})}{\sqrt{T_\beta(\mathbf{X},\mathbf{X})}\,\sqrt{T_\beta(\mathbf{Y},\mathbf{Y})}}\,.$$

Of course, $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if $R_\beta(\mathbf{X},\mathbf{Y}) = T_\beta(\mathbf{X},\mathbf{Y}) = 0$. Thanks to the choice of the weight function, $T_\beta(\mathbf{X},\mathbf{Y})$ has an explicit form: assuming that $(\mathbf{X}_i,\mathbf{Y}_i)$, $i = 1,2,\ldots$, are iid copies of $(\mathbf{X},\mathbf{Y})$, we have

$$T_\beta(\mathbf{X},\mathbf{Y}) = \mathbb{E}[|\mathbf{X}_1-\mathbf{X}_2|^\beta|\mathbf{Y}_1-\mathbf{Y}_2|^\beta] + \mathbb{E}[|\mathbf{X}_1-\mathbf{X}_2|^\beta]\mathbb{E}[|\mathbf{Y}_1-\mathbf{Y}_2|^\beta] - 2\,\mathbb{E}[|\mathbf{X}_1-\mathbf{X}_2|^\beta|\mathbf{Y}_1-\mathbf{Y}_3|^\beta]\,,$$

and $R_\beta(c\mathbf{X},c\mathbf{Y}) = R_\beta(\mathbf{X},\mathbf{Y})$ for $c\in\mathbb{R}$, i.e., $R_\beta$ is scale-invariant.

With the final form, the distance covariance for vectors is easily extended into that for random fields. Let $(X,Y)$ be a pair of random fields on $B \subset \mathbb{R}^d$ and define the norm $\|f\|_2 = (|B|^{-1}\int_B f^2(\mathbf{u})d\mathbf{u})^{1/2}$. The distance covariance $T_\beta(X,Y)$, $\beta\in(0,2)$, between two random fields $X,Y$ on some bounded Borel set $B\subset\mathbb{R}^d$ of finite positive Lebesgue measure is defined by

$$\begin{aligned} T_\beta(X,Y) = {} & \mathbb{E}[\|X_1-X_2\|_2^\beta\,\|Y_1-Y_2\|_2^\beta] + \mathbb{E}[\|X_1-X_2\|_2^\beta]\,\mathbb{E}[\|Y_1-Y_2\|_2^\beta] \\ & - 2\,\mathbb{E}[\|X_1-X_2\|_2^\beta\|Y_1-Y_3\|_2^\beta]\,, \end{aligned}$$

where $(X_i,Y_i)$, $i=1,2,\ldots$, are iid copies of $(X,Y)$, and the distance correlation $R_\beta(X,Y)$ is defined correspondingly. We apply the statistic $R_\beta(X,Y)$ to an independence for random fields.

Usually a sample of whole paths of $(X, Y)$ is rarely at our disposal, and we observe $(X, Y)$ on a lattice with equidistant grid sizes or observe $(X, Y)$ at random locations. Accordingly asymptotic behavior of the test statistic depends on each of sampling schemes. We provide asymptotic theory for the sample distance correlation in both situations and show bootstrap consistency. The latter fact allows one to build a test for independence of $X$ and $Y$ based on the considered discretizations of these fields. We illustrate the performance of the bootstrap test in a simulation study involving fractional Brownian and infinite variance stable fields. The independence test is applied to Japanese meteorological data, which are observed over the entire area of Japan.

### Main contents

**1**. Distance covariance for random fields on a lattice in $[0, 1]^d$.
We provide asymptotic theory for the lattice case on $B = [0, 1]^d$ for increasing intensity $p$ which is the total (deterministic) number of grid points. We show that the discretized sample distance correlation converges to non-discretized whole path sample distance correlation as $p_n \to \infty$ depending on the sample size $n \to \infty$. Therefor they have the same asymptotic distribution. We show the bootstrap consistency for this sampling scheme.

**2**. Distance covariance for random fields at random locations.
We define $T_\beta(X^{(p)}, Y^{(p)})$ and $R_\beta(X^{(p)}, Y^{(p)})$ for non-lattice based discretizations $X^{(p)}, Y^{(p)}$ of $X, Y$ on $B$. We choose a random number $N_p$ of random locations $(\mathbf{U}_i)$ where the processes $X, Y$ are observed. Typically, these locations are uniformly distributed on $B$ and $N_p \overset{a.s.}{\to} \infty$ as $p$ increases with the sample size $n$ to infinity. We show that the sample distance correlation at random locations well approximate the whole path sample distance correlation as $p_n \to \infty$. Moreover we show that $n\, T_{n,\beta}(X^{(p)}, Y^{(p)})$ conditional on $(N^{(p)})_{p>0}$ has the same weak limit as $n\, T_{n,\beta}(X, Y)$. We provide consistency of a suitable bootstrap procedure in this case.

**3**. A Monte Carlo study.
We conduct a Monte Carlo study of the finite sample behavior of the sample distance correlation for $\beta = 1$. We consider a fractional Brownian sheet and, as a heavy-tailed alternative, we choose symmetric 1.8-stable Lévy sheets. The observations are given either on a lattice or at random locations in $[0, 1]^2$. In addition, we illustrate the performance of the bootstrap procedure for the test for independence based on distance correlation in the cases of fixed locations on a lattice and of randomly scattered locations. We focus on independent pairs $X, Y$, Brownian or 1.8-stable sheets.

**4**. An application to Japanese meteorological data.
We apply our results to Japanese meteorological data. We choose the 3 most fundamental factors: *temperature (temp), precipitation (prec) and wind speed (wind)* which have been observed for a long time and over a wide range of Japan. We focus on the distance correlations for the pairs (prec & temp), (prec & wind), (temp & wind) and conduct the bootstrap tests for pair-wise independence.

## References


[1] Dehling, H.G., Matsui, M., Mikosch, T., Samorodnitsky, G. and Tafakori, L. (2020) Distance covariance for discretized stochastic processes. *Bernoulli* **26**, 2758–2789.

[2] Matsui, M., Mikosch, T. and Samorodnitsky, G. (2017) Distance covariance for stochastic processes. *Probab. Math. Statist.* **37**, 355–372.

[3] Matsui, M., Mikosch, T., Roozegar, R. and Tafakori, L. (2021) Distance covariance for random fields. *arXiv:2107.03162.*

# Time Series Quantile Regressions by using Random Forest
# (ランダムフォレストを用いた時系列分位点回帰)

Ryotaro Shibuki* , Tomoshige Nakamura† , Hiroshi Shiraishi‡

In this paper, we discuss an estimation procedure of conditional quantile by using random forests in time series setting. Our study is an extension of the quantile random forest (QRF) by Meinshausen (2006), the generarized random forest (GRF) by Athey et. al (2019) and random forest in time series setting by Davis and Nielsen (2020).

**Model** Let $(\varepsilon_t)_{t\geq 1}$be a sequence of i.i.d. random variables with $\mathbb{E}[\varepsilon_t]=0$ and $\mathbb{E}[\varepsilon_t^2]<\infty$, and fix an integer $p\geq 1$. Given a measurable function $g:\mathbb{R}^p\to\mathbb{R}$, define the process $(Y_t)_{t\geq 1}$ recursively by

$$Y_t = g(X_t)+\varepsilon_t, \quad \boldsymbol{X}_t=(Y_{t-1},\ldots,Y_{t-p}). \tag{1}$$

In addition to the initial data $\eta=(Y_0,Y_{-1},\cdots,Y_{1-p})$, suppose that we have $T$ observations $Y_1,\ldots,Y_T$ from the model (1) available and that we group them in input-output pairs, $\mathcal{D}_T=\{(\boldsymbol{X}_1,Y_1),\ldots,(\boldsymbol{X}_T,Y_T)\}$. For each fixed value $\tau\in(0,1)$, we seek forest-based (function) estimator of $q_*^\tau:\mathbb{R}^p\to\mathbb{R}$ defined by a solution of local estimating equation of the form

$$\Psi^\tau(q_*^\tau,\boldsymbol{x}) := \mathbb{E}[\psi_{q_*^\tau}^\tau(Y_t)|\boldsymbol{X}_t=\boldsymbol{x}]=0, \quad \text{for all } \boldsymbol{x}\in\mathbb{R}_t^p \tag{2}$$

where $\psi_q^\tau(y)=\tau-\mathbf{1}_{\{y\leq q\}}$. Let $q_*^\tau\equiv(q_*^\tau(\boldsymbol{x}))_{\boldsymbol{x}\in\mathbb{R}^p}$ be the solution of (2) under the model (1). We assume that there exists $q_*^\tau(\boldsymbol{x})$ for all $\boldsymbol{x}\in\mathbb{R}^p$ and $\tau\in(0,1)$.

**Double sample** We next define our random forests following Athey et. al (2019). Our random forests consists of the double sample trees, which are regression trees based on two subsamples $\mathcal{I}_s$ and $\mathcal{J}_s$ from sample $\mathcal{D}_T$.

**Definition 1.** *(Double Sample) Suppose that sample $\mathcal{D}_T$ is available and the sub-sample size $s=s(T)$ with $s\leq T$ is provided. Let*

$$\mathcal{A}_s := \left\{ A=A^{\mathcal{I}}\cup A^{\mathcal{J}})\subset\{1,2,\ldots,T\} \middle| A^{\mathcal{I}}\cap A^{\mathcal{J}})=\emptyset,\ |A^{\mathcal{I}}|=\left\lfloor\frac{s}{2}\right\rfloor,\ |A^{\mathcal{J}}|=\left\lceil\frac{s}{2}\right\rceil \right\}$$

*For any $A=A^{\mathcal{I}}\cup A^{\mathcal{J}})\in\mathcal{A}_s$, we define two sub-samples $\mathcal{I}_s$ and $\mathcal{J}_s$ by $\mathcal{I}_s=\mathcal{D}_{A^{\mathcal{I}}},\mathcal{J}_s=\mathcal{D}_{A^{\mathcal{J}}}$ where $\mathcal{D}_{\boldsymbol{\theta}\cdot}=\{(\boldsymbol{X}_t,Y_t)\}_{t\in A^{\cdot}}$.*

**Splitting rule** We next define splitting rule in order to construct the double-sample regression trees following Wager and Athey (2018).

**Definition 2.** *(Splitting rule) Given subsample $\mathcal{J}_s$ in Definition 1, we define a sequence of partitions $\mathcal{P}_0,\mathcal{P}_1,\ldots$ by starting form $\mathcal{P}_0=\{\mathbb{R}^p\}$ and then, for each $\ell\geq 1$, construct $\mathcal{P}_\ell$ from $\mathcal{P}_{\ell-1}$ by replacing one set (parent node) $P\in\mathcal{P}_{\ell-1}$ by (child node) $C_1:=\{\boldsymbol{x}=(x_1,\ldots,x_p)\in P\subset\mathbb{R}^p : x_\xi\leq\zeta\}$ and $C_2:=\{\boldsymbol{x}=(x_1,\ldots,x_p)\in P\subset\mathbb{R}^p : x_\xi>\zeta\}$, where the split direction $\xi\in\{1,\ldots,p\}$ is randomly chosen (i.e., random split) .*

*Graduate School of Science and Technology, Keio University
†Faculty of Science and Technology, Keio University
‡Faculty of Science and Technology, Keio University

**Double-sample regression trees** A given partition $\Lambda$ of $\mathbb{R}^p$ is called "recursive" if $\Lambda = \mathcal{P}_\ell$ for some $\ell \geq 0$, where $\mathcal{P}_0, \ldots, \mathcal{P}_\ell$ are obtained as above. Note that the splitting rules determining how to choose node, direction and position of a split may depend on the data $\mathcal{D}_T$, the double sampling procedure $A \in \mathcal{A}_s$ and the sequence of independent random splittings $\boldsymbol{\xi} = \{\xi_i\}_{i=1,\ldots,\ell}$ with $\xi_i \overset{i.i.d.}{\sim} \Xi$. By using the recursive partition $\Lambda$, we define our double-sample regression trees.

**Definition 3.** *Given a recursive partition $\Lambda(A, \boldsymbol{\xi}, \mathcal{D}_T) = \{L_1, \ldots, L_{|\Lambda|}\}$ and a fixed $\boldsymbol{x} \in \mathbb{R}^d, q \in \mathbb{R}, \tau \in (0,1)$, our double-sample regression tree $\mathcal{T}(q, \boldsymbol{x}; A, \boldsymbol{\xi}, \mathcal{D}_T)$ is defined by*

$$\mathcal{T}(q, \boldsymbol{x}; A, \boldsymbol{\xi}, \mathcal{D}_T) = \sum_{t \in A^{\mathcal{I}}} \frac{\mathbf{1}_{\{\boldsymbol{X}_t \in L_A(\boldsymbol{x})\}}}{|L_A(\boldsymbol{x})|} \psi_q^\tau(Y_t)$$

*where $L_A(\boldsymbol{x}) = \{\boldsymbol{X}_t : \boldsymbol{X}_t \in \tilde{L}(\boldsymbol{x})\} \cap \mathcal{I}_s$ and $\tilde{L}(\boldsymbol{x}) \in \Lambda(A, \boldsymbol{\xi}, \mathcal{D}_T)$ is a leaf containing $\boldsymbol{x}$ (i.e., $\boldsymbol{x} \in \tilde{L}_A(\boldsymbol{x})$).*

**Random Forests** According to Wager and Athey (2018), the predictor $\mathcal{T}$ defined by Definition 3 is called "$k$- PNN predictor" if the assumption (A-3) is satisfied. Then, we define our random forests following Athey et. al (2019).

**Definition 4.** *For a fixed $\boldsymbol{x} \in \mathbb{R}^d, q \in \mathbb{R}, \tau \in (0,1)$, our forest score is defined by*

$$\Psi_T^\tau(q, \boldsymbol{x}) := \frac{1}{|\mathcal{A}_s|} \sum_{A \in \mathcal{A}_s} \mathcal{T}(q, \boldsymbol{x}; A, \boldsymbol{\xi}, \mathcal{D}_T) =: \sum_{t=1}^T \alpha_t(\boldsymbol{x}) \psi_q^\tau(Y_t)$$

*where $\alpha_t(\boldsymbol{x}) = \frac{1}{|\mathcal{A}_s|} \sum_{A \in \mathcal{A}_s} \alpha_{A,t}(\boldsymbol{x})$ and $\alpha_{A,t}(\boldsymbol{x}) = \frac{\mathbf{1}_{\{\boldsymbol{X}_t \in L_A(\boldsymbol{x})\}}}{|L_A(\boldsymbol{x})|}$.*

**Quantile estimator** By using the above random forest, we can define an estimator of the conditional quantile as follows.

**Definition 5.** *For each $\tau \in (0,1)$ and given $X_t = x$, we define an estimator of $q_*^\tau(x)$ by*

$$\hat{q}_T^\tau(x) \in \arg\min_{q \in \mathbb{R}} \left\{ \left\| \sum_{t=1}^T \alpha_t(x) \psi_q^\tau(Y_t) \right\|_2 \right\}.$$

**Theorem 1.** *Under some regularity conditions and subsample size $s(T)$ satisfies $s(T)/T \to 0$ and $s(T) \to \infty$ as $T \to \infty$. For each $\tau \in (0,1)$ and $\boldsymbol{x} \in \mathbb{R}^p$, any sequence of estimators $\hat{q}_T^\tau(\boldsymbol{x})$ converges in probability to $q_*^\tau(\boldsymbol{x})$, that is,*

$$\hat{q}_T^\tau(\boldsymbol{x}) \xrightarrow{p} q_*^\tau(\boldsymbol{x}) \quad \text{as } T \to \infty.$$

## 参考文献


An, H. Z., and F. C. Huang. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinical*, 6(4), 943-956.

Athey, P., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2), 1148-1178.

Davis, Richard A. and Nielsen, Mikkel S. (2020). Modeling of time series using random forests: theoretical developments. *Electronic Journal of Statistics*, 14(2), 3644-3671.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research (JMLR)*, 7, 983-999.

Wager, S. and Athey, P. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.

# Analysis of Spatially Resolved Spectral Map of a Galaxy by the High-Dimensional Statistical Method

Tsutomu T. TAKEUCHI[1,2], Kai T. KONO[1], Kazuyoshi YATA, Makoto AOSHIMA[3], Aki ISHII[4], Kento EGASHIRA[5], Kouichiro NAKANISHI[6], Suchetha COORAY[1,†], Kotaro KOHNO[7]

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

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

3. Institute of Mathematics, University of Tsukuba, Ibaraki 305-8571, Japan

4. Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Chiba 278-8510, Japan

5. Graduate School of Science and Technology, University of Tsukuba, Ibaraki 305-8571, Japan

6. The ALMA Project, National Astronomical Observatory of Japan, Mitaka, Tokyo 181-8588, Japan

7. Institute of Astronomy, the University of Tokyo, Mitaka, Tokyo 181-0015, Japan

† JSPS Research Fellow (DC1)

## 1. Galaxy Formation and Evolution

A galaxy is a huge agglomeration of stars, interstellar medium (ISM: gas and dust), and dark matter (DM), a complex system with a complicated interaction between each component. There are a few billions of galaxies in the observable Universe, which delineate the appearance of the visible Universe at optical wavelengths. However, galaxies have formed from a tiny (order of $\sim 10^{-5}$) fluctuation of matter (mainly DM) in the early Universe, when the age of the Universe was only 380,000 yr. The initial Gaussian fluctuations of DM start to grow by gravitational interactions, finally to form virialized structures called dark halos. The dark halos approach each other and finally merge to form larger halos. The formation proceeds from smaller to larger structures. This is the so-called hierarchical structure formation, currently the most reliable scenario of the structure formation in the Universe. During the merging of dark halos, the baryonic gas falls into the gravitational potential wells of DM and is compressed there to form first stars and galaxies. Finally, some galaxies merge and form larger galaxies. Present-day large galaxies (up to $M_{\rm baryon} \sim 10^{12}\ M_{\odot}$) have formed in the merger process. Strong merging process is often accompanied by an effective compression of gas, inducing burst of star formation. Therefore, galaxy evolution is a highly complex process that depends on the environment of galaxies (number density of ambient galaxies and gas), as well as their internal processes. The formation and evolution of galaxies are regarded as one of the most important phenomena in the history of the Universe ranging of 13.8 billion years.

## 2. Evolution of the Interstellar Medium and Star Formation in Galaxies

### 2.1 Star formation in the interstellar medium

The most important internal physical process that drives the galaxy evolution is the star formation. The star formation proceeds as follows: 1) the interstellar medium (ISM) gravitationally contracts, 2) the ISM cools down and changes its phase as ionized, atomic, and then molecular gas, 3) high-density

molecular gas clumps are formed, and 4) the nuclear fusion ignites. Since stars form in molecular clouds, it is important to elucidate the physical state of the molecular clouds for the understanding of galaxy evolution. Observationally, an empirical relation between the star formation rate (SFR) and the gas mass density is known (Kennicutt-Schmidt law: Schmidt 1959; Kennicutt 1989, 1998). The understanding of the fundamental processes in the ISM is also expected to be a clue to explain the relation.

### 2.2 Spectroscopic observation of the ISM

Direct laboratory experiments are impossible for most of the problems in astronomy. Instead, only a unique method to obtain the physical information of the ISM in remote objects. Particularly, since the electromagnetic emission of astronomical molecules are mainly emitted as the radio emission lines, the astrophysics of the molecular clouds has been developed in radio astronomy. The largest and most efficient observational facility to observe molecular emission lines is the Atacama Large Millimeter/Submillimeter Array (ALMA).

Modern astronomical spectroscopic facilities as ALMA can provide tremendous amount of information on the atoms, molecules, ions, and dust in the ISM. However, spectroscopic observations are very time-consuming in general, and mapping an extended object by observing many positions on the sky is not easy. As a result, the number of independent sample size (number of observed positions) $n$ is much smaller than the wavelength/frequency dimension $d$, i.e., $n \ll d$. Such data are called high-dimensional low-sample size (HDLSS) data. Such a problem was regarded as an ill-posed problem in traditional astronomy and has been simply given up to study further. It has long been believed that we must make $d$ smaller than $n$ by throwing away the information of the data, in order to analyze such type of problem. Obviously a method to make a maximal use of the information of the data is desirable.

### 2.3 High-dimensional statistics

In a research field other than astronomy, for example in genomics, it is not rare to handle data with the dimension of the base sequence $d \sim 10^5$ while the sample size $n \sim 100$. A new statistical method to deal with the HDLSS data has been developed in recent decade, which is referred to as **the high-dimensional statistical method.** The high-dimensional statistics is continuously providing new findings (e.g., Aoshima 2018).

In this study, we apply the high-dimensional statistics to the spectroscopic mapping data obtained by ALMA. As mentioned above, the observational cost of spectroscopic mapping is very expensive, and for the ALMA observation, the sample size is at most $n \sim 200$, while the frequency dimension is $d \sim 2000$. Thus, the spectroscopic map of the ALMA is typically HDLSS. In traditional astronomy, they extracted some emission lines that are already known to be useful and used them to classify objects. However, ALMA revealed that molecular emissions appear to be significantly different even between neighboring molecular clouds in the same galaxy (Fig. 1), and the limitation of the

traditional method was clarified. The excessively huge amount of information obscures the relation between the evolution of the molecular cloud and star formation. The high-dimensional statistics is expected to make it possible to classify the molecular clouds by using all the emission lines, which has been considered to be forlorn.

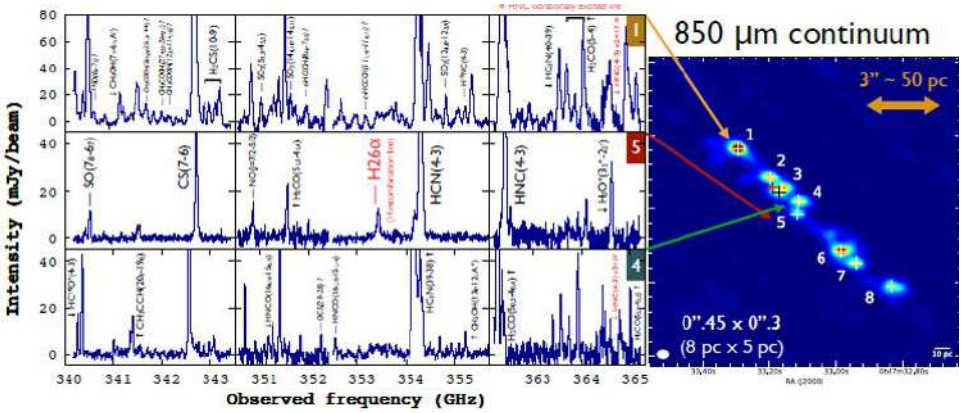


**Figure 1:** The spectral map of the central region of a starburst galaxy NGC 253 by ALMA (adopted from Ando et al. 2017). Left: spectra of star forming regions, Right: corresponding regions in NGC 253.

## 3. Application of the High-Dimensional PCA

The spatial dimension of the map of the central region of NGC 253 is 231, and the spectral dimension is 2248, namely $n$ = 231 and $d$ = 2248, typical HDLSS data. We applied the high-dimensional statistical analysis to the original ALMA spectral mapping data of NGC253 (Fig. 2).

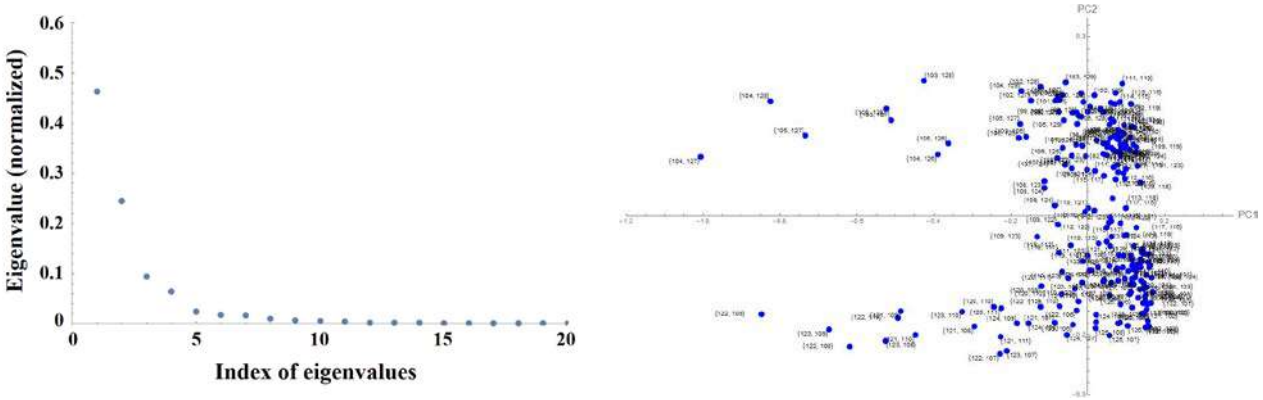


**Figure 2:** The result of the high-dimensional PCA to the NGC 253. Left: eigenvalue distribution, Right: distribution of the PC1 and 2.

It is remarkable that the complicated molecular spectra are characterized by first a few PCs. The PC1 and 2 are particularly dominant, and show a butterfly-like pattern. The physical meaning of these features are simple. The molecular clouds are coherently rotating around the center of the galaxy, which causes a systematic Doppler shift of the spectral lines. Thus, PC1 represents the total intensity of the spectral lines, and PC2 shows the Doppler shift (blueshift and redshift) of the lines.

Then we further proceeded the PCA by eliminating the effect of the Doppler shift to explore more subtle features in the spectra. The result of the PCA to the Doppler-corrected spectra is presented in

Fig. 3. New PC2 is much smaller than the original one, and there is no butterfly pattern in the PC distribution. It shows that the systemic rotation is well eliminated. By examining the spatial distribution of the PCs on the map of the star forming regions, we discovered that the new PC2 and PC3 represent the effect of local expansion of these regions.

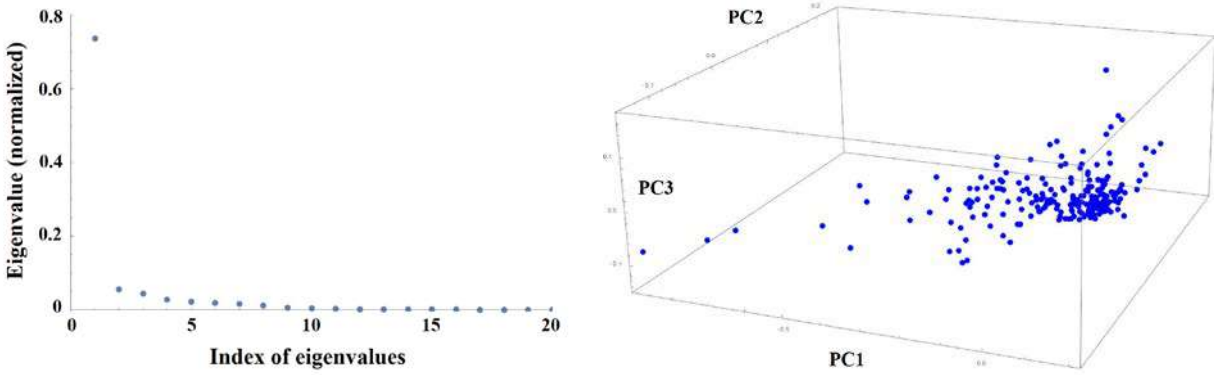


**Figure 3:** The result of the high-dimensional PCA to the Doppler-corrected spectral map of NGC 253. Left: eigenvalue distribution, Right: distribution of the PC1, 2, and 3.

## 4. Conclusion

Galaxies ubiquitously exist in the present-day Universe, but they have been formed from a tiny fluctuation of matter in the early Universe. Evolution of galaxies is mainly driven by the star formation, a transition from ISM to stars. Various phases of the ISM are related, and the evolution of the ISM is a key to complete the understanding of the galaxy evolution. Spectroscopic observations are of vital importance to extract and interpret the information of matter in galaxies. Spectroscopic mapping and similar methods are fundamentally important to reveal the ISM physics, but the data are high-dimensional low sample size. We summarize the conclusions from this study.

1. We applied the high-dimensional PCA on the NGC 253 spectral map. ALMA mapping data are typically HDLSS in general, and in this case $n = 231$ and $d = 2228$.
2. Very large variety in the molecular line spectra of NGC253 map can be described only by two PCs. Each PC consists of ~ 20 elements, much fewer than $d$. Because these elements may be a part of same features, the key features may be reduced to several.
3. The high-dimensional PCA successfully chose two PCs that reproduce the general properties of the ALMA spectroscopic map of NGC253.
4. The controlling feature was HCN(4-3) rotational lines. PC1 describes the total intensity of the lines, and PC2 represents the Doppler shift caused by the systemic rotation.
5. After correcting the Doppler shift due to the systemic rotation, we could obtain information on the smaller-scale velocity field described by PC2 (new) and PC3. These are caused by outflow of starburst regions.

We stress that this result is not obtained by choosing a handful of features by hand, but by making use of the full information of the high-dimensional data.

# Statistical Inference for Glaucoma Detection [1]

Ngai Hang Chan
Department of Statistics
The Chinese University of Hong Kong
Shatin, NT, Hong Kong
*nhchan@sta.cuhk.edu.hk*

**Abstract**

This talk will review some of the statistical techniques used in the analysis and detection of one of the most commonly encountered eye diseases, Glaucoma. By means of some of the recent artificial intelligence algorithms, ophthalmologists are employing modern machine learning technologies such as CNN and its variates to assist in early detection of eye symptoms in Glaucoma studies. This talk will address some of the related issues and discuss the potential of statistical ideas in these studies. Some examples will be given.

[1]September 3–5 , 2021, Time Series Conference, Niigata University, Niigata, Japan. Research supported in part by grants from HKSAR-RGC-TBS and HKSAR-RGC-GRF.

# Mixtures of Nonlinear Poisson Autoregressions

Paul Doukhan[1] Konstantinos Fokianos [2] Joseph Rynkiewicz[3]

[1]AGM, UMR 8088, University of Cergy-Pontoise
e-mail: doukhan@u-cergy.fr

[2]Department of Mathematics & Statistics, University of Cyprus
e-mail: fokianos@ucy.ac.cy

[3]SAMM, Université de Paris 1
e-mail: joseph.rynkiewicz@univ-paris1.fr

## 1 Introduction

We consider models for count time series, which allow for the mean process to change according to the values of an unobservable discrete random variable. Such models are directly related to Markov switching AR models (see Hamilton (1994, Ch.22), Cappé et al. (2005) and Frühwirth-Schnatter (2006, Ch. 11-12)). This class of processes is defined by regime specific models where transition among different regimes is determined by the state of an unobserved Markov chain. The conditional distribution of the process, given the past and the specific regime, is assumed to be Poisson with a time-varying mean modelled by a non-linear autoregressive infinite order model. The Poisson distributional assumption has been employed for modeling count time series by several authors including Rydberg and Shephard (2000), Davis et al. (2003), Ferland et al. (2006) and Fokianos et al. (2009), among others. The aim of this contribution is to extend this framework to Markov switching Poisson processes which can deal with complex dynamics and take into account several issues found in count data including proper modeling of non-linearities, overdispersion, multimodal conditional distribution and outliers. In the context of Gaussian time series, these issues were examined by Le et al. (1996) and Wong and Li (2000), among others.

Count time series analysis is a research topic receiving considerable attention over the last years, see Kedem and Fokianos (2002, Sec 4 & 5) and the recent edited volume by Davis et al. (2016) for several additional references. INARCH (INteger ARCH) and INGARCH (INteger GARCH) processes have been found useful in applications because they allow for estimation, model assessment and forecasting by employing existing statistical software. They belong to a broad class of models that falls under the exponential family framework; see Douc et al. (2017) who provide a unified point of view for time series generalized linear models. This study includes models and associated inference for mixtures of count time series models, a topic which has not attracted a lot of attention. Related early work was given by Albert (1991). Subsequently Carvalho and Tanner (2005, 2007) proposed a mixture-of-experts approach to model nonlinearities in count time series models. These authors studied maximum likelihood estimation, investigated identifiability and asymptotic normality of the estimates, and employed the AIC and BIC for selecting the number of mixture components in the model. However, they imposes strict conditions for developing asymptotic inference and their study focused explicitly on log-linear models for count time series. A mixture of linear models was considered by Zhu et al. (2010) but the authors did not provide theoretical evidence about its properties. More recently, Berentsen et al. (2018) applied a Markov-switching Poisson log-linear model autoregressive model to a study of corporate defaults. For the case of continuous valued time series, studies of mixture models were given by Jacobs et al. (1991), Le et al. (1996), Francq and Roussignol (1998) and Francq and Zaköian (2001), among others. In addition, Wong and Li (2000, 2001)

have used a two-component mixture model to extend AR and conditional ARCH models, respectively. Stability properties of such models have been studied by Saikkonen (2007) and some further work is given by Kalliovirta et al. (2015). In the context of GARCH models see Francq et al. (2001) while $t$-distributed autoregressive models were considered by Wong et al. (2009). Mixtures of AR models, under a Bayesian framework, have been considered by Wood et al. (2011), among others.

The main goals of this work is to fill several gaps in the literature of mixture models for count time series models. Precisely,

1. We introduce Markov switching non linear autoregressive Poisson models and study their properties by considering two cases: (a) For the mixture setup, we show that infinite order models are weakly dependent. (b) For the Markov switching case, we prove that finite order models are geometric $\beta$-mixing.

2. We study the statistical properties of the maximum likelihood estimator for the case of finite order autoregression of Markov switching models.

3. We show that the marginal likelihood ratio test for testing the number of hidden regimes converges to a Gaussian process. This fact implies that the BIC estimator is consistent estimator for selecting the true number of regimes.

4. We provide a counterexample which shows that estimation of the true number of regimes, under a misspecified GARCH type count time series model, is not possible.

5. We apply the theoretical results to the weekly number of E.coli cases and compare mixture models to models with interventions.

# References


Albert, P. S. (1991). A two–state Markov mixture model for a time series of epileptic seizure counts. *Biometrics 47*, 1371–1381.

Berentsen, G. D., J. Bulla, A. Maruotti, and B. Støve (2018). Modelling corporate defaults: A Markov-switching Poisson log-linear autoregressive model. *ArXiv e-prints*. https://arxiv.org/abs/1804.09252.

Cappé, O., E. Moulines, and T. Rydén (2005). *Inference in Hidden Markov Models*. New York: Springer.

Carvalho, A. X. and M. A. Tanner (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *Canad. J. Statist. 33*, 97–113.

Carvalho, A. X. and M. A. Tanner (2007). Modelling nonlinear count time series with local mixtures of Poisson autoregressions. *Comput. Statist. Data Anal. 51*, 5266–5294.

Davis, R. A., W. T. M. Dunsmuir, and S. B. Streett (2003). Observation-driven models for Poisson counts. *Biometrika 90*, 777–790.

Davis, R. A., S. H. Holan, R. Lund, and N. Ravishanker (Eds.) (2016). *Handbook of Discrete-Valued Time Series*. Handbooks of Modern Statistical Methods. London: Chapman & Hall/CRC.

Douc, R., K. Fokianos, and E. Moulines (2017). Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models. *Electron. J. Stat. 11*, 2707–2740.

Ferland, R., A. Latour, and D. Oraichi (2006). Integer–valued GARCH processes. *Journal of Time Series Analysis Analysis 27*, 923–942.

Fokianos, K., A. Rahbek, and D. Tjøstheim (2009). Poisson autoregression. *Journal of the American Statistical Association 104*, 1430–1439.

Francq, C. and M. Roussignol (1998). Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics 32*, 151–173.

Francq, C., M. Roussignol, and J.-M. Zakoïan (2001). Conditional heteroskedasticity driven by hidden Markov chains. *Journal of Time Series Analysis 22*, 197–220.

Francq, C. and J.-M. Zaköian (2001). Stationarity of multivariate Markov-switching ARMA models. *J. Econometrics 102*, 339–364.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press, Princeton, NJ.

Jacobs, R., M. Jordan, S. Nowlan, , and G. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*, 79–87.

Kalliovirta, L., M. Meitz, and P. Saikkonen (2015). A Gaussian mixture autoregressive model for univariate time series. *J. Time Series Anal. 36*, 247–266.

Kedem, B. and K. Fokianos (2002). *Regression Models for Time Series Analysis*. Hoboken, NJ: Wiley.

Le, N. D., R. D. Martin, and A. E. Raftery (1996). Modelling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association 91*, 1504–1515.

Rydberg, T. H. and N. Shephard (2000). A modeling framework for the prices and times of trades on the New York stock exchange. In W. J. Fitzgerlad, R. L. Smith, A. T. Walden, and P. C. Young (Eds.), *Nonlinear and Nonstationary Signal Processing*, pp. 217–246. Cambridge: Isaac Newton Institute and Cambridge University Press.

Saikkonen, P. (2007). Stability of mixtures of vector autoregressions with autoregressive conditional heteroskedasticity. *Statist. Sinica 17*, 221–239.

Wong, C. S., W. S. Chan, and P. L. Kam (2009). A Student $t$-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika 96*, 751–760.

Wong, C. S. and W. K. Li (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society B62*, 95–115.

Wong, C. S. and W. K. Li (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association 96*, 982–995.

Wood, S., O. Rosen, and R. Kohn (2011). Bayesian mixtures of autoregressive models. *J. Comput. Graph. Statist. 20*, 174–195.

Zhu, F., Q. Li, and D. Wang (2010). A mixture integer-valued ARCH model. *J. Statist. Plann. Inference 140*, 2025–2036.

# ブリッジ推定量におけるチューニングパラメーターの選択について

高崎経済大学・経済学部　　宮田 庸一

2021 年 8 月 15 日

$'$ 記号は, 行列の転置を表すものとする. 観測ベクトル $(Y_i, \boldsymbol{x}_i')'$ は, 以下のスパースな線形モデルに従うとする.

$$\begin{aligned} Y_i &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + u_i, \quad (i = 1, ..., n) \\ &= \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_{k_n}^* x_{k_n,i} + 0 \cdot x_{k_n+1,i} + \cdots + 0 \cdot x_{p_n i} + u_i, \end{aligned}$$

ただし, $\boldsymbol{x}_i = (x_{1i}, \cdots, x_{p_n,i})'$ は $p_n$ 次元の説明変数を表すベクトルとし, $Y_i$ は被説明変数とする. また $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, .., \beta_{k_n}^*, 0, \cdots, 0)' \in \mathbb{R}^{p_n+1}$ は真のパラメーターのベクトルとし, $\beta_j^* \neq 0$ $(j = 1, ..., k_n)$, および $\beta_j^* = 0$ $(j = k_n + 1, .., p_n)$ であるとする. $u_i$ は期待値 $E(u_i) = 0$, 分散 $Var(u_i) = \sigma_0^2$ となる撹乱項とする. ここで, 切片以外の真のパラメーターベクトルにおける非ゼロの要素の添え字の集合を $S_{0,n} = \{j \in \{1, ..., p_n\} | \beta_{0,j} \neq 0\}$ とし, その要素の個数を $k_n := |S_{0,n}|$ で表すとことにする. ここでは $k_n$ と説明変数の個数 $p_n$ は, 標本の大きさ $n$ に依存してもよいことにする. このモデルにより生成された標本 $(Y_i, \boldsymbol{x}_i')'$ $(i = 1, ..., n)$ に対して, 以下線形モデルを当てはめることを考える.

$$Y_i = \beta_0 + \beta_{j_1} x_{j_1 i} + \cdots + \beta_{j_k} x_{j_k i} + \epsilon_i, \qquad (i = 1, ...n),$$

ただし, $j_1, ..., j_k \in \{1, ..., p_n\}$, $k = 1, ..., p_n$ とする. このような説明変数の選択問題において, $p_n$ の次元がそれほど大きくない場合には, AIC, BIC などの情報量規準を用いて総当たり法を行うのが標準的なアプローチであろう. しかしその一方で, $p_n$ の次元が高くなるにと, 評価すべきモデルの数が指数的に増加するため, 総当たり法を行うのが困難になる. このような問題に対する一つの標準的なアプローチは, Tibshirani (1996) による LASSO(Least absolute shrinkage and selection operator) に代表される罰則付き最小二乗推定量を用いることである. 具体的には, $\boldsymbol{y} = (Y_1, ..., Y_n)'$, $\boldsymbol{X}_1 = (x_{ji}) : n \times p_n$ 行列, $\mathbf{1}_n = (1, ..., 1)'$, $\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_n & \boldsymbol{X}_1 \end{pmatrix}$ とおき, 損失関数

$$L_{0,n}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \|\boldsymbol{\beta}\|_1 \tag{1}$$

を最小にする $\boldsymbol{\beta}$ の値 $\hat{\boldsymbol{\beta}}_L$ を求めればよい. ただし, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ とする. この推定量は LASSO 推定量と呼ばれ, $\lambda_n \to \infty$ $(n \to \infty)$ とその他の条件のもとで漸近正規性を持つことが知

られている. 一方で, LASSO 推定量をモデル選択手法の観点から見たときには, いくつかの問題が生じる. それは収束レートが $\sqrt{n}$ である漸近正規性が成り立つためのチューニングパラメーター $\lambda_n$ のオーダーの仮定の下では, 真のモデルを選ぶ確率は漸近的に 1 にならないことが知られている. この問題を改善するために, 式 (1) における罰則項を, 非凸の形のものに置き換えた以下の損失関数を考え, これを最小にする推定量 $\hat{\boldsymbol{\beta}}$ を求める:

$$L_n(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|^\gamma, \tag{2}$$

ただし $0 < \gamma < 1$ とする. この推定量はブリッジ推定量と呼ばれている. 当然のことながら, ブリッジ推定においても適切な $\lambda_n$ の選び方が重要になる. Umezu et al. (2019) においては, $p_n$ は $n$ に依存しない条件, しかし線形モデルより広い一般化線形モデルの下での AIC 型の情報量規準を与えている. 一方で, Wang et al. (2009) では, 線形モデルの下で, モデル選択に関して一致性を持つ BIC 型の情報量規準を与えている. Huang et al. (2008) では, オラクル性と呼ばれる, 推定量 $\hat{\boldsymbol{\beta}}$ が漸近正規性, 漸近有効性を持ち, なおかつ説明変数に関しても一致性を持つような条件を与えている. 例えば $\gamma = 1/2$ として, $k_n = k$(定数), $p_n = \log n$ とおいた場合, $\lambda_n$ のオーダーに関する条件は, $\lambda_n/(n^{1/4}(\log n)^{3/4}) \to \infty$ $(n \to \infty)$, $\lambda_n = o(n^{1/2})$ となるが, そのような $\lambda_n$ の選択肢は, $\lambda_n = 3n^{3/8}$ でも良いし, $\lambda_n = 10n^{3/8}$ でもよいことになる. すなわち Huang et al. (2008) の結果は $\lambda_n$ の選択に関しては, 何も述べていないことになる. このため, 今回の講演では, Huang et al. (2008) の条件の下で $\lambda_n$ を選ぶための規準を, 一般化周辺尤度に対する近似として与えられることを示し, シミュレーションによりそのパフォーマンスを検証した. ここで用いた一般化周辺尤度とは, Yamanishi (1998) により与えられた, 重みつき対数尤度に $(-1)$ を乗じたものを損失関数とした拡張型確率的コンプレキシティに対応するものである.

## 参考文献


Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist. 36*(2), 587–613.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Umezu, Y., Y. Shimizu, H. Masuda, and Y. Ninomiya (2019). AIC for the non-concave penalized likelihood method. *Ann. Inst. Statist. Math. 71*(2), 247–274.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*(3), 671–683.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theory 44*(4), 1424–1439.

# ３次元分割表における条件付き独立性検定統計量の変換統計量について

北海道教育大・札幌　種市信裕
北海道教育大・釧路　関谷祐里
数学利用研究所　外山淳

## 1. ３次元分割表における条件付独立性モデル.

３次元の $J \times K \times L$ 分割表において多項分布モデルを考える. $(j,k,l)$ セルの観測度数を表す確率変数を $X_{jkl}$ とする. ただし，$X_{jkl}$ は $\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} X_{jkl} = n$ を満たす非負整数の値を取り，自然数 $n$ は定数とする. また，$(j,k,l)$ セルのセル確率を $p_{jkl}$ とする. このとき，確率変数ベクトル $X = (X_{111}, \ldots, X_{JKL})'$ が多項分布 $\mathcal{M}_{JKL}(n,p)$ に従う場合について考える. ここで，$p = (p_{111}, \ldots, p_{JKL})'$ である. このとき，この $J \times K \times L$ 分割表のセル確率に関する条件付き独立性の仮説

$$H_0 : p_{jkl} = \frac{p_{j\cdot l}p_{\cdot kl}}{p_{\cdot\cdot l}} \quad (j = 1,\ldots,J,\ k = 1,\ldots,K,\ l = 1,\ldots,L)$$

を検定するための $\phi$-divergence に基づく検定統計量は

$$C_\phi = 2n\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\hat{p}_{jkl}\phi\left(\frac{\tilde{p}_{jkl}}{\hat{p}_{jkl}}\right)$$

である. ここで，$\tilde{p}_{jkl} = X_{jkl}/n$, $\hat{p}_{jkl} = X_{j\cdot l}X_{\cdot kl}/(nX_{\cdot\cdot l})$, であり，$\phi$ は $(0,\infty)$ において定義された実凸関数であり，$\phi(1) = \phi'(1) = 0$ と $\phi''(1) = 1$ を満たすものとする.

## 2. 独立性検定統計量の分布の漸近展開に基づく近似.

$H_0$ のもとでの検定統計量 $C_\phi$ の分布の漸近展開に基づく次のような近似を考える. $\Pr\{C_\phi \le x|H_0\} \approx W_1 + W_2$, ここで，$W_1$ は多変量エッジワース展開に基づく項，$W_2$ は不連続性を考慮に入れた離散項である. $W_2$ は，多項分布の適合度検定におけるピアソンのカイ二乗統計量の分布に対して，Yarnold [4] により与えられた漸近展開に基づく近似式で導入された項に対応する. $W_1$ に関して以下の定理が成り立つ.

**定理.** (Taneichi et al. [3])

関数 $\phi(t)$ が 6 回微分可能で，$\phi^{(6)}(t)$ が $t = 1$ で連続ならば，$W_1$ は次のように評価される.

$$W_1 = \Pr\{\chi^2_M \le x\} + \frac{1}{n}\sum_{j=0}^{3} v_j^\phi \Pr\{\chi^2_{M+2j} \le x\} + O(n^{-2}), \tag{1}$$

ここで，$M = (J-1)(K-1)L$, $v_0^\phi = \gamma_0$, $v_1^\phi = (\gamma_1+\gamma_2) + 2\phi'''(1)\gamma_2 + \phi^{(4)}(1)\gamma_3 + \{\phi'''(1)\}^2\gamma_4$, $v_2^\phi = 2(-\gamma_2+\gamma_3) - 2\phi'''(1)(\gamma_2+\gamma_4) - \phi^{(4)}(1)\gamma_3 - 2\{\phi'''(1)\}^2\gamma_4$, $v_3^\phi = \gamma_4\{1+\phi'''(1)\}^2$, $\gamma_0 = -\frac{1}{12}\Gamma_1 - \frac{1}{12}\Gamma_2 + \frac{1}{12}(\Gamma_3+\Gamma_4)$, $\gamma_1 = \frac{1}{4}JK\Gamma_1 + \frac{1}{4}\Gamma_2 - \frac{1}{4}(K\Gamma_3 + J\Gamma_4)$, $\gamma_2 = \frac{1}{8}J^2K^2\Gamma_1 + \frac{1}{8}\Gamma_2 - \frac{1}{8}(K^2\Gamma_3 + J^2\Gamma_4)$, $\gamma_3 = \left(-\frac{1}{2}JK - \frac{1}{8} + \frac{1}{4}J + \frac{1}{4}K\right)\Gamma_1$

$-\frac{1}{8}\Gamma_2+\frac{1}{4}(K\Gamma_3+J\Gamma_4)-\frac{1}{8}(\Gamma_3+\Gamma_4)$, $\gamma_4=\left(\frac{1}{8}J^2K^2+\frac{3}{4}JK+\frac{1}{3}-\frac{1}{2}J-\frac{1}{2}K\right)\Gamma_1$ $+\frac{5}{24}\Gamma_2-\frac{1}{8}(K^2\Gamma_3+J^2\Gamma_4)-\frac{1}{4}(K\Gamma_3+J\Gamma_4)+\frac{1}{6}(\Gamma_3+\Gamma_4)$,

$$\Gamma_1=\sum_{\ell=1}^{L}p_{\cdot\cdot\ell}^{-1},\ \Gamma_2=\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=1}^{L}q_{jkl}^{-1},\ \Gamma_3=\sum_{j=1}^{J}\sum_{\ell=1}^{L}p_{j\cdot\ell}^{-1},\ \Gamma_4=\sum_{k=1}^{K}\sum_{\ell=1}^{L}p_{\cdot k\ell}^{-1},\ q_{jkl}=\frac{p_{j\cdot l}p_{\cdot kl}}{p_{\cdot\cdot l}}$$

である.

**3. 漸近展開の連続項に基づく変換統計量.**

本報告では, 少ない標本数でも $C_\phi$ より速く極限分布に近づく検定統計量の構築を目指す. 導出された離散項 $W_2$ の評価式は非常に複雑であることもあり $C_\phi$ の分布に対して，(1) 式で与えられる $W_1$ の近似式のみを用いて，小標本におけるカイ二乗近似の改良を考える. (1) 式にバートレット修正および改良変換の構築と漸近展開式との関係の理論 (e.g., Fujikoshi [2], Cordeiro and Ferrari [1]) を適用した. (1) 式において, $\phi'''(1)=-1$ かつ $\phi^{(4)}(1)=2$ が成り立つ場合には $v_1^\phi=-v_0^\phi$ かつ $v_2^\phi=v_3^\phi=0$ が成り立つ. このことは, バートレット修正が可能なことを意味する. よってこの場合には, $C_\phi$ にバートレット修正を施した変換統計量 $(C_\phi)_B=\{1+2v_0^\phi(nM)^{-1}\}C_\phi$ を構築できる. (1) 式においてその他の場合には, 改良変換統計量 ([2])

$$\begin{aligned}(C_\phi)_I &= (n\alpha+\beta)^2\log\left[1+\frac{1}{(n\alpha)^2}\left\{C_\phi+\frac{1}{n\alpha}((C_\phi)^2+\gamma(C_\phi)^3)\right.\right.\\ &\qquad\left.\left.+\frac{1}{(n\alpha)^2}\left(\frac{1}{3}(C_\phi)^3+\frac{3\gamma}{4}(C_\phi)^4+\frac{9\gamma^2}{20}(C_\phi)^5\right)\right\}\right],\end{aligned}$$

ただし, $\alpha=-M(M+2)\{2(v_2^\phi+v_3^\phi)\}^{-1}$, $\beta=-(M+2)v_0^\phi\{2(v_2^\phi+v_3^\phi)\}^{-1}$, $\gamma=v_3^\phi\{(M+4)(v_2^\phi+v_3^\phi)\}^{-1}$, および, バートレット型変換統計量 ([1])

$$(C_\phi)_{CF}=\left\{1+\frac{2v_0^\phi}{nM}+\frac{2(v_0^\phi+v_1^\phi)}{nM(M+2)}C_\phi+\frac{2(v_0^\phi+v_1^\phi+v_2^\phi)}{nM(M+2)(M+4)}(C_\phi)^2\right\}C_\phi$$

を構築できる.

## 参考文献


[1] Cordeiro, G. M. and Ferrari, S. L. P.: *Biometrika*, **78**, (1991), 573–582.

[2] Fujikoshi, Y.: *Journal of Multivariate Analysis*, **72**, (2000), 249–263.

[3] Taneichi, N., Sekiya, Y. and Toyama, J.: *Journal of Multivariate Analysis*, **171**, (2019), 193-208.

[4] Yarnold, J. K.: *Ann. Math. Statist.*, **43**, (1972), 1566–1580.