高次元複雑データの統計モデリング
報告書

●基盤研究(A)15H01678「大規模複雑データの理論と方法論の総合的研究」
研究代表者:青嶋誠(筑波大学)

●開催責任者:廣瀬慧

●日時:8 月 29 日 ～ 8 月 30 日

●場所:九州大学伊都キャンパス ウエスト 1 号館 C501 大講義室

●内容・目的:

近年, 様々な分野で高次元かつ複雑な様相を呈するデータが取得されています. そのようなデータに対して適切に解析を行うためには, 複雑な構造を捉える統計モデリングおよびその理論・計算アルゴリズムの構築が必要となります. そこで, 高次元複雑データに対する統計解析の最新の研究をお知らせし, 今後の統計モデリングの方向性について考えていくため, 科研費シンポジウム「高次元複雑データの統計モデリング」を開催します.

科研費シンポジウム「高次元複雑データの統計モデリング」
講演プログラム

8月29日（木）

時間	名前	所属	タイトル	座長
10：00〜10：20	オープニング (青嶋誠, 廣瀬慧)			
10：20〜11：00	矢田 和善 石井　晶 青嶋　誠	筑波大学 東京理科大学 筑波大学	Geometrical quadratic discriminant analysis for high-dimension, strongly spiked eigenvalue models	廣瀬 慧
11：00〜11：40	新村 秀一	成蹊大学	高次元遺伝子解析の呪いからの解放1　-統計が1970年からこの問題を解決できなかった理由-	
11：40〜13：00	休憩			
13：00〜13：40	川口 淳 田尻 涼	佐賀大学	脳画像における高次元データ解析法	二宮 嘉行
13：40〜14：20	朴 喜椄	広島大学	Sparse Tensor subspace method for biological modulators selection	
14：20〜14：40	休憩			
14：40〜15：20	二宮 嘉行	統計数理研究所	因果推論におけるセミパラメトリックアプローチのための情報量規準	矢田 和善
15：20〜16：00	作村 建紀 柳本 武美	法政大学 統計数理研究所	方向データに対する von Mises 分布の母数推定について	
16：00〜16：30	廣瀬 雅代	九州大学	A Practicable Estimation of Mean Squared Prediction Error in Small Area Estimation	
16：30〜17：00	総合討論			
18：30〜20：30	懇親会@大名つつじ庵			

科研費シンポジウム「高次元複雑データの統計モデリング」
講演プログラム

8月30日（金）

	名前	所属	タイトル	座長
10：20〜11：00	野村 俊一	統計数理研究所	前震の統計的判別に基づく地震予測	高井 啓二
11：00〜11：40	鈴木 拓海	東京大学	Penalized least squares approximation methods	
11：40〜13：00	休憩			
13：00〜13：40	高井 啓二	関西大学	欠測データを用いたフィッシャースコアリング法	廣瀬 雅代
13：40〜14：20	森川 耕輔	大阪大学	Bayesian Fractional Imputation With Nonignorable Nonresponse Data	
14：20〜14：40	休憩			
14：40〜15：20	和田 裕一郎	名古屋大学	スペクトラル埋め込みを利用した深層クラスタリング	増田 弘毅
15：20〜16：00	小林 景	慶應義塾大学	距離空間の変形と埋め込みを用いたデータ解析	
16：00〜16：20	クロージング			

Geometrical quadratic discriminant analysis for high-dimension, strongly spiked eigenvalue models

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

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

1 Introduction

We consider a classifier for high-dimensional data under the strongly spiked eigenvalue (SSE) model. We create a new classification procedure on the basis of the high-dimensional eigenstructure. We propose a quadratic classification procedure by using a data transformation. Suppose we have two classes $\pi_i,\ i=1,2$, and define independent $p \times n_i$ data matrices, $\boldsymbol{X}_i = [\boldsymbol{x}_{i1},...,\boldsymbol{x}_{in_i}],\ i=1,2$, from $\pi_i,\ i=1,2$, where $\boldsymbol{x}_{ij},\ j=1,...,n_i$, are independent and identically distributed (i.i.d.) as a p-dimensional distribution with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i\ (\geq \boldsymbol{O})$. We assume that $\limsup_{p\to\infty} ||\boldsymbol{\mu}_i||^2/p < \infty$ for $i=1,2$, where $||\cdot||$ denotes the Euclidean norm. Also, we assume that $\mathrm{tr}(\boldsymbol{\Sigma}_i)/p \in (0,\infty)$ as $p\to\infty$ for $i=1,2$. Here, for a function, $f(\cdot)$, "$f(p)\in(0,\infty)$ as $p\to\infty$" implies $\liminf_{p\to\infty} f(p) > 0$ and $\limsup_{p\to\infty} f(p) < \infty$. We assume $n_i \geq 3,\ i=1,2$. The eigen-decomposition of $\boldsymbol{\Sigma}_i$ is given by $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i\boldsymbol{\Lambda}_i\boldsymbol{H}_i^T = \sum_{s=1}^{p}\lambda_{s(i)}\boldsymbol{h}_{s(i)}\boldsymbol{h}_{s(i)}^T$, where $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{1(i)},...,\lambda_{p(i)})$ having $\lambda_{1(i)} \geq \cdots \geq \lambda_{p(i)}(\geq 0)$ and $\boldsymbol{H}_i = [\boldsymbol{h}_{1(i)},...,\boldsymbol{h}_{p(i)}]$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{X}_i - [\boldsymbol{\mu}_i,...,\boldsymbol{\mu}_i] = \boldsymbol{H}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{Z}_i$ for $i=1,2$. Then, $\boldsymbol{Z}_i$ is a $p\times n_i$ sphered data matrix from a distribution with the zero mean and identity covariance matrix. Let $\boldsymbol{Z}_i = [\boldsymbol{z}_{1(i)},...,\boldsymbol{z}_{p(i)}]^T$ and $\boldsymbol{z}_{j(i)} = (z_{j1(i)},...,z_{jn_i(i)})^T,\ j=1,...,p$, for $i=1,2$. We assume that the fourth moments of each variable in $\boldsymbol{Z}_i$ are uniformly bounded for $i=1,2$. Let $\boldsymbol{z}_{oj(i)} = \boldsymbol{z}_{j(i)} - (\bar{z}_{j(i)},...,\bar{z}_{j(i)})^T,\ j=1,...,p;\ i=1,2$, where $\bar{z}_{j(i)} = n_i^{-1}\sum_{k=1}^{n_i} z_{jk(i)}$. We also assume that $P\Big(\liminf_{p\to\infty}||\boldsymbol{z}_{o1(i)}|| \neq 0\Big) = 1$ for $i=1,2$. Let $\boldsymbol{x}_0$ be an observation vector of an individual belonging to $\pi_i\ (i=1,2)$. We assume $\boldsymbol{x}_0$ and $\boldsymbol{x}_{ij}$s are independent. We estimate $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ by $\bar{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i}\boldsymbol{x}_{ij}/n_i$ and $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_{in_i})^T/(n_i-1)$. As for the HDLSS data, Aoshima and Yata [1] considered distance-based classifiers and Aoshima and Yata [1] gave the misclassification rate adjusted classifier for multiclass, high-dimensional data whose misclassification rates are no more than specified thresholds under the following condition for eigenvalues:

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

Aoshima and Yata [3] called (1.1) the "non-strongly spiked eigenvalue (NSSE) model". Also, Aoshima and Yata [3] gave the following eigenvalue model called the "strongly spiked eigenvalue (SSE) model":

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

Aoshima and Yata [4] considered a liner classifier under the SSE model 1.2 by using a data transformation from the SSE model to the NSSE model. They gave a consistency property of the classifier and discussed the asymptotic normality when $p\to\infty$ and $n_i\to\infty\ (i=1,2)$. On the other hand, Ishii [5] proposed a linear classification procedure which has the consistency property even when n_i's are very small under the following eigenvalue condition:

$$\frac{\sum_{s=2}^{p}\lambda_{s(i)}^2}{\lambda_{1(i)}^2} = o(1) \text{ as } p\to\infty \text{ for } i=1,2. \tag{1.3}$$

Note that (1.3) implies the conditions that $\lambda_{2(i)}/\lambda_{1(i)}\to 0$ and $\lambda_{1(i)}^2/\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\to 1$ as $p\to\infty$.

2 New geometrical quadratic discriminant analysis under the SSE model

Aoshima and Yata ([2]) gave a geometrical quadratic discriminant analysis (GQDA) under the NSSE model (1.1). We consider a new GQDA for (1.3). We assume the following condition:

(A-i) $\boldsymbol{h}_{1(1)} = \boldsymbol{h}_{1(2)}$ $(= \boldsymbol{h}_1$, say) and $\lambda_{1(1)}/\lambda_{1(2)} \in (0, \infty)$ as $p \to \infty$.

Note that (A-i) is much milder than $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Aoshima and Yata [4] considered a distance-based classifier by using a data transformation from the SSE model to the NSSE model. They gave the consistency property for the classifier and discussed the asymptotic normality when $p \to \infty$ and $n_i \to \infty$ $(i = 1, 2)$. On the other hand, Ishii [5] gave a distance-based classifier by using the data transformation when $p \to \infty$ while n_i's are fixed. In this talk, we create a new quadratic classifier by using the data transformation. We construct the following new GQDA:

$$\begin{aligned}\widetilde{G}_{\text{DT}}(\boldsymbol{x}_0) =& p\frac{||\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_{1n_1}||^2 - \{(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_{1n_1})^T\tilde{\boldsymbol{h}}_{1(2)}\}^2}{\text{tr}(\boldsymbol{S}_{1n_1}) - \tilde{\lambda}_{1(1)}} \\ & - p\frac{||\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_{2n_2}||^2 - \{(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_{2n_2})^T\tilde{\boldsymbol{h}}_{1(1)}\}^2}{\text{tr}(\boldsymbol{S}_{2n_2}) - \tilde{\lambda}_{1(2)}} \\ & - \frac{p}{n_1} + \frac{p}{n_2} - p\log\left\{\frac{\text{tr}(\boldsymbol{S}_{2n_2}) - \tilde{\lambda}_{1(2)}}{\text{tr}(\boldsymbol{S}_{1n_1}) - \tilde{\lambda}_{1(1)}}\right\}. \end{aligned} \tag{2.1}$$

Then, one classifies $\boldsymbol{x}_0$ into π_1 if $\widetilde{G}_{\text{DT}}(\boldsymbol{x}_0) < 0$ and π_2 otherwise. Here, $\tilde{\lambda}_{1(i)}$'s and $\tilde{\boldsymbol{h}}_{1(i)}$'s are the noise reduction (NR) estimators of $\lambda_{1(i)}$'s and $\boldsymbol{h}_{1(i)}$'s.

Theorem 2.1. *Assume (A-i), (1.3) and some regularity conditions. For the classifier given by (2.1), we have that as* $p \to \infty$

$$e(1) \to 0 \text{ and } e(2) \to 0. \tag{2.2}$$

Here, $e(i)$ *denotes the error of misclassifying an individual from* π_i *into* π_j *for* $i, j = 1, 2$ *and* $i \neq j$.

We gave the performances of $\widetilde{G}_{\text{DT}}(\boldsymbol{x}_0)$ by simulation studies and real data examples.

References

[1] Aoshima, M., Yata, K.: A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. Ann. Inst. Statist. Math. **66**, 983-1010 (2014).

[2] Aoshima, M., Yata, K.: Geometric classifier for multiclass, high-dimensional data. Sequential Anal. (Special Issue: Celebrating Seventy Years of Charles Stein's 1945 Seminal Paper on Two-Stage Sampling) **34**, 279-294 (2015).

[3] Aoshima, M., Yata, K., 2018. Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Stat. Sin.* 28, 43-62.

[4] Aoshima, M., Yata, K.: Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models. Ann. Inst. Statist. Math. **71**, 473-503 (2019).

[5] Ishii, A.: A classifier under the strongly spiked eigenvalue model in high-dimension, low-sample-size context. Commun. Stat. Theory Methods. (2019), in press.

高次元遺伝子解析の呪いからの解放

成蹊大学名誉教授　　新村秀一

「高次元 Microarray データを用いて癌遺伝子の特定と癌の亜種を見つける研究」が 1970 年頃から行われてきた(Golub, 1999)。これらの研究で用いられたデータが公開されているので、統計に限らず機械学習（AI)、パターン認識、Bio 工学の新テーマとして研究されてきたが、いずれの研究も成功していない(判別分析の **Problem5**)。しかし、仮に症例数が n=100 で遺伝子数が p=10,000 の発現量データとすれば、2 群判別が最も適した手法である。筆者は 2015 年 10 月 28 日から 12 月 20 日の僅か 54 日間で、簡単にこの問題を解決した。用いたデータは、1999 年から 2004 年の間に米国の 6 医学研究プロジェクトが論文を発表し、研究に用いた公開データである。これらは癌と健常、あるいは 4 種の異なった癌の 2 クラスである。

結果は非常に単純である。6 種のデータは線形分離可能なデータ(Linearly Separable Data, **LSD**)である（判別分析の **Fact3**)。この重要な事実であり信号が、これまでの研究で誰も指摘していない。さらに、筆者が開発した Matryoshka Feature Selection Method **(Method2)**で簡単に線形分離可能な含まれる遺伝子数が n 個以下の遺伝子の k 組の部分空間(SMall Matryoshka, **SM**)と最小誤分類数(Minimum Number of MisclassificationSM**NM**)が 1 以上の遺伝子の雑音部分空間に分割できた(**Fact4**)。各 SM は統計分析が容易な小標本であるが、ロジスティック回帰以外の統計手法（一元配置の分散分析、 t 検定、相関分析、クラスター分析、PCA）で線形分離可能な事実が示されなかった(**Problem6**)。そこで、MNM 基準による改定 IP-OLDF（**RIP**）の判別スコア(RIP Discriminant Score, **RipDS**)を変数とし、n 症例*k 次元の**信号データ**(k<=n)を作成した。これを上記の統計手法で分析し「癌の遺伝子診断の統計分析法を世界で初めて提案」できた。今年 5 月に Springer から **Springer2** を出版した。

以上の研究が簡単にできたのは、大学卒業以来行ってきた判別分析の新理論(**Springer1**）が 2015 年に完成し、新理論がその応用問題として 1970 年頃から未解決の Microarray を用いた癌の遺伝子解析（**Problem5**）を簡単に解決できた。本来であれば癌の遺伝子研究の専門家でない筆者が「癌の遺伝子診断」までを行うことは適していない。しかし、癌は遺伝子の病気であり、高次元の Microarray 空間で 2 群が完全に分かれていて、さらに MNM ＝ 0 である k 組の SM に分割できる。そして RipDS で信号データを作ることで、上記の統計手法で線形分離可能であり有効と考えられる Malignancy Indexes が数多く発見できた。しかし、これ等のどの Malignancy Indexes が医学的に役に立つか否かは医学専門家の検証が必要である。

残念ながら Golub らの研究後に、「**NIH が乳がん以外の癌に関して** Microarray による**研究は成果が出ないと判断し、医学研究が終わったようである**」。このため、いかに医学専門家の検証につなげるかを 2016 年から模索している。しかし統計や工学研究者は、NIH の報告を知らずに研究を続けているのは一般的に問題であろう。また、データが LSD であるのに、そのデータを学習標本に用いた AI 研究が LSD の事実を指摘しない点だけが、まだ説明できていない。

大学卒業以来の研究テーマである判別分析の新理論を確立し、その応用として「高次元 Microarray の癌の遺伝子解析と診断」にはじめて成功した。そこでこれまでの研究を見直した結果、LSD である高次元データは、ケース数 n 個以下の遺伝子の k 組の小標本に必ず分割できるという事実が統計にとって一番重要と考えた。すなわち、我々は「**高次元データの呪いから数理計画法(MP)の LP と IP で定式化した LDF で解放される**（2 次計画法 QP で定式化した SVM ではできない)**。そして、分割された SM を統計分析するとこれまで見えてこなかった新しい癌の遺伝子診断の世界が広がる**」。

以上の重要なテーマを以下の 4 回のシンポジュームで報告したい。

1)　九州大学：「高次元遺伝子解析の呪いからの解放 1　-統計が 1970 年からこの問題を解決できなかった理由-」

2) 新潟大学：「高次元遺伝子解析の呪いからの解放 2　-癌の遺伝子診断-」

3) 東京工業大学：「高次元遺伝子解析の呪いからの解放 3　-機械学習などの工学研究の問題点-」

4) 秋田大学：「高次元遺伝子解析の呪いからの解放 4　-高次元データの分割法の最新結果-」

予稿集は、以下の 6 章からなり、1 章と 2 章は共通の基礎知識、3 章から 6 章は 4 大学での発表に対応していて、各大学の予稿はその章しか含んでいないで、必要であれば他大学の予稿を参考にしてほしい。

1 章では、Springer1 と「新村(2010).　最適線形判別関数. 日科議連」の中から、統計的判別関数がなぜ癌の遺伝子解析に役に立たなかったかの理由を報告する。すなわち LSD 判別は、MNM 基準による改定 IP‐OLDF(**RIP**)とハードマージン最大化 SVM（**H-SVM**）でしか理論的に正しくできないことが原因である。回帰係数や判別係数を 0 にすることで、Problem5 に対応できると考える LASSO 研究の間違いを指摘する

2 章では、1970 年頃から解決できなかった Problem5 の結果を紹介する。

3 章では、高次元の Microarray がなぜ n 個以下の MNM＝0 である k 組の SM に簡単に分割できるかを数理計画法(MP)の基礎知識（新村(2010).　数理計画法による問題解決法. 日科議連」）と連立方程式の解の基礎知識の簡単な組み合わせで説明する。そして、統計的判別関数と 2 次計画法 QP で定式化された H－SVM が高次元の LSD である Microarray を k 組の SM に分割できない理由を説明する。

4 章では、本研究における 8 種類の LDF の役割を概観する。Microarray が LSD である Fact3 は、RIP、H-SVM、Revised LP-OLDF と SVM4($C=10^4$)で発見できる。そして RIP と Revised LP-OLDF が高次元の呪いから研究者を解放し、SVM ができない理由を示す。高次元 Microarray は k 組の n 個以下の遺伝子の部分空間の SM に分割できる。これらは小標本であり統計手法で簡単に分析し癌の遺伝子診断が行えると考えた。しかし、ロジスティック回帰だけが全ての SM がＮＭ＝0 で、LSD であることが分かる。しかし他の統計手法で LSD の事実が得られなかった（**Problem6**）。試行錯誤の末、信号データを作成してこれを解決し、癌の遺伝子診断が可能になった。しかし LSD である Microarray データを学習に用いているのに、なぜ AI 研究は LSD の事実を発見できないかを参加者と議論したい。6 月開催の IEEE の機械学習の国際会議で良い情報が得られればそれも報告する。

5 章では、世界で初めて成功した癌の遺伝子診断の結果の概略を Springer 2 から説明する。

6 章では、Method 2 が Microarray データだけでなく、6 変数の普通車と小型車の 2 群判別にも適用できることを示す。即ち本研究は、LSD は必ずより小さい SM や最小次元の Basic Gene Sets(**BGS**)に分割できることが今後の統計にとって重要なことを示す。そして Method2 が求めた SM をさらに BGS（iPS 研究の山中 4 因子と同じ概念）に分割すると、多くの場合に 2 個の BGS が含まれることが分かった。

筆者の方法によって、今後高次元の他の LSD であっても、容易に統計分析の研究対象になる。本研究は、質が高く、2 群が LSD であるという検証しやすいデータを用いたことで、LINGO[2]と JMP[1]の組み合わせで初めて役に立つ研究を退官後に完成できたことは医学データを研究対象としたことが幸運であったと考える。

1　新村秀一　(2004).『JMP活用　統計学とっておき勉強法』. 講談社.

2　新村秀一 (2010). 『最適線形判別関数』. 日科技連出版.

3　新村秀一 (2011). 『数理計画法による問題解決法』. 日科技連出版.

4　Shinmura S (2000b). Optimal Linear Discriminant Function using Mathematical Programming. Disertation, Okayama Univ.

5　Springer1: Shinmura S (2016). The New Theory of Discriminant Analysis after R Fisher, Springer. DOI: 10.1007/978-981-10-2164-0

6　Springer2: Shinmura S (2019a) High Dimensional Microarray Data Analysis – Cancer Gene Diagnosis and Malignancy Indexes by Microarray. Springer.

脳画像における高次元データ解析法

佐賀大学 医学部
川口淳　田尻涼

・概要

近年の医学ビッグデータ解析は，これまで個々に解析されていた複数のデータセットを統合的に解析することによって，疾患の特徴づけなどを多様な角度から同時に行っている．脳画像解析ではマルチモダルと呼ばれ，脳の形態と機能などの側面からの脳病態が評価される．

本研究では多種モダリティ脳画像間の関連性を明らかにする方法を提案した．この方法ではモダリティの情報のみならず，アウトカムの情報を組み込んで解釈のしやすい結果を導くように工夫した．またスパース推定を用いアウトカムに関連する変数の特定を行った．

・適用データ

US-ADNI2 にて PET と MRI、DTI を撮影した 42 名より正常群、AD 群それぞれ 10 名、計 20 名をランダムにサンプリングした。この 20 名の MRI と PET、DTI 画像に対して教師付きマルチブロック主成分法を適用し、性能の評価を行った．本データは申請を通せば使えるオープンデータである．

・前処理

sMRI 画像は SPM12 を用いてセグメンテーションを行い、標準脳に対して解剖学的標準化を行った．これにより灰白質画像(GM: Gray matter)、白質画像(WM: White matter)、脳脊髄液画像(CSF: Cerebrospinal fluid)の 3 種画像を得る．DTI 画像は fsl を用いて FA(Fractional Anisotropy)画像を作成、これに対して解剖学的標準化を行った．PET(使用薬剤: FDG)画像は SPM12 を用いて、検査中に複数回撮影される画像を 1 画像にするために画像の加算を行い、解剖学的標準化を行った．以上 3 モダリティ、5 種の画像を用いて解析を行う．

・結果

教師付きマルチブロック主成分法にて、$\mu = 0.5$とし、5 成分を求めた．各モダリティの各成分に対する寄与を見るため以下に各成分のスーパーウェイトを記す．

	GM	WM	CSF	FA	PET
成分 1	0.43	0.24	0.43	0.71	0.27
成分 2	0.39	0.26	0.28	0.83	0.14
成分 3	0.37	0.21	0.37	0.80	0.22
成分 4	0.29	0.23	0.30	0.55	0.69
成分 5	0.21	0.10	0.30	0.31	0.87

表１　各成分のスーパーウェイト

成分１～３では FA 画像が非常に大きな重みをもっており、それに付随する形で構造情報として灰白質､脳脊髄液画像が比較的大きな重みとなっている. 成分 4 では FA に代わって PET の割合が大きくなっており、神経情報と機能情報をバランスよく含んでいると解釈できる. 成分 5 では PET の割合が極めて大きくなり、機能情報を主に含んだ成分になっている.　また、以下にブロックウェイトとして第 1 成分の灰白質と FA の重みを元画像に表示したものを示す.

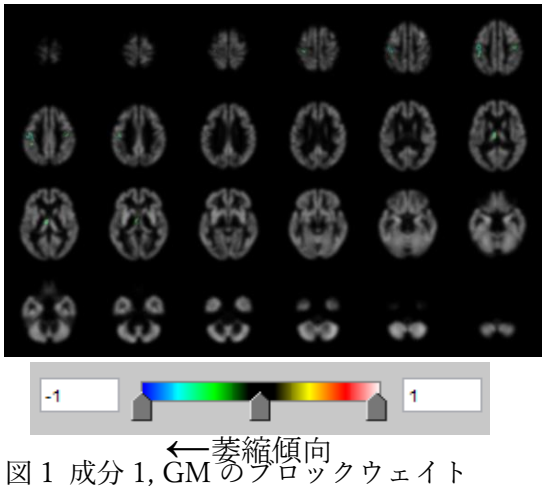


図 1 成分 1, GM のブロックウェイト

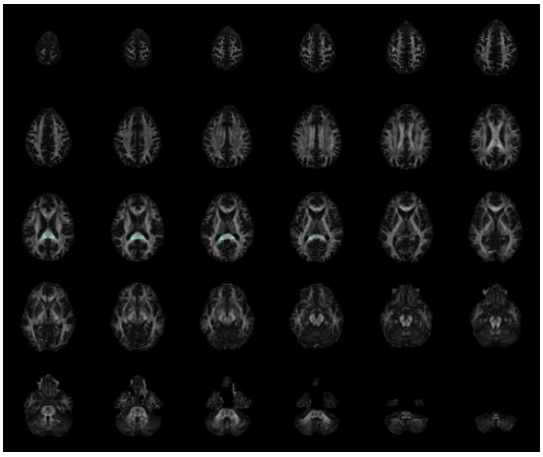


図 2 成分 1, FA のブロックウェイト

得られたスーパースコアの妥当性を検証のためロジスティック回帰を行った. 変数選択は stepwise 法を用い、Null モデルを初期モデルとし、AIC 最小となるモデルを求めた結果、次のモデルが選択された.

$$\text{logit}(\Pr) = \beta_0 + \beta_1 comp1 + \beta_2 comp2 + \beta_3 comp4 + \beta_4 comp5$$

パラメータ	β_0	β_1	β_2	β_3	β_4
推定値	-1.67	0.212	0.0875	0.199	0.549
SE	3.49	0.248	0.870	0.259	0.713
p 値	0.633	0.395	0.314	0.444	0.441

AIC=17.3

表２　ロジスティック回帰の各パラメータ値

各成分のパラメータにおいて 5 %を切ったものはなかった. これは 20 例での検証のため、例数不足の可能性がある. しかし、AIC の観点では得られた 5 つスコアのうち 4 つのパラメータが AD に対して良い予測スコアとなっていることが示された. 従って、教師データに対応する部位の抽出に成功したと考えられる.

Sparse Tensor Subspace method for biological modulator selection

Heewon Park（広島大学情報科学部）
小西貞則（中央大学理工学部）

We are often interested in analysis of multiple datasets comprising the same variables as measured in different groups, and we call this type of data as tensor data. In this study, we consider that *"the multiple datasets have a common eigenvector structure but with different sets of eigenvalues"* (Pepler, 2014). Flury (1984) focused that *"the covariance matrices of different groups have a common basic structure, even though the underlying covariance matrices are not exactly identical across all groups"*, and generalized a technique for PCA to identify a common structure of multiple groups. Wang et al. (2011) proposed a method for common component analysis (CCA) to extract common low-dimensional subspace, that accurately describes all datasets of multiple groups. However, the existing methods for CCA provide fully dense common loadings, thus tensor subspaces are constructed by all variables. It implies that tensor subspace construction procedures can be disturbed by noisy features, because noisy features are inevitably included in datasets. Moreover, the fully dense loadings lead to difficulty in interpretation of CCA results, like in PCA (Al-Kandari and Jolliffe, 2005).

To resolve these issues, we incorporate sparsity into CCA, and propose sparse CCA to construct sparse common subspace (i.e., sparse tensor subspace) of datasets from multiple groups. In order to perform sparse CCA, we first focus that *"ordinary CCA can be implemented by eigen analysis of a matrix (i.e., common loadings of multiple datasets are estimated as eigenvectors of a matrix* $\boldsymbol{M}$*"*. We then consider that the common loadings of a matrix can be also obtained by SVD of square root of the matrix $\boldsymbol{M}$. In order to incorporate sparsity into CCA, we introduce the use of sparse PCA (Zou et al.,2006) with non-centered matrix, i.e., the common loadings of multiple datasets can be estimated by sparse PCA of the square root of the Gram matrix, which is a non-centered data matrix. Incorporation of sparsity into CCA leads to effective analysis of multiple high-dimensional datasets, because the high-dimensional data inevitably contain noisy features. Furthermore, our method provides efficient interpretation results for the constructed common subspace (i.e., we can select crucial common-features of multiple datasets).

We apply the proposed method to construct tensor subspace of drug sensitivity-specific gene regulatory networks, i.e., drug sensitive (resistance) subspace. It can be expected that the reconstructed drug sensitive and resistant networks on the common subspace can effectively describe characteristics of drug sensitive and resistant networks, respectively. Thus, we consider candidate drug identification based on the similarity test between the reconstructed drug sensitive

and resistance tensors on common subspace. For biological modulator selection, we also propose a permutation-based statistical test to evaluate whether the connectivity of genetic network varies with modulators. The similarity test between two network tensors is conducted based on the proposed similarity test on not original data space but the constructed drug sensitivity-specific subspace. We can see through the simulation studies and real data analysis that the proposed method can effectively perform for common structure identification from multiple datasets, and the effectiveness leads to reliable results to biological modulator selection (i.e., candidate drug discovery).

因果推論におけるセミパラメトリックアプローチのための情報量規準

二宮 嘉行 （統計数理研究所）

1 準備

周辺構造平均モデルは，潜在結果変数の周辺期待値に対するモデルとして提案されている．今，処置が H 種類あり，h 番目の処置を受けたときの潜在結果変数を $\boldsymbol{y}^{(h)}$ $(\in \mathbb{R}^m)$，その処置を受けると 1 で受けないと 0 になるランダムな割り当て変数を $t^{(h)}$ とする $(h \in \{1,2,\ldots,H\},\ \sum_{h=1}^{H} t^{(h)} = 1)$．そして，各潜在結果変数に線形回帰モデルを想定した周辺構造平均モデル

$$\boldsymbol{y} = \sum_{h=1}^{H} t^{(h)} \boldsymbol{y}^{(h)} = \sum_{h=1}^{H} t^{(h)} \left(\boldsymbol{X}^{(h)} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \right)$$

を考える．右辺の $\boldsymbol{X}^{(h)}$ $(\in \mathbb{R}^{m \times p})$ は独立変数行列，$\boldsymbol{\varepsilon}$ $(\in \mathbb{R}^m)$ は平均が $\boldsymbol{0}_m$ で分散共分散行列が $\sigma^2 \boldsymbol{I}_m$ の誤差変数ベクトルである．ここで，$\boldsymbol{0}_m$ は m 次元零ベクトル，$\boldsymbol{I}_m$ は m 次元単位行列である．左辺の $\boldsymbol{y}$ は観測される結果変数となることに注意する．このモデルでは，$t^{(h)} = 0$ なる $H-1$ 個の潜在結果変数 $\boldsymbol{y}^{(h)}$ が欠測していると考えられる．そして，一般には $\mathrm{E}[\boldsymbol{y}^{(h)}] \neq \mathrm{E}[\boldsymbol{y}^{(h)} \mid t^{(h)} = 1]$ であるため，観測値のみから単純に $\boldsymbol{X}^{(h)}\boldsymbol{\beta}$ を推定するとバイアスが生じてしまう．本発表では，このバイアスの除去が可能となるように，$\boldsymbol{y}^{(h)}$ と $t^{(h)}$ の交絡変数ベクトル $\boldsymbol{z}$ $(\in \mathbb{R}^s)$ が観測されていることを想定する．

こういったモデルと変数に対し，通常置かれる条件を仮定する．まず $\boldsymbol{X}^{(h)}$ についてだが，一般的な設定を扱うために交絡変数 $\boldsymbol{z}$ を含んでもよい (ランダムであることを許す) ことにする．また，式における表現の煩雑さを軽減するため，本質的ではないがこの独立変数は $\mathrm{E}[\sum_{h=1}^{H} \boldsymbol{X}^{(h)\mathrm{T}} \boldsymbol{X}^{(h)}] = \boldsymbol{I}_p$ となるように基準化されているものとする．この基準化の有無によらず，以降で結果として導かれる基準は同じものとなる．次に，上述のバイアスの除去を可能とするためのものである，弱く無視できる割り当て条件

$$\boldsymbol{y}^{(h)} \perp t^{(h)} \mid \boldsymbol{z} \qquad (h \in \{1,2,\ldots,H\})$$

を仮定する．この条件における $\boldsymbol{y}^{(h)}$ は $\boldsymbol{\varepsilon}$ に変えられることに注意する．さて，このモデルにしたがう N 個のサンプルがあるとし，第 i サンプルの変数には添え字 i を付けることにする．そして，$\tilde{\boldsymbol{y}}^{(h)} = (\boldsymbol{y}_1^{(h)\mathrm{T}}, \boldsymbol{y}_2^{(h)\mathrm{T}}, \ldots, \boldsymbol{y}_N^{(h)\mathrm{T}})^{\mathrm{T}}$，$\boldsymbol{T}^{(h)} = \mathrm{diag}\{t_i^{(h)} \boldsymbol{I}_m\}$，$\tilde{\boldsymbol{X}}^{(h)} = (\boldsymbol{X}_1^{(h)\mathrm{T}}, \boldsymbol{X}_2^{(h)\mathrm{T}}, \ldots, \boldsymbol{X}_N^{(h)\mathrm{T}})^{\mathrm{T}}$，$\tilde{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \boldsymbol{\varepsilon}_2^{\mathrm{T}}, \ldots, \boldsymbol{\varepsilon}_N^{\mathrm{T}})^{\mathrm{T}}$ とし，まとめて

$$\tilde{\boldsymbol{y}} = \sum_{h=1}^{H} \boldsymbol{T}^{(h)} \tilde{\boldsymbol{y}}^{(h)} = \sum_{h=1}^{H} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{X}}^{(h)} \boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \right)$$

と表すことにする．ただし，サンプルは独立，つまり

$$(t_i^{(h)}, \boldsymbol{X}_i^{(h)}, \boldsymbol{\varepsilon}_i, \boldsymbol{z}_i) \perp (t_j^{(h)}, \boldsymbol{X}_j^{(h)}, \boldsymbol{\varepsilon}_j, \boldsymbol{z}_j) \qquad (i \neq j,\ h \in \{1,2,\ldots,H\})$$

であることを仮定する．これより当然 $\boldsymbol{y}_i \perp \boldsymbol{y}_j$ $(i \neq j)$ である．また，通常の回帰モデルのように $\boldsymbol{X}_i^{(h)}$ と $\boldsymbol{\varepsilon}_i$ は独立であると仮定しておく．

潜在結果変数 $\boldsymbol{y}^{(h)}$ と交絡変数 $\boldsymbol{z}$ の関係を正しくモデリングできていれば，無視できる割り当て条件のもとで観測値のみから $\boldsymbol{y}^{(h)}$ の周辺期待値を一致推定することができるが，一般にこのモデリングは難しい．そこで近年は，この正しいモデリングを必ずしも必要としない，傾向スコア $e_i^{(h)} \equiv \mathrm{P}(t_i^{(h)} = 1 \mid \boldsymbol{z}_i)$ を用いたセミパラメトリックアプローチを用いられることが多い．本発表では，このアプローチにおいて代表的な IPW 推定を扱う．この推定法では，観測値に傾向スコアの逆数を重みとしてかけることで疑似的に欠測値を復元し，その後に通常の推定をおこなう．具体的には，重み行列 $\boldsymbol{W}^{(h)} \equiv \mathrm{diag}(t_i^{(h)} \boldsymbol{I}_r / e_i^{(h)})$ を用いて重み付き二乗損失関数を

$$\sum_{h=1}^{H} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)} \boldsymbol{\beta} \right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)} \boldsymbol{\beta} \right) \tag{1}$$

と定義し，これを $\boldsymbol{\beta}$ について最小化することで IPW 推定量

$$\hat{\boldsymbol{\beta}}^{\mathrm{IPW}} \equiv \left\{ \sum_{h=1}^{H} \tilde{\boldsymbol{X}}^{(h)\mathrm{T}} \boldsymbol{W}^{(h)} \tilde{\boldsymbol{X}}^{(h)} \right\}^{-1} \sum_{h=1}^{H} \tilde{\boldsymbol{X}}^{(h)\mathrm{T}} \boldsymbol{W}^{(h)} \tilde{\boldsymbol{y}}$$

を与える．IPW 推定量は，弱く無視できる割り当て条件のもとで一致性をもつ．

2　モデル選択基準

データに欠測がある場合の MSE として自然に考えられるものを二種類考える．まず一種類目として mean weighted squared error を

$$\begin{aligned}
\mathrm{MwSE} &= \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right] \\
&= \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}\right)\right] \\
&\quad - \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right] \\
&\quad + 2\sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right]
\end{aligned}$$

と定義する．ここで，通常の C_p 基準の導出にならい，定義後には三つの項に分解している．ここでの重み付き二乗和は，観測値に重みを付けて復元したデータの期待値とその推定値の差の二乗和とみなせ，その重みは IPW 推定でも用いられているものである．実際，分解後の第一項目は (1) の期待値をとったものである．つまり，推定量の導出と推定量の誤差評価に同種の損失関数で考えており，その点で自然である．次に，二種類目として mean unweighted squared error を

$$\begin{aligned}
\mathrm{MuSE} &= \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right] \\
&= \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}\right)\right] \\
&\quad - \sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right] \\
&\quad + 2\sum_{h=1}^{H} \mathrm{E}\left[\left(\tilde{\boldsymbol{y}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)^{\mathrm{T}} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}} - \mathrm{E}\left[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}\right]\right)\right]
\end{aligned}$$

と定義する．ここでは観測値そのもののデータの期待値とその推定値の差の二乗和を考えている．観測されたものに対する推定精度の向上により注力したいという考え方のもとでは，こちらの損失関数の方が自然といえる．いうまでもなく，$\boldsymbol{W}^{(h)}$ や $\boldsymbol{T}^{(h)}$ は非負定値対称行列であり，両 MSE ともノルムになっている．通常の C_p 基準の導出にならい，分解したものの第一項目の期待値を外し，モデルに依存しない第二項目を無視し，$\mathrm{E}[\tilde{\boldsymbol{y}}^{(h)} \mid \tilde{\boldsymbol{X}}^{(h)}] = \tilde{\boldsymbol{X}}^{(h)}\boldsymbol{\beta}$ として第三項目を漸近評価したものを因果推論における C_p 基準とする．漸近評価では，期待値の中身に対して漸近的に主項となるものを取り出し，その期待値を陽に求める．そして，MwSE あるいは MuSE から導かれる基準をそれぞれ $\mathrm{w}C_p$ あるいは $\mathrm{u}C_p$ と表すことにする．

Theorem 1. 傾向スコアが既知であるケースで IPW 推定量を用いたときの C_p 基準は

$$\begin{aligned}
\mathrm{w}C_p &= \sum_{h=1}^{H} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}\right)^{\mathrm{T}} \boldsymbol{W}^{(h)} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}\right) + 2\sum_{h=1}^{H} \mathrm{E}\left[\frac{1}{e^{(h)}}\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{X}^{(h)}\boldsymbol{X}^{(h)\mathrm{T}}\boldsymbol{\varepsilon}\right], \\
\mathrm{u}C_p &= \sum_{h=1}^{H} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}\right)^{\mathrm{T}} \boldsymbol{T}^{(h)} \left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}\right) + 2\sigma^2 p
\end{aligned}$$

で与えられる．

$\mathrm{w}C_p$ のペナルティ項は陽に計算できるとは限らないが，$\sum_{i=1}^{N}\sum_{h=1}^{H} t_i^{(h)}(\boldsymbol{y}_i^{(h)} - \boldsymbol{X}_i^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}})^{\mathrm{T}}\boldsymbol{X}_i^{(h)}\boldsymbol{X}_i^{(h)\mathrm{T}}(\boldsymbol{y}_i^{(h)} - \boldsymbol{X}_i^{(h)}\hat{\boldsymbol{\beta}}^{\mathrm{IPW}})/(Ne_i^{(h)2})$ のように一致推定量を容易に与えることができる．

方向データに対する
von Mises 分布の母数推定について

作村 建紀 [*]　　柳本武美 [†]

1　はじめに

方向に関する分布として，最も有名なものに von Mises 分布がある．この分布は，指数分布族に属しており，方向データに関する代表的な分布に位置する．この分布は 2 つのパラメータを持っており，それぞれ方向パラメータ，集中度パラメータとして知られている．

方向データに関する研究は，例えば，風向きの分析 (Mardia & Jupp, 1972) などがあり，近年，広い分野で応用が期待されている．また，von Mises 分布に関するパラメータ推定の研究としては，Schou (1978) による周辺最尤推定量の提案などが挙げられるが，その他の例は少ない．

一方，指数型分布族において自然母数の事後平均は最適性を保つ (Yanagimoto & Ohnishi, 2009)．これを von Mises 分布に適用し，提案推定量とすることを考えると，最適性の観点から，提案推定量は良い性能を示すことが期待される．

2　von Mises 分布

標本 $X_1, \ldots, X_n$ は，次の確率密度関数を持つ von Mises 分布 $\mathrm{VM}(\eta, \tau)$ $(\eta \in (-\pi, \pi),\ \tau > 0)$ からの無作為標本とする．

$$p(x; \eta, \tau) = \frac{1}{2\pi I_0(\tau)} \exp\left\{\tau \cos(x - \eta)\right\},\ x \in (-\pi, \pi)$$

ここで，関数 $I_0(\cdot)$ は第 1 種変形ベッセル関数である．ここで，$C(\boldsymbol{x}) = \sum_{i=1}^n \cos(x_i)$, $S(\boldsymbol{x}) = \sum_{i=1}^n \sin(x_i)$ とし，$C(\boldsymbol{x}) = R(\boldsymbol{x}) \cos \bar{x}$, $S(\boldsymbol{x}) = R(\boldsymbol{x}) \sin \bar{x}$ を満たす合成ベクトル $R(\boldsymbol{x})$ と標本平均方向 $\bar{x}$ を考えると，

$$R(\boldsymbol{x}) = \sqrt{C^2(\boldsymbol{x}) + S^2(\boldsymbol{x})}, \tag{1}$$

であり，$R(\boldsymbol{x}) \neq 0$ のとき $\bar{x} = \tan^{-1}(S(\boldsymbol{x})/C(\boldsymbol{x}))$ である．

3　提案推定量

Yanagimoto and Ohnishi (2009) により，最適性が保証された推定量である自然母数の事後平均を，von–Mises 分布の母数 (η, τ) の推定に対する有効な推定量として提案する．von Mises 分布の自然母数は，$\boldsymbol{\theta} = (\theta_1, \theta_2) = (\tau \cos \eta, \tau \sin \eta)$ であり，対応する十分統計量は $\boldsymbol{x} = (\cos x, \sin x)$ である．事前分布には参照事前密度を用いる (Garvan & Ghosh, 1997; Robert, 2007).

$$\pi_R(\eta, \tau) \propto \left\{1 - \frac{A(\tau)}{\tau} - A^2(\tau)\right\}^{1/2}.$$

ここで，$A(x) = I_1(x)/I_0(x)$ である．これはパラメータ η に依存しないことから，以降，$\pi_R(\tau)$ と表す．自然母数 θ_1 の事後平均は，

$$\hat{\theta}_1 = \cos \bar{x} \int \tau \pi_m(\tau|\boldsymbol{x}) A(\tau R(\boldsymbol{x})) d\tau,$$

となる．同様に，

$$\hat{\theta}_2 = \sin \bar{x} \int \tau \pi_m(\tau|\boldsymbol{x}) A(\tau R(\boldsymbol{x})) d\tau,$$

が得られる．これらは，パラメータ η, τ に関して不変であるから，$\boldsymbol{\theta} = (\theta_1, \theta_2) = (\tau \cos \eta, \tau \sin \eta)$ より，求める提案推定量は，

$$\hat{\eta} = \tan^{-1}\left(\frac{\hat{\theta}_2}{\hat{\theta}_1}\right) = \bar{x}, \tag{2}$$

$$\hat{\tau} = \sqrt{\hat{\theta}_1^2 + \hat{\theta}_2^2} = \int \tau \pi_m(\tau|\boldsymbol{x}) A(\tau R(\boldsymbol{x})) d\tau, \tag{3}$$

となる．

4　推定量の比較

提案推定量の性能を調査するために，既存推定量との比較実験を行う．既存推定量としては，最尤推定量（MLE），周辺最尤推定量（MMLE）(Schou, 1978)，モデルパラメータの事後平均（PM）を対照とする．

ここで，τ に関する MLE は以下で得られる．

$$\hat{\tau}_{ML} = A^{-1}\left(\frac{R(\boldsymbol{x})}{n}\right)$$

また，Schou (1978) は，τ に関して，次式を満たす MMLE, $\hat{\tau}_{MML}$ を提案している．

$$nA(\hat{\tau}_{MML}) = R(\boldsymbol{x}) A(\hat{\tau}_{MML} R(\boldsymbol{x})),\ \ R(\boldsymbol{x}) > \sqrt{n}$$

また，$0 \leq R(\boldsymbol{x}) \leq \sqrt{n}$ のとき，$\hat{\tau}_{MML} = 0$ である．さらに，τ の PM は次式で与えられる．

$$\hat{\tau}_{PM} = \int \tau \pi_m(\tau \mid \boldsymbol{x}) d\tau$$

[*] 法政大学理工学部, 〒184-8584 東京都小金井市梶野町 3–7–2
[†] 統計数理研究所, 〒190-0014 東京都立川市緑町 10–3

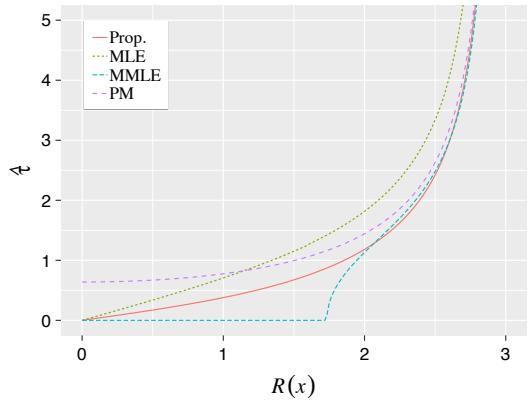


Figure1: サンプルサイズ $n = 3$ における，$R(x)$ に対する τ の推定量の振る舞い: 実線が提案推定量（Prop.），点線が MLE，破線が PM，長破線が MMLE を指す．

4.1 推定量の振る舞い

集中度パラメータ τ のそれぞれの推定量は，$R(x)$ に対応して推定される．そこで，まず，$R(x)$ に応じた推定量 $\hat{\tau}$ の振る舞いについて検討する．サンプルサイズを $n = 3$ と固定し，$R(x)$ を変化させたときの推定量を exact に求めた結果を Figure 1に示す．図より，以下のことが言える．MMLE は $\tau = \sqrt{3}$ で折れ点がある．また，PM は $R(x) = 0$ のときに $\hat{\tau} \neq 0$ である．$R(x)$ が大きくなると，提案推定量と MMLE は非常に近い値を取り，PM もまた近づいている．MLE は提案推定量よりも常に大きい値を取る．

4.2 リスク比較

提案推定量の性能をリスク比較によって検証する．リスクは頻度論の観点から計算する．つまり，

$$R(\hat{\tau}) = \int L(\hat{\tau}, \tau) p(x|\tau) dx$$

ここで，$p(x|\tau)$ は $\eta = \bar{x}$ とした von Mises 分布の標本密度であり，$L(\hat{\tau}, \tau)$ は損失関数を表し，ここでは，e システムにおける KL ダイバージェンス, KL_e と，m システムにおける KL ダイバージェンス, KL_m，二乗損失, MSE を用いる．

既存推定量を対照とした提案推定量とのリスク比較を，シミュレーションによって行う．シミュレーションにおいて，方向パラメータ $\eta = 0$ に固定し，サンプルサイズ $n = \{10, 30, 100\}$, 集中度パラメータ $\tau = \{0, .3, 1, 3, 10\}$ の構成で，それぞれにおいて 10,000 サンプルを生成して行う．結果を Table 1 に示す．

Table 1 より，提案推定量は，MMLE を対照とすると，$n = 10$ において，$\tau = 0$ では MMLE に劣るものの，$\tau = 0.3$ および $\tau = 1$ では優越しており，$\tau \geq 3$ でほぼ同等である．また，$n = 30, 100$ の場合においては，$\tau = 1$ のときの結果は逆転してるものの，ほぼ同じ傾向にある．また，MLE と比べると，提案推定量はほとんどの条件のもとで優越している．PM と比較すると，特に $\tau = 0$ に近いほど，提案推定量のほうが良い傾向にある．これらは，von Mises Fisher 分布などの関連分布における推定量の構成に関して，その基礎となる結果である．さらに，ここでは無情報事前分布を仮定してリスク比較を行ったが，情報のある事前分布での選択やその推定量の構成についても有益な結果であると期待できる．

Table1: 損失 KL_e のもとでのリスク比較

n	τ	Prop.	MMLE	MLE	PM
10	0	0.5586	0.5165	1.043	0.9011
	0.3	0.6298	0.6686	1.055	0.8968
	1	0.9858	1.169	1.065	0.9540
	3	1.104	1.065	1.100	1.075
	10	1.055	1.052	1.087	1.049
30	0	0.5175	0.4923	1.009	0.8787
	0.3	0.6925	0.8573	1.014	0.8517
	1	1.064	1.045	1.000	1.018
	3	1.033	1.023	1.028	1.024
	10	1.021	1.019	1.032	1.019
100	0	0.5065	0.4822	1.004	0.8743
	0.3	0.9277	1.122	1.019	0.9126
	1	1.031	1.018	1.000	1.020
	3	1.012	1.009	0.9991	1.009
	10	1.022	1.020	1.025	1.020

References

Garvan, C. W. & Ghosh, M. (1997, Dec.). Noninformative priors for dispersion models. *Biometrika*, *84*(4), 976-982.

Mardia, K. V. & Jupp, P. E. (1972). *Directional statistics*. John Wiley & Sons, LTD.

Robert, C. (2007). *The bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Schou, G. (1978). Estimation of the concentration parameter in von mises-fisher distributions. *Biometrika*, *65*(2), 369-377.

Yanagimoto, T. & Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, *39*, 3064-3075.

A Practicable Estimation of Mean Squared Prediction Error in Small Area Estimation

Masayo Y. Hirose
The Institute of Mathematics for Industry, Kyushu University

Abstract

The empirical best linear unbiased prediction (EBLUP) estimator is utilized for efficient inference in various research areas, especially for small-area estimation. In order to measure its uncertainty, we generally need to estimate its mean squared prediction error (MSPE). In this study, we seek a MSPE estimation method achieving several desired properties under aggregated level model.

1 Introduction

There has been high demand for reliable statistics on smaller geographic areas and sub-populations where large samples are not available. Even in such a situation, an explicit model-based approach can achieve more accurate estimates by borrowing strength from related areas.

The Fay–Herriot model (Fay and Herriot, 1979), in particular, is widely used as an aggregated level model for small-area inference as follows:

For $i = 1, \ldots, m$,

$$\begin{aligned} &\text{Level 1 }: y_i|\theta_i \overset{ind}{\sim} N(\theta_i, D_i); \\ &\text{Level 2 }: \theta_i \overset{ind}{\sim} N(x_i'\beta, A). \end{aligned} \tag{1}$$

The level-1 model takes into account the sampling distribution of the direct estimator y_i for the ith small area. The true small-area mean for the ith area, denoted by θ_i, is linked to the area-specific auxiliary variables $x_i = (x_{i1}, \cdots, x_{ip})'$ in the level-2 model. In practice, the coefficient vector β in $\mathbb{R}^p$ and the model variance parameter A in this linking model are unknown. The sampling variance D_i is often assumed to be known.

Since A and β are unknown in practice, the empirical best linear unbiased predictor (EBLUP) of θ_i is generally used for small-area inference. It would also be quite important to measure the MSPE of EBLUP as its uncertainty. For small-area inference, its MSPE needs to be estimated with high accuracy.

In this study, we consider what a desirable properties of practicable MSPE estimation. We also compare the performance of several MSPE estimators in our investigated class and others through a Monte Carlo simulation study. This study has been published in Hirose (2019).

Acknowledgement

This research was supported by Grant-in-Aid for Research Activity Start-up, JSPS Grant Number 26880011.

References

[1] Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269-277

[2] Hirose, M.Y. (2019). A Class of General Adjusted Maximum likelihood Methods for Desirable Mean Squared Error Estimation of EBLUP under the Fay-Herriot Small Area Model. *Journal of Statistical Planning and Inference*, **199**, 302-310.

前震の統計的判別に基づく地震予測

統計数理研究所　野村俊一

任意の地震群は，群内の地震のマグニチュード差や時空間的距離の分布形に見られる特徴及び地域性に基づき，前震群である確率を本震が起こる前に評価することができる．図 1 に示す Single-link 法[1]に従い地震群を構成し，構成した地震群ごとに群内の最大地震を本震，本震の前と後の群内地震をそれぞれ前震，余震と定義する．そして，活動中の地震群が 30 日以内に本震を起こす前震群である確率を評価するための統計モデルを構築し，その識別性能を検証する．

地震群のマグニチュード差や時空間的距離に関する特徴は，地震群内の地震数 N と最大マグニチュード M_1 に依存して変化するものの，その 2 つが同条件の下では前震群と他の地震群の間に相対的な傾向の違いが現れる[2]．そこで，群内の最大マグニチュード M_1 と二番目に大きいマグニチュード M_2 の差 ΔM，地震群の継続期間 T（日），群内の平均震央間距離 D（km），地震群の中心経度 X，緯度 Y の位置情報（地域性）を前震識別のための特徴量として，ある地震群について，群内の最新地震から 30 日以内の本震発生確率を本稿での前震確率と定義し，次式に基づいて評価する:

$$\mathrm{logit}\, p = g(X, Y) + f_1(N, M_1, \Delta M) + f_2(N, M_1, T) + f_3(N, M_1, D) + \alpha_0 + \alpha_s$$

ここで，$\mathrm{logit}\, p = \log\{p/(1-p)\}$ は p の取る値を $0 < p < 1$ の範囲へ限定するロジット関数であり，このとき右辺の各項は特徴量による前震確率の利得あるいは前震群と他の地震群の対数オッズ比と解釈される．右辺第 1 項 $g(X, Y)$ は地域性のみに依存したベースラインであり薄板スプライン関数により推定される．また，$f_1(N, M_1, \Delta M), f_2(N, M_1, T), f_3(N, M_1, D)$は各特徴の組合せに基づくオッズの自然対数を，それぞれ 3 次スプライン関数により表現した項である．また α_0 は定数項であり，さらに，特異な地震群の影響による推定の不安定化を抑えるために，以上の項に含まれない地震群固有の変量効果 α_s を加えた．

1926 年 1 月 1 日から 1999 年 12 月 31 日までの気象庁震源カタログによる上の評価式の推定結果を図 2 に示した．図 2(d)の地域差の推定結果からは，主に伊豆半島沖と東北沖において前震確率が高くなっている．また，図 2(a)からはマグニチュード差 ΔM が小さいほど前震確率が高く，図 2(b)からは継続期間 T が短いほど前震確率が高く，図 2(c)からは平均震央間距離が長いほど若干前震確率が高くなっており，この傾向は地震数 N によって変化していることがわかる．なお，図 2 に共通する特徴として，最大マグニチュード M_1 が小さいほど前震確率が高くなっているが，これは，M_1 が小さいほど対象とする本震マグニチュードの下限が下がるためである．

以上で推定された前震確率の評価式を，2000 年 1 月 1 日から 2017 年 10 月 31 日までの同カタログに適用し，群内地震数が $N = 2, 4, 8$ に達した各地震群について前震確率を評価した．地震群ごとの前震確率の評価値を 10%区切りの階級に分けて集計し，各階級中の実際の前震割合すなわち適中率と比較した結果を表 1 に示す．適中率は，該当の地震群数が少ない階級を除いて，各階級の前震確率評価値と概ね整合的である．また，60%超の高い前震確率を得たのは伊豆半島沖の地震群であり，実際に M6 前後の本震を起こしている．ただし，たとえば 2016 年熊本地震の前震系列のように群内地震数が 10 を超えてくると，事例数の低下により前震確率の評価は不安定なものとなる．今後は下限マグニチュードの引き下げや海外のカタログの活用を検討して事例数の充足を図りたい．

参考文献

[1] Frohlich, C., and Davis. S. D. (1990). Single-link cluster analysis as a method to evaluate spatial and temporal properties of earthquake catalogues, *Geophys. J. Int.*, 100, pp.19-32.

[2] Ogata, Y., Utsu ,T. and Katsura, K. (1995). Statistical features of foreshocks in comparison with other earthquake clusters, *Geophys. J. Int.*, 121, pp.233-254.

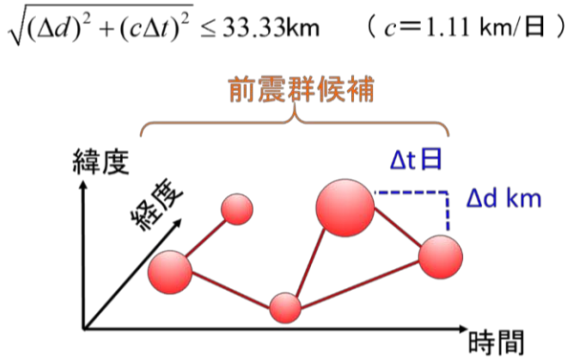


図 1.　Single-link 法による前震群候補の構成方法.

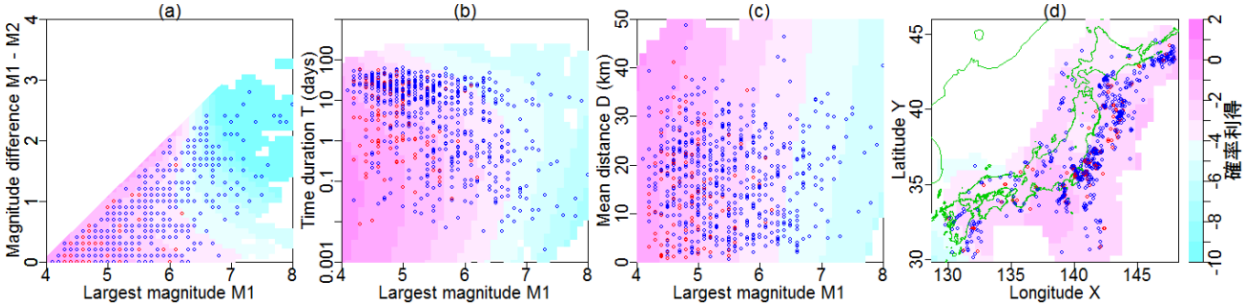


図 2.　地震数 N = 4 の前震群（赤点）と他の地震群（青点）の分布の違い．両者の分布の違いから推定された確率利得（対数オッズ比）をカラースケールで示している．

表 1.　2000 年以降の前震確率評価値の度数分布と適中率（＝前震群数／全地震群数）.

前震確率の評価値		0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	合計
N = 2	全地震群数	463	483	227	64	14	2	0	0	1253
	前震群数	32	74	47	20	4	2	0	0	179
	適中率	7%	15%	21%	31%	29%	100%	N/A	N/A	14%
N = 4	全地震群数	151	98	54	22	14	5	1	2	347
	前震群数	14	16	10	8	2	1	1	2	54
	適中率	9%	16%	19%	36%	14%	20%	100%	100%	16%
N = 8	全地震群数	82	17	10	2	6	5	2	0	124
	前震群数	7	3	4	1	3	2	2	0	22
	適中率	9%	18%	40%	50%	50%	40%	100%	N/A	18%

Penalized least squares approximation methods

鈴木　拓海

東京大学大学院数理科学研究科

2019 年 8 月 30 日

- $\mathcal{B} = (\Omega, \mathcal{F}, \mathbf{F}, P), \mathbf{F} = (\mathcal{F}_t)_{t\in\mathbb{R}_+}$: a stochastic basis
- $\Theta \subset \mathbb{R}^{\mathsf{p}}$：パラメータ空間
  - Θ は有界開集合
- θ^*：真値，$\theta^* \in \Theta$ かつ 0 の成分が p_0 個
  - $\theta^*_1 \neq 0, ..., \theta^*_{\mathsf{p}^0} \neq 0$，$\theta^*_{\mathsf{p}^0+1} = ... = \theta^*_{\mathsf{p}} = 0$
- r_T：T の関数．多くの場合は $r_T^{-1} = \sqrt{T}$ を考える．
- ベクトル $v \in \mathbb{R}^{\mathsf{p}}$ をサブベクトル $v_{\mathcal{J}^1} \in \mathbb{R}^{\mathsf{p}^0}$ および $v_{\mathcal{J}^0} \in \mathbb{R}^{\mathsf{p}-\mathsf{p}^0}$ を用いて次のように表すことにする．

$$v = \begin{bmatrix} v_{\mathcal{J}^1} \\ v_{\mathcal{J}^0} \end{bmatrix}$$

同様に $\mathsf{p} \times \mathsf{p}$ 行列 M を

$$M = \begin{bmatrix} M_{\mathcal{J}^{11}} & M_{\mathcal{J}^{10}} \\ M_{\mathcal{J}^{01}} & M_{\mathcal{J}^{00}} \end{bmatrix}$$

のように表すことにする．

Objective function

$$Q_T^{(q)}(\theta) = (\theta - \tilde{\theta})'\hat{G}(\theta - \tilde{\theta}) + \sum_{j=1}^{\mathsf{p}} \kappa_T^j |\theta_j|^q,$$

- $\tilde{\theta}$：初期推定量
  - 例えばあるロス関数 $\mathcal{L}_T(\theta)$ に対して，$\tilde{\theta} \in \mathrm{argmin}_\theta \mathcal{L}_T(\theta)$ なるもの
- $\hat{G} \in \mathbb{R}^{\mathsf{p}\times\mathsf{p}}$：（ランダムな）正定値対称行列
- $\kappa_T^j = \alpha_T |\tilde{\theta}_j|^{-\gamma}$
  - γ：定数，$\gamma > -(1-q)$
    - 例えば，$\gamma = 1$.
  - $(\alpha_T)_T$：deterministic な数列，$r_T^{-(2-q+\gamma)}\alpha_T \to \infty$，　$r_T^{-1}\alpha_T = o(1)$
    - 例えば，$\alpha_T = r_T^2$.

penalized LSA (Least Squares Approximation) 推定量 $\hat{\theta}^{(q)}$

$$\hat{\theta}^{(q)} \in \underset{\theta\in\bar{\Theta}}{\mathrm{argmin}}\ Q_T^{(q)}(\theta)$$

仮定 1

- $r_T^{-1}(\tilde{\theta} - \theta^*) = O_p(1)$.
- $\hat{G} \to^p G$
  - $G \in \mathbb{R}^{\mathsf{p}\times\mathsf{p}}$：ある（ランダムな）正定値対称行列

Theorem 1 (r_T^{-1}-consistency and selection consistency)

仮定 *1* のもとで，

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*) = O_p(1), \quad P[\hat{\theta}^{(q)}_{\mathcal{J}^0} = 0] \to 1$$

が成り立つ．

仮定 2

$r_T^{-1}(\tilde{\theta} - \theta^*) \to^{d_s} G^{-\frac{1}{2}}\zeta$

- ζ：p 変量標準正規確率変数，$\mathcal{G}$ と独立
- $\to^{d_s}$：$\mathcal{G}$-stable convergence
- $\mathcal{G}$：σ-field，$\sigma(G) \subset \mathcal{G} \subset \mathcal{F}$

Theorem 2 (Asymptotic normality)

$\mathfrak{G} = \begin{bmatrix} I_{\mathsf{p}^0} & (G_{\mathcal{J}^{11}})^{-1}G_{\mathcal{J}^{10}} \end{bmatrix}$ とおく．このとき，仮定 *1* のもとで，

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} - \mathfrak{G}\{r_T^{-1}(\tilde{\theta} - \theta^*)\} \to^p 0.$$

特に，仮定 *2* のもとで，

$$r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)_{\mathcal{J}^1} \to^{d_s} \mathfrak{G}G^{-\frac{1}{2}}\zeta \sim \mathrm{MN}_{\mathsf{p}^0}(0, (G_{\mathcal{J}^{11}})^{-1}).$$

Definition 3

確率過程 $X = \{X_T\}_T$ が $L^{\infty-}$-有界とは，$\sup_T E[|X_T|^p] < \infty, \forall p \geq 1$ のときにいう．

仮定 3

$\{\hat{G}\}_T$，$\{\hat{G}^{-1}\}_T$ および $\{r_T^{-1}(\tilde{\theta} - \theta^*)\}_T$ は $L^{\infty-}$-有界．

Theorem 4

1. 仮定 *3* のもとで $\{r_T^{-1}(\hat{\theta}^{(q)} - \theta^*)\}_T$ は $L^{\infty-}$-有界．
2. 任意の $L > 0$ に対し，定数 C_L が存在し，

$$P[\hat{\theta}^{(q)}_{\mathcal{J}^0} = 0] \geq 1 - C_L r_T^{2L}$$

が任意の $T > 0$ に対して成り立つ．

P-O 推定量

1. r_T^{-1}-一致性を持つ初期推定量 $\tilde{\theta}$ を用意する．
2. $\tilde{\theta}$ を用いて，単位行列 I_{p} を係数行列とする penalized LSA 推定量 $\hat{\theta}^{(q)}_{I_{\mathsf{p}}}$ を求める：

$$\hat{\theta}^{(q)}_{I_{\mathsf{p}}} \in \operatorname*{argmin}_{\theta \in \hat{\Theta}} \sum_{j=1}^{\mathsf{p}} \Big((\theta_j - \tilde{\theta}_j)^2 + \kappa_T^j |\theta_j|^q \Big). \tag{2.1}$$

3. 新しいロス関数 $\mathbb{L}_T(\theta)$ を用いて，P-O 推定量 $\tilde{\theta}$ を以下で定義する．

$$\tilde{\theta} \in \operatorname*{argmin}_{\theta \in \hat{\Theta}} \mathbb{L}_T(\theta)$$

ここで，$\hat{\Theta} = \{\theta \in \Theta; \theta_j = 0, j \in \hat{\mathcal{J}}^0\}$，$\hat{\mathcal{J}}^0 = \{j = 1, ..., \mathsf{p}; \hat{\theta}^{(q)}_{I_{\mathsf{p}},j} = 0\}$ とした．

P-O 推定量は，変数選択については penalized LSA 推定量と同等の性質を持つ一方，有効推定量となっている．さらに，係数行列を単位行列にしたことにより，(2.1) を見て分かるように，目的関数の最適化問題が 1 次元の最適化問題に帰着されている．一般に変数選択の問題は高次元になることが多いため，これは非常に有用である．

次のような intensity $\lambda(t, \theta)$ をもつ point process N を考える．

$$\lambda(t, \theta) = \exp\Big(\sum_{j \in \mathbf{J}} \theta_j X_t^j \Big) \tag{3.1}$$

ここで，$X = (X_t^j)_{t \in [0,T], j = 1, ..., \mathsf{p}}$ は左連続な p-次元確率過程とする．
モデル (3.1) に対して，擬似尤度解析を考える．擬似対数尤度関数を

$$\ell_T(\theta) = \sum_{\alpha \in \mathbf{I}} \int_0^T \log\left(\lambda^\alpha(t, \theta)\right) dN_t^\alpha - \sum_{\alpha \in \mathbf{I}} \int_0^T \lambda^\alpha(t, \theta) dt. \tag{3.2}$$

とおくと，$\mathcal{L}_T(\theta) = -\ell_T(\theta)$ はロス関数とみなせる．また，擬似最尤推定量 $\tilde{\theta}$ を $\ell_T(\theta)$ を最大化するものとして与えると，$\tilde{\theta}$ は仮定 1-3 を満たす．

covariate process $X = (X_t)_{t \in [0,T]}$ は 20-次元 OU 過程で以下の SDE を満たすものを考える．

$$dX_t^i = -\mathsf{a}_i X_t^i dt + 0.4 dW_t^i, \quad X_0 = 0, \quad t \in [0, T].$$

ここで，$\mathsf{a}_i, i = 1, \cdots, 20$ は

$$\begin{aligned}
\mathsf{a}_1 &= \mathsf{a}_6 &&= \mathsf{a}_{11} = \mathsf{a}_{16} = 0.15, \\
\mathsf{a}_2 &= \mathsf{a}_7 &&= \mathsf{a}_{12} = \mathsf{a}_{17} = 0.2, \\
\mathsf{a}_3 &= \mathsf{a}_8 &&= \mathsf{a}_{13} = \mathsf{a}_{18} = 0.25, \\
\mathsf{a}_4 &= \mathsf{a}_9 &&= \mathsf{a}_{14} = \mathsf{a}_{19} = 0.3, \\
\mathsf{a}_5 &= \mathsf{a}_{10} &&= \mathsf{a}_{15} = \mathsf{a}_{20} = 0.35
\end{aligned}$$

とし，$W = (W^i)_{i=1,\cdots,20}$ は 20-次元標準 Wiener 過程とする．データは intensity $\lambda(t, \theta^*)$ をもつ point process $N = (N_t)_{t \in [0,T]}$ のサンプルパスとする．また，真値 θ^* は

$$\theta^* = [2, -1, 1, -0.5, -1.5, 1.5, 0.5, 0.75, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]'$$

とする．

Table: Results of the variable selection under $T = 50, 100, 200, 400$.

	(γ, r, q)	$T = 50$	$T = 100$	$T = 200$	$T = 400$
%(p-LSA)	(1, 1.2, 0.3)	32.1	70.4	96.8	99.9
% (unified LASSO)	(1, 1.2, 1)	5.9	17.4	47.1	79.6
% (Bridge type)	(0, 1, 0.3)	8.9	21.9	52.3	80.3

Table: The summary of results for the simulation under $T = 200$.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
true	2	-1	1	-0.5	-1.5	1.5	0.5	0.75	0	0
initial	1.9938	-0.9936	0.9941	-0.5003	-1.4909	1.4978	0.5009	0.7490	0.0001	-0.0002
	(0.0722)	(0.0830)	(0.0858)	(0.0889)	(0.0962)	(0.0745)	(0.0851)	(0.0880)	(0.0908)	(0.0974)
p-LSA	1.9918	-0.9872	0.9877	-0.4758	-1.4877	1.4946	0.4750	0.7383	-0.0001	-0.0008
	(0.0728)	(0.0840)	(0.0868)	(0.1035)	(0.0965)	(0.0748)	(0.1049)	(0.0904)	(0.0265)	(0.0340)
P-O	1.9959	-0.9971	0.9964	-0.4979	-1.4931	1.4998	0.4946	0.7523	-0.0003	-0.0007
	(0.0565)	(0.0682)	(0.0713)	(0.0892)	(0.0823)	(0.0602)	(0.0915)	(0.0730)	(0.0228)	(0.0307)
%(p-LSA)	100.0	100.0	100.0	98.8	100.0	100.0	98.4	100.0	99.4	99.2

	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}	θ_{17}	θ_{18}	θ_{19}	θ_{20}
true	0	0	0	0	0	0	0	0	0	0
initial	0.0014	-0.0038	0.0007	0.0026	0.0063	-0.0015	0.0042	0.0013	-0.0048	-0.0014
	(0.0760)	(0.0794)	(0.0872)	(0.0947)	(0.0971)	(0.0701)	(0.0802)	(0.0896)	(0.0914)	(0.0941)
p-LSA	0.0000	0.0003	0.0007	0.0002	-0.0004	0.0000	0.0011	-0.0003	0.0013	0.0004
	(0.0000)	(0.0185)	(0.0252)	(0.0240)	(0.0248)	(0.0000)	(0.0194)	(00239)	(0.0247)	(0.0208)
P-O	0.0000	-0.0003	0.0007	0.0000	-0.0003	0.0000	0.0007	-0.0004	0.0010	0.0004
	(0.0000)	(0.0152)	(0.0221)	(0.0188)	(0.0259)	(0.0000)	(0.0134)	(0.0214)	(0.0199)	(0.0213)
%(p-LSA)	100.0	99.7	99.4	99.4	99.5	100.0	99.7	99.5	99.7	99.5

欠測データを用いたフィッシャースコアリング法

関西大学 高井啓二

本発表では，欠測データにもとづいて最尤推定値を計算する手法として不完全データのフィッシャースコアリング法 (Incomplete-data Fisher Scoring; IFS) を提唱する．IFS の導出方法，収束の性質，そして実際の適用例を示す．

従来，欠測データにもとづいて最尤推定値を計算するには，EM アルゴリズムが標準的に用いられてきた．現在では，EM アルゴリズムは単に欠測データにもとづいてパラメータを推定するだけでなく，潜在変数モデルのパラメータの推定や一般の推定関数のパラメータ推定にも利用されてきている．現在のところ，欠測データにもとづいてパラメータを推定する際には，EM アルゴリズムは事実上の第一選択候補となっている．

EM アルゴリズムの利点は，その更新式の導出がしやすいこと，そして更新式が単純な形になることである．これは実装のしやすさや，更新式の解釈のしやすさに直結している．しかし，それは EM アルゴリズムが最も優れてたアルゴリズムであることを意味してはいない．EM アルゴリズムは，従来から収束が遅いことが指摘されてきた．ここで言う収束の遅さとは，収束までの繰り返し回数のことである．EM アルゴリズムは，他のアルゴリズムに比べて収束までに要する繰り返しの計算回数が多くなる傾向にある．このような問題点を解決すべく，様々な加速法が試みられてきた．例えば，doubling step 法や，Aitken 加速などである．一般にこういった方法は，本質的には目的関数の二階微分の情報を使うことになる．それゆえ，最初から目的関数の二階微分あるいはそれに相応するような情報を使うことが合理的であろう．そういった考え方にもとづいて，EM アルゴリズムの情報とその他の情報 (例えば勾配) を利用する hybrid EM アルゴリズムや，EM のアイディアを用いながらも EM アルゴリズムを使わない更新式 (疑ニュートン法や経験的なフィッシャー情報行列を用いたフィッシャースコアリング法) が開発されてきた．

本研究で提唱する IFS は，前者と後者の中間に位置付けることができる．IFS はステップ幅を調整することによる加速法である．具体的には，パラメータ $\boldsymbol{\alpha}$ の現在の推定値 $\boldsymbol{\alpha}^{(t)}$ から新たな推定値 $\boldsymbol{\alpha}^{(t+1)}$ への更新を

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + s\frac{1}{n}\left(J_{\mathrm{com}}(\boldsymbol{\alpha}^{(t)})\right)^{-1}\nabla\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)})$$

で行なう．ここで，n はサンプルサイズ，s はステップ幅，$J_{\mathrm{com}}(\boldsymbol{\alpha})$ は完全データのフィッシャー情報行列である．IFS の特筆すべき性質は，例えば正規分布の場合には $s = 1$ とすると，(同一の初期値であれば)EM アルゴリズムと全く同一の推定値の列を生成するということである．このことは，$s > 1$ とすることにより，EM アルゴリズムよりも速い収束が見込めるということでもある．そして，ステップ幅 s を適切に調節することにより，より高速な収束を実現できるであろう．

EM が持っていた目的関数の単調増加性については IFS も有している．IFS による尤度の更新は，

$$\begin{aligned}\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t+1)}) - \ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)}) &\approx \nabla\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)})'(\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)}) \\ &= \frac{s}{n}\nabla\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)})'\left(J_{\mathrm{com}}(\boldsymbol{\alpha}^{(t)})\right)^{-1}\nabla\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)})\end{aligned}$$

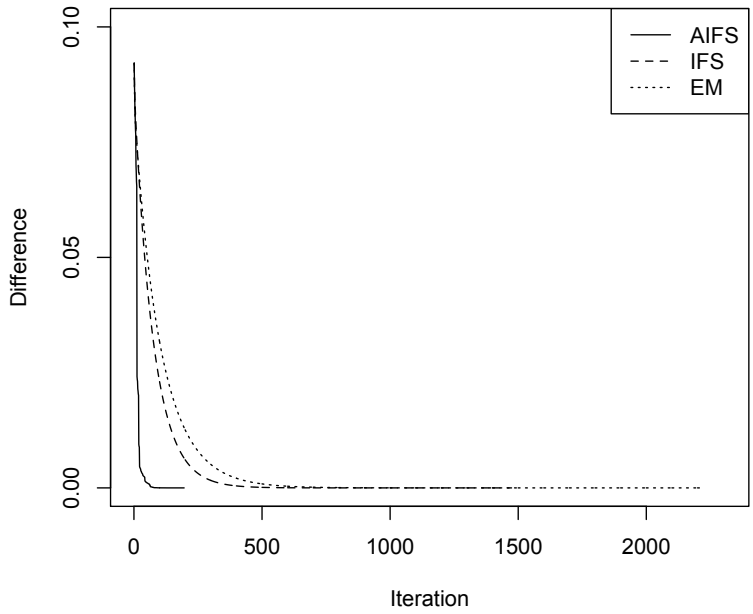


図 1: 収束の過程

で与えられる．最後の項の $J_{\mathrm{com}}(\boldsymbol{\alpha})$ が正定値であれば，$\ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t+1)}) - \ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)}) > 0$ は明らかである．一般的な統計モデルにおいて，フィッシャー情報行列が正定であることを仮定することは決して制限的ではないであろう．従って，一般的な統計モデルに対して (ステップ幅 s が適切である限りは)，目的関数の単調増加性を担保できる．

本研究では，この直感に理論的な裏づけを与えた．具体的には，(1) IFS を単調に増加させるステップ幅 $(s > 0)$ が必ず存在すること，(2) IFS の収束の速さは $\nabla^2 \ell_{\mathrm{obs}}(\boldsymbol{\alpha}^{(t)}) J_{\mathrm{com}}(\boldsymbol{\alpha}^{(t)})^{-1}$ の固有値に依存すること，である．

最後に，IFS を単変量混合ポアソン分布のパラメータ推定に適用した結果を手短に報告する．1910 年から 1912 年の London Times 紙の死亡人数のデータ (Hasselblad, 1969; Titterington, 1984) には，二つのポアソン分布の混合分布が良く適合することが知られている (Lange, 1995)．この分布のパラメータ推定に IFS を用いてみる．図 1 は，EM アルゴリズムによる推定過程，ステップ幅を 2 固定した IFS の推定過程，最適なステップ幅にした IFS の推定過程を示している (それぞれ図中では EM，IFS，AIFS と表記)．横軸は繰り返し回数であり，縦軸は収束先の尤度とその時点での尤度の差である．どの方法も単調に収束していることが分かる．収束までに，EM アルゴリズムは 2000 回以上，ステップ幅を固定した IFS は 1500 回程度，ステップ幅を調整した IFS は 200 回程度の繰り返し計算が必要となる．ステップ幅を調整した IFS は EM アルゴリズムの 1/10 ほどの繰り返し回数となって，適切なステップ幅を用いることで加速できることが分かる．

Bayesian Fractional Imputation With Nonignorable Nonresponse Data

Kosuke Morikawa[1,2], Yoshikazu Terada[1,3], Jae Kwang Kim[4]

Graduate School of Engineering Science, Osaka University[1]
Earthquake Research Institute, The University of Tokyo[2]
RIKEN AIP[3]
Department of Statistics, Iowa State University[4]

08/30/2019
シンポジウム「高次元複雑データの統計モデリング」
@九州大学伊都キャンパス

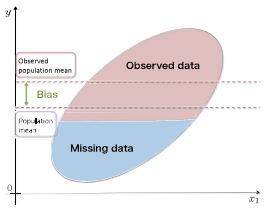
Missing data problems

- Missing data has become a major problem in statistical analysis, especially in social science, epidemiology, marketing, and so on.
  - Nonresponse
  - Causal inference
- A report on missing data was issued in National Research Council
  - ⇒ It has been a big problem in the world
- Our main goal is to obtain an estimator which is close to that with complete data

KLIPS data

- Korea Labor and Income Panel Survey (KLIPS) data
  - Sample size: 2506
  - Y: Income in year 2008 (missing)
  - X_1: Income in year 2007
  - X_2: gender (1 or 2)
  - X_3: age (1, 2 or 3)
- If the data Y of lower-income workers are likely to miss, the average income should overestimate the population mean income
- How can we estimate the mean income in 2008 by using only observed data?

Scatter plot of (X_1, Y)



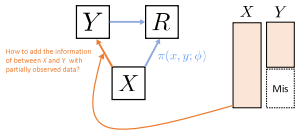
What are Nonignorable Nonresponse Data ?

- Nonresponse Data
  - Y: response variable (missing)
  - X : covariate vector
  - R : response indicator of Y
- Nonignorable Data (NMAR)
  - Response mechanism $P(R = 1 \mid x, y) = \pi(x, y)$ depends on Y
  - If the mechanism does NOT depend on Y, it is called ignorable or MAR

Problems and Goal of this talk

- Problems.
  - In estimating NMAR data, an outcome model $f(y \mid x)$ is to be modeled as well as a response model $\pi(x, y; \phi)$
- Goal. We propose,
  (i). (Frequnetist) an empirical likelihood (EL) type semiparametric estimator.
  (ii). (Bayesian) a multiple imputation (MI) method with the EL estimator without specifying any outcome model.
- In common, we do not require $f(y \mid x)$ distribution

Conceptual graph of the problem



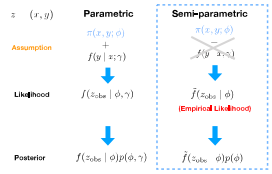
Parametric assumption on $\pi(x, y; \phi)$

- Logistic model is considered in this talk:

$$\pi(x, y; \phi) = \frac{1}{1 + \exp(\phi_0 + \phi_1 x + \phi_2 y)}$$

  - $\phi_1 \neq 0 \Rightarrow$ NMAR
  - $\phi_2 = 0 \Rightarrow$ MAR
- Maximum likelihood can NOT be used without specifying $f(y \mid x)$ (Greenlees et al., 1982, JASA)

How to conduct Bayesian estimation for semiparametric model?



Outline

Introduction
Empirical Likelihood (EL)
Multiple Imputation (MI)

Proposed method
EL in NMAR
MI in NMAR

Numerical Experiment

Outline

Introduction
Empirical Likelihood (EL)
Multiple Imputation (MI)

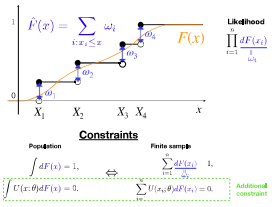
Proposed method
EL in NMAR
MI in NMAR

Numerical Experiment

Characteristics of EL

- Semiparametric method (Owen, 1988, BMK; Owen, 1990, AoS; Qin & Lawless 1994, AoS)
- For statistical tests of EL estimators, variance estimation is unnecessary (Wilks' theorem)
- Target parameter θ: a unique solution to $E\{U(\theta; X, Y)\} = 0$
  - e.g. $\theta = E(Y) \Rightarrow U(\theta) = \theta - Y$.
    θ : regression coefficient $\Rightarrow U(\theta) = A(X)(Y - X\theta)$
    $A(X) = X^\top$: LSE
    $A(X) = \partial(X\theta)/\partial\theta^\top V^{-1}(X)$: GEE

Illustration of EL



Qin and Lawless (1994, AoS)'s EL estimator

Maximum Empirical Likelihood Estimator (MELE)

- $\hat{\theta} = \arg\max_{\theta} \arg\max_{\omega_1, \ldots, \omega_n} \prod_{i=1}^{n} \omega_i$
  $= \arg\max_{\theta} \prod_{i=1}^{n} \hat{\omega}_i(\theta)$ ← profile likelihood
- subject to
  $\sum_{i=1}^{n} \omega_i = 1, \quad \omega_i \geq 0, \quad \sum_{i=1}^{n} \omega_i U(x_i; \theta) = 0.$

Note.

- $\hat{\omega}_i(\theta)$ has an explicit form by using the Lagrange multiplier
- $\prod_{i=1}^{n} \hat{\omega}_i(\theta)$ can be regarded as a (pseudo) likelihood

No plug-in estimator is required in statistical test

Wilks' theorem in EL (Theorem 2 in Qin & Lawless, 1994)

- Empirical weight: $\hat{\omega}_i(\theta) = \frac{1}{n\{1 + \lambda^\top U(x_i; \theta)\}}$
- Empirical likelihood: $L(\theta) = \prod_{i=1}^{n} \hat{\omega}_i(\theta)$
- Under some regularity conditions,
  $$-2 \log \frac{L(\theta_0)}{L(\theta)} \xrightarrow{D} \chi^2(d),$$
  - d: dimension of θ

EL estimator in NMAR

- Qin et al. (2002, JASA)'s estimator:
  - $f(y \mid x)$ is unnecessary
  - Consistency and asymptotic normality (CAN)
- Two shortcomings:
  1. Efficiency is unknown (asymptotic variance is too complicated)
  2. Wilks' theorem may not hold

Multiple imputation –1/2–

- Create multiple complete datasets by imputing missing values with some reasonable ones (Rubin, 1978, 1987; Kim & Shao, 2013)
  - It is easy and convenient for users to deal with missing data
  - Variance estimation can be conducted by the Rubin's rule
- Let $z = (x, y)$, $\xi = (\phi, \gamma)$; γ is a parameter prescribing $f(y \mid x; \gamma)$

(P-step) For $j = 1, \ldots, M$, $\xi^{(j)} \sim p(\xi \mid z_{\text{obs}}) \propto p(z_{\text{obs}} \mid \xi) p(\xi)$

(I-step) For $i = 1, \ldots, n_0$, $y_i^{(j)} \sim p(y \mid x, r = 0; \xi^{(j)})$

where $p(\xi)$ is a prior distribution of ξ

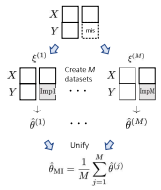
Multiple imputation –2/2–

- Compute interesting parameter $\hat{\theta}^{(j)}$ with the j-th complete dataset $(j = 1, \ldots, M)$

Rubin's rule

$$\hat{V}(\hat{\theta}_{\text{MI}}) = W_M + \left(1 + \frac{1}{M}\right) B_M,$$

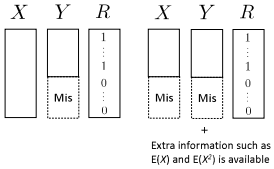
- $W_M \quad \frac{1}{M} \sum_{j=1}^{M} \hat{V}(\hat{\theta}^{(j)})$
- $B_M = \frac{1}{M-1} \sum_{j=1}^{M} \left(\hat{\theta}^{(j)} - \hat{\theta}_{\text{MI}}\right)^{\otimes 2}$



Multiple imputation in NMAR

- In P-step, some outcome model $f(y \mid x; \gamma)$ is to be specified
  - Otherwise, posterior can not be computed
  - In Im and Kim (2017, JKSS), $f_1(y \mid x) = f(y \mid x, r = 1)$ is specified instead of f
- Semiparametric Bayesian estimation is required (Chernozhukov and Hong, 2003, JoE; Lazar, 2003, BMK; Yin, 2009, BA; Chaudur et al., 2017, JRSSB)

Dataset our method can be applied



Outline

Introduction
Empirical Likelihood (EL)
Multiple Imputation (MI)

Proposed method
EL in NMAR
MI in NMAR

Numerical Experiment

Motivation

- Let $g(x)$ be any integrable function of x.
  - e.g. $g(x) = \{1, x, x^2, \ldots\}$
- By Bayes' formula, it always holds that

$$E\{g(X)\} = \frac{E_1\{\pi^{-1}(X, Y) g(X)\}}{E_1\{\pi^{-1}(X, Y)\}} = E_1\left\{\frac{p g(X)}{\pi(X, Y)}\right\}$$
$$\Rightarrow E_1\left\{\frac{p g(X)}{\pi(X, Y)} - E\{g(X)\}\right\} = 0,$$

where $p = P(R = 1)$ and $E_1(\cdot) = E(\cdot \mid r = 1)$.

Proposed estimator for $\zeta = (\theta, \phi)$

Maximum Empirical Likelihood Estimator (MELE):

$$\hat{\zeta} = \arg\max_{\zeta} \arg\max_{\omega_1, \ldots, \omega_{n_1}} \prod_{r_i=1} \omega_i$$

s.t.

$$\sum_{r_i=1} \omega_i = 1, \quad \omega_i \geq 0$$
$$\sum_{r_i=1} \omega_i \left\{\bar{p} \frac{g(x_i)}{\pi(x_i, y_i; \phi)} - \bar{g}_n\right\} = 0.$$
$$\sum_{r_i=1} \omega_i \bar{p} \frac{U(x_i, y_i; \theta)}{\pi(x_i, y_i; \phi)} = 0.$$

where $\bar{g}_n \quad n^{-1} \sum_{i=1}^{n} g(x_i)$.

On Empirical Weights $\hat{\omega}_i$

- By using the Lagrange multiplier,

$$\hat{\omega}_i(\zeta) = \frac{1}{n_1\{1 + \lambda^\top h(x_i; \zeta)\}},$$

  - $h(x; \zeta) = [\{\bar{p} g(x)^\top \pi^{-1}(x_i, y_i; \zeta) - \bar{g}_n^\top\}, \{\bar{p} U(\theta)^\top \pi^{-1}(x_i, y_i; \zeta)\}]^\top$.
  - λ satisfies
  $$\sum_{r_i=1}^{n} \frac{h(x_i; \zeta)}{1 + \lambda^\top h(x_i; \zeta)} = 0$$
- $\hat{\zeta} = \arg\max_{\zeta} \prod_{r_i=1} \hat{\omega}_i(\zeta)$

Theorem 1. CAN of MELE

Theorem 1.

Under some regularity conditions, MELE $\hat{\zeta}$ has CAN and the asymptotic distribution can be written as

$$\sqrt{n}(\hat{\zeta} - \zeta_0) \xrightarrow{D} \mathcal{N}(0,\ \Sigma),$$

- $\Sigma^{-1} = E\left\{\frac{\partial \boldsymbol{k}(\zeta)}{\partial \zeta^\top}\right\}\left\{E\left(\boldsymbol{k}(\zeta)^{\otimes 2}\right)\right\}^{-1}\left[E\left\{\frac{\partial \boldsymbol{k}(\zeta)}{\partial \zeta^\top}\right\}\right]^\top$
- $\boldsymbol{k}(r, \boldsymbol{z}; \zeta) = \left[\left\{\frac{r}{\pi(\boldsymbol{z};\phi)} - 1\right\} g(x)^\top,\ \frac{r}{\pi(\boldsymbol{z};\phi)} U(x, y; \theta)^\top\right]^\top$.

Remark

- Dimension of g must be larger than that of ϕ
  - When $\dim(g) = \dim(\phi)$, the estimator is exactly the same as GMM-based estimator (Kott & Chang, 2010, JASA)
- There exists a "best" g-function which minimizes the asymptotic variance of $\hat{\zeta}$ (Morikawa & Kim, 2017).
  - The lower bound is called the semiparametric efficiency bound
- Our estimator is applicable for the data in which both x and y are missing, but we know $E(X)$ and $E(X^2)$ such as the Census Bureau

Empirical Likelihood Ratio

- Empirical Likelihood:

$$L(\phi) \prod_{r_i=1} \hat{\omega}_i(\phi) \qquad \sum_{r_i=1} \log\left[n_1\{1 \mid \lambda(\phi)^\top h(x_i;\phi)\}\right]$$

- Likelihood Ratio Statistics:

$$R(\phi) = -2\log\frac{L(\phi_0)}{L(\hat{\phi})}$$

  - ϕ_0 : true value (or null hypothesis: $\phi = \phi_0$)

Theorem 2. Wilks' theorem

Theorem 2.

Let ζ be a d-dimensional parameter. Then, we have the following results:

- Test $H_0 : \zeta = \zeta_0$.
  - $\hat{\zeta}_{0L}$ be MELE

$$R(\hat{\zeta}_{\mathrm{EL}}) \xrightarrow{D} \chi^2(d). \tag{1}$$

- Test $\tilde{H}_0 : \zeta^{(1)} = \zeta_0^{(1)}$,
  - $\zeta^{(1)}$ is a sub-vector of $\zeta = (\zeta_1^\top, \zeta_2^\top)^\top$
  - $\hat{\zeta}_{0L} = (\zeta_{10}^\top, \hat{\zeta}_{2,\mathrm{EL}}^\top)^\top$ be MELE with $\zeta = (\zeta_{10}, \zeta_2)$ and fixed ζ_{10}

$$R_2(\hat{\zeta}_{\mathrm{EL}}, \tilde{\zeta}_{\mathrm{EL}}) \xrightarrow{D} \chi^2(d - d_1), \tag{2}$$

where

$$R_2(\zeta_1, \zeta_2) = -2\log\frac{L(\zeta_1)}{L(\zeta_2)}$$

Semiparametric test of response mechanism: MAR v.s. NMAR

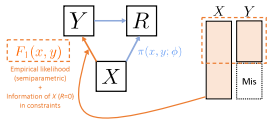
- Consider a simple situation that we are only interested in ϕ
- Logistic response model:

$$\pi(x, y; \phi) \qquad \frac{1}{1 + \exp(\phi_0 + \phi_1 x + \phi_2 y)}$$

  - $\hat{\phi}_{2\mathrm{L}}$: MELE with $\phi = (\phi_0, \phi_1, \phi_2)$
  - $\hat{\phi}_{21}$: MELE with $\phi = (\phi_0, \phi_1, 0)$
- If the null hypothesis was correct, we have

$$R_2(\tilde{\phi}_{\mathrm{EL}}, \hat{\phi}_{\mathrm{EL}}) \xrightarrow{D} \chi^2(1),$$

Conceptual graph of the problem



Multiple Imputation in NMAR –P-step–

- Consider an EL

$$\prod_{r_i=1} \omega_i(\phi)$$

s.t.

$$\sum_{r_i=1} \omega_i \quad 1, \quad \omega_i > 0$$

$$\sum_{r_i=1} \omega_i \left\{\hat{p}\frac{q(x_i)}{\pi(x_i, y_i; \phi)} - \hat{y}_H\right\} = 0.$$

- We can regard this likelihood as a pseudo likelihood!!
- Empirical Posterior Distribution (Lazar, 2003, BMK):
  Posterior of ϕ can be sampled from a distribution proportional to

$$\underbrace{p(\phi)}_{\text{prior on } \phi} \prod_{r_i=1} \omega_i(\phi)$$

Multiple Imputation in NMAR –P-step–

- Recently, Chauduri et al. (2017, JRSSB) shows that Hamiltonian Monte Carlo (HMC) method is effective to sample from the empirical posterior distribution
- An R package dlhmc is available

Multiple Imputation in NMAR –I-step–

- In a similar way to Kim & Yu (2011, JASA), we have

$$f(y \mid x, r = 0) = \frac{f(y, x \mid r \quad 1)O(x, y)}{\int f(y, x \mid r = 1)O(x, y)dy},$$

  - $O(x, y) = \pi^{-1}(x, y) - 1$
- Kernel density estimator with EL weights (Chen, 1997, AJS):

$$\hat{f}(y, x \mid r = 1) = \frac{1}{h^{p+1}} \sum_{r_i=1} w_i(\phi) K_x\left(\frac{x - x_i}{h}\right) K_y\left(\frac{y - y_i}{h}\right),$$

  - x: p-dimensional covariate
  - h: bandwidth
  - K_x, K_y: kernel function of x and y

Multiple Imputation in NMAR –I-step– MCMC is not needed!!

Proposition 1.

Assume that

- Response model: $\pi(x, y; \phi) = \frac{1}{1 - \exp(\phi_0 + \phi_1^\top x + \phi_2 y)}$
- K_y is the density of the standard normal distribution

Then, the imputation model becomes a normal mixture model:

$$\hat{f}_0(y \mid x; \phi) = \sum_{i=1}^{n_1} \varpi_i(x)\mathcal{N}(y_i + h^2\varphi_2, h^2),$$

where $\mathcal{N}(\mu, \sigma^2)$ is the density function of normal distribution with mean μ and variance σ^2,

$$\varpi_i(x) \quad \frac{d_i(x)}{\sum_i d_i(x)} \quad \text{and} \quad d_i(x) \quad \omega_i(\phi) K_x\left(\frac{x - x_i}{h}\right) \exp(\phi_2 y_i).$$

Summary of proposed MI

- P-step
  - Sample parameters from the empirical posterior distribution with HMC

$$\phi^{(j)} \sim p(\phi \mid z_{\mathrm{obs}}) \propto p(\phi) \prod_{r_i=1} \omega_i(\phi) \quad (j = 1, \ldots, M)$$

- I-step
  - For each missing data $\{i : r_i = 0\}$, impute them by

$$y_i^{(j)} \sim \hat{f}_0(y \mid x; \phi^{(j)}) = \sum_{i=1}^{n_1} \varpi_i(x)\mathcal{N}(y_i + h^2\phi_2^{(j)}, h^2),$$

- The rest is same as the classical MI

Outline

Introduction
Empirical Likelihood (EL)
Multiple Imputation (MI)

Proposed method
EL in NMAR
MI in NMAR

Numerical Experiment

Setup –1/2–

- $X \sim U(-1, 1)$; $Y \mid (x, r = 1) \sim N(x - 0.121,\ 1/\sqrt{3}^2)$
- Response mechanism:

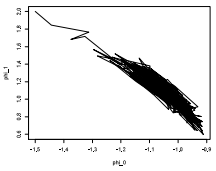
$$\pi(x, y) \qquad \frac{1}{1 + \exp(-0.95 + 1.0y)}$$

- Target parameter
  - $E(Y)$
  - β_1 : $E(Y \mid x) \quad \beta_0 \mid \beta_1 x$

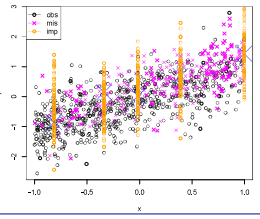
Setup –2/2–

- $n = 1,000$
- #Imputation: 10, 100 or 1,000
- #repricaton: 1,000
- $q(x) \quad (1,\ x,\ x^2,\ x^3)$
- Prior: $p(\phi_0) = \mathcal{N}(\log\{(1 - \hat{p})/\hat{p}\},\ 100^2) \quad p(\varphi_1) = \mathcal{N}(0,\ 100^2)$
  - Assume MCAR as the prior

Trajectory of sampling parameters at P-step



Imputed data in a dataset with $M = 100$



Results

Table: Results of numerical simulation.

	$E(Y)$			β_1		
M	10	100	1,000	10	100	1,000
BIAS	0.024	0.024	0.024	0.001	0.001	0.001
SE	0.037	0.034	0.034	0.043	0.042	0.042
RMSE	0.198	0.190	0.189	0.212	0.209	0.208
TEST	0.931	0.961	0.966	0.942	0.943	0.947

TEST: Rate of the confidence interval with significance level $\alpha = 0.05$ including the true value

Conclusion

- We proposed two semiparametric methods to analyze NMAR data
  1. (Frequentist) Empirical Likelihood (EL)
  2. (Bayesian) Multiple Imputation (MI)
- Partial information can be effectively incorporated
- Theoretical results for MI are needed
  - In what conditions, Rubin's rule holds? bandwidth?
- The assumption on the response model can be relaxed such as

$$\pi(x, y; \phi) = \frac{1}{1 - \exp\{h(x) + \phi y\}},$$

where $h(x)$ is an arbitrary function (Shao & Wang, 2016, BMK)

Spectral Embedded Deep Clustering

Yuichiro WADA[1], Shugo MIYAMOTO[3], Takumi NAKAGAWA[2], Léo Andéol[4,5], Wataru KUMAGAI[5], and Takafumi KANAMORI[2,5]

[1]Nagoya University, Furocho, Chikusaku, Nagoya 464-8601 Japan
[2]Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552 Japan
[3]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan
[4]Sorbonne Université, 4 place Jussieu 75005 Paris, France
[5]RIKEN AIP, Nihonbashi 1-chome Mitsui Building, 15th floor,1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027 Japan

1 Abstract

We propose a new clustering method based on a deep neural network. Given an unlabeled dataset and the number of clusters, our method directly groups the dataset into the given number clusters in the original space. We use a conditional discrete probability distribution defined by a deep neural network as a statistical model. Our strategy is, first to estimate the cluster labels of unlabeled data points selected from high density region, and then to conduct semi-supervised learning to train the model by using the estimated cluster labels and the remaining unlabeled data points. Lastly, by using the trained model, we obtain the estimated cluster labels of all given unlabeled data points. The advantage of our method is that it does not require key condition. Existing clustering methods with deep neural networks based assumed that the cluster-balance of given dataset is uniform. Moreover, it also can be applied to various data domains as long as the data is expressed by a feature vector. In addition, it was observed that our method was robust against outliers. Therefore, the proposed method is expected to averagely perform better than previous methods. We conducted numerical experiments on five commonly used datasets to confirm the effectiveness of the proposed method.

2 Proposed Method

Given an unlabeled dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ ($\mathbf{x}_i \in \mathbb{R}^D$) and the number of clusters C, our proposed deep clustering named SEDC (*Spectral Embedded Deep Clustering*) groups $\mathbf{X}$ into C clusters. Since this grouping is achieved by obtaining the estimated cluster labels of $\mathbf{X}$, our goal can be replaced by estimating the cluster labels up to permutation of labels. In SEDC, the estimated cluster label of each $\mathbf{x}_i \in \mathbf{X}$ is defined by $\mathrm{argmax}_{j=1,\ldots,C} p_{\theta^*}(y = j|\mathbf{x}_i)$, where θ^* is the trained set of parameters. The classifier $p_\theta(y|\mathbf{x})$ is parameterized by a fully connected deep neural network. The training scheme of this classifier is as follows: we firstly only estimate the cluster labels of selected unlabeled data points by using only $\mathbf{X}$ (this part is done by SGSC (*Selective Geodesic Spectral Clustering*) algorithm.), and then conduct semi-supervised learning to train the classifier. Regarding with this semi-supervised learning, we use the estimated cluster labels of selected unlabeled data points and the remaining unlabeled data points, which are treated as the given true cluster labels and unlabeled data points respectively. The objective function of this semi-supervised learning is as follows:

$$\mathcal{R}_{\mathrm{VAT}}(\theta; \mathbf{X}) + \frac{\lambda_1}{h} \sum_{i=1}^{h} \mathbb{KL}\left[p_\theta(y|\mathbf{x}_{(i)}) \| q_{(i)}\right] + \lambda_2 \mathbb{H}(\mathbf{Y}|\mathbf{X}), \tag{1}$$

where the first, second and third terms are VAT loss, pseudo empirical loss with estimated cluster probabilities ($q_{(i)}$) of selected data points based on KL divergence [1] and conditional entropy loss [1], respectively. λ_1 and λ_2 are hyperparameters that range over positive numbers. y and h mean the cluster label and the number of selected data points, respectively. Eq.(1) will be minimized.

Table 1: The mean clustering accuracy (ACC) of Eq.(2) and standard deviation are shown. Five popular clustering methods and our proposed method were tested on five datasets. MNIST and Reuters are real-world datasets. FC (Four Clusters), TM (Two Monns) and TR (Three Rings) are artificial datasets. For each method, Average means the averaged ACC over the five datasets. The experiments were conducted seven times on each pair of method and dataset.

Method	MNIST	Reuters-10k	FC	TM	TR	Average
k-means [4]	0.53	0.53(0.04)	0.60(0.05)	0.64(0.04)	0.35(0.03)	0.53
SC [5]	0.72	0.62(0.03)	0.80(0.04)	0.85(0.03)	0.96(0.03)	0.79
IMSAT [2]	**0.98**	0.71(0.05)	0.70(0.04)	0.66(0.05)	0.34(0.01)	0.68
DEC [7]	0.84	**0.72(0.05)**	0.72(0.04)	0.67(0.03)	0.48(0.04)	0.69
SpectralNet [6]	0.83	0.67(0.03)	0.79(0.03)	0.87(0.02)	**0.99(0.01)**	0.83
SEDC	0.89	**0.73(0.05)**	**0.95(0.03)**	**0.96(0.02)**	**0.99(0.00)**	**0.90**

3 Numerical Experiment and Conclusion

Table 1 shows the clustering accuracy of each methods. Our method is SEDC. The evaluation metric is as follows:

$$\mathrm{ACC} = \max_{\tau} \frac{\sum_{i=1}^{n} \mathbf{1}[y_i = \tau(\hat{y}_i)]}{n}, \tag{2}$$

where $\{y_i\}_{i=1}^n$ and $\{\hat{y}_i\}_{i=1}^n$ be its true cluster label set and estimated cluster label set, respectively. The number of data points is denoted by n. τ ranges over all permutations of cluster labels, and $\mathbf{1}[\cdots]$ is the indicator function. The optimal assignment of τ can be computed using the Kuhn-Munkres algorithm [3].

With respect to the conclusion, we propose a deep clustering method named SEDC. Given an unlabeled dataset and the number of clusters, the method groups the dataset into the given number clusters. Regarding its advantages, it does not require an additional condition except two fundamental assumptions: smoothness and manifolds assumptions. In addition, SEDC also can be applied to various data domains since it does not have preferred data domains, as long as raw data is transformed to feature vectors. Furthermore, the performance of SEDC can be robust against existence of outliers. According to these advantages, our proposed method can be expected to averagely perform better than previous deep clustering methods.Therefore, we think our method can be a competitive candidate for users in some practical clustering scenarios where prior knowledge of the given unlabeled dataset is limited.

References

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[2] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *International Conference on Machine Learning*, pages 1558–1567, 2017.

[3] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[4] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[5] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[6] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*, 2018.

[7] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

科研費シンポジウム「高次元複雑データの統計モデリング」2019年8月30日発表「距離空間の変形と埋め込みを用いたデータ解析」報告書

慶應義塾大学　小林　景

1　発表概要

本発表ではまず，リーマン多様体や距離空間上のFréchet平均，extrinsic平均およびFréchet関数に関する既存研究を概説した．これらはベクトルデータで我々が通常用いている「平均」を，線形とは限らない距離空間（例えば球面）上に一般化する必要があるときに自然に導かれる統計量である．次に，これらの研究分野の展開の一つとして，[1] において著者らが提案した距離空間の変形を用いたデータ解析手法およびその理論を簡単に紹介した．

2　Fréchet平均（intrinsic平均）とExtrinsic平均

まず，距離空間上の確率分布もしくは標本分布に対するFréchet平均およびextrinsic平均について説明した．また漸近理論を用いることにより，Fréchet平均に対する仮説検定（$H_0:\ \mu = \mu_0$ vs $H_1:\ \mu \neq \mu_0$）や2標本のFréchet平均の差異に関する検定，もしくは$\varphi(\mu)$の信頼区間の構成といった通常のユークリッド空間上の統計学的手法をリーマン多様体上に一般化することが可能となることを説明した．

次に，Fréchet平均以外の距離空間上の「平均」の候補として，Fréchet平均とくらべて一般に算出が容易なextrinsic平均を紹介した．extrinsic平均は，ユークリッド空間に埋め込まれた距離空間に対して，ユークリッド距離を用いるFréchet平均として定義される．発表においてはextrinsic平均の射影定理を紹介した後，その算出および漸近的議論がFréchet平均にくらべて容易になる理由を説明した．

extrinsic平均は射影定理を用いると算出や漸近理論が平易化されるという大きなメリットがあるが，そもそも距離空間がユークリッド空間に埋め込めないような場合には用いることができない．特に，今回紹介する研究で提案するようにデータ空間の距離を変換する際には，ユークリッド空間への埋め込みはすぐに成り立たなくなる．このことを見るために，距離空間のユークリッド空間への埋め込み可能性について，Schoenbergによるよく知られた結果を紹介した．

3　Fréchet平均の一意性とCAT(k)空間

一般にFréchet平均集合もextrinsic平均集合もサイズが2以上になり得るため，平均の一意性は成立しない．extrinsic平均の場合は埋め込まれたユークリッド空間でのFréchet平均が一意に定まることから，一意性は直交射影の一意性，ひいてはユークリッド空間への埋め込まれ方に依存する．一方，Fréchet平均の一意性はデータ空間自身の距離から定まる曲率により特徴づけされる．発表においては，まずデータ空間がリーマン多様体であるときの曲率とFréchet平均の一意性の関係を簡単に紹介した後，より一般にLength space上の曲率（CAT(k)性）とFréchet平均の一意性について知られている結果についても紹介した．

一方，距離グラフ，多面体的複体，錐などは一般に微分可能な多様体ではなく，これまでの議論を当てはめることができない．そこで，上記のような幾何学的対象やリーマン多様体を含む Length space に理論を一般化した既存研究を紹介した．さらに，断面曲率 k 以下のリーマン多様体に対応する CAT(k) 空間の定義と性質を紹介した．

4 β 距離変換

次に，[1] において著者らが提案したデータ空間の距離変換手法について簡単に説明した．距離の変換は，曲率，特に CAT(k) 性に着目して提案された α 距離変換と β 距離変換とよばれる 2 段階の変換の合成として定義される．α 距離変換については簡単に説明した後，今回の発表の主題である β 距離変換の定義および計量錐への埋め込みとの関係を説明した．

まず，以下のように距離 d と凹関数 $g_\beta(\cdot)$ を合成することにより，距離 d_β を定義する．

$$d_\beta(x_0, x_1) = g_\beta(d(x_0, x_1)) \ \ (\beta > 0), \quad \text{ただし}$$

$$g_\beta(z) = \begin{cases} \sin(\frac{\pi z}{2\beta}), & \text{for } \ 0 \le z \le \beta, \\ 1, & \text{for } \ z > \beta \end{cases}$$

である．β の変化により，Fréchet 関数 $F(y) = \sum_i d_\beta(x_i, y)^\gamma$ の極小点（Karchar 平均）の数を調整できる．Fréchet 関数の極小点をクラスター中心の候補とすることにより，クラスタリングに応用することができる．これは Fréchet 関数が常に凸関数となるユークリッド距離やノルム空間では不可能な解析手法である．

一方，β 距離変換による曲率の変化は，α 距離変換と異なり直接的には議論できない．これは，β 距離変換をはじめとして，Length space(M, d) を凹関数 φ を用いて変換した距離空間 $(M, \varphi(d))$ は一般に Length space とならず，CAT(k) 性が定義できないためである．しかし β 距離変換特有の性質として，埋め込まれた「計量錐」の曲率の変化を評価することが可能である．$\mathcal{X}$ 上の距離 d_β は，埋め込まれた動径 $\beta/2$ の計量錐 $\tilde{\mathcal{X}}$ での extrinsic 距離となることから，$\mathcal{X}$ 上の確率分布や標本分布の Fréchet 平均は，計量錐に埋め込んだ場合の extrinsic 平均として解釈できる．また β が小さくなると計量錐の曲率が小さくなる傾向があることを示す定理を紹介した．

最後に，まとめと課題として，ユークリッド空間と計量錐の extrinsic 平均の違いについて紹介した．

参考文献

[1] Kobayashi, K., Wynn, H.P. (2019), Empirical geodesic graphs and CAT(k) metrics for data analysis, to appear in *Statistics and Computing*, https://doi.org/10.1007/s11222-019-09855-3.