# シンポジウム「統計科学の革新にむけて」
# 報告書

主催：科学研究費・基盤研究（A)
「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），課題番号：20H00576）

・開催日時：2021 年 1 月 22 日（金）～ 1 月 23 日（土）
・会場：金沢大学サテライトプラザ三階集会室
・開催責任者：星野 伸明（金沢大学経済学経営学系）

統計シンポジウム「統計科学の革新にむけて」プログラム

主催：科学研究費・基盤研究（A）
「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）, 課題番号：20H00576）

・開催日時：2021年1月22日（金）～1月23日（土）
・会場：金沢大学サテライトプラザ三階集会室
・開催責任者：星野伸明（金沢大学経済学経営学系）

| | | 座長 | 講演者 | 講演者所属 | 共同研究者 | 演題 |
|---|---|---|---|---|---|---|
| 1/22(金) | 13:00 | 内藤貫太（千葉大学） | 佐々木拓真 | 筑波大学 | 矢田和善（筑波大学）・青嶋誠（筑波大学） | 高次元相互共分散行列の特異値分解とその応用 |
| | 13:40 | | 江頭健斗 | 筑波大学 | 矢田和善（筑波大学）・青嶋誠（筑波大学） | 高次元における DWDとWDWDのバイアス補正とその比較 |
| | 14:20 | | 間野修平 | 統計数理研究所 | 廣瀬雅代（九州大学） | 漸近不偏推定量のベイズ的構成 |
| | 15:00 | 休憩 | | | | |
| | 15:20 | 永井勇（中京大学） | 丸山祐造 | 神戸大学 | | 平均ベクトルの推定における分散未知のもとでの許容的でミニマクスな推定量 |
| | 16:00 | | 下野寿之 | 統計数理研究所 | | Multiplicative Decompositions of Stochastic Distributions and Their Applications |
| | 16:40 | 終了 | | | | |

| | | 座長 | 講演者 | 講演者所属 | 共同研究者 | 演題 |
|---|---|---|---|---|---|---|
| 1/23(土) | 9:30 | 間野修平（統計数理研究所） | Junichi Hirukawa | Niigata University | Kou Fujimori (Shinshu University) | Weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation |
| | 10:10 | | 松井宗也 | 南山大学 | 末石直也（神戸大学） | Approximated likelihood estimation of stable laws/OU processes |
| | 10:50 | | 栗木哲 | 統計数理研究所 | 松原隆彦(高エネルギー加速器研究機構 ) | 弱非ガウス確率場に対する期待ミンコフスキー汎関数 |
| | 11:30 | 休憩 | | | | |
| | 13:00 | 松井宗也（南山大学） | 永井勇 | 中京大学 | | 外れ値がある場合でも安定した推定を行う罰則付推定法 |
| | 13:40 | | 上本拓弥 | 千葉大学 | 内藤貫太（千葉大学） | Support vector regression with penalized likelihood |
| | 14:20 | | 三次琢巳 | 千葉大学 | 内藤貫太（千葉大学） | 一般化線形モデルにおける擬Huber関数によるロバスト推測 |
| | 15:00 | 休憩 | | | | |
| | 15:20 | 栗木哲（統計数理研究所） | 内藤貫太 | 千葉大学 | | Regression with localized functional Bregman divergence |
| | 16:00 | | 牧草夏実 | 千葉大学 | | 再生核ヒルベルト空間におけるMaximum Variance Discrepancyの実際的挙動 |
| | 16:40 | | 西田喜平次 | 兵庫医療大学 | 内藤貫太（千葉大学） | Density Estimation via Stagewise Algorithm with a Dictionary Having Various Bandwidths |
| | 17:20 | 終了 | | | | |

# 高次元相互共分散行列の特異値分解とその応用

佐々木拓真 (筑波大学数理物質科学研究科)
矢田和善 (筑波大学数理物質系)
青嶋　誠 (筑波大学数理物質系)

ゲノム科学・情報工学・金融工学などの現代科学の一つの特徴は，データがもつ次元数の膨大さにある．このようなゲノムデータには，次元数が数万にものぼる一方で，標本数は 100 にも満たないという事例が多く見られる．これは，いわゆるビッグデータの一種で，データの次元数 $p$ と標本数 $n$ に

$$p \gg n\,(\text{もしくは，}\ p > n)$$

といった大小関係をもつ．これが高次元データの一つの特徴であり，大小関係を強調して高次元小標本データとよぶこともある．

高次元データにおいては，変数群間に高い相関がしばしば見られ，その相関に強く寄与している変数を見つけることが重要となる．2 つの変数群の相関構造を推測する手法として，古くから正準相関分析（[3]）が知られている．高次元のもと共分散行列に正則性が仮定できない場合，Hardoon et al. [2] において正則化法などを使った正準相関分析法が紹介されている．一方で，相互共分散行列の特異値分解によって，変数群間の構造を推測することができる．本講演では，相互共分散行列の特異値分解に着目し，変数群間の構造を，高次元における特異値のスパイク性に基づき推定した．

まず母集団分布に $p$ 次元の分布を考え, $n$ 個のデータ $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出する．ただし,

$$\boldsymbol{x}_j^T = (\boldsymbol{x}_{1j}^T, \boldsymbol{x}_{2j}^T), \quad j = 1, ..., n$$

とし, $\boldsymbol{x}_{ij} \in \mathrm{R}^{p_i}$ とする．さらに， $\boldsymbol{x}_{ij}$ は共分散行列 $\boldsymbol{\Sigma}_i$ をもつとする．ただし，$p_1 < n < p_2$ とし，$\limsup_{p\to\infty} p_1 < \infty$ と仮定する．ここで， $\mathrm{Cov}(\boldsymbol{x}_{ij}) = \boldsymbol{\Sigma}_*$ とおく．相互共分散行列 $\boldsymbol{\Sigma}_*$ の特異値を $\lambda_{*1} \geq \cdots \geq \lambda_{*p_1} \geq 0$ とし，

$$\boldsymbol{\Sigma}_* = \sum_{j=1}^{p_1} \lambda_{*j} \boldsymbol{u}_j \boldsymbol{v}_j^T$$

と特異値分解する．ただし，$\boldsymbol{u}_j$ は左特異ベクトル，$\boldsymbol{v}_j$ は右特異ベクトルである．

Aoshima and Yata [1] では，Yata and Aoshima [4] で開発されたクロスデータ行列法を用いて $\mathrm{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Sigma}_*^T)$ の不偏推定量を与え，その推定量の高次元漸近正規性を証明することで，$\boldsymbol{\Sigma}_*$ の構造に関する検定手法を与えた．一方で，Yata and Aoshima [5, 6] では，拡張クロスデータ行列法を提案し，高次元のもと $\boldsymbol{\Sigma}_*$ の構造に関する高精度な推定法を与えた．

本講演では，まず従来の標本相互共分散行列による特異値の推定量が，高次元のもと不一致性をもつことを示した．その解決策として，拡張クロスデータ行列法を用いた特異値・特異ベクトルの新たな推定法を提案し，高次元のもと一致性をもつことを示した．

## 参考文献


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

[2] Hardoon, D., Szedmak, S., Shawe-Taylor, J. (2006). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, **16**, **12**, 2639-2664.

[3] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, **28**, 321-377.

[4] Yata, K., Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, **101**, 2060-2077.

[5] Yata, K., Aoshima, M. (2013). Correlation tests for high-dimensional data using extended cross-data-matrix methodology, *Journal of Multivariate Analysis*, **117**, 313-331.

[6] Yata, K., Aoshima, M. (2016). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology, *Journal of Multivariate Analysis*, **151**, 151-166.

# Bias corrections of DWD and WDWD in high-dimensional settings
(高次元における DWD と WDWD のバイアス補正とその比較)

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

[a]**Degree Programs in Pure and Applied Sciences, Graduate School of Science and Technology, University of Tsukuba**
[b]**Institute of Mathematics, University of Tsukuba**

In this talk, we consider two-class linear discriminant analysis for the HDLSS data. Suppose we have two independent and $d$-variate populations, having unknown mean vectors and unknown covariance matrices. Let us have trainning data sets from the both populations. Let the total sample size $N$.

In the HDLSS context, Chan and Hall[6], Hall et al.[8] and Aoshima and Yata[2] considered distance-based classifiers. Aoshima and Yata[4] considered a distance-based classifier based on a data transformation technique. Aoshima and Yata[1, 3] considered geometric classifiers based on a geometric representation of HDLSS data. Aoshima and Yata[5] considered quadratic classifiers in general and discussed an optimality of the classifiers under high-dimension, non-sparse settings. In the field of machine learning, there are many studies for classification (supervised learning). A typical method is the support vector machine (SVM) developed by Vapnik[13]. Chan and Hall[6], Hall et al.[7] and Nakayama et al.[10, 11] investigated asymptotic properties of the support vector machine (SVM) in the HDLSS context. Nakayama et al.[10, 11] pointed out the strong inconsistency of the SVM when $n_i$s are imbalanced. They proposed bias-corrected SVMs and showed its superiority to the SVM. On the other hand, Marron et al.[9] pointed out that the SVM causes data piling in the HDLSS context. Data piling is a phenomenon that the projection of a trainning data to the normal direction vector of a separating hyperplane is same for each class. In order to avoid the data piling problem of the SVM, Marron et al. proposed distance weighted discrimination (DWD). Whereas the SVM finds the optimal hyperplane by maximizing the minimum distances from each class to the hyperplane, the DWD finds a proper hyperplane by minimizing the sum of reciprocals of the distance from each data point to the hyperplane. The DWD cares all the data vectors that are not always used in the SVM. Unfortunately, the DWD is designed for balanced trainning data sets. For imbalanced trainning data sets, Qiao et al.[12] developed weighted DWD (WDWD) that imposes different weights on two classes. However, the WDWD is sensitive for a choice of weights.

In this talk, we investigated the DWD and the WDWD theoretically in the HDLSS context where $d \to \infty$ while $N$ is fixed. We gave asymptotic properties of the DWD and showed that the DWD includes a huge bias caused by heterogeneity of covariance matrices as well as sample imbalance. Then, we proposed a bias corrected-DWD (BC-DWD) and showed that the BC-DWD can enjoy consistency properties about misclassification rates. Finally, we gave asymptotic properties of the WDWD.

# References


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

[2] Aoshima, M., Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66, 983–1010.

[3] Aoshima, M., Yata, K. (2015). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis*, 34, 279–294.

[4] Aoshima, M., Yata, K. (2019a). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. *Annals of the Institute of Statistical Mathematics*, 71, 473-503.

[5] Aoshima, M., Yata, K. (2019b). High-dimensional quadratic classifiers in non-sparse settings. *Methodology and Computing in Applied Probability*, 21, 663–682.

[6] Chan, Y.-B., Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96, 469–478.

[7] Hall, P., Marron, J.S., Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society, Series B*, 67, 427–444.

[8] Hall, P., Pittelkow, Y., Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society, Series B*, 70, 159–173.

[9] Marron, J.S., Todd, M.J., Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102, 1267–1271.

[10] Nakayama, Y., Yata, K., Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *Journal of Statistical Planning and Inference*, 191, 88–100.

[11] Nakayama, Y., Yata, K., Aoshima, M. (2020). Bias-corrected support vector machine with Gaussian kernel in high-dimension, low-sample-size settings. *Annals of the Institute of Statistical Mathematics*, 72, 1257–1286.

[12] Qiao, X., Zhang, H. H., Liu, Y., Todd, M.J., Marron, J.S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association*, 105, 401–414.

[13] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory (second ed.)*. New York: Springer.

# 漸近不偏推定量のベイズ的構成

統計数理研究所 間 野 修 平
smano@ism.ac.jp
共同研究者 九州大学 廣 瀬 雅 代
masayo@imi.kyushu-u.ac.jp

未知の確率測度 $p$ と規格化しないモデル $q(x;\xi)$ について Kullback-Leibler (KL)-divergence は

$$D\{p\|q(x;\xi)\} = \int_{\mathcal{X}} q(x;\xi)dx - 1 + \int_{\mathcal{X}} p(x)\log p(x)dx - \int_{\mathcal{X}} p(x)\log q(x;\xi)dx$$

である．$p$ の経験分布と規格化されたモデル $q_0(x;\xi)$ を脱規格化したモデル $q(x;\xi) = q_0(x;\xi)z(\xi),\ z(\xi) := \int_{\mathcal{X}} q(x;\xi)dx > 0$ の KL-divergence の最小化は，事前分布の対数を $\tilde{l}(\xi) = n\{1 - z(\xi) + \log z(\xi)\} \leq 0$ とした MAP 推定量

$$\hat{\xi}(x_1,...,x_n) = \operatorname*{argmin}_{\xi}\left\{-\frac{1}{n}l(\xi;x_1,...,x_n) + \tilde{l}(\xi)\right\}$$

を与える．MLE は

$$\hat{\xi}_0(x_1,...,x_n) = \operatorname*{argmin}_{\xi}\left\{-\frac{1}{n}\sum_{i=1}^{n}\log q_0(x_i;\xi)\right\} = \operatorname*{argmin}_{\xi}\left\{-\frac{1}{n}l(\xi;x_1,...,x_n)\right\},$$

である．規格化されたモデル上で $\mathbb{E}_{\xi}f(\hat{\xi}) = f(\xi),\ \forall\xi$ を満たすように事前分布を定める．モデル多様体は $C^{\infty}$ 多様体 $\mathcal{M}$, Fisher 計量 $g$，skewness テンソル $S$ の組 $(\mathcal{M}, g, S)$ で表される．$\alpha$ ラプラス作用素を $\Delta^{(\alpha)}f := \nabla^{(\alpha)i}\nabla_i^{(\alpha)}f$ で定義する．

**定理 1** (HM arXiv:2011.14747)**.** 正則条件の下，$U$ 推定可能函数 $f \in C^3(\mathbb{R}^d)$ と事前分布の対数 $\tilde{l}(\xi) = O(1) \in C^4(\mathbb{R}^d)$ が

$$\Delta^{(-1)}f + 2\langle\partial\tilde{l}, \partial f\rangle = o(n^{-1})$$

を満たすとき，$\hat{\xi}$ を母数 $\xi$ の MAP 推定量として，$f(\hat{\xi})$ は $f(\xi)$ の 2 次漸近不偏推定量である．特に，完備十分統計量があれば，$f(\hat{\xi})$ は UMVUE に $O(n^{-1})$ まで一致する．

**注意 1.** MLE の典型的バイアスは $O(n^{-1})$ である．MAP 推定量は MLE に対し偏りを改善し，二乗損失について漸近的に同等である．

**系 1.** 1 次元モデル多様体を考える．正則条件の下，$U$ 推定可能函数 $f \in C^3(\mathbb{R})$, $f'(\xi) > 0$ について，事前分布が

$$e^{\tilde{l}(\xi)} \propto \frac{\{g(\xi)\}^{1/4}}{\sqrt{f'(\xi)}}e^{\frac{1}{4}\int^{\xi}S_1(\tilde{\xi})d\tilde{\xi}}$$

を満たせば，$\hat{\xi}$ を $\xi$ の MAP 推定量として推定量 $f(\hat{\xi})$ は 2 次漸近不偏で，完備十分統計量があれば UMVUE に $O(n^{-1})$ まで一致する．

**系 2.** $\alpha$ $(\neq 0)$ 平坦モデル多様体と $\alpha$ アファイン座標系 $\xi$ について，$\hat{\xi}$ を $(\alpha-1)/2$ 平行体積要素の密度を事前分布とする MAP 推定量として，$\alpha$ 調和函数 $f$ の推定量 $f(\hat{\xi})$ は 2 次漸近不偏で，完備十分統計量があれば UMVUE に $O(n^{-1})$ まで一致する．

Estimand $f(\xi)$ と事前分布の対数 $\tilde{l}(\xi)$ はモデル多様体の余次元 1 の葉層を定める．$\tilde{l}$ が定める葉層を $f$ が定める葉層と等しくとれば，事前分布の構成を 1 次元の積分に帰着できる．

**命題 1.** 正則条件の下，$U$ 推定可能函数 $f \in C^3(\mathbb{R}^d)$, $f'(\gamma) > 0$, $\gamma \in C^3(\mathbb{R}^d)$ について，事前分布が

$$e^{\tilde{l}(\gamma)} \propto \frac{1}{\sqrt{f'(\gamma)}} \exp\left(-\frac{1}{2}\int^{\gamma} \frac{\Delta^{(-1)}\tilde{\gamma}}{\langle \partial\tilde{\gamma}, \partial\tilde{\gamma}\rangle} d\tilde{\gamma}\right)$$

を満たせば，MAP 推定量 $\hat{\xi}$ について $f(\gamma(\hat{\xi})) = f \circ \gamma(\hat{\xi})$ は 2 次漸近不偏で，完備十分統計量があれば UMVUE に $O(n^{-1})$ まで一致する．

**例 1.** 線形混合効果モデル

$$x_{ij}|z_i \sim N(z_i, d), \quad j \in \{1, ..., n_i\}, \quad z_i \overset{iid}{\sim} N(0, a),$$

$i \in \{1, ..., m\}$ を考える．平均ベクトルの最良線形不偏予測量は

$$\hat{z}_i = \{1 - b^{(i)}(a, d)\}\bar{x}_i, \quad b^{(i)}(a, d) := \frac{d/n_i}{d/n_i + a}$$

であり，縮小因子は $b^{(i)}(a, d)$ である．Hirose-Lahiri (2020) は $d$ が既知，つまり 1 次元の場合の縮小因子の漸近 2 次不偏推定量を導出した．モデル多様体は象限 $\{(a, d) \in \mathbb{R}^2_{>0}\}$ で $(-1)$ 平坦，$(a, d)$ は $(-1)$ アファイン座標系である．1 次元の場合の類推で事前分布を $\tilde{l}^{(i)}(a, d) = \log(d/n_i + a)$ ととれば良いと予想でき，計算で確認できる．事前分布の選択は一意ではない．縮小因子 $b^{(i)}$ と事前分布 $\tilde{l}^{(i)}$ は異なる $(-1)$ 測地葉層を定める．$b^{(i)}(\gamma) = (1 + n_i\gamma)^{-1}$, $\gamma = a/d > 0$ であるが，$\tilde{l}^{(i)}$ が定める葉層を $b^{(i)}$ が定める葉層と等しくとると命題 1 より

$$e^{\tilde{l}(\gamma)} \propto \frac{d/n_i + a}{d} \left\{\prod_{j=1}^{m} \frac{e^{(1+n_j a/d)^{-1}}}{1 + n_j a/d}\right\}^{1/(2n)}$$

を得るが，これは類推で求めた事前分布よりも複雑である．

本講演では時間の制約上偏微分方程式の解法については触れなかった．論文 (arXiv:2011.14747) では，命題 1 の $\gamma$ を 0 測地線とした葉層や等質空間上の積分について，指数型分布族として分散既知の多変量正規分布 (Euclid 空間)，群分布族として位置尺度族 (双曲空間) を例として説明している．

# 平均ベクトルの推定における分散未知のもとでの許容的でミニマクスな推定量

神戸大学・経営学部・丸山　祐造

$X \sim N_p(\theta, I_p/\eta)$ の平均ベクトル $\theta$ の推定を考える．ここで $\eta$ は未知であり，また $\eta$ に関連する統計量として $S \sim \chi^2_n/\eta$ があるとする．推定量の精度は二乗損失関数 $\eta\|\delta-\theta\|^2$ で測られる．ただし，推定量 $\delta(X,S)$ は確率変数なのでリスク関数 $E\left[\eta\|\delta(X,S)-\theta\|^2\right]$ によって推定量の優劣を決める．この統計モデルは正規線形回帰モデル

$$y = A\beta + \epsilon, \qquad y \in \mathbb{R}^N,\ \beta \in \mathbb{R}^q,\ \epsilon \sim N_N(0, I_N/\eta)$$

の正準形である．ここで $p=q$，$n=N-q$ として以下のような対応を考えれば良い．

- パラメータの対応 $\theta \Leftrightarrow (A^{\mathrm{T}}A)^{1/2}\beta$
- 確率変数の対応 $X \Leftrightarrow (A^{\mathrm{T}}A)^{1/2}\hat{\beta}$ 最小二乗推定量
- 確率変数の対応 $S \Leftrightarrow \|(I-A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}})y\|^2$ 残差平方和
- 損失関数の対応 $\eta\|\delta-\theta\|^2 \Leftrightarrow \eta(\hat{\beta}-\beta)^{\mathrm{T}}A^{\mathrm{T}}A(\hat{\beta}-\beta) = \eta\|A\hat{\beta}-A\beta\|^2$ predictive loss

本講演では特に一般化ベイズ推定量（広義事前分布に対するベイズ推定量）に関する性質，許容性とミニマクス性に興味がある．特に許容性については，関連する概念とともに以下のように定義される．

- $\delta$ が $\delta_0$ を優越する（改良する） if

$$R(\theta,\eta,\delta) \le R(\theta,\eta,\delta_0) \ \text{ for all values of } \ (\theta,\eta)$$
$$R(\theta,\eta,\delta) < R(\theta,\eta,\delta_0) \ \text{ for at least one value of } \ (\theta,\eta)$$

- 他の推定量によって優越される推定量は非許容的である．
- 他のどのような推定量によっても優越されない推定量は許容的である．

許容的な推定量が欲しいだけなら，狭義事前分布のもとでのベイズ推定量でよい．しかし，（最尤法，不偏性，不変性，ミニマクス性の保持などの考察から）自然な推定量が与えられたとき，しばしば一般化ベイズ推定量である．例えば，広義事前分布 $\pi(\theta)=1$ に関する一般化ベイズ推定量 $X$ を考える．これは最尤推定量，不偏推定量でもあり，定数リスクを持つミニマクス推定量である．その自然さ故に許容性を保有することが強く期待される．

この問題設定において $p=1,2$ のとき，$X$ は許容的であるが，$p \ge 3$ で許容的になることがスタイン現象として知られている．このことは広義事前分布が許容性を導く場合と非許容性を導く場合があることを示唆していており，その境界に理論的興味が生じる．

$\eta$ が既知の一般化ベイズ推定量の許容性は Brown (1971) で研究された．ラフに言えば，広義事前分布が与えられたとき，それに対する一般化ベイズ推定量が許容的か非許容的か判別できる定理が提示された．

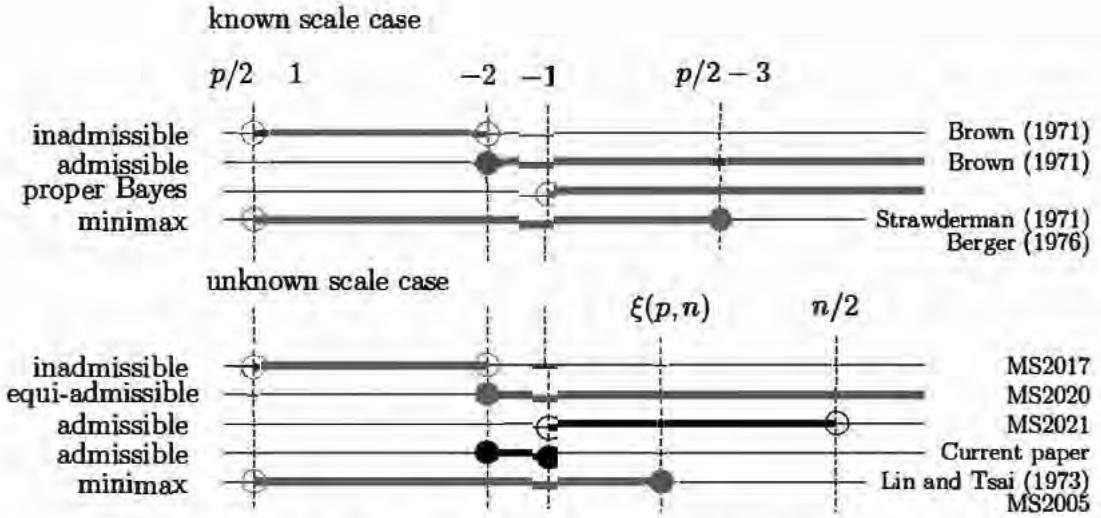



図 1　Ranges of $a$ for admissibility/inadmissibility and minimaxity

Strawderman prior は許容性，非許容性の境界付近でパラメタライズされた狭義広義いずれも含む事前分布であり，図 1 のような結果が知られている．

一方 $\eta$ が未知の一般化ベイズ推定量の許容性は対応する結果が得られてなかった．一般化ベイズ推定量が許容的であるための十分条件 Blyth (1951) は，広義事前分布に収束する適切な分布列が必要であり，その構成が難しかったからである．

今回得られた結果（図 1）は次のように要約される．「Strawderman prior 型の事前分布 $G(\|\mu\|)$（狭義・広義の両方を含む）が分散既知の場合に許容的な（一般化）ベイズ推定量を導く事前分布ならば，分散未知の場合に広義事前分布

$$\frac{1}{\eta} \times \eta^{p/2} G(\eta^{1/2}\|\theta\|)$$

に対する一般化ベイズ推定量は許容的である．またその中にミニマクス性を併せ持つ一般化ベイズ推定量が存在することを示した」

# Multiplicative Decomposition of Stochastic Distributions and Their Applications

Toshiyhuki Shimono*

**Multiplying the absolute values from the $t$-distribution of $\nu = 2$ produces the Bradley-Terry model.** Observed values are often affected by stochastic modification from the latent variables to the observable variables, and a kind of multiplicative relations among the familiar probability distributions may play major roles. Consider the Bradley-Terry model (1952) where each player $i$ has its latent value $\pi_i > 0$ to "beat" another player $j$ (to be indicated as $i \succ j$) with the probability $\pi_i/(\pi_i + \pi_j)$. Suppose $i \succ j$ happens when and only when $\pi_i \times |v_i| \geq \pi_j \times |v_j|$ with independent variates $v_i \sim D_i$ and $v_j \sim D_j$ where $D_i$ and $D_j$ are probability distributions to be determined. The necessary and sufficient condition upon the pair of $D_i$ and $D_j$ is the condition $|v_1/v_2| \sim F(2,2)$ where $F(2,2)$ is the Snedecor-Fisher distribution with the paired parameters of 2 and 2 degrees of freedom. Note that $F(2,2)$ has the density probability function $1/(1+x)^2$ for $x \geq 0$ and its variates can be generated by $1/u - 1$ with $u$ being uniformely distribued variates on $[0,1]$. Interestingly, if $D_i$ and $D_j$ are the identical distribution, a solution of $D_i$ or $D_j$ is the Student's $t$-distribution with 2 degrees of freedom, to be denoted as $T(2)$. Otherwise, $D_1$ and $D_2$ can be a pair of $T(1)$ and $\sqrt{F(2,2)}$, where $T(1)$ is the Cauchy distribution, $\sqrt{V}$ denotes the distribution of $\sqrt{v}$ with the variates $v \sim V$.

The above can be formulated as $F(2,2) = |T(2)| \times |1/T(2)|$ and $F(2,2) = |T(1)| \times \sqrt{F(2,2)}$, where $1/V$, $|V|$ and $V_1 \times V_2$ mean the distributions of $1/v$, $|v|$ and $v_1 \times v_2$, respectivly, with the independent variates $v \sim V$, $v_1 \sim V_1$ and $v_2 \sim V_2$. These relations can be proven by considering the $m$-th order moment of any distribution $V$ where $m \in \mathbb{R}$, not constrained to $m \in \mathbb{N}$, which is denoted to be $M\langle V \rangle$ as a function of $m$. Note that $M\langle V_1 \times V_2 \rangle = M\langle V_1 \rangle M\langle V_2 \rangle$. One can calculate that $M\langle |T(1)| \rangle = \Gamma(1/2 + m/2)\Gamma(1/2 - m/2)/\pi$, $M\langle |T(2)| \rangle = \sqrt{2^m}\Gamma(1/2+m/2)\Gamma(1-m/2)/\sqrt{\pi}$, $M\langle F(2,2) \rangle = \Gamma(1+m/2)\Gamma(1/2+m/2)\Gamma(1/2-m/2)\Gamma(1-m/2)/\pi$ where $\Gamma(\cdot)$ is the gamma function, which lead to the proof.

* The Institute of Statistical Mathematics, tshimono@05.alumni.u-tokyo.ac.jp

# Weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation

Junichi Hirukawa and Kou Fujimori
Niigata University and Shinshu University

**ABSTRACT**

An integral transformation which changes a fractional Brownian motion to a process with independent increments has been given. A representation of a fractional Brownian motion through a standard Brownian motion on a finite interval has also been given. On the other hand, it is known that the partial sum of the discrete time fractionally integrated process ($I(d)$ process) weakly converges to a fractional Brownian motion in infinite interval representation. In this talk we derive the weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation.

## 1 Introduction

Stochastic analysis for FBM has been developed by Decreusefond and Üstünel (1997) using Malliavin calculus. Norros et al. (1999) showed that many basic results can be obtained more directly with rather elementary arguments and computations. Norros et al. (1999) considered a normalized fractional Brownian motion (FBM) $(Z_t)_{g\geq 0}$ with self-similarity parameter $H \in (0,1)$. Mandelbrot and Van Ness (1968) defiend the process more constructively as the integral

$$Z_t - Z_s = c_H\left(\int_s^t (t-u)^{H-1/2}\,dW_u + \int_{-\infty}^t \left\{(t-u)^{H-1/2} - (s-u)^{H-1/2}\right\} dW_u\right),$$

where $W_t$ is the standard Brownian motion. The normalization $E\left(Z_1^2\right) = 1$ is achieved with the choice

$$c_H = \left(\frac{2H\Gamma\left(\frac{3}{2} - H\right)}{\Gamma\left(H + \frac{1}{2}\right)\Gamma\left(2 - 2H\right)}\right)^{1/2},$$

where $\Gamma(\cdot)$ denotes the Gamma function.

### 1.1 The fundamental martingale $M$

Norros et al. (1999) considered the following process. Let $w(t,s)$ be the function

$$w(t,s) = \begin{cases} c_1 s^{1/2-H}\,(t-s)^{1/2-H}, & \text{for } s \in (0,t), \\ 0, & \text{for } s \notin (0,t), \end{cases}$$

where

$$c_1 = \left\{2HB\left(\frac{1}{2} - H, H + \frac{1}{2}\right)\right\}^{-1}$$

and $B$ is the beta funtion

$$B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}.$$

Then, the centered Gaussian process

$$M_t = \int_0^t w(t,s)\, dZ_s$$

has independent increments and variance function

$$E\left(M_t^2\right) = c_2^2 t^{2-2H},$$

where

$$c_2 = \frac{c_H}{2H(2-2H)^{1/2}}.$$

In particular, $M$ is a martingale.

## 2 Weak convergence of $I(d)$ process

Now, we obtain the following functional central limit result for $I(d)$ process

$$\frac{1}{\sigma n^{d+1/2}}\widetilde{Z}_{[nt]} = \frac{1}{\sigma n^{d+1/2}}\sum_{s=1}^{[nt]} v_{s-1}^{1/2} DW_s + \frac{1}{\sigma n^{d+1/2}}\sum_{s=1}^{[nt]-1}\left\{\sum_{u=s}^{[nt]-1}\theta_{u,u+1-s}v_{s-1}^{1/2}\right\}DW_s$$
$$\Rightarrow \frac{1}{\Gamma(d)}\int_0^t s^{-d}\left\{\int_s^t (u-s)^{d-1}u^d du\right\}dW(s) := \int_0^t dZ(s) = Z(t).$$

# References


Mandelbrot B. B. and van Ness J. W. (1968). Fractional Brownian motions, fractional noises and applications *SIAM Review,* **10**, 422–437.

Norros, I. and Valkeila, E. and Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions *Bernoulli* **4**, 571–587.

Pipiras, V. and Taqqu, Murad S. (2001). Are classes of deterministic integrands for fractional Brownian motion on an interval complete? *Bernoulli* **7**, 873–897.

Tanaka, K. (2013). Distributions of the maximum likelihood and minimum contrast estimators associated with the fractional Ornstein-Uhlenbeck process *Stat. Inference Stoch. Process.* **16**, 173–192.

# Applications of Distance Correlation to Time Series：距離の相関係数の時系列解析への応用

南山大学　松井 宗也

## 研究内容

近年、「距離の相関係数」(Distance Correlation) という 2 つの確率ベクトルの依存関係を測る指標が注目されている。ここで 2 つのベクトルの次元は任意で異なっていてもよい。2005 年に Székely, G.j. により提案されて以来盛んに研究されている（詳しい定義等は [Székely et al.(2007)] や [Székely and Rizzo(2009)] 参照）。注目すべき特徴は、ケンドールやスピアマンの順位相関係数といった従来の指標では捉えることの難しい非線形な相関も検知できることである。そもそも 2 つの確率変数が独立であるとは、その同時特性関数がそれぞれの周辺特性関数の積として書けることと同値である。「距離の相関係数」の基本的なアイデアは、同時特性関数と独立な場合のそれ（各々の周辺関数の積）の距離をみて相関を測るものである。2 つの確率ベクトル $X \in \mathbb{R}^p$ と $Y \in \mathbb{R}^q$（次元は任意で異なっていてもよい）の同時特性関数を $\varphi_{X,Y}(s,t)$ とおき、またそれぞれの周辺特性関数を $\varphi_X(s)$、$\varphi_Y(s)$ とおく（ここで $p, q \in \mathbb{N}$）。また重み付き測度（重み関数（正）かけるルベーグ測度等）を $\mu(s,t)$ とおく。すると「距離の相関係数」は

$$\int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)|^2 \mu(ds, dt)$$

と重み付けした $L_2$ 距離で定義される。この形を見れば統計量は独立ならばその時に限り 0 となることがわかる。特に Székely,G.j. らは具体的に重み関数を $\mu(ds, dt) = c_{p,q}|s|^{-\alpha-p}|t|^{-\alpha-q} ds\, dt$ とおくことで統計量を

$$\begin{aligned} T(X, Y; \mu) \;&=\; E[|X - X'|^\alpha |Y - Y'|^\alpha] + E[|X - X'|^\alpha]\, E[Y - Y'|^\alpha] \\ &\quad -2\, E[|X - X'|^\alpha |Y - Y''|^\alpha] \end{aligned}$$

と陽な形に表した。ここで $(X', Y')$ は $(X, Y)$ の iid コピーで、$Y''$ は全ての変数と独立な $Y$ のコピーである。

本研究はこの「距離の相関係数」を 1 次元と多次元の定常な時系列へ応用するものである。主結果は論文 [Davis et al.(2018)] にまとめてある。主な目的は

１．2 つの時系列が独立かどうかを検定する、あるいは「距離の相関」で 2 つの時系列の相互依存関係を測る。

２．タイムラグをとった系列の「距離の相関」をみることで、1 つの時系列が系列相関を持つかどうか検定する。

ことである。時系列解析では、多次元の系列相関を測る指標として（多次元の）自己相関関数が一般的に知られている。ここで言う相関はピアソンの積率相関係数である。非常に便利な指標であるが、非線形時系列モデルの依存関係を捉えるのが難

しいことが知られている。非線形時系列は計量ファイナンス分野において近年よく研究されていて、例えばその代表例として $GARCH(p,q)$ モデルが挙げられる。このモデルは、単に原系列の自己相関をみても依存関係を検出できず、原系列を変換したも（例えば絶対値や 2 乗を考えたもの）の自己相関をみて初めて依存関係が捉えられる。

本研究のアイデアは、そのピアソン流の相関を「距離の相関係数」に置き換えることで依存関係をより良く検出しようというものである。論文では新しく「自己距離の相関関数」という指標を自己相関関数に代わりうるものとして提案した。そして、ミキシング条件 (strong mixing) のもとで、標本「自己距離の相関関数」に関連する漸近論を導出した。まず、緩やかモーメント条件のもとで統計量の一致性を示した。その後いくつか追加的な条件を与え、依存関係がある場合と独立な場合の両方で統計量の漸近分布を理論的に導出した。独立な場合とそうでない場合は収束のオーダーが異なることも示した。検定統計量は特に独立な場合は、独立なカイ２乗変数の線形無限和で表され、一般に上側確率を求めるのが難しいことが知られている。これに対し、ブートスラップ法を用いた数値実験で漸近分布がうまく近似できることも示した。

２つ目の研究である 1 系列における「自己距離の相関関数」に関しては、具体的なモデルとして $AR(p)$ モデルを考えた。モデルの適合度をみるために、パラメータ推定後に得られる残差に「自己距離の相関関数」を適応した。すると興味深いことに、iid 系列に「自己距離の相関関数」を適応した場合と比較して明らかに異なることもわかった（これは自己相関関数と同様の現象である)。その他、誤差項に裾の重い分布を仮定した場合は Székely, G.j. のオリジナルな定義では漸近論がうまく機能しないこともわかった。パラメータの推定も考えたもとでは、モーメント制約がより厳しくなるためと考えられる。我々が提案した他の重み関数によるものは、裾の厚い誤差項に関してもうまく対応することも確認できた。

実証研究も行い、アマゾンの株価収益率データや風速データなどで提案した方法の応用を試みた。

キーワード：Auto- and cross-distance correlation function, testing independence, strong mixing, ergodicity, Fourier analysis, $U$-statistics, AR process, residuals

# 参考文献

[Davis et al.(2018)] Davis, R.A., Matsui, M., Mikosch, T. and Wan, P. (2018) Applications of distance correlation to time series. *Bernoulli* **24**, 3087–3116.

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

[Székely et al.(2007)] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

[Székely and Rizzo(2009)] Székely, G.J. and Rizzo, M.L. (2009) Brownian distance covariance. *Ann. Appl. Stat.* **3**, 1236–1265.

# 弱非ガウス確率場に対する期待ミンコフスキー汎関数

栗木哲　(情報・システム研究機構 統計数理研究所 数理・推論研究系)[*1]
松原隆彦　(高エネルギー加速器研究機構 素粒子原子核研究所・理論センター)

$X(t)$ を $E \subset \mathbb{R}^n$ 上の滑らかなサンプルパスを持つ実数値確率場とする．$X(t)$ の閾値 $v$ に対する上側レベル集合 $E_v = \{t \in E \mid X(t) \geq v\}$ をエクスカーション集合という．$B^n$ を $n$ 次元単位球とし，$E_v$ と半径 $\rho$ の球 $\rho B^n$ とのミンコフスキー和 (距離 $\rho$ のチューブ近傍) $\mathrm{Tube}(E_v, \rho) = E_v + \rho B^n$ を考える．その体積は $\rho$ が小さい範囲で $\rho$ の多項式

$$\mathrm{Vol}_n(\mathrm{Tube}(E_v, \rho)) = \sum_{j=0}^{n} \omega_{n-j} \rho^{n-j} \mathcal{L}_j(E_v) = \sum_{j=0}^{n} \rho^j \binom{n}{j} \mathcal{M}_j(E_v)$$

($\omega_d = \mathrm{Vol}_d(B^d)$) となることが知られている．係数 $\mathcal{M}_j(E_v)$, $\mathcal{L}_j(E_v)$ をそれぞれ $E_v$ のミンコフスキー汎関数，リプシッツ・キリング曲率という．これらは $E_v$ の体積や表面積を特別な場合として含む幾何量である．特に $\mathcal{L}_0(E_v) = \chi(E_v)$ はオイラー標数である．

$X(t)$ が平均 0 の等方的ガウス確率場の場合，$\mathcal{L}_j(E_v)$ の期待値は古くから陽に知られている．平均 0 の等方的ガウス確率場は，距離のみで定義される 2 点相関関数 (共分散関数) $\mathbb{E}[X(t_1)X(t_2)] = \rho\big(\|t_1 - t_2\|^2/2\big)$ で特徴付けられる．ここでは等方的弱非ガウス確率場を表現するために $N$ 点相関関数 ($N$ 次キュムラント関数)

$$\mathrm{cum}(X(t_1), \ldots, X(t_N)) = \nu^{N-2} \kappa^{(N)}\big(\|t_1 - t_1\|^2/2, \|t_1 - t_3\|^2/2, \ldots, \|t_{N-1} - t_N\|^2/2\big)$$

($\nu \ll 1$ は非ガウス性を表すパラメータ) を導入し，$\mathbb{E}[\mathcal{L}_j(E_v)]$ の $\nu \downarrow 0$ のときの摂動展開を論じる．一般性を失うことなく $\rho(0) = \mathbb{E}[X(t)^2] = 1$ とする．

$$\begin{bmatrix} k+d \\ k \end{bmatrix} = \frac{\Gamma(\frac{k+d+1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{k+1}{2})\Gamma(\frac{d+1}{2})}$$

とおく．

**定理 1.** 任意の正整数 $s$ に対して，正則条件の下で，$k = 0, \ldots, n$ について

$$\mathbb{E}[\mathcal{L}_k(E_v)] = \omega_{n-k}^{-1} \binom{n}{k} \mathbb{E}[\mathcal{M}_{n-k}(E_v)] = \sum_{d=0}^{n-k} \begin{bmatrix} k+d \\ k \end{bmatrix} \mathcal{L}_{k+d}(E)\, \Xi_d(v) + o(\nu^{s-2}), \qquad (1)$$

ただし

$$\begin{aligned}
\Xi_n(x) =& \gamma^{n/2}(2\pi)^{-n/2}\phi(x) \times \Big(H_{n-1}(x) + \nu\Delta_{1,n}(x) + \nu^2\Delta_{2,n}(x)\Big) + o(\nu^2),\\
\Delta_{1,n}(x) =& \tfrac{1}{2}\gamma^{-2}\kappa_{11}n(n-1)H_{n-2}(x) - \tfrac{1}{2}\gamma^{-1}\kappa_1 n H_n(x) + \tfrac{1}{6}\kappa_0 H_{n+2}(x),\\
\Delta_{2,n}(x) =& \Big(-\tfrac{1}{6}\gamma^{-3}(3\widetilde{\kappa}^a_{111} + \widetilde{\kappa}^d_{111}) + \tfrac{1}{8}\gamma^{-4}\kappa_{11}^2(n-7)\Big)n(n-1)(n-2)H_{n-3}(x)\\
&+ \Big(\tfrac{1}{8}\gamma^{-2}\big(\widetilde{\kappa}^{aa}_{11}(n-2) + 4\widetilde{\kappa}^a_{11}(n-1)\big)\\
&\quad - \tfrac{1}{4}\gamma^{-3}\kappa_1\kappa_{11}(n-1)(n-4)\Big)nH_{n-1}(x)\\
&+ \Big(-\tfrac{1}{4}\gamma^{-1}\widetilde{\kappa}_1 + \tfrac{1}{24}\gamma^{-2}\big(3\kappa_1^2(n-2) + 2\kappa_0\kappa_{11}(n-1)\big)\Big)nH_{n+1}(x)\\
&+ \Big(\tfrac{1}{24}\widetilde{\kappa}_0 - \tfrac{1}{12}\gamma^{-1}\kappa_0\kappa_1 n\Big)H_{n+3}(x) + \tfrac{1}{72}\kappa_0^2 H_{n+5}(x).
\end{aligned}$$

[*1] e-mail: kuriki@ism.ac.jp

ここで $\phi(x)$ は標準ガウス密度関数，$H_j(x)$ はエルミート多項式，$\gamma = -\rho'(0)$, また

$$\kappa^{(N)}_{(i_{12},\dots,i_{N-1,N})} = \left( \prod_{1\le a<b\le N} \left(\frac{\partial}{\partial x_{ab}}\right)^{i_{ab}} \right) \kappa^{(N)}(x_{12},\dots,x_{N-1,N}) \Big|_{(x_{12},\dots,x_{N-1,N})=(0,\dots,0)}$$

とおくとき，$\kappa_0 = \kappa^{(3)}_{(0,0,0)}$, $\kappa_1 = \kappa^{(3)}_{(1,0,0)}$, $\kappa_{11} = \kappa^{(3)}_{(1,1,0)}$, $\widetilde{\kappa}_0 = \kappa^{(4)}_{(0,0,0,0,0,0)}$, $\widetilde{\kappa}_1 = \kappa^{(4)}_{(1,0,0,0,0,0)}$, $\widetilde{\kappa}^a_{11} = \kappa^{(4)}_{(1,1,0,0,0,0)}$, $\widetilde{\kappa}^{aa}_{11} = \kappa^{(4)}_{(1,0,0,0,0,1)}$, $\widetilde{\kappa}^d_{111} = \kappa^{(4)}_{(1,1,1,0,0,0)}$, $\widetilde{\kappa}^a_{111} = \kappa^{(4)}_{(1,1,0,0,1,0)}$.

式 (1) は，Gaussian kinematic formula [3] のひとつの一般化である．

証明は，Kac-Rice 公式 (Morse の定理の積分形) を用いて表した $E_v$ のオイラー標数 $\mathcal{L}_0(E_v)$ を，原点における確率場およびその 1,2 階導関数 $(X(0), \nabla X(0), \nabla^2 X(0))$ $(1+n+n(n+1)/2$ 次元ベクトル) のグラムシャリエ展開のもとで期待値をとることで $\mathbb{E}[\mathcal{L}_0(E_v)]$ を得る．また等方性の仮定より，$\mathbb{E}[\mathcal{L}_k(E_v)]$ $(k \ge 1)$ は $\mathbb{E}[\mathcal{L}_0(E_v)]$ から Crofton 公式を通して得られる．

摂動項の評価のためには，$A = (a_{ij})$ を $n \times n$ GOE ランダム行列の $\sqrt{2}$ 倍，すなわち対角成分と上三角成分が独立にガウス分布 $a_{ii} \sim N(0,2)$, $a_{ij}\,(= a_{ji}) \sim N(0,1)$ $(i<j)$ に従うとき，その固有多項式を含むモーメントの評価が必要となる．$A$ のモーメント母関数は $\mathbb{E}[e^{\mathrm{tr}(\Theta A)}] = e^{\mathrm{tr}(\Theta^2)}$ であることに注意する．対称行列 $A = (a_{ij})$ の微分作用素 $D_A$ を，その $(i,j)$ 要素が

$$(D_A)_{ij} = \frac{1+\delta_{ij}}{2}\frac{\partial}{\partial a_{ij}} \quad (i \le j)$$

である行列作用素と定義する．

**補題 1.** $m = \sum_{i=1}^{\ell} c_i$ とおく．

$$\begin{aligned}(-1/2)^{m-\ell}(n)_m H_{n-m}(x) &= \mathbb{E}\big[\mathrm{tr}(D_A^{c_1})\cdots\mathrm{tr}(D_A^{c_\ell})\det(xI_n + A)\big] \\ &= \det(xI + D_\Theta)\big(e^{\mathrm{tr}(\Theta^2)}\mathrm{tr}(\Theta^{c_1})\cdots\mathrm{tr}(\Theta^{c_\ell})\big)\Big|_{\Theta=0}.\end{aligned}$$

とくに $\ell = m = 0$ のとき，

$$H_n(x) = \mathbb{E}[\det(xI_n + A)] = \det(xI_n + D_\Theta)e^{\mathrm{tr}(\Theta^2)}\Big|_{\Theta=0}.$$

宇宙論においては宇宙場 (宇宙マイクロ波背景放射，宇宙の大規模構造，銀河サーベイにおける弱重力レンズ効果など) のモデルの適合度の検定統計量としてミンコフスキー汎関数が用いられている．本発表では，それらの問題背景について併せて説明した．詳細は [1], [2] を参照のこと．

## 参考文献


[1] Kuriki, S. and Matsubara, T. (2020). Perturbation of the expected Minkowski functional for weakly non-Gaussian isotropic fields on a bounded domain, arXiv:2011.04953 [math.ST] https://arxiv.org/abs/2011.04953

[2] Matsubara, T. and Kuriki, S. (2020). Weakly non-Gaussian formula for the Minkowski functionals in general dimensions, arXiv:2011.04954 [astro-ph.CO] https://arxiv.org/abs/2011.04954

[3] Taylor, J. E. and Adler, R. J. (2009). Gaussian processes, kinematic formulae and Poincaré's limit, *Ann. Probab.*, **37** (4), 1459–1482.

# 外れ値がある場合でも安定した推定を行う罰則付推定法

永井 勇[†]

[†] 中京大学 教養教育研究院

2021年1月23日
**統計科学の革新に向けて**



## 概要

- 線形回帰モデルでよく使われる推定量の問題点
  説明変数の多重共線性・目的変数の外れ値に弱い
  - 多重共線性を回避する手法の一つ；
    Ridge 回帰 (Hoerl & Kennard, 1970)

**本講演の研究の主目的**

多変量線形回帰モデルで
多重共線性に対しても，外れ値に対しても，
頑健な推定量の構築

- 多変量線形回帰モデルでの Ridge 回帰 (Yanagihara & Satoh, 2010) を拡張

1. 予測平均二乗誤差が従来のより小さくできる
2. 導入したパラメータの最適な値が
   陽に求まる推定量の提案



## 目次





## ここからの話





## 扱うモデル; 多変量線形回帰モデル

モデル 多変量線形回帰モデル; $\boldsymbol{Y} = \boldsymbol{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi} + \boldsymbol{\mathcal{E}}$

既知のものなど $\boldsymbol{Y}$; $n \times p$ 目的変数行列,
$\boldsymbol{A}$; $n \times k$ 説明変数行列,
$\boldsymbol{1}_n$: $n$ 次元の1ベクトル

未知のものなど $\boldsymbol{\mu}$; $p$ 次元未知ベクトル,
$\boldsymbol{\Xi}$; $k \times p$ 次元未知行列,
$\boldsymbol{\mathcal{E}}$; $n \times p$ 誤差行列

仮定 $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$ ($\boldsymbol{0}_k$ は $k$ 次元0ベクトル), $\mathrm{rank}(\boldsymbol{A}) = k$,
$E[\boldsymbol{\mathcal{E}}] = \boldsymbol{0}_n\boldsymbol{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$, $\mathrm{rank}(\boldsymbol{\Sigma}) = p$,
$n - k - 1 > 0$, $\boldsymbol{\mu} \neq \boldsymbol{0}_p$

推定量 $\boldsymbol{\mu}$ や $\boldsymbol{\Xi}$ の最小二乗推定量 (LSE);

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

問題点 $\boldsymbol{A}$ の多重共線性・$\boldsymbol{Y}$ の外れ値 に弱い



## 従来の推定量の問題点

- LSE などの従来の推定量の問題点;
  $\boldsymbol{A}$ の多重共線性や $\boldsymbol{Y}$ の外れ値に弱い
- $\boldsymbol{A}$ に多重共線性がある
  (説明変数間に相関の高い組がある)

→ $\hat{\boldsymbol{\Xi}}$ が信頼できない

⇒ Ridge 回帰による推定法 (Yanagihara & Satoh, 2010),
一般化 Ridge 回帰による推定法 (Yanagihara, Nagai & Satoh, 2009)

- $\boldsymbol{Y}$ に外れ値がある

→ $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Xi}}$ はその影響を大きく受ける

今回 Ridge 回帰による推定法 & 外れ値への対応

⇒ **両方の問題へ対応した頑健な推定量を提案**



## ここからの話





## 二つの問題点に対応する推定量と拡張

- $\boldsymbol{A}$ の多重共線性 ⇒ Ridge 回帰による推定

$$\hat{\boldsymbol{\Xi}}(\theta) = (\boldsymbol{A}'\boldsymbol{A} + \theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y},$$

ここで $\theta$ ($\geq 0$) はパラメータ.

- $p = 1$ の場合の $y$ の外れ値に頑健な推定量
  (Jimichi, 2016)

$$(\boldsymbol{1}_n'\boldsymbol{1}_n + \lambda)^{-1}\boldsymbol{1}_n'\boldsymbol{y}.$$ ($\boldsymbol{y}$ は $p = 1$ の際の目的変数ベクトル)

拡張 ⇒ $\boldsymbol{Y}$ の外れ値への頑健な推定量 ($p \geq 1$ のケース) へ

$$\hat{\boldsymbol{\mu}}(\lambda) = (\boldsymbol{1}_n'\boldsymbol{1}_n + \lambda)^{-1}\boldsymbol{1}_n'\boldsymbol{Y},$$

ここで $\lambda$ ($\geq 0$) はパラメータ.

⇒ $\hat{\boldsymbol{\Xi}}(\theta)$ と $\hat{\boldsymbol{\mu}}(\lambda)$ により
$\boldsymbol{A}$ の多重共線性と $\boldsymbol{Y}$ の外れ値に強い推定が可能
(LSE より予測平均二乗誤差 (後述) を小さくできる)



## 推定量の評価; 予測平均二乗誤差

- $\hat{\boldsymbol{\mu}}(\lambda)$, $\hat{\boldsymbol{\Xi}}(\theta)$ の評価基準
- 推定量の評価基準; 予測平均二乗誤差 (PMSE)

$$\mathrm{PMSE}[\hat{\boldsymbol{Y}}] \overset{\mathrm{def}}{=} E_{\boldsymbol{U}}[E_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{U} - \hat{\boldsymbol{Y}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{U} - \hat{\boldsymbol{Y}})'\}]]$$

ここで $E_{\boldsymbol{Z}}[\cdot]$ は $\boldsymbol{Z}$ に関する期待値,
$\boldsymbol{U}$ は $\boldsymbol{Y}$ と独立に同一の分布から得られる行列,
$\hat{\boldsymbol{Y}}$ は $\boldsymbol{Y}$ などから作られた予測値

意味 今のデータ $\boldsymbol{Y}$ などから作った予測値 $\hat{\boldsymbol{Y}}$ と
同じモデルから得られた新しいデータ $\boldsymbol{U}$ との
基準化した残差平方和の平均

- $\hat{\boldsymbol{\mu}}(\lambda)$, $\hat{\boldsymbol{\Xi}}(\theta)$ のパラメータ $\lambda$, $\theta$ の最適化

⇔ $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ が小さくなるように選ぶ
ここで $\hat{\boldsymbol{Y}}(\lambda, \theta) = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}(\lambda)' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}(\theta)$.

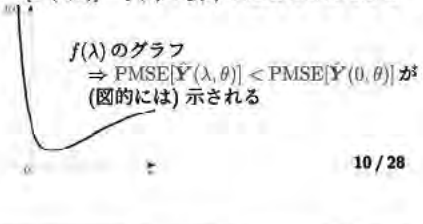


## $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$ ? (1/2)

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$
  となる $\lambda \geq 0$ や $\theta \geq 0$ は存在するのか？
  意味 $\lambda$, $\theta$ を上手く選べば，他よりも提案した推定量が良いか？

注意 この不等式は Ridge 回帰による推定で分かっている

→ $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] = f(\lambda) + g(\theta) +$ 定数 の形にできる




$f(\lambda)$ のグラフ
⇒ $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, \theta)]$ が
(図的には) 示される



## $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$ ? (2/2)

問題 $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$
となる $\lambda \geq 0$ や $\theta \geq 0$ は存在するのか？

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$ は Ridge 回帰による推定より

→ $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] = f(\lambda) + g(\theta) +$ 定数 の形

→ $\partial f(\lambda)/(\partial\lambda)|_{\lambda=0} \leq 0$ であればいい

実際 $\partial f(\lambda)/(\partial\lambda)|_{\lambda=0} = -2p/n < 0$

**$\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, \theta)] < \mathrm{PMSE}[\hat{\boldsymbol{Y}}(0, 0)]$
となる $\lambda > 0$, $\theta > 0$ が存在する.**

**LSE より Ridge 回帰による推定量のほうが良い.
両者より提案する推定量 ($\hat{\boldsymbol{\mu}}(\lambda)$, $\hat{\boldsymbol{\Xi}}(\theta)$) のほうが良い.**

⇒ $\lambda$, $\theta$ の最適化



## PMSE に基づく最適化のアイデア

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ が小さくなるように $\lambda$ と $\theta$ を選ぶ

注意 $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ には $\boldsymbol{U}$ や $\boldsymbol{\Sigma}$ (未知) などがある

- $\lambda$ と $\theta$ の最適化のアイデア

1. $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ の推定量の最小化

名称 $C_p$ 型情報量規準による最適化 ($C_p$ 型最適化)
(Yanagihara & Satoh (2010) など)

2. $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\lambda$ や $\theta$ を求め，
   未知の部分 ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ など) に推定量を代入

名称 Plug-in による最適化 (PI 型最適化)
(Nagai, Yanagihara & Satoh (2012) など)

今回 $\lambda$ に対して両方の最適化法

- $\theta$: 最適な $\lambda$ を用いて $C_p$ 型最適化



## ① $\lambda$ の $C_p$ 型最適化 (1/5)

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$
  $= E[\mathrm{tr}\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta))\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta))'\}]$
  $+ 2p \times \mathrm{tr}\{\boldsymbol{G}(\lambda) + \boldsymbol{H}(\theta)\} +$ 定数,
  ここで $\hat{\boldsymbol{Y}}(\lambda, \theta) = (\boldsymbol{G}(\lambda) + \boldsymbol{H}(\theta))\boldsymbol{Y}$,
  $\boldsymbol{G}(\lambda) = \boldsymbol{1}_n(\boldsymbol{1}_n'\boldsymbol{1}_n + \lambda)^{-1}\boldsymbol{1}_n'$, $\boldsymbol{H}(\theta) = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'$.

→ ここをどう推定するか？ ($\boldsymbol{\Sigma}$; 未知)

- $C_p$ 規準 ($\lambda$, $\theta$ に関する項のみ);

$$C_p(\lambda, \theta) = 1 \times \mathrm{tr}\left\{\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)\boldsymbol{S}^{-1}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)'\right\} + 2p \times \mathrm{tr}\{\boldsymbol{G}(\lambda) + \boldsymbol{H}(\theta)\},$$

$\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \boldsymbol{G}(0) - \boldsymbol{H}(0)\}\boldsymbol{Y}/(n - k - 1)$
($\det(\boldsymbol{S}) \neq 0$ を仮定).



## ① $\lambda$ の $C_p$ 型最適化 (2/5)

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] \neq E[C_p(\lambda, \theta)] +$ 定数

→ バイアス補正

仮定追加 $n - k - p - 2 > 0$, $\boldsymbol{\mathcal{E}}$ の各行 $\sim N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$

→ $MC_p$ 規準 ($\lambda$, $\theta$ に関する項のみ);

$$MC_p(\lambda, \theta) = c \times \mathrm{tr}\left\{\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)\boldsymbol{S}^{-1}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)'\right\} + 2p \times \mathrm{tr}\{\boldsymbol{G}(\lambda) + \boldsymbol{H}(\theta)\},$$

ここで $c = 1 - \dfrac{p + 1}{n - k - 1}$.

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)] = E[MC_p(\lambda, \theta)] +$ 定数
- $C_p(\lambda, \theta)$; $c$ のところが 1

⇒ $C_p(\lambda, \theta)$, $MC_p(\lambda, \theta)$ の最小化により最適化



## ① $\lambda$ の $C_p$ 型最適化 (3/5)

- $C_p(\lambda, \theta)$, $MC_p(\lambda, \theta)$ を用いて最適化

→ $\lambda$ と $\theta$ に関連する項はほぼ共通の形

→ $GC_p(\lambda, \theta|\alpha) = \alpha \times \mathrm{tr}\left\{\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)\boldsymbol{S}^{-1}\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\lambda, \theta)\right)'\right\} + 2p \times \mathrm{tr}\{\boldsymbol{G}(\lambda) + \boldsymbol{H}(\theta)\}$

- $GC_p(\lambda, \theta|1) = C_p(\lambda, \theta)$
- $GC_p(\lambda, \theta|c) = MC_p(\lambda, \theta)$

→ $(\hat{\lambda}(\alpha), \hat{\theta}(\alpha)) = \arg\min_{\lambda \geq 0, \theta \geq 0} GC_p(\lambda, \theta|\alpha)$ とすると，

$\hat{\lambda}(1), \hat{\theta}(1) \Leftrightarrow C_p(\lambda, \theta)$ に基づく最適化の結果

$\hat{\lambda}(c), \hat{\theta}(c) \Leftrightarrow MC_p(\lambda, \theta)$ に基づく最適化の結果

→ $GC_p(\lambda, \theta|\alpha)$ を最小化

## ① $\lambda$の$C_p$型最適化 (4/5)

- $(\hat{\lambda}(\alpha), \hat{\theta}(\alpha)) = \arg\min_{\lambda\geq 0, \theta\geq 0} GC_p(\lambda, \theta|\alpha)$ を求める
  - $\alpha = 1 \Leftrightarrow C_p(\lambda, \theta)$ での最適化
  - $\alpha = c \Leftrightarrow MC_p(\lambda, \theta)$ での最適化
- $GC_p(\lambda, \theta|\alpha) = F(\lambda|\alpha) + G(\theta|\alpha) +$ 定数 とできる,

  ここで $F(\lambda|\alpha) = \dfrac{-\alpha(n+2\lambda)}{(n+\lambda)^2}\mathbf{1}_n'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{Y}'\mathbf{1}_n + \dfrac{2pn}{n+\lambda}$,

  $G(\theta|\alpha) = \alpha\mathrm{tr}((\boldsymbol{H}(\theta) - 2\boldsymbol{I}_n)\boldsymbol{H}(\theta)\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{Y}')$
  $+2p\mathrm{tr}(\boldsymbol{H}(\theta))$.
  - $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ より

→ $\hat{\lambda}(\alpha) = \arg\min_{\lambda\geq 0} F(\lambda|\alpha)$, $\hat{\theta}(\alpha) = \arg\min_{\theta\geq 0} G(\theta|\alpha)$
- $\hat{\theta}(\alpha)$ は陽に求まらない
- $\hat{\lambda}(1), \hat{\theta}(1) \Leftrightarrow C_p(\lambda, \theta)$ での最適化結果
  $\hat{\lambda}(c), \hat{\theta}(c) \Leftrightarrow MC_p(\lambda, \theta)$ での最適化結果



## ① $\lambda$の$C_p$型最適化 (5/5)

- $(\hat{\lambda}(\alpha), \hat{\theta}(\alpha)) = \arg\min_{\lambda\geq 0, \theta\geq 0} GC_p(\lambda, \theta|\alpha)$ を求める
  - $\alpha = 1 \Leftrightarrow C_p(\lambda, \theta)$ での最適化
  - $\alpha = c \Leftrightarrow MC_p(\lambda, \theta)$ での最適化

→ $\hat{\lambda}(\alpha) = \arg\min_{\lambda\geq 0} F(\lambda|\alpha)$ となった
- $\hat{\theta}(\alpha)$ は陽に求まらない

⇒ $\hat{\lambda}(\alpha)$ は陽に求まる;

$$\hat{\lambda}(\alpha) = \begin{cases} \dfrac{np}{\alpha t - p} & (\alpha t > p \text{ の場合}) \\ \infty & (\text{それ以外の場合}) \end{cases},$$

ここで $t = n\hat{\boldsymbol{\mu}}'\boldsymbol{S}^{-1}\hat{\boldsymbol{\mu}}$.
- $\hat{\theta}(\alpha) = \arg\min_{\theta\geq 0} G(\theta|\alpha)$ で $\theta$ の最適化
- 多変量線形回帰モデルにおける一般化リッジ回帰による推定 (Yanagihara ら, 2009) の最適なパラメータに対応

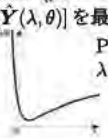


## ② $\lambda$のPI型最適化 (1/4)

今まで PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ の推定量 $(C_p(\lambda, \theta),\ MC_p(\lambda, \theta))$ を構築 → 最小化により最適化
- 最適な $\lambda$ は陽に求まる
- 最適な $\theta$ は陽に求まらない

← 最適な $\theta$ は $G(\theta|\alpha)$ の最小化
$(C_p(\lambda, \theta|\alpha),\ MC_p(\lambda, \theta|\alpha))$ の $\theta$ に関する部分)

今から PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\lambda$, $\theta$ を求める
→ 未知の部分 $(\boldsymbol{\mu}, \boldsymbol{\Sigma}$ など) に推定量を代入
- PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\lambda$ は存在しそう



PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ の
$\lambda$ に関する項のグラフ



## ② $\lambda$のPI型最適化 (2/4)

- $(\lambda^*, \theta^*) = \arg\min_{\lambda\geq 0, \theta\geq 0} \mathrm{PMSE}[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を求める
- PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)] = f(\lambda) + g(\theta) +$ 定数,

  ここで $f(\lambda) = \dfrac{n^2}{(n+\lambda)^2}\{p - (n+2\lambda)\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}$,

  $g(\theta) = p\mathrm{tr}(\boldsymbol{H}(\theta)^2) + \mathrm{tr}\{(\boldsymbol{H}(\theta) - 2\boldsymbol{I}_n)\boldsymbol{H}(\theta)\boldsymbol{A}\Xi\boldsymbol{\Sigma}^{-1}\Xi'\boldsymbol{A}'\}$.
  - $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ より

→ $\lambda^* = \arg\min_{\lambda\geq 0} f(\lambda)$, $\theta^* = \arg\min_{\theta\geq 0} g(\theta)$
注意; $g(\theta)$ は未知の $\Xi, \Sigma$ がある
- $\lambda^*$ は未知の値を含みつつ陽に求まる;
  - $\theta^*$ は陽に求まらない

→ $\lambda^*(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \dfrac{p}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$

注意 $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ は未知



## ② $\lambda$のPI型最適化 (3/4)

- PI型最適化の結果; $\hat{\lambda} = \lambda^*(\hat{\boldsymbol{\mu}}, \boldsymbol{S})$
  - PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\lambda$ は陽に求まる
  - → 未知の部分 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に推定量を代入したもの
  - PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\theta$ は陽に求まらない

→ $\boldsymbol{Y}$ に外れ値がある → $\hat{\boldsymbol{\mu}}$ が信頼できない
⇒ $\boldsymbol{\mu}$ にそのまま代入するだけでいいの？
- Nagai ら (2012) での最適化法のアイデアの一つと同様

→ $\boldsymbol{\mu}$ の代わりに $\hat{\boldsymbol{\mu}}(\hat{\lambda}) = (\mathbf{1}_n'\mathbf{1}_n + \hat{\lambda})^{-1}\mathbf{1}_n'\boldsymbol{Y}$ を使う

**代入を繰り返す**

1. $\hat{\lambda}^{[0]} = 0$ とする
2. $\hat{\lambda}^{[s]} = \dfrac{p}{\hat{\boldsymbol{\mu}}(\hat{\lambda}^{[s-1]})'\boldsymbol{S}^{-1}\hat{\boldsymbol{\mu}}(\hat{\lambda}^{[s-1]})}$ $(s = 1, 2, \ldots)$



## ② $\lambda$のPI型最適化 (4/4)

- $\hat{\lambda}^{[0]} = 0 \to \hat{\lambda}^{[1]} \to \cdots$ として得られる数列 $\{\hat{\lambda}^{[s]}\}$ を考える

**$\{\hat{\lambda}^{[s]}\}$ の性質**

1. $\hat{\lambda}^{[s-1]} < \hat{\lambda}^{[s]}$ $(s = 1, 2, \ldots)$
2. $\hat{\lambda}^{[s]} < n$ ($\hat{\lambda}^{[1]} \leq n/4$ の場合, $s = 1, 2, \ldots$)

⇒ $\hat{\lambda}^{[s]}$ は (状況によっては) $s \to \infty$ で収束する

収束先 → $\hat{\lambda}^{[\infty]} = \begin{cases} \dfrac{n\left\{t - 2p - \sqrt{t(t-4p)}\right\}}{2p} & (t \geq 4p \text{ の場合}) \\ \infty & (\text{それ以外の場合}) \end{cases}$,

ここで $t = n\hat{\boldsymbol{\mu}}'\boldsymbol{S}^{-1}\hat{\boldsymbol{\mu}}$.



## $\lambda$の最適化のまとめ

- ① $C_p$ 型最適化・② PI型最適化ともに,
  最適な $\lambda$ は陽に求まる

① $C_p(\lambda, \theta)$, $MC_p(\lambda, \theta)$ に基づく $\lambda$ の最適化結果;

$$\hat{\lambda}(\alpha) = \begin{cases} np/(\alpha t - p) & (\alpha t > p \text{ の場合}) \\ \infty & (\text{それ以外}) \end{cases}$$

ここで $t = n\hat{\boldsymbol{\mu}}'\boldsymbol{S}^{-1}\hat{\boldsymbol{\mu}}$.
- $\alpha = 1 \Leftrightarrow C_p(\lambda, \theta)$ での最適化
- $\alpha = c \Leftrightarrow MC_p(\lambda, \theta)$ での最適化

② PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ の最小化に基づく $\lambda$ の最適化結果;

$$\hat{\lambda} = p/\hat{\boldsymbol{\mu}}'\boldsymbol{S}^{-1}\hat{\boldsymbol{\mu}} = np/t,$$

$$\hat{\lambda}^{[\infty]} = \begin{cases} n\{t - 2p - \sqrt{t(t-4p)}\}/2p & (t \geq 4p \text{ の場合}) \\ \infty & (\text{それ以外}) \end{cases}$$



## ここからの話





## 数値実験による比較 (1/2)-設定-

- $\boldsymbol{Y} \sim N_{n\times p}(A\Xi, \Sigma_p(\rho_y) \otimes \boldsymbol{I}_n)$ (反復ごとに生成)
  $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}_k(\rho_a)^{1/2}$, $\boldsymbol{U}$ は各成分が $U(-1, 1)$ の乱数 (反復中固定) の $n \times k$ 行列, $\Sigma_r(\rho)$ は $(i, j)$ 成分が $\rho^{|i-j|}$ の $r \times r$ 行列, $\Xi$ の成分は Nagai ら (2012) の参照, $\boldsymbol{Y}$ の $(n \times \gamma)$ 行に $\delta$ を加える
  - $\gamma$ は外れ値の割合, $\delta$ は外れ具合
  - 最適化や予測の際は $\boldsymbol{A}$ は基準化

基準 $\mathrm{tr}\{(\hat{\boldsymbol{Y}} - E[\boldsymbol{Y}])\boldsymbol{\Sigma}_p(\rho_y)^{-1}(\hat{\boldsymbol{Y}} - E[\boldsymbol{Y}])'\}$ の $10^4$ 回反復の平均
- $\hat{\boldsymbol{Y}}$; 各手法でパラメータの最適化をした際の予測値

- 比較手法; $\mu$ の推定量と $\Xi$ の推定量の組み合わせ

従来法 LSE; $\lambda = 0$ & $\theta = 0$, Ridge; $\lambda = 0$ & $\arg\min_{\theta\geq 0} MC_p(0, \theta)$

提案法 $\hat{\lambda}(1)$ & $\arg\min_{\theta\geq 0} C_p(\hat{\lambda}(1), \theta)$, $\hat{\lambda}(1)$ & $\arg\min_{\theta\geq 0} MC_p(\hat{\lambda}(1), \theta)$
$\hat{\lambda}(c)$ & $\arg\min_{\theta\geq 0} C_p(\hat{\lambda}(c), \theta)$, $\hat{\lambda}(c)$ & $\arg\min_{\theta\geq 0} MC_p(\hat{\lambda}(c), \theta)$
$\hat{\lambda}$ & $\arg\min_{\theta\geq 0} C_p(\hat{\lambda}, \theta)$, $\hat{\lambda}$ & $\arg\min_{\theta\geq 0} MC_p(\hat{\lambda}, \theta)$
$\hat{\lambda}^{[\infty]}$ & $\arg\min_{\theta\geq 0} C_p(\hat{\lambda}^{[\infty]}, \theta)$, $\hat{\lambda}^{[\infty]}$ & $\arg\min_{\theta\geq 0} MC_p(\hat{\lambda}^{[\infty]}, \theta)$



## 数値実験による比較 (2/2)-結果-LSEの結果との比

| $(n, p, \delta) = (30, 3, 10)$ の場合のみ | | | | Ridge | 提案手法 1 | | | | 提案手法 2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\lambda = 0$ | $\hat{\lambda}(1)$ | | $\hat{\lambda}(c)$ | | $\hat{\lambda}$ | | $\hat{\lambda}^{[\infty]}$ | |
| $k$ | $\gamma$ | $\rho_a$ | $\rho_y$ | $MC_p$ | $C_p$ | $MC_p$ | $C_p$ | $MC_p$ | $C_p$ | $MC_p$ | $C_p$ | $MC_p$ |
| 5 | 0.05 | 0.5 | 0.5 | 37 | 18 | 14 | 16 | 13 | 23 | 20 | 14 | 11 |
| | | | 0.99 | 83 | 65 | 63 | 54 | 62 | 70 | 68 | 61 | 59 |
| | | 0.99 | 0.5 | 38 | 18 | 15 | 16 | 14 | 23 | 20 | 14 | 11 |
| | | | 0.99 | 75 | 57 | 54 | 55 | 53 | 61 | 59 | 53 | 50 |
| | 0.25 | 0.5 | 0.5 | 70 | 44 | 43 | 40 | 38 | 49 | 47 | 26 | 24 |
| | | | 0.99 | 90 | 66 | 65 | 62 | 60 | 70 | 69 | 49 | 48 |
| | | 0.99 | 0.5 | 70 | 44 | 43 | 40 | 38 | 49 | 48 | 26 | 25 |
| | | | 0.99 | 86 | 62 | 60 | 57 | 56 | 66 | 65 | 46 | 43 |
| 10 | 0.05 | 0.5 | 0.5 | 25 | 16 | 12 | 15 | 11 | 19 | 15 | 14 | 10 |
| | | | 0.99 | 62 | 56 | 51 | 55 | 50 | 58 | 53 | 53 | 46 |
| | | 0.99 | 0.5 | 24 | 15 | 11 | 14 | 10 | 18 | 14 | 13 | 9 |
| | | | 0.99 | 56 | 52 | 47 | 51 | 46 | 54 | 49 | 50 | 45 |
| | 0.25 | 0.5 | 0.5 | 53 | 37 | 34 | 33 | 30 | 40 | 37 | 25 | 22 |
| | | | 0.99 | 75 | 60 | 56 | 55 | 51 | 63 | 59 | 48 | 44 |
| | | 0.99 | 0.5 | 53 | 37 | 34 | 33 | 30 | 40 | 37 | 26 | 23 |

- $\hat{\lambda}^{[\infty]}$ (PMSE$[\hat{\boldsymbol{Y}}(\lambda, \theta)]$ を最小にする $\lambda$ を更新した収束先) が最適



## まとめ

モデル 多変量線形回帰モデル

問題点 従来の推定量は多重共線性・外れ値に弱い
→ 多重共線性 ⇐ (多変量) Ridge回帰による推定量
外れ値 ⇐ 新たなパラメータ ($\lambda$) を導入

最適化 予測平均二乗誤差 (PMSE) を基に二通り;
① PMSE の推定量を作って最小化
② PMSE を最小にする値の未知の項に代入など

結果 新たに導入したパラメータ ($\lambda$) の最適な値が陽に求まる
- LSE より Ridge回帰による推定より提案した推定量が良い (⇔ PMSE を小さくできる)
  - 理論的にも数値実験的にも
- 数値実験では ② の更新の収束先が最良



## 参考文献


1. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.
2. Jimichi, M. (2016). *Shrinkaga Regression Estimators and Their Feasibilities*, Kwansei Gakuin University Press.
3. Nagai, I., Yanagihara, H. and Satoh, K. (2012). Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods. *Hiroshima Math. J.*, **42**, 301–324.
4. Yanagihara, H. and Satoh, K. (2010). A unbiased $C_p$ criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**, 1226–1238.
5. Yanagihara, H., Nagai, I. and Satoh, K. (2009). A bias-corrected $C_p$ criterion for optimizing ridge parameters in multivariate generalized ridge regression. *Japanese J. Appl. Statist.*, **38**, 151–172 (in Japanese).




終

# Support vector regression with penalized likelihood

千葉大・融合理工学府　上本 拓弥
千葉大・理学研究院　内藤 貫太

**はじめに：**目的変数 $Y \in \mathbb{R}$，$p$ 次元の説明変数 $\underline{X} \in \mathbb{R}^p$ のデータ $(Y_1, \underline{X}_1), \ldots, (Y_n, \underline{X}_n)$ について，$\underline{X}$ を用いて $Y$ を予測するサポートベクター回帰を考える．モデルは，パラメータ $\underline{\eta}=\begin{pmatrix} b & \underline{\beta}^T \end{pmatrix}^T \in I_b \times B \subset \mathbb{R} \times \mathbb{R}^p$ によって規定される線形関数 $\underline{x} \mapsto \underline{\eta}^T \underline{\tilde{x}}$ である．ただし，$\underline{\tilde{x}} = \begin{pmatrix} 1 & \underline{x}^T \end{pmatrix}^T$ である．サポートベクター回帰とは，損失関数として $\varepsilon$-insensitive loss $|z|_\varepsilon = \max\{0, |z| - \varepsilon\}$，$z \in \mathbb{R}$，$\varepsilon > 0$ を用いて，

$$C \sum_{i=1}^{n} |Y_i - \underline{\eta}^T \underline{\tilde{X}}_i|_\varepsilon + \frac{1}{2} ||\underline{\beta}||^2 \tag{1}$$

の最小化によりパラメータ $\underline{\beta}$，$b$ を推定する手法である (詳細は [1]，[2]，[3] を参照)．ただし，$C > 0$ は損失項と罰則項を調整するパラメータ，$||\cdot||$ は通常のユークリッドノルムである．ここで，パラメータ $\varepsilon$ と $C$ は学習する前に与えるパラメータであり，この $\varepsilon$ と $C$ に学習の精度が依存することが知られており，その与え方は重要である．本発表では，目的変数がラプラス分布から連想される確率密度関数に従うと仮定し，罰則付き尤度を導入することにより $C$ の推定を $\varepsilon$ の推定に集約させ，回帰係数と共に $\varepsilon$ をデータから推定する手法を提案した．

**罰則付き尤度の導入：**データ $(Y_1, \underline{X}_1), \ldots, (Y_n, \underline{X}_n)$ が独立で同一の分布 $f(y, \underline{x}) = p(y|\underline{x})q(\underline{x})$ に従うとする．ただし，$f$ は $(Y, \underline{X})$ の同時確率密度関数，$p(\cdot|\underline{x})$ は $\underline{X} = \underline{x}$ を与えた下での $Y$ の条件付き確率密度関数，$q$ は $\underline{X}$ の確率密度関数である．本発表では，$\underline{X} = \underline{x}$ を与えた下での $Y$ の条件付き確率密度関数が

$$p(y_i|\underline{x}_i) = p_{\underline{\theta}_0}(y_i|\underline{x}_i) = \frac{1}{4\varepsilon_0} \exp\left\{ -\frac{1}{\varepsilon_0} \left| y_i - \underline{\eta}_0^T \underline{\tilde{x}}_i \right|_{\varepsilon_0} \right\} \tag{2}$$

であると仮定する．ここで，$\underline{X}_i = \underline{x}_i$ を与えた下での条件付き期待値と条件付き分散はそれぞれ $E[Y_i|\underline{X}_i = \underline{x}_i] = \underline{\eta}_0^T \underline{x}_i$，$Var[Y_i|\underline{X}_i = \underline{x}_i] = 8\varepsilon_0^2/3$ となる．罰則付き最尤法によりパラメータ $\begin{pmatrix} \underline{\eta}_0^T & \varepsilon_0 \end{pmatrix}^T = \underline{\theta}_0 \in \Theta \subset \mathbb{R}^{p+1} \times \mathbb{R}$ の推定を行う．データを $(y_i, \underline{x}_i)$，$i = 1, \ldots, n$ と表すと，罰則付きの負の対数尤度は

$$-\log \prod_{i=1}^{n} f\left(y_i, \underline{x}_i\right) + \frac{1}{2} \left\| \underline{\beta} \right\|^2 = n\ell_n(\underline{\theta}) + \text{Constant}$$

となる．ただし，

$$\ell_n(\underline{\theta}) = \log\varepsilon + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\varepsilon}\left|y_i - \underline{\eta}^T\underline{\tilde{x}}_i\right|_\varepsilon + \frac{1}{2n}\left\|\underline{\beta}\right\|^2$$

である．よって，考える最小化問題は

$$\min_{\underline{\theta}\in\Theta}\quad\left[n\log\varepsilon + \frac{1}{\varepsilon}\sum_{i=1}^{n}\left|y_i - \underline{\eta}^T\underline{\tilde{x}}_i\right|_\varepsilon + \frac{1}{2}\left\|\underline{\beta}\right\|^2\right] \tag{3}$$

となる．このように，条件付き密度 (2) の導入により，(1) における学習で必要なパラメータ $C$, $\varepsilon$ を尤度の観点からデータに基づき “自動的に” 求めることが可能となる．よって，$\underline{\theta}_0$ の推定量 $\hat{\underline{\theta}}$ は

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}\in\Theta}\ \ell_n(\underline{\theta}) \tag{4}$$

として構築される．また，

$$\ell(\underline{\theta}) = \mathbb{E}\left[\log\varepsilon + \frac{1}{\varepsilon}\left|Y - \underline{\eta}^T\underline{\tilde{X}}\right|_\varepsilon\right]$$

とすると，

$$\underline{\theta}_0 = \arg\min_{\underline{\theta}\in\Theta}\ \ell(\underline{\theta})$$

と特徴づけられる．

**推定量の計算方法と漸近的性質：**最適化問題 (3) の解を実際に求める計算方法と，(4) で構築される推定量 $\hat{\underline{\theta}}$ について，適当な仮定の下，漸近一致性と漸近正規性が成り立つことを報告した．さらに，それらを確認するシミュレーション結果や実データへの適用結果についても報告した．

**参考文献**


[1] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer, New York.

[2] Vapnik, V. N. (1998). *Statistical Learning Theory*. John Wiley & Sons, New York.

[3] Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. 2nd Edition. Springer, New York.

# 一般化線形モデルにおける
# 擬 Huber 関数によるロバスト推測

千葉大・融合理工学府　三次 琢巳
千葉大・理学研究院　内藤 貫太

**はじめに**：一般化線形モデルにおいて，擬 Huber 関数を用いたパラメータのロバストな推定方法について考察を与える．一般化線形モデルとは，線形回帰モデルにおける目的変数の分布を正規分布から指数型分布族に拡張した回帰モデルであり，正規分布以外にも二項分布，ポアソン分布など多くの代表的な分布を仮定することが可能となる回帰モデルである．ここで，用いるデータに外れ値が含まれている場合，一般化線形モデルにおける尤度に基づく方法で推定されたパラメータは，外れ値による影響を強く受けてしまうことが知られている．そこで一般化線形モデルにおけるパラメータのロバストな推定方法として，データに対して Huber 関数を用いた重みづけを行うことで，外れ値からの影響を少なくする推定方法が Cantoni and Ronchetti (2001) によって提案されている．外れ値の影響を緩和していた Huber 関数の代わりに，Huber 関数を微分可能な関数で近似した擬 Huber 関数を用いることで，Huber 関数を用いた推定方法と比較し外れ値による影響をどれほど緩和できているか調べた結果について報告した．

**一般化線形モデル**：目的変数 $Y_i \in \mathbb{R}$ と説明変数 $\boldsymbol{x}_i = (x_1 \ \cdots \ x_p)^T \in \mathbb{R}^p$ のデータ $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ に対して，一般化線形モデルを考える．一般化線形モデルとは，目的変数が指数型分布族に属し，期待値構造が，パラメータ $\boldsymbol{\beta} \in \mathbb{R}^p$ とリンク関数 $g(\cdot)$ を用いて $g(E[Y]) = g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$ と仮定するモデルである ([1] 参照).

**擬似尤度推定**：擬似尤度推定とは，一般化線形モデルにおける最尤推定を拡張した推定方法である．擬似尤度推定では，一般化線形モデルと同様の期待値構造を仮定するが，分布について指数型分布族を仮定せず，目的変数 $Y$ の期待値と分散の構造を $E[Y] = \mu$, $Var[Y] = \mathrm{V}(\mu)$ と仮定するモデルである ([1] 参照).

**ロバスト推定**：一般化線形モデルにおける尤度に基づく方法で推定されたパラメータは，解析するデータに外れ値が含まれている場合，外れ値による影響を強く受けてしまうことが知られている．このような外れ値の影響を緩和するため，[2] では，Huber 関数を用いたロバストな推定方法が提案されている．ここで Huber 関数はチューニングパラメータ

$c > 0$ に対し

$$\psi_c^{\mathrm{H}}(x) = \begin{cases} x & , |x| \leq c, \\ c\ \mathrm{sign}(x) & , |x| > c \end{cases}$$

と定義される．導かれる推定方程式は

$$U(\boldsymbol{\beta})^{\mathrm{H}} = \sum_{i=1}^{n} \psi_c^{\mathrm{H}} \left( \frac{y_i - \mu_i}{\sqrt{\mathrm{V}(\mu_i)}} \right) \frac{w(\boldsymbol{x}_i)}{\sqrt{\mathrm{V}(\mu_i)}} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

と与えられる ([2] 参照)．

**擬 Huber 関数**：擬 Huber 関数とは，Huber 関数を微分可能な関数で近似した関数であり，Huber 関数と同様にチューニングパラメータ $c > 0$ に対して

$$\psi_c^{\mathrm{pH}}(x) = \frac{cx}{\sqrt{c^2 + x^2}}$$

と定義される ([3] 参照)．

**擬 Huber 関数による推定アルゴリズム**：画像解析や信号処理では，Huber 関数の代わりに擬 Huber 関数を用いられることがある．一方，擬似尤度に基づく統計推測において，Huber 関数を擬 Huber 関数に置き換えた方法については議論が見当たらなかった．擬 Huber 関数を用いた場合，推定方程式は

$$U(\boldsymbol{\beta})^{\mathrm{pH}} = \sum_{i=1}^{n} \psi_c^{\mathrm{pH}} \left( \frac{y_i - \mu_i}{\sqrt{\mathrm{V}(\mu_i)}} \right) \frac{w(\boldsymbol{x}_i)}{\sqrt{\mathrm{V}(\mu_i)}} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

と与えられる．

**比較**：Huber 関数を用いた推定方法と，擬 Huber 関数を用いた推定方法を，有効性の観点から理論比較した．その後シミュレーションを用いて推定量の精度や，ロバスト性の比較を行い，実データに適用した結果について報告を行った．

# 参考文献


[1] 汪金芳 (2016). 一般化線形モデル．朝倉書店．

[2] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022-1030.

[3] Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*, Cambridge University Press.

# Regression with localized functional Bregman divergence

Kanta NAITO
Department of Mathematics and Informatics, Chiba University, Japan

**Setting:** Let $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n) \sim_{i.i.d.} f(y, \boldsymbol{x}) = p(y|\boldsymbol{x})q(\boldsymbol{x})$, where $(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^d$, $f$ is the joint density of $(Y, \boldsymbol{X})$, $p$ is the conditional density of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$, and $q$ is the density of $\boldsymbol{X}$. Let $\boldsymbol{t} \in \mathbb{R}^d$ be a target point at which we want to estimate the value of regression function $\mu(\boldsymbol{t}) = E[Y|\boldsymbol{X} = \boldsymbol{t}]$. Our parametric model for $\mu$ is having the form

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = G^{-1}\left(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}\right),$$

where $\widetilde{\boldsymbol{x}}^T = [1\ \boldsymbol{x}^T] \in \mathbb{R}^{d+1}$ is the vector of explanatory variables, $\boldsymbol{\theta} = [\theta_0\ \theta_1\ \cdots\ \theta_d]^T \in \Theta \subset \mathbb{R}^{d+1}$ is the parameter vector, and $G$ is the link function.

**The functional Bregman divergence:** Fix a strictly convex function $U$. Then the discrepancy between $\mu(\cdot)$ and its parametric model $m(\cdot, \boldsymbol{\theta}) = m_\theta(\cdot)$ can be measured by the functional Bregman divergence defined as

$$\begin{aligned} D_{U^*}(u(m_\theta), u(\mu)) &= \int_{\mathbb{R}\times\mathbb{R}^d} \left[U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta}))) - y \cdot u(m(\boldsymbol{x}, \boldsymbol{\theta}))\right] f(y, \boldsymbol{x}) dy d\boldsymbol{x} \\ &\quad + \int_{\mathbb{R}^d} \left[-U^*(u(\mu(\boldsymbol{x}))) + \mu(\boldsymbol{x}) u(\mu(\boldsymbol{x}))\right] q(\boldsymbol{x}) d\boldsymbol{x}, \end{aligned} \tag{1}$$

where $u = U'$: the derivative of $U$, $U^*$ is the convex conjugate of $U$: $U^*(s) = \sup_{z\in\mathbb{R}}\{zs - U(z)\}$. The usual parametric regression can be carried out by using a certain estimator $\hat{\boldsymbol{\theta}}$ of the true value of $\boldsymbol{\theta}$. Necessary tools for this estimation scheme are

$$\rho(y, \boldsymbol{x}, \boldsymbol{\theta}) = U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta}))) - y \cdot u(m(\boldsymbol{x}, \boldsymbol{\theta})), \tag{2}$$

$$\psi(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) = \{m(\boldsymbol{x}, \boldsymbol{\theta}) - y\} \frac{u'(m(\boldsymbol{x}, \boldsymbol{\theta}))}{G'(m(\boldsymbol{x}, \boldsymbol{\theta}))} \widetilde{\boldsymbol{x}}, \tag{3}$$

by which the estimator $\hat{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}_*$ of $\boldsymbol{\theta}$ can be defined as

$$\boldsymbol{\theta}_* = arg\min_{\theta\in\Theta} \int_{\mathbb{R}\times\mathbb{R}^d} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x}), \tag{4}$$

$$\hat{\boldsymbol{\theta}} = arg\min_{\theta\in\Theta} \int_{\mathbb{R}\times\mathbb{R}^d} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) dF_n(y, \boldsymbol{x}), \tag{5}$$

where $F_n$ is the empirical distribution function based on $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$, $F$ is the cumulative distribution function with its density $f$. Note that $\boldsymbol{\theta}_*$ in (4) is the minimizer of (1), and the minimizer of the empirical version of (1) is nothing other than $\hat{\boldsymbol{\theta}}$ in (5). The regression function estimator can be obtained by plugging $\hat{\boldsymbol{\theta}}$ into $\boldsymbol{\theta}$ in $m(\cdot, \boldsymbol{\theta})$:

$$\hat{\mu}_G(\boldsymbol{x}) = m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}). \tag{6}$$

**The localized functional Bregman divergence:** We organize a scheme of estimation of $\boldsymbol{\theta}$ depending on $\boldsymbol{t}$ locally. Necessary functions are listed as follows:

$$\rho(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) = K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h}\right)\rho(y, \boldsymbol{x}, \boldsymbol{\theta}), \tag{7}$$

$$\psi(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\rho(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}). \tag{8}$$

Here $K(\boldsymbol{z})$ is a smooth unimodal integrable function symmetric around $\boldsymbol{z} = \mathbf{0}_d$ satisfying $K(\mathbf{0}_d) = 1$, and $h > 0$ is the scalar bandwidth which controlls the degree of localization, here $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$. We notice that (7) and (8) are localized version of (2) and (3) respectively, with the use of the kernel $K$. Using these functions, we define the true parameter $\boldsymbol{\theta}_*(\boldsymbol{t})$ at $\boldsymbol{t}$ and its estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ as follows:

$$\begin{aligned}\boldsymbol{\theta}_*(\boldsymbol{t}) &= arg\min_{\theta\in\Theta}\int_{\mathbb{R}\times\mathbb{R}^d}\rho(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta})dF(y, \boldsymbol{x}),\\ \hat{\boldsymbol{\theta}}(\boldsymbol{t}) &= arg\min_{\theta\in\Theta}\int_{\mathbb{R}\times\mathbb{R}^d}\rho(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta})dF_n(y, \boldsymbol{x}).\end{aligned}$$

This local estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ of $\boldsymbol{\theta}_*(\boldsymbol{t})$ also suggests us to make a regression estimator defined as

$$\hat{\mu}_L(\boldsymbol{x}) = m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{9}$$

which we call the *local estimator* of $\mu(\boldsymbol{x})$, because the involved estimator of parameter is determined locally. Since $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ can vary depending on $\boldsymbol{x}$, $\hat{\mu}_L$ would be expected to be more flexible than $\hat{\mu}_G$. On the other hand, we call the estimator $\hat{\mu}_G(\boldsymbol{x})$ in (6) the *global estimator* of $\mu(\boldsymbol{x})$.

The following topics were reported at the symposium:

1. Asymptotic evaluation of the risk difference between the global estimator $\hat{\mu}_G$ in (6) and the local estimator $\hat{\mu}_L$ in (9), under the situation both $n \to \infty$ and $h \to \infty$.

2. The local estimator $\hat{\mu}_L$ in (9) asymptotically improves the risk of the global estimator $\hat{\mu}_G$ in (6), provided that the link function $G$ equals to $u = U'$.

3. A robustfing the above methodology by utilizing an another feature of the functional Bregman divergence.

4. The localization can be applied also in this robust setting, and the the risk difference between the global estimator $\hat{\mu}_G$ and the local estimator $\hat{\mu}_L$ can asymptotically be evaluated.

5. Under the use of the pseudo-Huber function as $U$, the local estimator $\hat{\mu}_L$ improves the risk of the global estimator $\hat{\mu}_G$, as the parameter $\delta$ involved in the pseudo-Huber function getting large.

6. Some simple numerical illustrations which confirm theoretical results as well as the practical performance of estimators.

# 再生核ヒルベルト空間における Maximum Variance Discrepancy の実際的挙動

千葉大・融合理工学府　牧草 夏実

## 1　はじめに

$P, Q$ をヒルベルト空間 $\mathcal{H}$ 上の確率分布とするとき，二標本 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} Q$ に基づく検定

$$\text{帰無仮説 } H_0 : P = Q \quad \text{vs.} \quad \text{対立仮説 } H_1 : P \neq Q$$

を考える．ユークリッド空間での二標本検定はすでに様々な検定方法が議論されているが，ヒルベルト空間に値をとる確率変数に対する二標本検定を考えることで，高次元データに対する二標本検定の議論を与える．高次元データに対するアプローチとして，Maximum Mean Discrepancy(MMD) に基づく二標本検定が [1] によりすでに議論されているが，この MMD と同様の考え方により，Maximum Variance Discrepancy(MVD) と呼ばれる新たな分布間の違いを測る指標を提案し，その検定について考える．

## 2　再生核ヒルベルト空間でのモーメントの定義

確率変数 $X \sim P$, $Y \sim Q$ を正定値カーネル $k$ によって，この $k$ に対応する再生核ヒルベルト空間 $H(k)$ 上に，それぞれ $k(\cdot, X)$，$k(\cdot, Y)$ により変換を行う．このとき，この $k(\cdot, X)$，$k(\cdot, Y)$ の分散 $\Sigma_k(P)$，$\Sigma_k(Q)$ は，それぞれヒルベルト空間 $H(k)^{\otimes 2} = H(k) \otimes H(k)$ での期待値 $\Sigma_k(P) = \mathbb{E}_{X\sim P}[(k(\cdot, X) - \mu_k(P))^{\otimes 2}]$，$\Sigma_k(Q) = \mathbb{E}_{Y\sim Q}[(k(\cdot, Y) - \mu_k(Q))^{\otimes 2}]$ によって定められている．ここで，$\mu_k(P)$，$\mu_k(Q)$ は $k(\cdot, X)$ の期待値 $\mu_k(P) = \mathbb{E}_{X\sim P}[k(\cdot, X)]$，$\mu_k(Q) = \mathbb{E}_{Y\sim Q}[k(\cdot, Y)]$ であり，$\otimes$ はテンソル積を表しており，任意の $f \in H(k)$ に対し，$f^{\otimes 2} = f \otimes f = \langle f, \cdot \rangle_{H(k)} f$ である．

## 3　検定統計量の構築

この $k(\cdot, X)$ と $k(\cdot, Y)$ の期待値の差

$$\sup_{||f||_{H(k)}=1} | \langle f, \mu_k(P) - \mu_k(Q) \rangle_{H(k)} | = ||\mu_k(P) - \mu_k(Q)||_{H(k)}$$

により，2 つの分布の違いを測るのが，Maximum Mean Discrepancy (MMD) と呼ばれるものである．同様の考え方により，$k(\cdot, X)$ と $k(\cdot, Y)$ の分散の差

$$\sup_{||A||_{H(k)^{\otimes 2}}=1} | \langle A, \Sigma_k(P) - \Sigma_k(Q) \rangle_{H(k)^{\otimes 2}} | = ||\Sigma_k(P) - \Sigma_k(Q)||_{H(k)^{\otimes 2}}$$

により 2 つの分布の違いを測る．この違い $\|\Sigma_k(P) - \Sigma_k(Q)\|^2_{H(k)^{\otimes 2}}$ は

$$\widehat{T}^2_{n,m} = \left|\left|\widehat{\Sigma}_k(P) - \widehat{\Sigma}_k(Q)\right|\right|^2_{H(k)^{\otimes 2}}$$

によって推定することができる．ただし，

$$\widehat{\Sigma}_k(P) = \frac{1}{n}\sum_{i=1}^{n}(k(\cdot, X_i) - \widehat{\mu}_k(P))^{\otimes 2}, \quad \widehat{\mu}_k(P) = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i),$$

$$\widehat{\Sigma}_k(Q) = \frac{1}{m}\sum_{i=1}^{m}(k(\cdot, Y_i) - \widehat{\mu}_k(Q))^{\otimes 2}, \quad \widehat{\mu}_k(Q) = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, Y_i)$$

である．

本発表では，この検定統計量 $\widehat{T}^2_{n,m}$ の漸近挙動と実際的側面について報告を行った．特に，帰無仮説 $H_0 : P = Q$ のもとで，退化 $V$ 統計量の結果に帰着させることで，$\widehat{T}^2_{n,m}$ の漸近分布が，独立な自由度 1 の $\chi^2$ 分布の重み付き無限和の形で得られること ([4] 参照)，その重みをデータに基づいて推定する方法について報告を行った．また，実際的側面として，推定した重み $\widehat{\lambda}_\ell$ に基づく近似分布 $\sum_{\ell=1}^{n-1} \widehat{\lambda}_\ell Z^2_\ell$ は帰無仮説のもとでの $(n+m)\widehat{T}^2_{n,m}$ の分布と比べて大きな分散を持っており，MMD でも同様の結果となっていることについて述べた．そして，MVD と MMD の帰無分布の分散をそれぞれ計算することによって，近似分布を修正し，その結果を報告した．また，この修正した近似分布を用いて，Type I error の確率と検出力に関するシミュレーションを行い，Type I error の確率が有意水準に近づいていくことと，MVD を用いた検定が MMD を用いた検定よりも検出力が大きいことについて報告を行った．

## 参考文献


[1] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B. and Smola, A. (2007). A kernel method for the two sample problem. *Advances in Neural Information Procesing Systems* **19**, *MIT Press*, 513–520.

[2] Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test, *Advances in Neural Information Processing Systems* **22**, *Curran Associates Inc.*, 673–681.

[3] Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer, New York.

[4] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

# 科研費シンポジウム「統計科学の革新にむけて」講演報告書
# 講演タイトル”Density Estimation via Stagewise Algorithm with a Dictionary Having Various Bandwidths”

## 1　講演情報


- 講演者　西田　喜平次

  兵庫医療大学 共通教育センター 講師

  〒 650-8530 兵庫県神戸市中央区港島 1-3-6, kiheiji.nishida@gmail.com
- 共同研究者　内藤 貫太

  千葉大学 大学院理学研究院 教授

  〒 263-8522 千葉県千葉市稲毛区弥生町 1-33, naito@math.s.chiba-u.ac.jp

- 日時 2021 年 1 月 23 日 16:40-17:20 (発表はオンライン)

## 2　講演内容

### 2.1　要旨


カーネル型密度関数推定法は, 実行に際してバンド幅の推定が必要となるが, 多次元の設定で洗練された推定方法として, Duong and Hazelton(2003, J. Nonpram. Stat.) の Direct Plug-in (DPI) 法が知られている. この方法は, バンド幅行列の非対角成分が非ゼロ値という, 強い設定下で実行されるもので, R にも実装されている方法である. 他方, Reduced Set Density Estimation(RSDE) 法 (Girolami and He 2003, IEEE PAMI) は, データ点毎に設定された重みパラメータ (=weight) の一部がゼロ値を取る事を許容の下で, バンド幅と weight の最適化を行う方法である. この方法では, weight がゼロとなるデータ点は使用せずに密度推定を行うことを意味するため, 密度推定の sparse 表現を可能にする方法である. 本講演は, 機械学習のアイデアを取り入れた, 逐次最小化アルゴリズム (以下, Stagewise Minimization Algorithm) を用いたカーネル型密度関数推定法の概要と理論的性質と, シミュレーションによる DPI および RSDE との効率性比較を紹介した.

## 2.2 提案アルゴリズムの概要

事前にアルゴリズムの最終ステージ数を決定する. そして, 簡便なバンド幅を持つ複数のカーネルから構成される「辞書」を定義する. アルゴリズムの各ステージでは, 前ステージで得られた推定量と, ある関数変換を施した上での「ワード」との凸結合を, すべてのワードについて構成し, それらの中で評価関数 (ここでは U-divergence 関数) が最適となるものを, 当該ステージの推定量と定義する. こうした再帰的操作を最終ステージまで実行し, 精度を逐次的に高める推定方法である.

特に辞書の構成に関して, 以下の特徴がある.

- できるだけ簡便な「辞書」を用いて, DPI や RSDE と対抗しうる推定量を構築する目的の下, 「バンド幅の辞書」はスカラー型バンド幅で構成した.
- 推定量の非漸近誤差限界を導出する目的で, 得られたデータの一部を辞書の構築に使用し (カーネルの平均として使用), 残りのデータを推定量の評価で使用するという設定で, アルゴリズムを構築した.

## 2.3 報告内容の細部

- 提案手法による推定量の非漸近誤差限界を定理として紹介した.
- できるだけ多様なデータに適用可能な, スカラー型バンド幅で構成される辞書の一例を紹介した.
- 提案手法による推定量の挙動をシミュレートし, 結果を紹介した.
  - MISE の観点から, 提案手法は DPI, および RSDE よりも優位になる場合があることを確認した.
  - 提案手法による推定は, RSDE と同様, 密度推定の sparse 表現を可能にするが, 密度関数の尾根にそったデータが選ばれる特徴があることを確認した.
  - 辞書に用いるデータ数と推定量の評価に用いるデータ数の配分比が, どのように推定結果へ影響するか調べた. 辞書に用いるデータ数の配分比を小さくしたほうが, MISE は改善される結果を得ている.
- 提案手法を実データ (UCI データリポジトリの Abalone データセット) に適用した推定結果を紹介した. 結果は概ね良好と判断した.