2025 年度科学研究費シンポジウム

データサイエンスの基盤を支える
次世代統計理論・方法論の挑戦と革新

- 日時：2025 年 12 月 1 日（月）～ 12 月 3 日（水）
- 会場：九州大学 西新プラザ
- 開催責任者：川野 秀一（九州大学）
- 内容・目的：計測技術の高度化と AI 技術の急速な進展により、さまざまな分野で大規模かつ複雑なデータが日々取得されています。これらの豊富なデータから真に価値ある知見を導き出すためには、データサイエンスの核となる統計科学の理論および方法論の発展が不可欠です。本シンポジウムでは、統計科学や機械学習の理論・方法論に関する研究から実社会への応用事例まで、データサイエンスの最前線を探究する幅広い研究発表を募ります。データサイエンス分野における最新の研究動向や各分野が直面している課題について知見を共有し、さらにはこの機会を通じて若手研究者の育成を図ることで、データサイエンス研究の更なる発展を目指します。

※ 本シンポジウムは、以下の科学研究費補助金の助成を受けて開催されます。
科学研究費補助金 基盤研究(A) 25H01107「大規模複雑データの理論と方法論の深化と展開」研究代表者：青嶋 誠（筑波大学）

2025 年度科研費シンポジウム
「データサイエンスの基盤を支える次世代統計理論・方法論の挑戦と革新」
講演プログラム

12 月 1 日（月）

12:55 ~ 13:00　オープニング

13:00 ~ 13:30　岩重 文也（広島大学），橋本 真太郎（広島大学）
　　Bayesian Computation for a Mixture of Finite Mixtures
13:30 ~ 14:10　草野 彰吾（熊本大学），内田 雅之（大阪大学）
　　高頻度データに基づく拡散過程に対する SEM の擬似 BIC
14:10 ~ 14:50　寺西 蓮（久留米大学），江村 剛志（広島大学）
　　M スプライン基底を用いた，B スプラインコピュラの構成

14:50 ~ 15:20　休憩

15:20 ~ 15:50　瀬谷 のどか（東京医科大学），田栗 正隆（東京医科大学），藤後 修（東京医科大学），鎌谷 研吾（統計数理研究所）
　　過去データを利用した共変量適応的ランダム化法と対応するブートストラップ法の提案
15:50 ~ 16:30　中北 誠（理化学研究所 AIP），鳥谷部 智規（金沢学院大学），中妻 照雄（慶應義塾大学），星野 崇宏（慶應義塾大学，理化学研究所 AIP）
　　Algorithms for Fast Gibbs Sampling in Hierarchical Bayesian Panel Modeling
16:30~ 17:10　柳川 堯（久留米大学）
　　RxS 分割表による医療データの総括・統合とその解析

12 月 2 日（火）

10:00 ~ 10:30　海野 哲也（筑波大学），矢田 和善（筑波大学），青嶋 誠（筑波大学）
　　高次元共分散構造の自動スパース推定とその一致性
10:30 ~ 11:10　小林 亮太（東京大学），藤田 葵（東京大学），中山 悠理（東京大学），山本 泰智（東京大学）
　　TopiCLEAR: 文書ベクトルのクラスタリングに基づくトピック抽出技術
11:10 ~ 11:50　森本 孝之（関西学院大学），赤間 陽二（東北大学），川崎 能典（統計数理研究所）
　　Analyzing Japanese Equity Returns with Equi-Correlation Structures

11:50 ~ 13:30　休憩

13:30 ~ 14:00　Ziyue Wang（東北大学），荒木 由布子（東北大学）
　　Spectral decomposition in dynamic systems of distributional data
14:00 ~ 14:40　永井 勇（中京大学）
　　GMANOVA モデルの最尤推定量における直接的な罰則のプラグイン型最適化
14:40 ~ 15:20　Xin Guan（東北大学），寺田 吉壱（大阪大学）
　　Regularized k-POD clustering for missing data

15:20 ~ 15:50　休憩

15:50 ~ 16:20　栗田 絵梨（東京理科大学）
　　単調欠測データに対する Mardia の正規性検定統計量について
16:20 ~ 17:00　二宮 嘉行（統計数理研究所），栁原 宏和（大阪公立大学）
　　高次元スパース回帰のための AIC の漸近的性質
17:00 ~ 17:40　吉田 朋広（東京大学）
　　ディープラーニングと確率過程の統計推測

12 月 3 日（水）

10:00 ~ 10:30　日野 雅喜（総合研究大学院大学），加藤 昇吾（統計数理研究所），江村 剛志（広島大学）
　　ゼロ過剰ガンマフレイルティによる二変量治癒コピュラモデル―治癒率と生存時間の従属関係
10:30 ~ 11:10　貝野 友祐（神戸大学）
　　微小撹乱パラメータを持つ線形放物型確率偏微分方程式モデルのパラメータ推定およびその応用
11:10 ~ 11:50　入江 薫（東京大学）
　　形状制約下での関数パラメータの事後分析

11:50 ~ 13:30　休憩

13:30 ~ 14:00　平木 大智（東京大学），大森 裕浩（東京大学）
　　Dynamic factor stochastic volatility in mean model
14:00 ~ 14:30　梶谷 文乃（広島大学），江村 剛志（広島大学）
　　ジョイントフレイルティコピュラモデルを用いた欠測を含む肺腺癌データの解析
14:30 ~ 15:10　栁原 宏和（大阪公立大学）

Modified AIC for canonical-link GLMs with known scale parameter

15:10 ~ 15:15　クロージング

※　発表者が複数人の場合、下線は登壇者を表す。

報告書

岩重 文也, 橋本 真太郎
広島大学大学院先進理工系科学研究科

本発表では，有限混合の混合（mixture of finite mixtures, MFM）に対して近年開発されたベイズ計算アルゴリズムを中心的に取り上げた．Argiento and De Iorio (2022) で開発された blocked Gibbs sampler と Frühwirth-Schnatter et al. (2021) で開発された telescoping sampler は同時期に提案されており，両者の論文の中で二つのアルゴリズムの関係性については述べられていなかった．本発表では，telescoping sampler の導出方法に基づくことで，本質的に二つのアルゴリズムが同等であることを確認した．また，telescoping sampler は混合重みの分布が混合数 M に依存する dynamic MFM への適用を念頭に提案されていること，blocked Gibbs sampler はディリクレ分布以外の混合重みを使えることに注目し，telescoping sampler を dynamic MFM に適用可能となるように拡張した．本研究の理論的な成果として，アルゴリズムを拡張する際に必要な交換可能な分割確率関数（EPPF）と M の完全条件付き分布を導出した：

命題 1. DMFM の EPPF $p(\mathcal{C})$ は次で与えられる:

$$p(\mathcal{C}) = \int_0^\infty \frac{u^{n-1}}{\Gamma(n)} \Omega(u;k) \mathrm{d}u,$$

ただし，

$$\Omega(u;k) = \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u;m)^m \left\{ \prod_{j=1}^{k} \kappa(n_j;u,m) \right\} q_M(m+k).$$

また，$U_n = u$ および $\mathcal{C}$ が与えられた下で M_{na} の分布は，

$$p(M_{na} = m \mid U_n = u, \mathcal{C}) \propto \frac{(m+k)!}{m!} \left\{ \prod_{j=1}^{k} \kappa(n_j;u,m) \right\} \psi(u;m)^m q_M(m+k).$$

ディリクレ分布以外の例として，正規化逆ガウス分布に基づく DMFM を提案した．計算アルゴリズムの導出のために逆ガウス分布のキュムラントが必要であるが，第二種の修正ベッセル関数を用いて陽に表せることを示した．

命題 2. $\kappa_{\mathrm{IGau}}(n;u) := (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}u^n} \psi_{\mathrm{IGau}}(u)$ は，

$$\kappa_{\mathrm{IGau}}(n;u) = \alpha^n \psi_{\mathrm{IGau}}(u) (1+2u)^{-n/2} \frac{K_{n-1/2}\left(\alpha\sqrt{1+2u}\right)}{K_{1/2}\left(\alpha\sqrt{1+2u}\right)},$$

ただし，K_m はオーダー m の第三種の修正ベッセル関数である．

数値実験と実データ解析を通して正規化逆ガウス分布に基づく MFM, DMFM を取り上げ，ディリクレ分布に基づく MFM, DMFM との性能比較を行った．数値実験では，混合重みが非常に小

さい混合成分を設定し，観測データ数が少ない場合に出現しにくい混合成分が存在する状況を想定した．このような状況下において，正規化逆ガウス分布に基づく MFM, DMFM がディリクレ分布に基づく手法よりクラスタリングの精度が良くなることを確認した．特に，観測データが限られている状況において精度の差が顕著に現れた．この結果は，正規化逆ガウス分布がディリクレ分布よりも情報量が少ないため，データ数が少ない場合でも安定した事後解析が可能であることによって説明できる．

実データ解析においては，甲状腺データを用いてクラスタ数と混合数の事後解析を行った．正規化逆ガウス分布を用いた場合，ディリクレ分布よりも妥当な結果を得ることができた．また，M の事前分布にバリエーションを与えたが，事前分布の選択に結果が依存しないことを確認した．

今後の展望として，より広いクラスで統一的に MFM や DMFM の性質を記述することが挙げられる．例えば，h の分布として，無限分解可能な分布を仮定し，クラスタ数の事前分布やクラスタサイズの事前分布を解析しやすい形で導出すること，normalized completely random measures に基づく Bayesian nonparametric prior (James et al., 2010) との理論的関係などが挙げられる．また，ディリクレ分布に基づく MFM は Pitman–Yor 過程の特殊ケースであり，そこから得られる分割の分布は Gibbs type である (Gnedin and Pitman, 2006)．しかし，ディリクレ分布を仮定しても，DMFM の分割の分布は Gibbs type ではなく，一般の単体上の分布による MFM も同様である．したがって，Gibbs type を含む形で，DMFM や MFM から導出される分割の分布を統一的に記述できるかは興味深い問いである．

参考文献

Argiento, R. and M. De Iorio (2022). Is infinity that far? a bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics 50*(5), 2641–2663.

Frühwirth-Schnatter, S., G. Malsiner-Walli, and B. Grün (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis 16*(4), 1279–1307.

Gnedin, A. and J. Pitman (2006). Exchangeable gibbs partitions and stirling triangles. *Journal of Mathematical sciences 138*(3), 5674–5685.

James, L. F., A. Lijoi, and I. Prünster (2010). On the posterior distribution of classes of random means.

高頻度データに基づく拡散過程に対する SEM の擬似 BIC

熊本大学 大学院先端科学研究部 草野 彰吾

大阪大学 大学院基礎工学研究科 内田 雅之

近年, ICT 技術の急速な発展により, 極めて短い時間間隔で観測される時系列データ（高頻度データ）が容易に取得できるようになった. 高頻度データの解析においては連続時間確率過程を用いたモデリングが有効であることから, 高頻度データに基づく連続時間確率過程の統計的推測に関する研究が盛んに行われている. 最近, Kusano and Uchida [2] は, 拡散過程に対する構造方程式モデリング（SEM）を提案し, 高頻度データを用いた SEM の実行を可能にした. SEM は潜在変数間の関係を統計的に検証する手法であるため, モデルに関する十分な事前情報が必要である. しかし, 背後の構造が複雑な場合や事前情報が不十分である場合には, モデルの候補が複数存在し, 候補モデルの中から最適なものを 1 つ選択する必要がある. そこで, Kusano and Uchida [3] では, 拡散過程に対する SEM においてモデル選択を行うために, 擬似尤度に基づく擬似赤池情報量規準を提案し, その数学的正当化を与えた. 一方で, この情報量規準はモデル選択の一致性をもたないことが知られている. そこで, 本研究では拡散過程に対する SEM の擬似ベイズ情報量規準を考える.

拡散過程モデルでは一般に尤度関数を陽に表現することが困難であるため, 通常の赤池情報量規準 (AIC) やベイズ情報量規準 (BIC) を直接適用することができない. この問題に対して, Uchida [4] は擬似尤度に基づいた AIC 型の情報量規準を, Eguchi and Masuda [1] は擬似尤度に基づいた擬似 BIC を提案した. この状況は本研究でも同様であるため, 拡散過程に対する SEM モデルに基づく擬似尤度 $\mathbf{L}_{m,n}(\theta_m)$ を用いて擬似 BIC を構成する. また, 擬似最尤推定量 $\hat{\theta}_{m,n}$ を

$$\mathbf{L}_{m,n}(\hat{\theta}_{m,n}) = \sup_{\theta_m \in \Theta_m} \mathbf{L}_{m,n}(\theta_m)$$

と定義する. ただし, $\theta_m \in \Theta_m \subset \mathbb{R}^{q_m}$ は拡散過程に対する SEM モデルにおける m 番目の候補モデルのパラメータである. また, Θ_m は凸でコンパクトなパラメータ空間で, 局所リプシッツ境界をもつとする. このとき, 周辺擬似尤度

$$h_{m,n} = \int_{\Theta_m} \exp\left\{\mathbf{H}_{m,n}(\theta_m)\right\} \pi_{m,n}(\theta_m) d\theta_m$$

の漸近展開に基づいて, 以下の 2 つの擬似 BIC を考える:

$$\mathbf{QBIC}^{(m)}_{1,n} = -2\mathbf{H}_{m,n}(\hat{\theta}_{m,n}) + \log\det n\tilde{\mathbf{\Gamma}}_{m,n}(\hat{\theta}_{m,n}),$$

$$\mathbf{QBIC}_{2,n}^{(m)} = -2\mathbf{H}_{m,n}(\hat{\theta}_{m,n}) + q_m \log n.$$

ただし, $\pi_{m,n}(\theta_m)$ は θ_m の事前分布で,

$$\mathbf{H}_{m,n}(\theta_m) = \log(2\pi h_n)^{np/2}\mathbf{L}_{m,n}(\theta_m), \quad J_{m,n} = \left\{ -n^{-1}\frac{\partial^2}{\partial\theta_m\partial\theta_m^\top}\mathbf{H}_{m,n}(\hat{\theta}_{m,n}) > 0 \right\}$$

であり,

$$\tilde{\mathbf{\Gamma}}_{m,n}(\hat{\theta}_{m,n}) = \begin{cases} -n^{-1}\dfrac{\partial^2}{\partial\theta_m\partial\theta_m^\top}\mathbf{H}_{m,n}(\hat{\theta}_{m,n}) & (\text{on } J_{m,n}), \\ \mathbb{I}_{q_m} & (\text{on } J_{m,n}^c) \end{cases}$$

である. まず, 正則条件の下で, これら 2 種類の情報量規準がモデル選択の一致性をもつことを示す. また, 候補モデルの中に誤特定モデルが含まれている場合についても考え, 候補モデルの中に正しく特定されたモデルが 1 つ以上含まれているならば, 正則条件の下で, 誤特定モデルを選択する確率が漸近的に 0 となることを示す. さらに, 数値シミュレーションを行い, 提案した 2 種類の情報量規準の漸近挙動を検証する.

References

[1] Eguchi, S. and Masuda, H. (2018). Schwarz type model comparison for LAQ models. *Bernoulli*, **24(3)**, 2278-2327.

[2] Kusano, S. and Uchida, M. (2024). Sparse inference of structural equation modeling with latent variables for diffusion processes. *Japanese Journal of Statistics and Data Science*, **7**, 101-150.

[3] Kusano, S. and Uchida, M. (2025). Quasi-Akaike information criterion of SEM with latent variables for diffusion processes. *Japanese Journal of Statistics and Data Science*, **8(1)**, 217-264.

[4] Uchida, M. (2010). Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Annals of the Institute of Statistical Mathematics*, **62**, 161-187.

M スプライン基底を用いた，B スプラインコピュラの構成

寺西 蓮[1]　**江村 剛志[2]（講演者）**

概要：B スプラインコピュラは，B スプライン基底関数の線形結合として表現されるコピュラである．B スプラインコピュラは相関構造を柔軟に表現できるが，ユーザーにとって計算上の困難がある．この研究の主な目的は，計算に便利な 5 つの M スプライン基底関数を使用して，B スプラインコピュラの特殊なケースを提案することである．提案する B スプラインコピュラの順位相関係数などの理論的特性も調べる．

キーワード： 非対称コピュラ，Bernstein コピュラ，有向従属，パラメトリックコピュラ，放射対称，裾従属，ノンパラメトリックコピュラ

1. はじめに

コピュラは 2 つの変数間の相関構造を記述するためのツールである（江村 2025）．コピュラに対するノンパラメトリック推定法は，統計学における基礎的な問題の 1 つである．最も古典的なノンパラメトリック推定量は，経験コピュラ（Empirical copulas）である(Deheuvels 1979)．ただし，経験コピュラの周辺分布は，離散型一様分布となることから，コピュラの条件の 1 つである，周辺分布の連続型一様性を満足しない．このため，経験コピュラは，コピュラの推定量であるものの，それ自身はコピュラにならない．

上述の経験コピュラの問題に対応したノンパラメトリック推定量がいくつか考案されている．経験チェス盤コピュラは，離散的な経験コピュラに多重線形補間を施し，連続にしたコピュラである(Durante and Sempi 2016; Genest et al. 2017)．Bernstein コピュラ(Sancetta and Satchell 2004)は，Bernstein 多項式によるコピュラの構成法である．この考え方は，後に経験ベータコピュラ(Segers et al. 2017)や B スプラインコピュラ(Shen et al. 2008)にも用いられている．

B スプラインコピュラ（Shen et al. 2008）は B スプライン基底関数の線形結合として定義される．さらに，Dou et al.（2021）は，B スプラインコピュラの理論的特性を導き出し，Dou et al.（2025）はパラメータを推定のための EM アルゴリズムを提案した．

B スプラインコピュラは相関構造を柔軟に表現できるが，多くのユーザーにとって計算上の困難がある．例えば，スプライン基底関数(ノット数を含む)の決定が必要であることや，多数のパラメータを持つこと，さらにパラメータ空間に制約があることが挙げられる．これ

[1] 久留米大学・バイオ統計センター

[2] 広島大学・情報科学部(Email:takeshiemura@gmail.com)

まで，B スプラインコピュラ関数を計算するためのソフトウェアは公開されていない．

本稿では，計算に便利な M スプライン基底関数を使用して，B スプラインコピュラの特殊なケースを提案する．特に，Emura et al.(2017)，Shih and Emura (2021)，Teranishi et al.(2025)でハザード関数のモデル化のために提案された，5 つの M スプライン基底関数を使用する．一般に，M スプライン基底関数は，積分すると 1 になるように B スプライン基底を標準化したものである．提案する B スプラインコピュラのケンドール順位相関やスピアマン順位相関などの理論的特性も調べる．加えて，提案するコピュラの数値例を与え，さまざまな従属構造がモデル化できることを示す．さらに，R パッケージ *splineCox* 内に新たに作成した R 関数 spline.copula を紹介する．提案する B スプラインコピュラの特別な場合は，ユーザーに対して簡便で実用的な計算ツールを提供することを目的としている．

参考文献:

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Bulletins de l'Académie Royale de Belgique*, 65(1), 274-292.

Dou, X., Kuriki, S., et al. (2021). Dependence properties of B-spline copulas. *Sankhya A*, 83, 283-311.

Dou, X., Kuriki, S., Lin, G. D., & Richards, D. (2025). EM estimation of the B-spline copula with penalized pseudo-likelihood functions. *Statistical Papers*, 66(1), 30.

Emura, T., Nakatochi, M., Murotani, K., & Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research*, 26(6), 2649-2666.

江村剛志 (2025).「コピュラ理論の基礎」，コロナ社

Genest, C., Nešlehová, J. G., & Rémillard, B. (2017). Asymptotic behavior of the empirical multilinear copula process under broad conditions. *Journal of Multivariate Analysis*, 159, 82-110.

Nelsen, R.B. (2006) *An Introduction to Copulas* (2nd ed.). Springer, New York.

Sancetta, A., & Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric theory*, 20(3), 535-562.

Segers, J., Sibuya, M., & Tsukahara, H. (2017). The empirical beta copula. *J Multivar Anal*, 155, 35-51.

Shen, X., Zhu, Y., & Song, L. (2008). Linear B-spline copulas with applications to nonparametric estimation of copulas. *Computational Statistics & Data Analysis*, 52(7), 3806-3819.

Shih, J. H., & Emura, T. (2021). Penalized Cox regression with a five-parameter spline model. *Communications in Statistics-Theory and Methods*, **50**(16), 3749-3768.

Teranishi, R., Furukawa, K., & Emura, T. (2025). A Two-Stage Estimation Approach to Cox Regression Under the Five-Parameter Spline Model. *Mathematics*, **13**(4).

過去データを利用した共変量適応的ランダム化法と対応するブートストラップ法の提案

瀬谷 のどか[1]　田栗 正隆[1]　藤後 修[1]　鎌谷 研吾[2]
[1] 東京医科大学　[2] 統計数理研究所

【概要】
共変量適応的ランダム化 (covariate adaptive randomization; CAR) は，それまでの割付と共変量の情報をもとに共変量の偶然の不均衡を抑えるようにランダム割付を行うデザインであり，単純なランダム割付より平均治療効果の推定効率 (推定精度) を高めるとして，臨床試験において広く用いられている．しかし，既存の CAR 法では，精度の高い推定と妥当な推論を両立することが難しい．本研究では，推定精度を高めるために，過去の臨床試験データやリアルワールドデータ (過去データ) を利用した CAR 法を提案する．さらに，CAR 法の下での推定精度を適切に評価するために，割付デザインを考慮した不等確率でサンプリングを行うブートストラップ法を提案する．

以下では，CAR 法とブートストラップ法のそれぞれについて，既存法の問題とそれを解決するための提案法のアイデアを簡単に説明する．当日は，提案法の詳細な説明に加え，その理論的な性質，数値実験による性能評価の結果と実際の臨床試験データへの適用結果も報告した．

【CAR 法】
被験者 $i = 1, \ldots, n$ の共変量を $X_i \in \mathbb{R}^p$，割付を $A_i \in \{0, 1\}$ (1: 新薬，0: 対照)，アウトカムを $Y_i \in \mathbb{R}$ とする．被験者 j の割付時に利用可能な試験内データは $\bar{X}_j := \{X_i\}_{i=1}^{j}$ と $\bar{A}_{j-1} := \{A_i\}_{i=1}^{j-1}$ のみである．X_i には連続共変量 (例: 年齢) と離散共変量 (例: 性別) の両方が含まれることが多いものの，層別ランダム化や最小化法といった伝統的な CAR 法では少数の離散共変量しか扱えないため，推定精度の向上に限界がある．Ma et al. (2022) は，X_i の各要素の 1 次項や 2 次項，交互作用項を含む $q > p$ 次元の共変量特徴 $\phi(X_i) : \mathbb{R}^p \to \mathbb{R}^q$ の不均衡が小さくなるように割付を行う CAR 法を提案した．具体的には，被験者 1 の割付確率を $\mathbb{P}[A_1 = 1 \mid X_1] = 0.5$，被験者 $j \geq 2$ の割付確率を

$$\mathbb{P}[A_j = 1 \mid \bar{A}_{j-1}, \bar{X}_j] = \begin{cases} \rho & \text{if } \mathrm{Imb}_j^{(1)} < \mathrm{Imb}_j^{(0)}, \\ 1 - \rho & \text{if } \mathrm{Imb}_j^{(1)} > \mathrm{Imb}_j^{(0)}, \\ 0.5 & \text{if } \mathrm{Imb}_j^{(1)} = \mathrm{Imb}_j^{(0)}, \end{cases} \tag{1}$$

とする CAR 法である．ここで，$0.5 < \rho \leq 1$ および

$$\mathrm{Imb}_j = \left\{ \sum_{i=1}^{j} (2A_i - 1)\phi(X_i) \right\}^T \left\{ \sum_{i=1}^{j} (2A_i - 1)\phi(X_i) \right\}$$

であり，$\mathrm{Imb}_j^{(a)}$ は $A_j = a$ とした場合の Imb_j である．この CAR 法の下で，平均治療効果の推定量 $\hat{\theta}_n := \sum_{i=1}^n A_i Y_i / \sum_{i=1}^n A_i - \sum_{i=1}^n (1-A_i) Y_i / \sum_{i=1}^n (1-A_i)$ の漸近分散が最小になるには，$\phi(X_i)$ が

$$g(X_i) := \{m^{(1)}(X_i) - \mathbb{E}[m^{(1)}(X_i)] + m^{(0)}(X_i) - \mathbb{E}[m^{(0)}(X_i)]\}/2 \in \{\beta^T \phi(X_i) \mid \beta \in \mathbb{R}^q\} \tag{2}$$

を満たす必要がある (Ma et al., 2022)．ここで，$m^{(a)}(X_i) = \mathbb{E}[Y_i \mid A_i = a, X_i]$ である．

ロジスティック回帰モデルのように $m^{(a)}(X_i)$ が $\phi(X_i)$ の要素の線形結合で表せない場合，一般に (2) 式は満たされないため，この CAR 法が $\hat{\theta}_n$ の漸近分散を最小にすることは難しい．さらに，この CAR 法の下では，一般に $\sum_{i=1}^n (2A_i - 1)\phi(X_i) = o_p(\sqrt{n})$ であり最適な $O_p(1)$ とはならないため (Ma et al., 2022)，現実的な中程度の n の場合には推定精度が低くなり得る．$\phi(X_i) = (1, g(X_i))^T$ と取ることができれば，これらの問題は解決し得るが，$g(X_i)$ に含まれる $m^{(a)}(X_i)$ は未知であり，また割付時に全ての被験者の X_i や Y_i を得ることができない制約から，試験内データによる $m^{(a)}(X_i)$ の推定は難しい．本研究では，試験開始前に過去データから $m^{(a)}(X_i)$ を予め学習し，それを用いて $g(X_i)$ の不均衡 $\sum_{i=1}^n (2A_i - 1) g(X_i)$ が小さくなるように割付を行う CAR 法を構築することで，これらの問題を解決する．

【ブートストラップ法】

(1) 式からも明らかなように，CAR 法の下では被験者同士のデータが独立にならないため，推論が難しい．Ma et al. (2022) のモデルベースの方法は，アウトカムモデルの正しい特定を必要とする．Shao et al. (2010) のブートストラップ検定は，区間推定には使えず，Zhang and Zheng (2020) の層別ブートストラップ法は，本研究で提案するような連続共変量を扱う CAR 法には使えないという難点がある．これら 2 つの既存のブートストラップ法における難点は，ブートストラップ標本において割付デザインを再現するために，元の割付を使用せずに割り付け直しをしていることに起因する．本研究では，この問題に対処するため，割り付け直しを伴わずに CAR 法に対応した尤度を得る新たなブートストラップ法を提案する．具体的には，単純なブートストラップ標本の尤度 $\prod_{i=1}^n f(Y_i^* \mid A_i^*, X_i^*) f(A_i^* \mid X_i^*) f(X_i^*)$ と CAR 法の下での尤度 $\prod_{i=1}^n f(Y_i^* \mid A_i^*, X_i^*) f(A_i^* \mid \bar{A}_{i-1}^*, \bar{X}_i^*) f(X_i^*)$ の比をとった $\prod_{i=1}^n f(A_i^* \mid \bar{A}_{i-1}^*, \bar{X}_i^*) / f(A_i^* \mid X_i^*)$ に比例する不等確率でサンプリングを行うブートストラップ法を提案する．

参考文献

Ma, W., Li, P., Zhang, L., and Hu, F. (2022). A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association*, 119(545):151–162.

Shao, J., Yu, X., and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, 97(2):347–360.

Zhang, Y. and Zheng, X. (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics*, 11(3):957–982.

Algorithms for Fast Gibbs Sampling in Hierarchical Bayesian Panel Modeling

Makoto Nakakita[1] Tomoki Toyabe[2] Teruo Nakatsuma[3]
Takahiro Hoshino[1,3]

Abstract

The ancillarity-sufficiency interweaving strategy (ASIS) has been proved to be a powerful tool for improving Markov chain Monte Carlo (MCMC) computation and widely applied for Bayesian estimation of dynamic linear models as well as panel data models. The previous studies, however, demonstrated the efficacy of ASIS only through applications to real-world data or numerical simulations. In this paper, we attempt to examine the performance of ASIS in the context of hierarchical Bayesian modeling of panel data in a more rigorous fashion. First we prove that ASIS can generate almost uncorrelated random sequences of individual effects in the panel data linear regression model in a simplified setting. Then we demonstrate the efficacy of ASIS in more general settings with Monte Carlo experiments.

Keywords: ASIS, hierarchical Bayesian modeling, MCMC, panel data

1 Introduction

Bayesian hierarchical models are increasingly popular in panel data analysis. However, MCMC convergence can be slow, especially in models with strong parameter dependence. The ancillarity–sufficiency interweaving strategy (ASIS) improves efficiency by combining centered and non-centered parameterizations.

[1]Center for Advanced Intelligence Project, RIKEN
[2]Department of Economics, Kanazawa Gakuin University
[3]Faculty of Economics, Keio University

2 Theoretical Contributions

We derive convergence rates under sufficient augmentation (SA) and ancillary augmentation (AA). By applying ASIS, the sample of global mean $\mu_\alpha^{(t)}$ become independent of the previous sample $\mu_\alpha^{(t-1)}$. ASIS exploits this property to achieve optimal convergence.

3 Simulation Study

Simulations across various (N, T) settings show that ASIS outperforms both SA and AA in terms of Monte Carlo standard error (MCSE) and autocorrelation. The predicted trade-off between SA and AA holds, and ASIS consistently yields the most efficient sampling.

4 Real Data Application

We analyze U.S. state-level panel data on cigarette consumption, controlling for income, price, and tax. ASIS outperforms SA and AA in estimating global means and regression coefficients. The empirical results reinforce ASIS's robustness in applied Bayesian analysis.

5 Conclusion

We provide the first theoretical justification for ASIS in hierarchical panel models and show its effectiveness across sample sizes. ASIS ensures efficient MCMC sampling, with applications in economics, health, and social sciences. Future work includes extensions to non-Gaussian and dynamic models.

Acknowledgments

This research was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP25K00626.

RxS 分割表による医療データの総括・統合とその解析

久留米大学バイオ統計センター
柳川　堯

次の、二つの話題を提供した.

1. 順序カテゴリーを効果指標とする臨床試験
2. 2 変量がともに順序カテゴリーである RxS 表の相関の検定

第 1 の話題: 順序カテゴリーを効果指標とする臨床試験

1.1 はじめに

認知症薬や疼痛薬などの臨床試験では、ベースラインの効果と比べて「著明効果あり」,「効果あり」, 「効果なし」等の順序カテゴリーで薬剤の臨床効果を, 表 1 のような 2xk 表に分類して評価する場合がある.

表 1 順序カテゴリーデータ

	C_1	C_2	…	C_k	計
Placebo 群	m_1	m_2	…	m_k	m
Test 薬剤群	n_1	n_2	…	n_k	n

$C_1, C_2, \ldots, C_k$は順序カテゴリー

1.2 問題の定式化

Placebo 群が従う分布関数を **F, F** に従う確率変数を X; Test 薬剤群が従う分布関数を **G, G** に従う確率変数を **Y** とする. このとき, 「薬剤効果あり」は「$\mathrm{G(x)>F(x)}$ for any $\mathbf{x}$」で表される. F, G を未知とした時, 従来 $H_0: F(x) = G(x)$ vs. $H_1: F(x) < G(x)$の検定には Wilcoxon 検定などのノンパラメトリック検定が適用されてきた.

しかし, 臨床試験では検定だけではなく薬剤の効果の推定が重んじられる. 表 1 から薬剤効果の推定を行うため本研究では F(Y)がベータ分布 Be(1, β)に従うことを仮定した。この仮定の下では

- 「$\mathrm{G(x)>F(x)}$ for any $\mathbf{x}$」が「β>1」と同値なこと、
- X=x での Test 薬剤の Placebo に対する薬剤効果が

  $\mathrm{G(x) - F(x) = 1 - t - (1 - t)^{\beta}}$, ただし, t=F(x)

で表されることを証明し、薬剤効果の新しい検定法と推定法を提案した. また, ベータ分布の適合度検定, および症例数の決定方式も提案した.

第 2 の話題：2 変量がともに順序カテゴリーである RxS 表の相関の検定

2.1 はじめに

✧ 臨床医学の分野でも多量のデータ(Real world data)が蓄積しており、ある疾患の患者に提供された様々な治療法の中で、どの治療法が有効であったかのエビデンスを Real world data から検証することが求められている (Evidence Based Medicine)

✧ 臨床データは、大小さまざまな施設（病院）に蓄積されているが,施設間での治療法・診断法にズレがある．例え同一施設内であっても，医師は患者個人の症状に最も適切と判断した治療法で治療を行うため，医師間によるズレがある.ズレがあるデータを単に集積・統合して解析すればよいとうわけには行かない.

✧ このような背景から得られるデータは，個々の値(実数値)よりも「無効」、「やや有効」、「かなり有効」、「著効」などのカテゴリーを用いて、まとめ直し、カテゴリーデータとして集積・統合して解析すれば，施設間差（ズレ）の影響が最小化でき，有効な evidence が取り出せる可能性が高い.

2.2 目的と概要

本研究では，表 2 に統合・集積された ordered categorical data にスポットライトを当て，A と B の関連性の検定法の開発を目的として、新しい検定法を提案した．とともに,応用例として，4 施設から集積・統合された基本理学療法の有効性の検証に適用した.

表 2 R x S 分割表(個数)

		B				
A		B_1	B_2	…	B_s	total
	A_1	n_{11}	n_{12}		n_{1s}	$n_{1.}$
	A_2	n_{21}	n_{22}		n_{2s}	$n_{2.}$
	…					
	A_r	n_{r1}	n_{r2}		n_{rs}	$n_{r.}$
	total	$n_{.1}$	$n_{.2}$		$n_{.s}$	N

$\{A_i\}$,$\{B_j\}$は，順序カテゴリー変数

$\{A_i\}$にはスコア c1,c2,…,cr (c1<c2<…<cr);

$\{B_j\}$にはスコア d1,d2,…,ds (d1<d2<…<ds)を割り付ける.

高次元共分散構造の自動スパース推定とその一致性

筑波大学・数理物質科学研究群　海野　哲也
筑波大学・数理物質系　矢田　和善
筑波大学・数理物質系　青嶋　誠

1　はじめに

本講演では，自動スパース推定を用いた高次元共分散構造に対するスパース推定について考えた．独立な要素数が $O(p^2)$ 個である共分散行列の推定は，特に高次元では推定精度・計算コストの両面で非常に困難な問題である一方，様々な応用分野において重要な課題である．共分散行列のスパース推定は，高次元における共分散構造の解釈性の向上という応用上重要な側面を有することに加え，共分散構造にスパース性が存在する場合には推定精度の向上にも寄与することから，これまでに多くの研究がなされてきた．共分散行列に対するスパース推定は，一般的に ℓ_1 罰則法に基づく手法と閾値法に基づく手法が知られている．例えば，Bickel and Levina [1] は閾値法によるスパース推定を提案し，その高次元における漸近的性質を解明した．さらに，Bien and Tibshirani [2] は ℓ_1 罰則に基づくスパース推定を提案した．しかしながら，これらの方法論で得られる推定量が一致性を有するには適切な正則化パラメータの選択が必要であり，最適なパラメータ探索のためのクロスバリデーションは，高次元において計算コストの増大や結果の安定性に関する問題が生じる．

スパース推定における正則化パラメータ選択に関する問題を回避した推定法として，最近，Yata and Aoshima [4] は正則化パラメータを必要としない，高次元主成分ベクトルの新たなスパース推定の方法論を提案し，高次元において高速かつ高精度に推定できることを示した．さらに，Umino et al. [3] は同手法を高次元相互共分散行列のスパース推定に応用した．本講演では，Umino et al. [3] で提案された手法を高次元共分散行列の非対角成分全体に拡張することで，正則化パラメータを必要としない高次元共分散行列のスパース推定法を提案し，その一致性について議論した．

2　標本共分散行列の漸近的性質

平均ベクトルに p 次元ベクトル $\boldsymbol{\mu}$，共分散行列に p 次の半正定値行列 $\boldsymbol{\Sigma}$ を持つ母集団から，$n(\geq 4)$ 個の p 次元データベクトル $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ を無作為に抽出したとする．$\boldsymbol{\Sigma}$ の (r,s) 成分を $\sigma_{(r,s)}$ と表記し，$\boldsymbol{\Sigma}$ から対角成分，非対角成分を抜き出した行列をそれぞれ $\boldsymbol{\Sigma}_D$，$\boldsymbol{\Sigma}_{ND}$ とする．すなわち，$\boldsymbol{\Sigma}_D = \mathrm{diag}(\sigma_{(1,1)}, \ldots, \sigma_{(p,p)})$ かつ $\boldsymbol{\Sigma}_{ND} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_D$ とする．さらに，$\|\cdot\|_F^2$ を行列のフロベニウスノルムとして，$\Delta = \|\boldsymbol{\Sigma}\|_F^2$, $\Delta_D = \|\boldsymbol{\Sigma}_D\|_F^2$ かつ $\Delta_{ND} = \|\boldsymbol{\Sigma}_{ND}\|_F^2$ とし，$\Delta_{ND} > 0$ であることを仮定する．

$\boldsymbol{\Sigma}$ の従来型の不偏推定量を $\boldsymbol{S} = \sum_{j=1}^n (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})^T/(n-1)$, $\overline{\boldsymbol{x}} = \sum_{j=1}^n \boldsymbol{x}_j/n$ とし，$\boldsymbol{\Sigma}$ と同様に $\boldsymbol{S} = (s_{(r,s)})$, $\boldsymbol{S}_D = \mathrm{diag}(s_{(1,1)}, \ldots, s_{(p,p)})$ かつ $\boldsymbol{S}_{ND} = \boldsymbol{S} - \boldsymbol{S}_D$ とする．このとき，次の結果が得られる．

命題 1. 適当な正則条件のもと，$p, n \to \infty$ で以下が成立する．

$$\frac{\|\boldsymbol{S}_D - \boldsymbol{\Sigma}_D\|_F^2}{\Delta_D} = o_P(1) \quad and \quad \frac{\|\boldsymbol{S}_{ND} - \boldsymbol{\Sigma}_{ND}\|_F^2}{\Delta_{ND}} = \frac{tr(\boldsymbol{\Sigma})^2}{n\Delta_{ND}}\{1 + o_P(1)\} + o_P(1).$$

上記の結果より，$\boldsymbol{S}_D$ は $\boldsymbol{\Sigma}_D$ の精度の良い推定量となるものの，$\boldsymbol{S}_{ND}$ には $\mathrm{tr}(\boldsymbol{\Sigma})^2/n = O(p^2/n)$ の巨大なノイズが発生し，

$$\frac{\|\boldsymbol{S}_{ND} - \boldsymbol{\Sigma}_{ND}\|_F^2}{\Delta_{ND}} = o_P(1)$$

の意味での一致性を有するには $p, n \to \infty$ のもと $\mathrm{tr}(\boldsymbol{\Sigma})^2/(n\Delta_{ND}) = o(1)$ なる条件が必要となる．さらに $\boldsymbol{S}_{ND}$ について，以下が成立する．

命題 2. 適当な正則条件のもと，$p, n \to \infty$ で以下が成立する．

$$\frac{tr(\boldsymbol{\Sigma}_{ND}\boldsymbol{S}_{ND}^T)}{\Delta_{ND}} = 1 + o_P(1).$$

上記の結果と命題 1 は，$\boldsymbol{S}_{ND}$ に対し，対角成分がすべて 0 かつ対称なランダム行列 $\boldsymbol{\Xi}_{ND} \in \mathbb{R}^{p\times p}$ が存在して

$$\begin{aligned}
&\boldsymbol{S}_{ND} = \boldsymbol{\Sigma}_{ND}\{1 + o_P(1)\} + \boldsymbol{\Xi}_{ND}, \\
&\mathrm{tr}(\boldsymbol{\Xi}_{ND}\boldsymbol{\Sigma}_{ND}^T) = 0 \quad \text{and} \quad \|\boldsymbol{\Xi}_{ND}\|_F^2 = \frac{\mathrm{tr}(\boldsymbol{\Sigma})^2}{n}\{1 + o_P(1)\}
\end{aligned}$$

を満たすことを意味している．この結果と Yata and Aoshima [4] で提案された自動スパース推定の考え方に基づけば，$\boldsymbol{S}_{ND}$ を補正した推定量 $\widetilde{\boldsymbol{\Sigma}}_{ND}$ を構築できる．講演では，自動スパース推定による推定量の構成方法とその一致性について紹介した．

さらに，共分散行列全体ではなく特定のブロック構造に関心がある場合におけるブロック共分散行列の推定も，応用上重要な課題である．講演では，$\boldsymbol{\Sigma}$ にブロック構造が仮定できる場合に，各ブロックに対して上記の自動スパース推定を適用し，計算コストを削減する方法論も提案した．また，数値実験を通じて従来法との比較を行い，提案手法の有効性を検証した．

参考文献

[1] Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604.

[2] Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

[3] Umino, T., Yata, K., and Aoshima, M. (2025). Automatic sparse estimation of the high-dimensional cross-covariance matrix. *Revised in Journal of Multivariate Analysis.*

[4] Yata, K. and Aoshima, M. (2025). Automatic sparse pca for high-dimensional data. *Statistica Sinica*, 35:1069–1090.

TopiCLEAR: 文書ベクトルのクラスタリングに基づくトピック抽出技術

小林 亮太 [1,2], 藤田葵 [1], 中山悠理 [1], 山本泰智 [1]
(1) 東京大学大学院 新領域創成科学研究科
(2) 東京大学 数理・情報教育研究センター
E-mail: r-koba@edu.k.u-tokyo.ac.jp

近年，X (旧 Twitter), Facebook, Reddit などのソーシャルメディアの普及により，人々はニュースや事件に対する意見や感想をリアルタイムに発信できるようになった．ソーシャルメディアの投稿 (ポストやツイートなど) を分析することで，多数の人々の興味・関心を分析することが可能となっている．そのため，ソーシャルメディアの投稿に見られる短文テキストの分析が重要な課題となっている．テキストデータから意味のあるテーマや話題 (トピック) を抽出するタスクは，トピック抽出 (Topic Extraction) と呼ばれ，Covid-19 ワクチン, 生成 AI，政治, 自然災害など，さまざまな話題に関する人々の興味・関心を分析する上で不可欠である．

トピックモデル [1] は，トピック抽出を行うための代表的な機械学習手法であり，テキストマイニングや自然言語処理分野で広く利用されている．代表的な手法である Latent Dirichlet Allocation (LDA) は，ニュース記事や学術論文などの文語的な文書分析で成功を収めてきた．しかし，トピックモデルは単語の出現頻度に依存するため，短文や口語的なテキストでは性能が低下する [2]．ソーシャルメディア投稿は，短さ，口語的表現，スペルミス，構文の不完全性などの特徴を持つため，既存のトピックモデルでは分析が困難である．近年，事前学習言語モデルの登場により，テキスト分析の方法論は大きく進化した．その中でも Sentence-BERT (SBERT) [3] は，文の意味的類似性を高精度に定量化できる点で画期的である．SBERT は，文章を高次元ベクトル空間に変換し，コサイン類似度に基づいて比較することで，質問応答や検索などのタスクで高い性能を示している．このような応用事例での有用性は，SBERT がトピック抽出など他のテキストマイニング課題にも有効である可能性を示唆している．

本講演では，テキストデータからトピック抽出を行う手法 TopiCLEAR (Topic extraction by CLustering Embeddings with Adaptive dimensional Reduction) [4] を提案する．提案手法 TopiCLEAR は以下のステップで構成される:

1. SBERT によるテキスト埋め込み： 各テキストを高次元ベクトル空間に変換.
2. 暫定クラスタリング： ガウス混合モデル (GMM) により初期クラスタを生成する.
3. クラスタ洗練： 線形判別分析 (LDA) と GMM を反復適用し，トピック分離度を最大化.

提案手法 TopiCLEAR の性能評価には，テキストデータと人間によって分類されたラベルを含む 4 種類のデータセット (20News, AgNewsTitle, Reddit, TweetTopic) を使用した．20News は標準的なニュース記事についてのデータセットである．AgNewsTitle はニュースタイトルについての短文データセット，Reddit と TweetTopic はソーシャルメディアのデータセットであるため，トピックを抽出するのが難しいデータセットになっている．評価指標には、クラスタリング性能を測る Adjusted Rand Index と Adjusted Mutual Information を採用し，人間によって分類されたラベルとの類似度を計算した．比較対象として，7 つの既存手法 (3 つのトピックモデル: LDA, BTM, ProdLDA，トピックモデルとテキスト埋め込みを融合させた手法: ETM, CTM，言語モデルに基づく手法: BERTopic, LLM: Gemini 2.5) を用いた．その結果，提案手法 TopiCLEAR はすべてのデータセットで既存手法よりも人間の分類に近いトピックを抽出できることが示された．さらに，TopiCLEAR はストップワード除去やステミングなどの前処理が不要であるため，実運用上の優位性を持つ．TopiCLEAR の python コードは GitHub (`https://github.com/aoi8716/TopiCLEAR`) で公開している．

謝辞

本研究は，JSPS 科研費 JP21H03559, JP22H03695, JP23K24950，AMED JP223fa627001，JST 創発 JPMJFR232O，プロアクティブ環境学国際卓越大学院プログラム (WINGS-PES) の支援を受けたものである．

参考文献

[1] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445, 2022.

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics.

[4] Aoi Fujita, Taichi Yamamoto, Yuri Nakayama, and Ryota Kobayashi. TopiCLEAR: Topic extraction by CLustering Embeddings with Adaptive dimensional Reduction. *arXiv preprint arXiv:2512.06694*, 2025.

Analyzing Japanese Equity Returns with Equi-Correlation Structures

森本孝之 (関西学院大学)，赤間陽二 (東北大学)，川崎能典 (統計数理研究所)

科研費シンポジウム「データサイエンスの基盤を支える次世代統計理論・方法論の挑戦と革新」
2025 年 12 月 2 日 九州大学 西新プラザ

1. ファクターリターンモデル

金融資産収益率の決定要因を説明する資産価格モデル (ファクターリターンモデル) は，過去さまざま提案されている．代表的なモデルとして，Sharpe の CAPM や Fama and French の 3 ファクターモデル (FF3)，更にそれらを 4 ファクター (FF4)，5 ファクター (FF5) に拡張したものがある．ここで言うファクターは線形回帰モデルにおける説明変数のことであり，いわゆる因子分析の意味でのファクターでないことは注意しておく．

これらのモデルの有効性は北米や欧州の金融市場で確認されている一方で，必ずしも日本では観察できないとする研究もある．また近年，多変量 GARCH モデルの一種から業種相関の指数系列を取り出し，資産価格モデルの説明力向上を狙った興味深い研究がある．本報告では，業種間相関だけでなく，業種内での銘柄相関の指数系列を取り出した上で，主成分系列に変換してモデルに投入することを提案し，既存のファクターリターンモデルとの組合せの中で，より説明力を持った定式化を探索した結果を報告する．文中で引用する文献に関しては Morimoto et al. (2025) を参照されたい．

2. 多変量 GARCH モデルからの情報抽出

N 個の資産価格系列 $\{p_{it}\}_{t=1}^{T}$ $(i = 1, \ldots, N)$ に対し, 時刻 t における資産 i の収益率を $r_{it} = \log p_{it} - \log p_{i,t-1}$, $i = 1, \ldots, N$ と書き，収益率ベクトルを $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{Nt})'$ とする．このとき条件付き共分散行列 $\boldsymbol{H}_t = \mathrm{E}[\boldsymbol{r}_t \boldsymbol{r}_t' | \mathcal{F}_{t-1}]$ をどうモデル化するかが問題となるが，$\boldsymbol{H}_t$ を対角行列 $\boldsymbol{D}_t = \mathrm{diag}(\sqrt{h_{1t}}, \ldots, \sqrt{h_{Nt}})$ と，時変相関行列 $\boldsymbol{R}_t$ とで $\boldsymbol{H}_t = \boldsymbol{D}_t \boldsymbol{R}_t \boldsymbol{D}_t$ と分解し，h_{it} と $\boldsymbol{R}_t$ のモデリングを別々に考えるのが基本的な方針である．

収益率ベクトルに対する条件付き分散行列は，パラメータを節約するために，一変量 GARCH モデル (しばしば GARCH(1,1) を仮定) を相関行列でつなぎ合わせて得る．つまり，各資産 i に関して $h_{it} = \omega_i + \alpha_i r_{i,t-1}^2 + \beta_i h_{i,t-1}$ を推定して h_{it} を得る．それを使って標準化残差 $\epsilon_{it} = r_{it}/\sqrt{h_{it}}$ が得られる．ベクトルで表現すれば $\boldsymbol{\epsilon}_t = \boldsymbol{D}_t^{-1} \boldsymbol{r}_t$ である．

標準化されていない相関プロセス $\boldsymbol{Q}_t$ を $\boldsymbol{Q}_t = (1 - a - b)\bar{\boldsymbol{Q}} + a(\boldsymbol{\epsilon}_{t-1} \boldsymbol{\epsilon}_{t-1}') + b\boldsymbol{Q}_{t-1}$ で定義する．ここで $\bar{\boldsymbol{Q}}$ は標準化残差の無条件共分散，a, b $(a \geq 0,\ b \geq 0,\ a + b < 1)$ は未知パラメータである．これを使って相関行列を $\boldsymbol{R}_t = \mathrm{diag}(\boldsymbol{Q}_t)^{-1/2} \boldsymbol{Q}_t \mathrm{diag}(\boldsymbol{Q}_t)^{-1/2}$ で与える．$\boldsymbol{R}_t$ は正定値で相関行列構造を満たす．これは Dynamic Conditional Correlation (DCC) モデルと呼ばれる． DCC の相関行列は，無制約のフル構造では $n(n-1)/2$ 個の時変相関を持つ．これを高次元 (数百次元) でも推定が容易になるよう，各 t で任意の 2 資産の相関がすべて同じ (等相関) という仮定を置く．これは Dynamic Equicorrelation (DECO) モデルと呼ばれる (Engle and Kelly, 2012)．DECO も $\boldsymbol{Q}_t$ に関して DCC 同様の逐次式を与える．パラメータが求まれば，各時点での $\boldsymbol{Q}_t$ の値も定まる．その (i, j) 要素 $q_{t,ij}$ を使って

$$\rho_t = \frac{2}{N(N-1)} \sum_{i<j} \frac{q_{t,ij}}{\sqrt{q_{t,ii} q_{t,jj}}}$$

により等相関 ρ_t を推定する．この集約は各時点 t で行われるので，結果的に ρ_t は時系列で得られる．

3. 業種間相関指数 (IEC) から業種内相関指数 (IIEC) へ

資産時系列に業種インデックス (東証 33 業種) の日次データを，2014 年 1 月 6 日から 2023 年 12 月 29 日までの期間で取る．この設定で推定される $\{\rho_t\}$ は，業種間等相関 (industry equicorrelation) と言える．時系列 ρ_t からここでは移動平均法でそのトレンド τ_t を推定し，トレンドからの乖離 $\xi_t = \rho_t - \tau_t$ を，Industry Equi-Correlation index (IEC 指数) と呼ぶ．IEC 指数は反景気循環的性質を持ち (countercyclical)，資産価格モデルの説明力向上に役立つことを Wang et al. (2020) は示した．

我々は Wang et al. (2020) の方法を拡張すべく，データを個別株に取り直して業種ごとに IEC 指数を抽出する．結果，業種内の等相関構造を集約した時系列 $\rho_{j,t}$ が 33 本できる ($j = 1, \ldots, 33$)．これを Intra-Industry Equi-Correlation index (IIEC 指数) と呼ぶ．ただ，IEC とは異なり IIEC は多変量時系列なので，$\rho_{j,t}$ の主成分 (時系列) を求め，トレンド除去後に資産価格モデルの説明変数として追加する．今回示す実証分析では第 1 主成分のみを使用し，PCA_t と言及する．

4. 実証分析の設定と評価指標

我々は，従来実証ファイナンスで受け入れられてきたファクターリターンモデルの追加的説明変数として IEC 指数 ξ_t や PCA_t を導入するが，ベースとなるファクターリターンモデルの定式化は極めて標準的である．即ち，推計式の右辺に来るマーケットインデックス，時価総額に基づくスプレッド，時価簿価比率に基づくスプレッド，前期騰落率から抽出するスプレッド，収益性に基づくスプレッド，投資の大きさに基づくスプレッドは，すべて Ken French のデータライブラリーで公開されているデータを用いる．左辺の収益率は，時価総額を軸として残りの 4 ファクターとのクロスセル ($5 \times 5 \times 4$) に落とし込んだ資産の平均収益率から構成される 100 個のテストポートフォリオリターンであるが，これも同ライブラリーで提供されている．

モデルの実証比較のための指標としては, Fama-MacBeth(FMB) 回帰, Hansen-Jagannathan(HJ) 距離, Gibbons-Ross-Shanken(GRS) 検定を採用する．FMB 回帰では，時系列回帰で資産ごとに得られた係数 (ファクターエキスポジャー) を，今度は t を止めてリターン $r_{i,t}$ を β_i に回帰する．これで各時点のリスクプレミアムが推定される．基本的に pricing error の小さいモデルが良いモデルという基準になるが，第 2 段階目の回帰の R^2, AIC, RMSE 等を重みづけた複合指標でモデルを評価する．HJ 距離は，pricing error を資産ベクトルの分散共分散行列を挟んだ 2 次形式で評価するものであり，HJ 距離が小さい方が平均分散フロンティアに近く，価格付けのパフォーマンスが良いと見なせる．GRS 検定は，ファクターリターンモデルの定数項を α_i と書くとき，「すべての資産で α_i がゼロ」という帰無仮説の同時検定を行うもので，帰無仮説が棄却されるとモデルが失敗していることを意味する．ファクターの Sharpe ratio を基準に，いずれかの α_i がデータから説明がつかない程大きいかどうかを検定する方法である．

5. 日本の株式市場での分析結果

今回の分析から以下の事柄が観察された．モデルに含めるファクター間の相関を観察すると，IEC も PCA も既存のファクターとの相関は低く，何らかの意味で新しい情報が抽出できていると思われる．IEC と PCA の相関は高く，第 1 主成分は IEC と類似の成分が出てきている可能性がある．

FMB 回帰の結果を見ると, FF4 あるいは FF5 に含めて推定された PCA 項だけが 5% 有意となった．最も R^2 が大きく, かつ RMSE も最小となったのは, FF5+IEC+PCA の定式化であったが, AIC で判断すると, FF4+IEC+PCA の定式化が最小 AIC を達成した．R^2, AIC, RMSE 等を重みづけた複合指標でモデルを評価すると，FF4+PCA, FF5+IEC+PCA，FF5+PCA の順で好成績を収めている．この 3 者はほぼ同等であり，4 位以下は大きく引き離される．概してモメンタム効果を入れると (FF4) モデルの適合度は向上するが，従来日本のデータで否定的な結論が得られている FF5 が選ばれているのはこれまでとは異なる実証結果である．

今回は全部で 16 個の定式化を比較したが，FMB 回帰の複合スコアでの順位，HJ 距離での順位，GRS 検定での順位を合算して総合ランキングをつけると (順位和が小さければ小さいほど良い)，FF5+IEC+PCA がベストモデルと結論された．

参考文献

Morimoto, T., Akama, Y. and Kawasaki, Y. (2025), Forecasting Japanese Equity Returns Using Equi-Correlation Structures and Component Selection. Available at SSRN: https://ssrn.com/abstract=5855708 or http://dx.doi.org/10.2139/ssrn.5855708.

Spectral Decomposition in Dynamic Systems of Distributional Data

Ziyue Wang , Yuko Araki
Graduate School of Information Sciences, Tohoku University

1 Introduction

In recent years, data arising from complex systems have increasingly been modeled or represented as distributions or other structured objects that evolve over time, reflecting advances in sensing, computation, and large-scale data collection. Such representations are useful in diverse fields such as economics, population studies, and machine learning, where understanding and forecasting the evolution of entire distributions is essential for explaining system behavior and detecting structural changes.

This study proposes a new statistical framework, called the *Koopman–Wasserstein* framework, for analyzing and forecasting the dynamics of distribution-valued data. By combining the spectral theory of the Koopman operator with Wasserstein geometry, the framework enables consistent representation and prediction of how probability distributions evolve over time.

Recent studies have modeled data as evolving probability distributions (Panaretos and Zemel (2019); Zhang et al. (2022); Chen et al. (2023)). This direction is motivated by applications where the shape of the distribution carries essential information, such as regional housing prices and demographic age–mortality patterns. Embedding such data in Euclidean space distorts the geometry defined by the Wasserstein distance. However, existing approaches, such as Wasserstein regression (Chen et al. (2023); Zhang et al. (2022)) and Fréchet regression (Petersen and Müller (2019)), which are based on Euclidean embeddings, focus on static distributions and do not model temporal dependence. Building on our previous study (Wang and Araki (2025)), which introduced a Koopman–Wasserstein framework for distributional time series forecasting, the present work extends the theory to spectral analysis of stochastic dynamics and provides new theoretical results on consistency and prediction accuracy.

2 Model Assumptions

Let $\{X_t\}_{t\geq 0}$ be a d-dimensional Itô diffusion:

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \qquad X_0 \sim \mu_0, \tag{1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are Lipschitz and twice continuously differentiable. Assume that the process is ergodic with a unique stationary density $p_s(x)$ satisfying $\int \|x\|^2 p_s(x)\,\mathrm{d}x < \infty$.

The associated Fokker–Planck operator is

$$\mathcal{L}^* = -\nabla\cdot\big(b\,\cdot\big) + \tfrac{1}{2}\nabla^2 : \big(\sigma\sigma^\top\cdot\big), \tag{2}$$

and admits a discrete spectrum $\{(\lambda_i, \varphi_i)\}$ in $L^2(p_s)$ under these regularity conditions.

The Koopman semigroup $\{K_t\}_{t\geq 0}$ acts on observables $f \in L^2(p_s)$ as

$$K_t f(x) = \mathbb{E}[f(X_t) \mid X_0 = x]\,, \qquad \mathcal{L}f = \langle b, \nabla f\rangle + \tfrac{1}{2}\mathrm{Tr}\big(\sigma\sigma^\top\nabla^2 f\big). \tag{3}$$

We assume reversibility (detailed balance), ensuring that $\mathcal{L}$ is self-adjoint in $L^2(p_s)$. These assumptions imply an orthonormal eigenbasis $\{\varphi_i\}$ representing the diffusion dynamics.

3 Estimation Method

We estimate the spectral structure of the Koopman operator using Extended Dynamic Mode Decomposition (EDMD; Williams et al. (2015)) with importance weighting. Let $\Psi(x) = (\psi_1(x), \ldots, \psi_J(x))^\top$ be bounded and linearly independent basis functions with $\psi_1 \equiv 1$. Let $\{z_k\}_{k=1}^M$ be samples from the stationary process and $\hat{p}_s$ a kernel estimator of p_s. Define weights

$$w_k = \frac{\hat{p}_s(z_k)}{\sum_{\ell=1}^M \hat{p}_s(z_\ell)}. \tag{4}$$

Then set

$$G_M = \sum_{k=1}^M w_k\,\Psi(z_k)\Psi(z_k)^\top, \tag{5}$$

$$A_M = \sum_{k=1}^M w_k\,\Psi(z_{k+1})\Psi(z_k)^\top, \tag{6}$$

and estimate

$$\widehat{K}_M = A_M\, G_M^{\dagger}, \tag{7}$$

where $G_M^{\dagger}$ denotes the Moore–Penrose pseudoinverse. We assume $\max_k w_k = O_{\mathbb{P}}(1/M)$ and $J = o(M^{1/2})$.

The estimated Koopman spectrum converges to the true spectrum, and the Wasserstein forecast achieves the optimal parametric rate. Proofs are omitted due to space limitations. The results may not hold for non-stationary or degenerate diffusions or when dictionary functions are unbounded.

4 Theoretical Results and Applications

We establish that, under the stated modeling assumptions, the spectrum of the associated Koopman operator converges to the true spectrum, and that the prediction error measured in the Wasserstein metric converges accordingly. We conduct experiments on U.S. housing price data and compare our approach with the WAR method; the results demonstrate that our method achieves smaller Wasserstein prediction errors and exhibits robustness in the presence of structural changes in the data.

For scenarios in which the model assumptions are violated—where the Koopman operator may be non-self-adjoint or non-normal, as in gradient-descent-driven training dynamics of neural networks—we analyze the approximate spectrum obtained via Hankel-DMD. Our empirical findings indicate that wider networks yield spectral points closer to the real axis, implying faster convergence, while the discrepancies between narrow networks are more pronounced than those between wider networks.

5 Discussion and Significance

The proposed Koopman–Wasserstein framework connects stochastic dynamics and distributional geometry through spectral representation. The proposed framework introduces a mathematically consistent way to represent, analyze, and forecast distributional dynamics by com-bining the Koopman operator with Wasserstein geometry.

This research lies at the intersection of several major trends in mathematics, statistics, and information science. Mathematically, it links operator theory and stochastic analysis with geometric statistics, providing a spectral foundation for analyzing distribution-valued processes. In statistics, it extends the scope of functional data analysis to non-Euclidean domains, offering a unified viewpoint for dynamic distributions. In information science, it contributes to interpretable modeling of complex systems, complementing data-driven machine learning methods with rigorous operator-theoretic structure.

The significance of this work is twofold. Theoretically, it provides the first consistent spectral estimator for distributional dynamics, unifying operator-theoretic and statistical perspectives. Practically, it enables interpretable, geometry-preserving forecasting of evolving distributions. The empirical results illustrate how the method captures both smooth and abrupt distributional changes. The Koopman–Wasserstein framework thus provides a new foundation for the statistical analysis of complex dynamic systems.

This research was supported by JSPS KAKENHI Grant Numbers 25H01464 and 23K28042.

References

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*, 47(2):691–719, 2019.

Ziyue Wang and Yuko Araki. Functional time series forecasting of distributions: A koopman–wasserstein approach. *Behaviormetrika*, 2025. doi: 10.1007/s41237-025-00278-1.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data-driven approximation of the koopman operator: extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6): 1307–1346, 2015.

Chao Zhang, Piotr Kokoszka, and Alexander Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(1):30–52, 2022.

GMANOVA モデルの最尤推定量における直接的な罰則のプラグイン型最適化

永井 勇 [1]

[1] 中京大学 教養教育研究院

2025 年 12 月 2 日
データサイエンスの基盤を支える
次世代統計理論・方法論の挑戦と革新
@九州大学 西新プラザ

本研究の概要

対象 経時測定データ; 各個体で時間と共に測定したデータ
目的 データに潜む時間による変動を捉える
前提 全ての個体の測定時点が共通
→ GMANOVA モデルが分析によく使われている

従来の推定手法で起きる問題

(1) 各個体の (測定時点と無関係な) 説明変数の間に相関の高い組がある場合, 推定量が不安定になる
(2) 時間による変動を捉えるために用いる関数が柔軟すぎる場合, 目的の時間による変動ではなく分析に用いるデータに過剰に適合する

提案手法; (1) と (2) を回避する手法と最適化

(1) と (2) を回避するための罰則付推定量と罰則パラメータの最適化の話
(このモデルで同様の考え方で構築される様々な罰則付推定量の話を含む)

目次

ここからの話

モデルと仮定

対象 経時測定データ ; n 個体で時点と共に測定したデータ
目的 データに潜む時間による変動 (経時変動) を捉える
前提 全ての個体で測定時点は共通 (p 回測定)
→ GMANOVA モデル (Potthoff & Roy, 1964) で分析される;
$\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$,

- $\boldsymbol{Y}$; 各行が各個体の経時測定データからなる $n \times p$ 行列
- $\mathbf{1}_r$; r 次元の 1 ベクトル, $\mathbf{0}_h$; h 次元ゼロベクトル
- $\boldsymbol{X}$; 経時変動を捉えるために使う $p \times q$ 行列 (解析者が決める)
- $\boldsymbol{A}$; $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ (各列で中心化後) の説明変数を表す $n \times k$ 行列
- $\boldsymbol{\mu}$, $\boldsymbol{\Xi}$; q 次元未知ベクトル, $k \times q$ 未知行列
- $\boldsymbol{\mathcal{E}}$; $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ の $n \times p$ 誤差行列, $\boldsymbol{\Sigma}$; $p \times p$ 未知正定値行列

→ $\boldsymbol{Y}$, $\boldsymbol{A}$, $\boldsymbol{X}$ などから $\boldsymbol{\mu}$, $\boldsymbol{\Xi}$ (必要なら $\boldsymbol{\Sigma}$) を推定
仮定 $\mathrm{rank}(\boldsymbol{A}) = k$, $\mathrm{rank}(\boldsymbol{X}) = q$, ($\boldsymbol{\Sigma}$の不偏推定量)$^{-1}$ が存在

$\boldsymbol{\mu}$, $\boldsymbol{\Xi}$ の推定と経時変動の推定について

例 $\boldsymbol{X} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{q-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^{q-1} \end{pmatrix}$ とする ($\mathrm{rank}(\boldsymbol{X}) = q$ の範囲内で)
($t_1, \ldots, t_p$ は経時測定データの測定時点 ($t_1 < \cdots < t_p$))
→ この $\boldsymbol{X}$ をモデル ($\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$) で用いると,
$\boldsymbol{\mu}$; 全ての個体で共通の多項式の各次数の係数と切片
$\boldsymbol{\Xi}$; 各説明変数に対応した多項式の各次数の係数と切片
→ $\boldsymbol{\mu}$, $\boldsymbol{\Xi}$ の推定 ⇔ 多項式の各次数の係数と切片を推定
⇔ $\boldsymbol{Y}$ の経時変動 ($= E[\boldsymbol{Y}]$) を ($\boldsymbol{X}$ の形で) 推定

- $\boldsymbol{X}$ の中の関数を変える (例; $\boldsymbol{X}$ の (i,j) 成分 $= \exp(-(t_i - j)^2)$)
  ⇔ {使う関数の重み付き和で $\boldsymbol{Y}$ の経時変動を推定する形 / $\boldsymbol{\mu}$, $\boldsymbol{\Xi}$ はそれぞれの関数への重みに対応}

よく使われる推定量と問題

- $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$ のモデルで $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定

→ $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_{n,p}(\mathbf{0}_n\mathbf{0}_p', \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$ の下での最尤推定量 (MLE) は $p \times p$ 正定値行列 $\boldsymbol{M}$を用いた次の形で $\tilde{\boldsymbol{\mu}}(\boldsymbol{\Sigma}), \tilde{\boldsymbol{\Xi}}(\boldsymbol{\Sigma})$となる;

$$\begin{cases} \tilde{\boldsymbol{\mu}}(\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n, \\ \tilde{\boldsymbol{\Xi}}(\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}. \end{cases}$$

注 実データの予測などでは, $\boldsymbol{M}$ に次の $\boldsymbol{S}$ や $\boldsymbol{I}_p$ を代入;
$\boldsymbol{S} = \boldsymbol{Y}'\{\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n'/n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}\boldsymbol{Y}/(n-k-1)$.

- $\boldsymbol{S}$ は $E[\boldsymbol{S}] = \boldsymbol{\Sigma}$ ($\boldsymbol{S}$ は $\boldsymbol{\Sigma}$ の不偏推定量) で $\mathrm{rank}(\boldsymbol{S}) = p$

これらの推定量の問題

(1) $\boldsymbol{A}$ の列の間 (説明変数の組) に相関の高い組がある場合, $\boldsymbol{\Xi}$ の推定量が不安定になる (多重共線性)
(2) $\boldsymbol{X}$ に使う関数が柔軟すぎる場合, $\boldsymbol{Y}$ に過剰に適合する ⇔ 経時変動を上手く捉えられない (過剰適合)

本研究の目的とアイデア

- $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$ のモデルで $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定

$$\begin{cases} \tilde{\boldsymbol{\mu}}(\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n, \\ \tilde{\boldsymbol{\Xi}}(\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}. \end{cases}$$

- 問題点; $\boldsymbol{A}$ に相関が高い列の組がある → 不安定になる
  $\boldsymbol{X}$ に使う関数が柔軟すぎる → 過剰適合する

本研究の目的
これらの問題点を回避したい

アイデア; リッジ型の推定量の拡張
不安定になる⇒Hoerl and Kennard (1970) の推定量の流用
過剰適合する⇒ Nagai (2011) の推定量の流用 +拡張

これらの問題を回避する一般的な罰則付推定量と
そこで導入したパラメータの最適化などについて

ここからの話

提案する推定量 (一般形) へ (1/2)

- $\boldsymbol{Y} = \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$ のモデルで,
  不安定性・過剰適合を回避しつつ $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定 (= 経時変動の推定)
  - 仮定; $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$ (各列で中心化されている), $\mathrm{rank}(\boldsymbol{A}) = k$, $\mathrm{rank}(\boldsymbol{X}) = q$, $E[\boldsymbol{\mathcal{E}}] = \mathbf{0}_n\mathbf{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$
  - $\boldsymbol{\mu}$ と $\boldsymbol{\Xi}$ の推定量; $\tilde{\boldsymbol{\mu}}(\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$, $\tilde{\boldsymbol{\Xi}}(\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X})^{-1}$.
    (実際に使う場合は $\boldsymbol{M}$ に $\boldsymbol{S}$ や $\boldsymbol{I}_p$ を代入.)
- Hoerl and Kennard (1970) などのアイデアを使うと

$$\begin{cases} \tilde{\boldsymbol{\mu}}(\lambda|\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n, \\ \tilde{\boldsymbol{\Xi}}(\theta, \lambda|\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A} + \theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}; \end{cases}$$

ここで $\theta \geq 0$, $\lambda \geq 0$ は罰則パラメータ.

この推定量の問題点
θ と λ の最適化 ↔ 二変数関数の最適化問題
(両方, 陽に求まらない)

提案する推定量 (一般形) へ (2/2)

- 同時最適化が必要な 2 個の罰則パラメータ θ, λ を用い,

$$\begin{cases} \tilde{\boldsymbol{\mu}}(\lambda|\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n, \\ \tilde{\boldsymbol{\Xi}}(\theta, \lambda|\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A} + \theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}. \end{cases}$$

アイデア; 一般化リッジ回帰のアイデアを導入すると？！
→ $\mathrm{diag}(\boldsymbol{b})$ をベクトル $\boldsymbol{b}$ を対角に並べたものとし, $\boldsymbol{Q}_M$ を $\boldsymbol{d}_M = (d_{M,1}, \ldots, d_{M,q})'$ として $\boldsymbol{Q}_M'\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{Q}_M = \mathrm{diag}(\boldsymbol{d}_M)$ となる直交行列, $\boldsymbol{T}$ を $\boldsymbol{T}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{T}$ が対角行列となる直交行列, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)'_{(\lambda_i \geq 0)}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'_{(\theta_i \geq 0)}$ を用い,

$$\begin{cases} \tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{Q}_M')^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n, \\ \tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{T}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{T}')^{-1}\boldsymbol{A}'\boldsymbol{Y} \\ \qquad \times \boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{Q}_M')^{-1}. \end{cases}$$

嬉 $\|(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}')\boldsymbol{M}^{-1/2}\|^2 + \|(\mathbf{1}_n\boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi})\boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})^{1/2}\|^2 + \|\mathrm{diag}(\boldsymbol{\theta})^{1/2}\boldsymbol{T}'\boldsymbol{\Xi}\boldsymbol{X}'\boldsymbol{M}^{-1/2}\|^2 + \|\mathrm{diag}(\boldsymbol{\theta})^{1/2}\boldsymbol{T}'\boldsymbol{\Xi}\boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})^{1/2}\|^2$

パラメータ最適化の目的の関数と疑問

- $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{M})$ や $\tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{M})$ の $\boldsymbol{\theta}$ ($= (\theta_1, \ldots, \theta_k)'$), $\boldsymbol{\lambda}$ ($= (\lambda_1, \ldots, \lambda_q)'$) をどの基準で最適化？

再 $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{M}) = (\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{Q}_M')^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{Y}'\mathbf{1}_n/n$,
$\tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{M}) = (\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{T}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{T}')^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{M}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{Q}_M')^{-1}$:

→ 予測平均二乗誤差 (PMSE) を小さくするように選ぶ;
$\mathrm{PMSE}[\hat{\boldsymbol{Y}}] \overset{\mathrm{def.}}{=} E_{\boldsymbol{Y}}[E_{\boldsymbol{Y}_F}[\mathrm{tr}\{(\boldsymbol{Y}_F - \hat{\boldsymbol{Y}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_F - \hat{\boldsymbol{Y}})'\}]]$,
ここで, $\boldsymbol{Y}_F$ は $\boldsymbol{Y}$ と独立に同一の分布から得られるもの, $\hat{\boldsymbol{Y}}$ は今の $\boldsymbol{Y}$ などからの予測値, $E_{\boldsymbol{W}}[\cdot]$ は確率変数 $\boldsymbol{W}$ に基づいた期待値.

目的の最適化; $\arg\min_{\theta_1,\ldots,\theta_k,\lambda_1,\ldots,\lambda_q} \mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})]$,
ここで $\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{1}_n\tilde{\boldsymbol{\mu}}'(\boldsymbol{\lambda}|\boldsymbol{M})\boldsymbol{X}' + \boldsymbol{A}\tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{M})\boldsymbol{X}'$.

→ 疑問; 経時変動を捉えたいのになぜ PMSE で選ぶ？
なぜ予測を良くするように $\boldsymbol{\theta}$ と $\boldsymbol{\lambda}$ を選ぶ？

パラメータの最適化へ向けて -変形-

- なぜ経時変動を捉えたいのに, PMSE を最小に？
  - なぜ $\arg\min_{\theta_1,\ldots,\theta_k,\lambda_1,\ldots,\lambda_q} \mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})]$ でパラメータ最適化？

疑問 PMSE を小さくする ⇔ 経時変動を上手く捉える
変形 $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})]$ の定数を除いた部分;
$E[\mathrm{tr}\{(\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - E[\boldsymbol{Y}])\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - E[\boldsymbol{Y}])'\}]$.

- $E_{\boldsymbol{Y}}[\cdot]$ を $E[\cdot]$ と略記

再掲 $\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathbf{1}_n\tilde{\boldsymbol{\mu}}'(\boldsymbol{\lambda}|\boldsymbol{M})\boldsymbol{X}' + \boldsymbol{A}\tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{M})\boldsymbol{X}'$
結局 $\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ と $E[\boldsymbol{Y}]$ (=経時変動) の差を測っている

疑問 (なぜ経時変動を捉えたいのに PMSE で最適化？) への回答
PMSE を小さくする ⇔ 経時変動を上手く捉える

→ **パラメータの最適化について**
経時変動などがあるが $\boldsymbol{\theta}$ と $\boldsymbol{\lambda}$ をどう最適化？

パラメータの最適化へ -どうやって最適化？ -

問題 PMSE を小さくする $\boldsymbol{\theta}$ と $\boldsymbol{\lambda}$ をどう得る？

- $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})] = E[\mathrm{tr}\{(\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - E[\boldsymbol{Y}])\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - E[\boldsymbol{Y}])'\}]$
  $+$ 定数 ($\boldsymbol{\theta}$と$\boldsymbol{\lambda}$に無関係な項).
- $E[\boldsymbol{Y}]$ (=経時変動 (未知)) や $\boldsymbol{\Sigma}$ (未知) が必要

① $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})]$ を推定する関数 (C_p 型情報量規準) を作って, その関数を最小にする $\boldsymbol{\theta}$ や $\boldsymbol{\lambda}$ を求める

- 今回の話だと, 以下のような形;
  $C_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathrm{tr}\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}))\boldsymbol{S}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}))'\} + 2\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}})\mathrm{tr}(\boldsymbol{G}_{\boldsymbol{\lambda},M})$,
  ここで, $\boldsymbol{H}_{\boldsymbol{\theta}} = \mathbf{1}_n\mathbf{1}_n'/n + \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{T}\mathrm{diag}(\boldsymbol{\theta})\boldsymbol{T}')^{-1}\boldsymbol{A}'$,
  $\boldsymbol{G}_{\boldsymbol{\lambda},M} = \boldsymbol{M}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X} + \boldsymbol{Q}_M\mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{Q}_M')^{-1}\boldsymbol{X}'\boldsymbol{M}^{-1/2}$ を用い, $\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{Y}\boldsymbol{M}^{-1/2}\boldsymbol{G}_{\boldsymbol{\lambda},M}\boldsymbol{M}^{1/2}$.
  再掲 $\boldsymbol{Q}_M$ は $\boldsymbol{Q}_M'\boldsymbol{X}'\boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{Q}_M$ が対角行列になる直交行列

② $\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\boldsymbol{\theta}, \boldsymbol{\lambda})]$ を最小にする $\boldsymbol{\theta}$ や $\boldsymbol{\lambda}$ を(未知の項を含む形で)導出し, 未知の項に推定量を代入

パラメータの最適化へ -今の推定量での問題とその回避-

問題 PMSE を小さくする $\boldsymbol{\theta}$ と $\boldsymbol{\lambda}$ を (①か②のどちらかを用いて) どう得る？
① PMSE を推定する関数を小さくする (C_p 型最適化)
② PMSE を小さくするものに推定量を代入する (Plug-in 型最適化)

→ **最適化してみたら起きた問題** (ご想像の通りだと思いますが)

- 最適な $\boldsymbol{\theta}$ のために $\boldsymbol{\lambda}$ が全部必要
- 最適な $\boldsymbol{\lambda}$ のために $\boldsymbol{\theta}$ が全部必要

→ 最適な $\boldsymbol{\theta}$ を求める 反復⇔ 最適な $\boldsymbol{\lambda}$ を求める
(収束するか否か・反復のため計算時間増 . . . の問題が！)

→ **回避のためのアイデア**
両方を一般化リッジ回帰の形にするから起きる！
⇒ 片方だけをリッジの形に制限して考える

パラメータの最適化へ -制限-

- $\boldsymbol{\theta}$ と $\boldsymbol{\lambda}$ の両方使うと大変

→ 片方をリッジ回帰の形にしてみる

→ パターンとしては以下の二つのパターンがある

(I) $\boldsymbol{\theta}$ と $\boldsymbol{\lambda} = \lambda \mathbf{1}_q$ $(\lambda \geq 0)$ にする

↔ $\tilde{\boldsymbol{\mu}}(\lambda \mathbf{1}_q | \boldsymbol{M})$ と $\tilde{\boldsymbol{\Xi}}(\boldsymbol{\theta}, \lambda \mathbf{1}_q | \boldsymbol{M})$ を使う

($\boldsymbol{X}$ に関する罰則の方だけを λ 一個 (リッジ) の形に)

(II) $\boldsymbol{\theta} = \theta \mathbf{1}_k$ $(\theta \geq 0)$ と $\boldsymbol{\lambda}$ にする

↔ $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda} | \boldsymbol{M})$ と $\tilde{\boldsymbol{\Xi}}(\theta \mathbf{1}_k, \boldsymbol{\lambda} | \boldsymbol{M})$ を使う

($\boldsymbol{A}$ に関する罰則の方だけを θ 一個 (リッジ) の形に)

⇒ さらなる分岐;

- $\boldsymbol{M} = \boldsymbol{I}_p$ か $\boldsymbol{M} = \boldsymbol{\Sigma}$ or $\boldsymbol{S}$

($\tilde{\boldsymbol{\mu}}(\mathbf{0}_q | \boldsymbol{I}_p)$, $\tilde{\boldsymbol{\Xi}}(\mathbf{0}_k, \mathbf{0}_q | \boldsymbol{I}_p)$ は Nagai (2011) で基にした推定量)

- 最適化法; ① C_p 型最適化 or ② Plug-in 型最適化

(これらの一部はすでに研究されている)

パラメータの最適化へ -どこができてるか-

- 片方をリッジ回帰の形にした場合で, どのパターンまでできているのか
- $\boldsymbol{M} = \boldsymbol{I}_p$ の場合

(ほぼ全部済み)	$\boldsymbol{\theta}$ と $\lambda \mathbf{1}_q$ の場合	$\theta \mathbf{1}_k$ と $\boldsymbol{\lambda}$ の場合
C_p 型最適化	Nagai (2011)	永井 (2020)
Plug-in 型最適化	計算済み (どこかで講演したはず…)	

- $\boldsymbol{M} = \boldsymbol{\Sigma}$ or $\boldsymbol{S}$ とした場合

(一部未完成)	$\boldsymbol{\theta}$ と $\lambda \mathbf{1}_q$ の場合	$\theta \mathbf{1}_k$ と $\boldsymbol{\lambda}$ の場合
C_p 型最適化	永井 (2018)	永井 (2025)
Plug-in 型最適化	計算中	今回

→ $\boldsymbol{M} = \boldsymbol{\Sigma}$ として $\theta \mathbf{1}_k$ と $\boldsymbol{\lambda}$ の最適化を考える

- $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{\Sigma})$, $\tilde{\boldsymbol{\Xi}}(\theta, \boldsymbol{\lambda}|\boldsymbol{\Sigma})$ を使った予測値の PMSE を最小にする
- パラメータを求めて未知の項に推定量を代入

パラメータの最適化 (1/6) -C_p 型最適化-

- C_p 型最適化 (永井, 2025) をまず見る

C_p 型情報量規準に基づく最適化 (詳細は永井 (2025))

$\operatorname{argmin}_{\lambda_1,\ldots,\lambda_q} C_p(\theta \mathbf{1}_k, \boldsymbol{\lambda})$ を (θ を固定した下で) それぞれ求めると,

$$\hat{\lambda}_i^C(\theta) = \begin{cases} 0 & (\text{if } \dot{C}_\theta > 0) \\ \infty & (\text{if } (\boldsymbol{Z}'\boldsymbol{H}_{\theta 1_k}\boldsymbol{Z})_{ii} \leq \operatorname{tr}(\boldsymbol{H}_{\theta 1_k})) \\ \dfrac{-d_i^2 \dot{C}_\theta}{2\{(\boldsymbol{Z}'\boldsymbol{H}_{\theta 1_k}\boldsymbol{Z})_{ii} - \operatorname{tr}(\boldsymbol{H}_{\theta 1_k})\}} & (\text{その他}) \end{cases},$$

ここで $(\boldsymbol{V})_{ii}$ は行列 $\boldsymbol{V}$ の第 (i,i) 成分を表し, $\boldsymbol{P}$ は $\boldsymbol{S}^{-1/2}\boldsymbol{X}$ の左特異ベクトルからなる直交行列, $\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{S}^{-1/2}\boldsymbol{P}'$, $\dot{C}_\theta = 2\{(\boldsymbol{Z}'\boldsymbol{H}_{\theta 1_k}\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{H}_{\theta 1_k}^2\boldsymbol{Z})_{ii} - \operatorname{tr}(\boldsymbol{H}_{\theta 1_k})\}/d_i$.

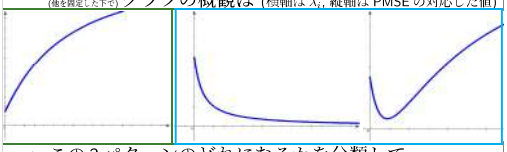
($\boldsymbol{P}$ は $\boldsymbol{D}_{\boldsymbol{S}} = \operatorname{diag}(\boldsymbol{d}_{\boldsymbol{S}})$ として $\boldsymbol{P}\boldsymbol{S}^{1/2}\boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{S}} = (\boldsymbol{D}_{\boldsymbol{S}}^{1/2}, \mathbf{0}_q \mathbf{0}_{p-q}')'$ となる直交行列.)

- $\hat{\boldsymbol{\lambda}}^C(\theta) = (\hat{\lambda}_1^C(\theta), \ldots, \hat{\lambda}_q^C(\theta))'$ として, $\operatorname{argmin}_{\theta \geq 0} C_p(\theta, \hat{\boldsymbol{\lambda}}^C(\theta))$ で θ の最適化

パラメータの最適化 (2/6) -PMSE を最小化する λ (1/2)-

目的 PMSE を最小にするパラメータは実際どんなもの？

- $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{\Sigma})$, $\tilde{\boldsymbol{\Xi}}(\theta 1_k, \boldsymbol{\lambda}|\boldsymbol{\Sigma})$ に基づく $\tilde{\boldsymbol{Y}}(\theta 1_k, \boldsymbol{\lambda})$ の PMSE の (θ を固定した下で) グラフの概観は (横軸は λ_i, 縦軸は PMSE の対応した値)



→ この 3 パターンのどれになるかを分類して, $\lambda_i \geq 0$ の範囲で PMSE を最小にする各 λ_i を求める.

- $\dot{P}_\theta \overset{\text{def.}}{=} \partial \mathrm{PMSE}[\tilde{\boldsymbol{Y}}(\theta \mathbf{1}_k, \boldsymbol{\lambda})]/(\partial \lambda_i)|_{\lambda_i = 0}$ とすると,

$\dot{P}_\theta > 0$ だと 左側, $\dot{P}_\theta < 0$ だと 中央か右側.

パラメータの最適化 (3/6) -PMSE を最小化する λ (2/2)-

目的 PMSE を最小にするパラメータは実際に求めると

$\arg\min_{\theta, \lambda_1, \ldots, \lambda_q} \mathrm{PMSE}[\tilde{\boldsymbol{Y}}(\theta 1_k, \boldsymbol{\lambda})]$ を θ を固定した下で求めた結果

$\mathrm{PMSE}[\hat{\boldsymbol{Y}}(\theta 1_k, \boldsymbol{\lambda})]$ を最小にする λ_i^* $(i = 1, \ldots, q)$ はそれぞれ,

$$\lambda_i^*(\theta) = \begin{cases} 0 & (\text{if } \dot{P}_\theta > 0) \\ \infty & (\text{if } (\boldsymbol{Q}_{\boldsymbol{\Sigma}}'\boldsymbol{K}_*'\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}_*\boldsymbol{Q}_{\boldsymbol{\Sigma}})_{ii} \leq 0) \\ \dfrac{-d_{\boldsymbol{\Sigma},i}\dot{P}_\theta}{2(\boldsymbol{Q}_{\boldsymbol{\Sigma}}'\boldsymbol{K}_*'\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}_*\boldsymbol{Q}_{\boldsymbol{\Sigma}})_{ii}} & (\text{その他}) \end{cases},$$

ここで $\boldsymbol{K}_* = \mathbf{1}_n \boldsymbol{\mu}' + \boldsymbol{A}\boldsymbol{\Xi}$, $\dot{P}_\theta = \partial \mathrm{PMSE}[\tilde{\boldsymbol{Y}}(\theta 1_k, \boldsymbol{\lambda})]/(\partial \lambda_i)|_{\lambda_i = 0}$
$= -2\{\operatorname{tr}(\boldsymbol{H}_{\theta 1_k}^2)/d_{\boldsymbol{\Sigma},i} + (\boldsymbol{Q}_{\boldsymbol{\Sigma}}'\boldsymbol{K}_*'\boldsymbol{H}_{\theta 1_k}(\boldsymbol{H}_{\theta 1_k} - \boldsymbol{I}_n)\boldsymbol{K}_*\boldsymbol{Q}_{\boldsymbol{\Sigma}})_{ii}\}$.

(再掲; $\boldsymbol{Q}_{\boldsymbol{\Sigma}}'\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{\Sigma}} = \operatorname{diag}(d_{\boldsymbol{\Sigma},1}, \ldots, d_{\boldsymbol{\Sigma},q})$, $\boldsymbol{H}_{\theta 1_k} = 1_n 1_n'/n + \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \theta \boldsymbol{I}_k)^{-1}\boldsymbol{A}'$,
さらに $(\boldsymbol{V})_{ii}$ は行列 $\boldsymbol{V}$ の (i,i) 成分を表している.)

($\lambda_i^*(\theta) = \infty$ になることはほぼないと思われる.)

パラメータの最適化 (4/6) -Plug-in 型最適化-

Plug-in 型最適化の結果; $\lambda_i^*(\theta)$ $(i = 1, \ldots, q)$ に色々代入

λ_i^* $(i = 1, \ldots, q)$ の未知の項に推定量を代入すると,

$$\hat{\lambda}_i(\theta) = \begin{cases} 0 & (\text{if } \hat{P}_\theta > 0) \\ \infty & (\text{if } (\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}'\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}\boldsymbol{Q}_{\boldsymbol{S}})_{ii} \leq 0) \\ \dfrac{-d_{\boldsymbol{S},i}\hat{P}_\theta}{2(\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}'\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}\boldsymbol{Q}_{\boldsymbol{S}})_{ii}} & (\text{その他}) \end{cases},$$

ここで $\boldsymbol{K} = \mathbf{1}_n \tilde{\boldsymbol{\mu}}'(\mathbf{0}_q | \boldsymbol{S}) + \boldsymbol{A}\tilde{\boldsymbol{\Xi}}(0, \mathbf{0}_q | \boldsymbol{S})$ ($\boldsymbol{K}$ の $\boldsymbol{\mu} \cdot \boldsymbol{\Xi}$ に推定量を代入),
$\hat{P}_\theta = -2\{\operatorname{tr}(\boldsymbol{H}_{\theta 1_k}^2)/d_{\boldsymbol{S},i} + (\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}'\boldsymbol{H}_{\theta 1_k}(\boldsymbol{H}_{\theta 1_k} - \boldsymbol{I}_n)\boldsymbol{K}\boldsymbol{Q}_{\boldsymbol{S}})_{ii}\}$.

(再掲; $\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{S}} = \operatorname{diag}(d_{\boldsymbol{S},1}, \ldots, d_{\boldsymbol{S},q})$, $\boldsymbol{H}_{\theta 1_k} = 1_n 1_n'/n + \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \theta \boldsymbol{I}_k)^{-1}\boldsymbol{A}'$.)

- $\hat{\boldsymbol{\lambda}}(\theta) = (\hat{\lambda}_1(\theta), \ldots, \hat{\lambda}_q(\theta))'$ として, $\operatorname{argmin}_{\theta \geq 0} C_p(\theta, \hat{\boldsymbol{\lambda}}(\theta))$ で θ の最適化.

パラメータの最適化 (5/6) -繰り返し代入 (1/2)-

- Plug-in して得られる最適化結果; $\hat{\boldsymbol{\lambda}}(\theta)$ $(= (\hat{\lambda}_1(\theta), \ldots, \hat{\lambda}_q(\theta))'$
- 未知の項に MLE などを入れている

→ **$\hat{\boldsymbol{\lambda}}(\theta)$ の問題点**

MLE などが不安定 ⇒ $\hat{\boldsymbol{\lambda}}(\theta)$ も不安定になるのでは？

→ **この問題点の解決法のアイデア**

Nagai, Yanagihara and Satoh (2012) の手法

- 最適化なパラメータ ($\lambda_i^*(\theta)$) の未知の項に MLE などを代入
  → そのパラメータを用いて推定量を更新
  → 更新した推定量を $\lambda_i^*(\theta)$ の未知の項に代入し, 推定量を再更新
  → 再更新した推定量を $\lambda_i^*(\theta)$ に代入し, 推定量を再々更新
  → また代入 …

⇒ 今回の $\hat{\boldsymbol{\lambda}}(\theta)$ で使った推定量も更新していく

パラメータの最適化 (6/6) -繰り返し代入 (2/2)-

Plug-in する推定量を更新してパラメータを更新する

$\hat{\boldsymbol{\lambda}}^{[s]}(\theta) \overset{\text{def.}}{=} (\hat{\lambda}_1^{[s]}(\theta), \ldots, \hat{\lambda}_q^{[s]}(\theta))'$ として, 以下で更新する;

$$\hat{\lambda}_i^{[s]}(\theta) = \begin{cases} 0 & (\text{if } \hat{P}_\theta^{[s-1]} > 0) \\ \infty & (\text{if } (\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}^{[s-1]\prime}\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}^{[s-1]}\boldsymbol{Q}_{\boldsymbol{S}})_{ii} \leq 0) \\ \dfrac{-d_{\boldsymbol{S},i}\hat{P}_\theta^{[s-1]}}{2(\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}^{[s-1]\prime}\boldsymbol{H}_{\theta 1_k}\boldsymbol{K}^{[s-1]}\boldsymbol{Q}_{\boldsymbol{S}})_{ii}} & (\text{その他}) \end{cases},$$

ここで, $\boldsymbol{K}^{[0]} = \mathbf{1}_n \hat{\boldsymbol{\mu}}'(\mathbf{0}_q | \boldsymbol{S}) + \boldsymbol{A}\tilde{\boldsymbol{\Xi}}(0, \mathbf{0}_q | \boldsymbol{S})$ (前の $\boldsymbol{K}$ と同じ),
$\ell \geq 1$ では $\boldsymbol{K}^{[\ell]} = \mathbf{1}_n \tilde{\boldsymbol{\mu}}'(\hat{\boldsymbol{\lambda}}^{[\ell]}(\theta) | \boldsymbol{S}) + \boldsymbol{A}\tilde{\boldsymbol{\Xi}}(0, \hat{\boldsymbol{\lambda}}^{[\ell]}(\theta) | \boldsymbol{S})$,
$\hat{P}_\theta^{[\ell]} = -2\{\operatorname{tr}(\boldsymbol{H}_{\theta 1_k}^2)/d_{\boldsymbol{S},i} + (\boldsymbol{Q}_{\boldsymbol{S}}'\boldsymbol{K}^{[\ell]\prime}\boldsymbol{H}_{\theta 1_k}(\boldsymbol{H}_{\theta 1_k} - \boldsymbol{I}_n)\boldsymbol{K}^{[\ell]}\boldsymbol{Q}_{\boldsymbol{S}})_{ii}\}$.

- s を固定して, $\operatorname{argmin}_{\theta \geq 0} C_p(\theta, \hat{\boldsymbol{\lambda}}^{[s]}(\theta))$ で θ の最適化
- $s = 1$ の場合は, MLE を代入したものに対応

ここからの話

まとめ

- 全個体で測定時点が共通の経時測定データの分析

→ $\boldsymbol{Y} = \mathbf{1}_n \boldsymbol{\mu}' \boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}}$ (GMANOVA モデル) が分析によく使われる

問題点 従来の推定量の問題点;
$\boldsymbol{A}$ の列に相関の高い組がある場合, 不安定になる.
$\boldsymbol{X}$ に用いる関数が柔軟すぎる場合, 過剰適合する.

本研究 これらの問題点を回避する罰則付推定量の提案

- Hoerl & Kennard (1970), Nagai (2011) のアイデアの拡張

→ 導入したパラメータ (θ, λ) の最適化が必要

→ 一部に一般化リッジ回帰による推定のアイデア導入

⇒ 1 個のパラメータ (θ) のみの最適化の形へ

課題 $\tilde{\boldsymbol{\mu}}(\boldsymbol{\lambda}|\boldsymbol{S})$ と $\tilde{\boldsymbol{\Xi}}(\theta, \boldsymbol{\lambda}|\boldsymbol{S})$ を用いた場合の最適な $\boldsymbol{\lambda}$ の導出
$\hat{\lambda}_i^{[s]}(\theta) \leq \hat{\lambda}_i^{[s+1]}(\theta)$ の証明など

参考文献

1. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.
2. Kokoszka, P. & Reimherr, M. (2017). *Introduction to Functional Data Analysis*, Chapman and Hall/CRC.
3. Nagai, I. (2011). Modified C_p criterion for optimizing ridge and smooth parameters in the MGR Estimator for the nonparametric GMANOVA model. *Open Journal of Statistics*, **1**, 1–14.
4. 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.
5. 永井 勇 (2018). Plug-in optimization method for generalized ridge regression for MLE in GMANOVA model. 多変量データ解析法における理論と応用.
6. 永井 勇 (2020). GMANOVA における直接的な罰則付推定法とその最適化. 2020 年度統計関連学会連合大会.
7. 永井 勇 (2025). Direct penalization method for MLE in the GMANOVA model and its optimization with C_p type criterion. データサイエンスにおける統計的理論の展開研究.
8. Potthoff, R. F. & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

終

Regularized k-POD Clustering for Missing Data

Xin Guan*[1] and Yoshikazu Terada[2]

[1]Graduate School of Information Sciences, Tohoku University
[2]Graduate School of Engineering Science, Osaka University

1 Introduction

Clustering is an important technique that groups data points without labels into several clusters. Notably, the k-means clustering is one of the most popular clustering methods, the main idea of which is to find k cluster centers and then cluster data points by assigning them to their nearest centers. Since the classical k-means clustering requires the data matrix to be complete, then in the case of missing data, directly conducting k-means on an incomplete data matrix is infeasible.

To deal with missingness for clustering, the traditional approach is to pre-process the incomplete data matrix by deletion or imputation to construct a new complete data matrix for conducting k-means clustering. However, these approaches do not work for large proportions of missingness or need a long computational time. Recently, the k-POD clustering was proposed by Chi et al. (2016) as a natural extension for k-means clustering to missing data, which can be applicable for even large missingness proportions within a short computational time. However, the k-POD clustering is not consistent with k-means in general, even under the missing completely at random mechanism (Terada & Guan 2025). That is, as $n \to \infty$, the estimated cluster centers of the k-POD clustering and k-means clustering would converge to different limits. Moreover, since the inconsistency is essentially due to the difference of loss functions between k-POD and k-means, it is challenging to propose a general de-biasing method.

In this work, our aim is to reduce the estimation bias of existing k-POD clustering, and we focus on a special case when there exist some features irrelevant to the cluster structure. To this end, we propose regularized k-POD clustering by introducing a regularization function of cluster centers to the loss of k-POD clustering, which shrinks cluster centers feature-wisely. This offers a significant advantage of reducing the bias of estimated cluster centers, particularly in irrelevant features. Our numerical experiments verify the effects of reducing estimation bias and improving clustering accuracy, and applications to real-world single cell RNA-sequencing data also show the better performance of the proposed method.

2 Methodology

2.1 Notations and preliminaries

Write $\mathrm{X} = (x_{ij})_{n\times p} \in \mathbb{R}^{n\times p}$ for the data matrix with n data points $\mathrm{X}_1, \ldots, \mathrm{X}_n$ in $\mathbb{R}^p$. The k cluster centers $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ are encoded by a matrix $\mathrm{M} = (\mu_{lj})_{k\times p} \in \mathbb{R}^{k\times p}$, where the l-th row represents the l-th cluster center. The membership between data points and cluster centers is denoted by a binary matrix $\mathrm{U} = (u_{il})_{n\times k} \in \{0,1\}^{n\times k}$, where $u_{il} = 1$ if and only if i-th data point X_i is assigned to l-th cluster. Since one data point is assigned to a unique cluster, it must satisfy that $\mathrm{U}\mathbf{1}_k = \mathbf{1}_n$, where $\mathbf{1}$ is the all-one vector.

For a complete data matrix X, the k-means clustering can be expressed as

$$\min_{\mathrm{U},\mathrm{M}} \quad \|\mathrm{X} - \mathrm{UM}\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, calculated as $(\sum_{i,j} a_{ij}^2)^{1/2}$ for $\mathrm{A} = (a_{ij})$. If there exist missing entries in X, the loss function cannot be directly calculated.

For an incomplete data matrix, the k-POD clustering records all observed positions in X by a set $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$, and introduces a mapping $\mathcal{P}$ onto the set Ω to replace the missing entries with zero. That is, $\mathcal{P}_\Omega : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$, and $(\mathcal{P}_\Omega(\mathrm{X}))_{ij} = x_{ij}$ if $(i,j) \in \Omega$, 0 otherwise. Then, the k-POD clustering is given by

$$\min_{\mathrm{U},\mathrm{M}} \quad \|\mathcal{P}_\Omega(\mathrm{X} - \mathrm{UM})\|_F^2. \tag{2}$$

*Corresponding author: guan.xin.c5@tohoku.ac.jp (XG)

2.2 Proposed method

Suppose that the data matrix $\mathrm{X} = (x_{ij})_{n\times p}$ is column-wised centered, that is, $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ for all $j = 1, \ldots, p$. Write $\mathrm{X}_i \in \mathbb{R}^p$ for the i-th data point $(i = 1, \ldots, n)$ and $\mathrm{X}_{(j)} \in \mathbb{R}^n$ for the j-th column of X $(j = 1, \ldots, p)$. Denote by Ω the set of observed positions of X and suppose that the number of clusters k is fixed.

We define the regularized k-POD clustering with respect to membership $\mathrm{U} \in \{0, 1\}^{n\times k}$, $\mathrm{U}\mathbf{1}_k = \mathbf{1}_n$, and cluster centers $\mathrm{M} \in \mathbb{R}^{k\times p}$ by

$$\min_{\mathrm{U},\mathrm{M}} \quad \|\mathcal{P}_\Omega(\mathrm{X} - \mathrm{UM})\|_F^2 + \lambda \cdot J(\mathrm{M}). \tag{3}$$

The first term is the loss of the k-POD clustering, and $J(\mathrm{M})$ is a regularization function with respect to M. To shrink the estimated cluster centers feature-wisely, we consider two types of $J(\mathrm{M})$:

$$\text{The } l_0 \text{ penalty}: \quad J_0(\mathrm{M}) = \sum_{j=1}^{p} \mathbb{1}(\|\mathrm{M}_{(j)}\| > 0)$$

$$\text{The group lasso penalty}: \quad J_1(\mathrm{M}) = \sum_{j=1}^{p} w_j \|\mathrm{M}_{(j)}\|,$$

where $\mathrm{M}_{(j)} = (\mu_{1j}, \ldots, \mu_{kj})^T$ denotes the j-th column of cluster centers M with μ_{lj} being the j-th component of the l-th cluster center $(l = 1, \ldots, k)$. The function $\mathbb{1}(\cdot)$ is the indicator function and w_j is the weight for $\mathrm{M}_{(j)}$. Both types of $J(\cdot)$ are column-wised, which means that all elements of $\mathrm{M}_{(j)}$, that is $\{\mu_{1j}, \ldots, \mu_{kj}\}$ would be shrunk together. The l_0 type $J_0(\cdot)$ constrains the number of non-zero columns of M, while the group lasso type $J_1(\cdot)$ constrains the weighted sum of l_2 norms of M in each feature. Therefore, with a suitable regularization parameter λ, the estimated cluster centers would be sparse in features.

We apply the majorization-minimization algorithm (MM algorithm) to solve Eq. (3). Specifically, given current $\mathrm{U}^{(t)}$ and $\mathrm{M}^{(t)}$, $t \in \mathbb{N}$, the $(t+1)$-th iteration consists of two steps. Step 1 imputes missing entries of X by the corresponding entries of multiplication matrix of current $\mathrm{U}^{(t)}$ and $\mathrm{M}^{(t)}$, so that we can get a new complete data matrix $\widehat{\mathrm{X}}^{(t+1)}$. Step 2 updates $\mathrm{U}^{(t+1)}$ and $\mathrm{M}^{(t+1)}$ by applying regularized k-means clustering on the imputed data matrix $\widehat{\mathrm{X}}^{(t+1)}$. Repeat the iteration until the loss converges.

3 Experiments

In this work, we compare the proposed method with other methods via numerical experiments, the results of which show a lower bias in estimating cluster centers as well as higher accuracy in clustering. Moreover, applications to real-world data also show the better performance of the proposed method. More details will be introduced in this talk.

References

Chi, J. T., Chi, E. C. & Baraniuk, R. G. (2016), 'k-pod: A method for k-means clustering of missing data', *The American Statistician* **70**(1), 91–99.

Terada, Y. & Guan, X. (2025), 'A note on the k-means clustering for missing data', *Transactions on Machine Learning Research* .
URL: *https://openreview.net/forum?id=pcqlTvePXS*

単調欠測データに対するMardiaの正規性検定統計量について

東京理科大学・理・院　栗田 絵梨

本報告では単調欠測データの下での多変量正規性検定問題について議論した．データが欠測していない完全データの下での多変量尖度の定義は Mardia (1970)，Srivastava (1984), Koziol (1989) などが与え，それぞれ多変量正規性検定のための統計量の提案を行なっている．その中で特に，Mardia が定義した多変量尖度の標本版を基に欠測データに対応できるように改良を行なった．本報告では Kurita and Seo (2022) で扱った 2-step 単調欠測データの下での多変量標本尖度から一般の k-step 単調欠測データの下での多変量標本尖度への拡張することで定義を与えた．そして，k-step 単調欠測データを 2-step ごとに分割し，Kurita and Seo (2022) で導出した期待値と分散をそれぞれ組み合わせることで，k-step 単調欠測データの下での多変量標本尖度の期待値と分散の漸近展開近似したものを得ることができる．最後に幾つかのパラメータの下で，モンテカルロ・シミュレーションで実験を行い，提案した検定統計量の有効性を示した．

本研究では 2-step 単調欠測データの下での多変量標本尖度の定義を拡張することで k-step 単調欠測データの下での多変量標本尖度の定義の導出および検定統計量を提案する． N_1 個の $p=(p_1+\cdots+p_k)$ 次元観測ベクトル $(\boldsymbol{x}_{1,i}^\top,\ldots,\boldsymbol{x}_{k,i}^\top)^\top,\ i=1,\cdots,N_1$ が $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ に従い，N_2 個の $(p_1+\cdots+p_{k-1})$ 次元観測ベクトル $(\boldsymbol{x}_{1,i}^\top,\ldots,\boldsymbol{x}_{k-1,i}^\top)^\top,\ i=N_1+1,\cdots,N_{1:2}$ が $N_{p_{1:k-1}}(\boldsymbol{\mu}_{(1\ldots k-1)},\boldsymbol{\Sigma}_{(1\ldots k-1)(1\ldots k-1)})$ に従い，そして N_k 個の p_1 次元観測ベクトル $\boldsymbol{x}_{1,i},\ i=N_{1:k-1}+1,\cdots,N$ が $N_{p_1}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11})$ に従っているとする．ただし $p=p_1+\cdots+p_k=p_{1:k}$, $N=N_1+\cdots+N_k=N_{1:k}$. と表記する．k-step 単調欠測データは以下のような形となっている．

$$
\begin{pmatrix}
\boldsymbol{x}_{1,1}^\top & \boldsymbol{x}_{2,1}^\top & \cdots & \boldsymbol{x}_{k,1}^\top \\
\vdots & \vdots & \vdots & \vdots \\
\boldsymbol{x}_{1,N_1}^\top & \boldsymbol{x}_{2,N_1}^\top & \vdots & \boldsymbol{x}_{k,N_1}^\top \\
\vdots & \vdots & & * \\
\boldsymbol{x}_{1,N_{1:k-2}+1}^\top & \boldsymbol{x}_{2,N_{1:k-2}+1}^\top & * & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\boldsymbol{x}_{1,N_{1:k-1}}^\top & \boldsymbol{x}_{2,N_{1:k-1}}^\top & \vdots & \vdots \\
\boldsymbol{x}_{1,N_{1:k-1}+1}^\top & * & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\boldsymbol{x}_{1,N}^\top & * & * & *
\end{pmatrix},
$$

ただし，“$*$”は欠測部分を表す．このとき k-step 単調欠測データ下での Mardia 型多変量標本尖度は

$$
b_{2,p}^{(k)}=\sum_{j=1}^{k}\left\{\frac{1}{N_{1:k+1-j}}\sum_{i=1}^{N_{1:k+1-j}}(U_{j,i}^{(k)})^2\right\}+\sum_{1\le m<l\le k}\left\{\frac{2}{N_{1:k+1-l}}\sum_{i=1}^{N_{1:k+1-l}}U_{m,i}^{(k)}U_{l,i}^{(k)}\right\}
$$

と与えることができる．ただし，

$$U_{1,i}^{(k)} = (\boldsymbol{x}_{1,i} - \widehat{\boldsymbol{\mu}}_1)^\top \widehat{\boldsymbol{\Sigma}}_{11}^{-1} (\boldsymbol{x}_{1,i} - \widehat{\boldsymbol{\mu}}_1),$$
$$U_{j,i}^{(k)} = (\boldsymbol{x}_{j\cdot 1:j-1,i} - \widehat{\boldsymbol{\mu}}_{j\cdot 1:j-1})^\top \widehat{\boldsymbol{\Sigma}}_{jj\cdot 1:j-1}^{-1} (\boldsymbol{x}_{j\cdot 1:j-1,i} - \widehat{\boldsymbol{\mu}}_{j\cdot 1:j-1}),\ j = 2, \ldots, k.$$

とする．このとき $\widehat{\boldsymbol{\mu}}_1$，$\widehat{\boldsymbol{\Sigma}}_{11}$，$\widehat{\boldsymbol{\mu}}_{j\cdot 1:j-1}$，$\widehat{\boldsymbol{\Sigma}}_{jj\cdot 1:j-1}$ はそれぞれ $\boldsymbol{\mu}_1$，$\boldsymbol{\Sigma}_{11}$，$\boldsymbol{\mu}_{j\cdot 1:j-1}$，$\boldsymbol{\Sigma}_{jj\cdot 1:j-1}$ の MLE である (Kanda and Fujikoshi (1998) 参照)．2-step 単調欠測データの下での多変量標本尖度 ($b_{2,p}^{(2)}$) と k-step 単調欠測データの下での多変量標本尖度 ($b_{2,p}^{(k)}$) を比較する．

$$b_{2,p}^{(2)} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(U_{1,i}^{(2)})^2}_{R_1^{(2)}} + \underbrace{\frac{1}{N_1}\sum_{i=1}^{N_1}(U_{2,i}^{(2)})^2}_{R_2^{(2)}} + \underbrace{\frac{2}{N_1}\sum_{i=1}^{N_1}U_{1,i}^{(2)}U_{2,i}^{(2)}}_{R_{12}^{(2)}}$$

$$b_{2,p}^{(k)} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(U_{1,i}^{(k)})^2}_{R_1^{(k)}\text{とおく}} + \cdots + \underbrace{\frac{1}{N_1}\sum_{i=1}^{N_1}(U_{k,i}^{(k)})^2}_{R_k^{(k)}\text{とおく}} + \underbrace{\frac{2}{N_{1:k-1}}\sum_{i=1}^{N_{1:k-1}}U_{1,i}^{(k)}U_{2,i}^{(k)}}_{R_{12}^{(k)}\text{とおく}} + \cdots + \underbrace{\frac{2}{N_1}\sum_{i=1}^{N_1}U_{k-1,i}^{(k)}U_{k,i}^{(k)}}_{R_{k-1\ k}^{(k)}\text{とおく}}$$

であるため，$k = 2$ のとき $b_{2,p}^{(k)} = b_{2,p}^{(2)}$ となる．$b_{2,p}^{(2)}$ の期待値の次元 p と個数 N を p_j，$N_{1:j}$ に置き換えることで $b_{2,p}^{(k)}$ の期待値を導出することができる．すなわち

$$\mathrm{E}[b_{2,p}^{(k)}] = p(p+2) - 2\left\{\sum_{j=1}^{k}\frac{1}{N_{1:j}}p_j(p_j+2) + \sum_{1\le m<l\le k}\frac{2}{N_{1:l}}p_m p_l\right\} + \mathrm{O}(N^{-\frac{3}{2}})$$

$$\mathrm{Var}[b_{2,p}^{(k)}] = 8\left[\sum_{j=1}^{k}\frac{1}{N_{1:j}}p_j(p_j+2) + \sum_{1\le m<l\le k}p_m p_l\left\{\left(\frac{1}{N_{1:l}} - \frac{1}{N_{1:m}}\right)p_l + \frac{2}{N_{1:l}}\right\}\right] + \mathrm{O}(N^{\frac{3}{2}})$$

ここで

$$m_2^{(k)} = p(p+2) - 2\left\{\sum_{j=1}^{k}\frac{1}{N_{1:j}}p_j(p_j+2) + \sum_{1\le m<l\le k}\frac{2}{N_{1:l}}p_m p_l\right\},$$

$$(\nu_2^{(k)})^2 = 8\left[\sum_{j=1}^{k}\frac{1}{N_{1:j}}p_j(p_j+2) + \sum_{1\le m<l\le k}p_m p_l\left\{\left(\frac{1}{N_{1:l}} - \frac{1}{N_{1:m}}\right)p_l + \frac{2}{N_{1:l}}\right\}\right]$$

とおく．
そして，多変量正規性検定統計量として

$$Z_{\mathrm{MM}}^{(k)*} = \frac{b_{2,p}^{(k)} - m_2^{(k)}}{\nu_2^{(k)}}.$$

を与える．このとき $N_i/N = \gamma_i \to \delta_i \in (0,1)$ である．ただし $i = 1, ..., k-1$ であり，これは漸近的に $N(0,1)$ であることを利用して検定を行う．
提案した検定統計量の正規近似に対する数値的精度を与えるために $k = 5$ の場合でモンテカルロ・シミュレーションを行なった．結果として，提案した k-step 単調欠測データの下での多変量標本尖度を用いた検定統計量は漸近的に標準正規分布に収束していることを確認することができた．
今後の課題としては本報告で提案した検定統計量のシミュレーションによる検出力を導出して比較を行うことや，今回導出しなかった $b_{2,p}^{(k)}$ の項に現れる $U_{k,i}^{(k)}$ と $U_{j,i}^{(k)}$ の共分散の導出が挙げられる．

高次元スパース回帰のための AIC 型情報量規準の漸近的性質

二宮 嘉行　　柳原 宏和

近年，サンプルサイズ n と目的変数の次元 p がともに大きく，その比が $p/n \to c \in (0,1)$ となるような設定における高次元統計理論の研究が盛んに行われている．説明変数の次元が大きいときの回帰分析も高次元と呼ばれるが，本発表ではこの p が大きいことを高次元と呼ぶことにする．このような double-asymptotic framework は，Marčenko and Pastur (1967) が random matrix theory (RMT) を用いて標本共分散行列の固有値分布の極限定理を導いたことに端を発し，その流れは Silverstein (1995) や Bai and Silverstein (2010) によって体系化された．double-asymptotic framework における統計理論は，Johnstone (2001) や Ledoit and Wolf (2004) などを通じて主成分分析や共分散推定における漸近的特性の理解を深め，Srivastava and Fujikoshi (2006) などを通じて高次元線形回帰分析の基盤ともなった．また，Meinshausen and Bühlmann (2006) や Zhu et al. (2012) は，理論を発展させるだけでなく，ゲノム解析における多遺伝子発現の同時予測や脳画像解析における多領域活動の統合的モデリングといった，重要な応用的課題に対して意義をもつことを示した．

上記の中で最も基本的といえる最尤推定を用いた高次元線形回帰であっても，非自明かつ意義のある興味深い理論研究が近年行われており，それは情報量規準の一致性に関する逆転現象である．古典的な固定次元の漸近理論においては，赤池情報量規準（AIC）は一致性をもたず過剰なモデルを選択する傾向がある一方，ベイズ情報量規準（BIC）は一致性をもち真のモデルを漸近的に確率 1 で選択することが示されている．しかし，double-asymptotic framework においては，これらの性質が必ずしも維持されない．Yanagihara et al. (2015) は，誤差項が正規分布にしたがうとき，AIC がむしろ一致性をもつ傾向があることを示した．また，Bai et al. (2022) は，誤差項が必ずしも正規分布にしたがわないような RMT が用いられる設定において，状況に応じて Cp 基準を含む AIC 型と BIC の選択挙動が逆転することを示している．このような結果は，情報量規準の設計すべき原理が，次元の増大速度と密接に関連していることを示唆しており，高次元統計におけるモデル選択理論の新たな展開を促している．

回帰におけるモデル選択といえば，どの説明変数を含めるかという変数選択が中心的課題である．そして，多項式回帰のように変数に自然な順序が存在するのでなければ，変数の組み合わせすべてがモデルの候補となり，変数の数が 20 近くにもなるとスパース推定に基づくアプローチが事実上の標準となる．なぜなら，網羅的に探索することが計算的に不可能となるし，統計的にも過剰適合のリスクが高まるからである．スパース推定において，チューニングパラメータの大きさは選択される変数の個数を直接的に支配し，過剰選択あるいは過少選択に影響するため，その決定は重要である．その決定方法としては，交差検証法に加えて情報量規準に基づくアプローチが盛んに検討されており，例えば AIC 型の基準としては Zou et al. (2007) による一般化 Cp 基準や Ninomiya and Kawano (2016) による AIC オリジナルの定義に基づいて導いたものがあり，

BIC 型の基準としては Chen and Chen (2008) や Wang et al. (2009) による extended BIC や modified BIC がある．また，Zhang et al. (2010) や Fan and Tang (2013) は，どちらも含む形式である GIC に対して漸近的性質を検証している．

本発表では，高次元線形回帰でスパース推定を用いた場合に対し，情報量規準の漸近的性質を導出する．特に，AIC 型の情報量規準が逆転現象により一致性をもつかどうか，を検証することを主目的とする．高次元線形回帰では，説明変数一つにつき目的変数の次元のパラメータがかかっているため，変数選択のためのスパース推定としてはグループ正則化を採用する．その中でも，モデル選択の一致性を議論するという目的を見据え，Zhang (2010) による minimax concave penalty (MC+) に基づいた Huang et al. (2012) の concave group selection を採用する．この手法は，グループ化されたパラメータベクトルに対して非凸ペナルティを課すことで，モデル選択の一致性と推定のバイアス低減の両立を可能にする点に特徴がある．それを実現するにはチューニングパラメータをうまく決定しなければならず，そのための情報量規準を構築するために目的変数が高次元でないときやスパース推定でないときの情報量規準の理論を組み込んでいく，というのが本発表の流れである．

スパース推定の selection consistency をもたらす情報量規準の議論に関する懸念は，結局のところ罰則項の係数をどう決めればよいかわからないことである．例えば，非スパース推定の際に selection consistency をもたらす BIC は Bayes factor から導かれる量であるため，罰則項の係数を 1 としていることにはある種の妥当性が備わっている．この BIC に合わせ，スパース推定の際も罰則項の係数を 1 とすることが通常であるが，その妥当性は必ずしもないということである．本発表では，この懸念に対する対応も二つ考える．一つ目は，Stein (1981)'s unbiased risk estimation 理論に基づかせることである．ノイズが正規分布であれば情報量規準の期待値が Kullback-Leibler ダイバージェンスに厳密になるようにするということであり，目的変数が高次元でないときの Zou et al. (2007) や Zhang (2010) が提案したような情報量規準にするということである．二つ目は，漸近損失有効性をもたせるようにすることである．目的変数が高次元でないならば情報量規準の罰則項の係数は 2，つまり非スパース推定のときと同じ結果であることを Zhang et al. (2010) は示している．

ディープラーニングと確率過程の統計推測

吉田朋広 (東京大学大学院数理科学研究科)

Muni Toke と Yoshida [2] は、リミットオーダーブックにおける注文フローの相対的な強度について、ratio model (Cox 型モデル) を用いたパラメトリック法を用いた。Cox 型モデルのベースラインハザードは、市場データにおける非定常な日中トレンドをキャンセルする利点がある。擬似尤度推定量の一致性と漸近正規性を示し、擬似尤度解析（Yoshida [4, 5]）に基づいて正当化された点過程に対する情報量規準によって共変量選択をおこなった。モデルをパリ証券取引所の実データに適用し、成行注文の高精度予測を実現し、従来の Hawkes モデルを上回る性能を示した。続いて、Muni Toke と Yoshida [3] は、ratio model を marked ratio model に拡張し、成行注文の階層構造を表現した。各成行注文は Bid/Ask で分類され、さらに価格変動を引き起こすかどうかに応じてアグレッシブと非アグレッシブに分類される。marked ratio model は、金融市場における成行注文の兆候とアグレッシブ性を予測する上で、Hawkes ベースモデルよりも優れた性能を示した。

しかし、[2, 3] におけるモデル選択の結果では、提案する共変量の組み合わせによって生成される多数のモデルの中で、情報量規準が比較的大きなモデルを好む傾向があり、モデルにさらに多くの共変量を取り入れる可能性が示唆されている。これが、ディープラーニングを用いて、非線形の相互作用のような、より多くの共変量を自動的に生成し、データの背後にある多様な非線形性に対するモデルの表現力を高める動機となっている。

α-ミキシング性のある共変量過程を入力とする強度を持つ点過程へのディープニューラルネットワークを考える。我々の汎用モデルには、Cox 型モデル、マーク付き点過程、および多変量点過程が含まれる。汎化誤差の収束率を与えるオラクル不等式が導出できる。シミュレーション実験は、マーク付き点過程が予測において単純な多変量モデルよりも優れていることを示している。marked ratio model をリミットオーダーブックの実データに適用した (Gyotoku et al. [1])。

オラクル不等式を導出するときに重要になるのは、α-ミキシング過程の汎関数に対する大偏差不等式である。ここでは点過程を扱ったが、オラクル不等式成立のためにはその構造は本質的ではなく、入力過程が α-ミキシングであれば、多様な確率過程にこの方法は適用可能である。

余裕があれば、確率微分方程式のディープラーニングに触れたい。

[1] Gyotoku, Y., Muni Toke, I., Yoshida, N. : Deep learning of point processes for modeling high-frequency data. arXiv preprint arXiv:2504.15944 (2025)

[2] Muni Toke, I., Yoshida, N.: Analyzing order flows in limit order books with ratios of Cox-type intensities. Quantitative Finance pp. 1–18 (2019)

[3] Muni Toke, I., Yoshida, N.: Marked point processes and intensity ratios for limit order book modeling. Japanese Journal of Statistics and Data Science **5**(1), 1–39 (2022)

[4] Yoshida, N.: Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Annals of the Institute of Statistical Mathematics **63**(3), 431–479 (2011)

[5] Yoshida, N.: Simplified quasi-likelihood analysis for a locally asymptotically quadratic random field. Annals of the Institute of Statistical Mathematics **77**(1), 1–24 (2025)

ゼロ過剰ガンマフレイルティによる二変量治癒コピュラモデル：治癒率と生存時間の従属関係

A bivariate cure copula model with zero-inflated gamma frailty: dependence between the two margins in cure rate and survival times

Masaki Hino[1,2], Shogo Kato[2], and Takeshi Emura[3]

[1] Department of Statistical Science, The Graduate University for Advanced Studies, SOKENDAI, Shonan village, Hayama, 240-0193, Kanagawa, Japan
hino.masaki@ism.ac.jp

[2] Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, 190-8562, Tokyo, Japan
skato@ism.ac.jp

[3] School of Informatics and Data Science, Hiroshima University, 1-3-2 Kagamiyama, Higashi-Hiroshima, 739-8511, Hiroshima, Japan
takeshiemura@gmail.com

Abstract. We propose a bivariate cure frailty copula model for survival data. Using a zero-inflated gamma frailty, the framework simultaneously accommodates a subpopulation of long-term survivors and a continuous positive frailty. Dependence between marginal cure rates is parameterized by an odds ratio, while dependence between uncured survival times is captured by a copula. We derive Kendall's tau for the proposed model to measure the degree of dependence. We also develop statistical inference methods based on maximum likelihood estimation. Simulation studies and an application to a real dataset demonstrate accurate and stable parameter estimation, highlighting the model's utility for analyzing paired time-to-event outcomes with cure fractions.

Keywords: Cure model · Mixture model · Gumbel copula · FGM copula · Kendall's tau

In survival analysis, a subset of individuals often never experience the event of interest during follow-up; their survival times are effectively infinite and they are treated as "cured." Cure models accommodate such long-term survivors by assigning a positive probability mass at infinity. Frailty models introduce unobserved random effects to represent heterogeneity or within-cluster correlation, and copula models deal with the dependence structure between event times. Existing bivariate approaches that combine these ideas (e.g., Rouzbahani et al., [2]) typically rely on discrete frailties, do not explicitly parameterize dependence between cure indicators, and lack closed-form expressions for dependence measures such as Kendall's tau. Fully unified frameworks that treat cure, frailty, and copula-based dependence simultaneously, with interpretable parameters for both cure status and event times, remain limited.

To address these issues, we proposed a bivariate cure frailty copula model based on a zero-inflated gamma frailty. For each margin $j = 1, 2$, the frailty is defined by

$$Z_j = (1 - X_j)W,$$

where $X_j \in \{0, 1\}$ is a cure indicator and W is a continuous positive frailty shared by both margins. When $X_j = 1$, the individual is cured on margin j and $Z_j = 0$; when $X_j = 0$, the individual is susceptible and $Z_j = W > 0$. The shared frailty W follows a gamma distribution with $E[W] = 1$ and $\mathrm{Var}(W) = \gamma$, so Z_j is zero-inflated but continuous on $(0, \infty)$. This avoids the somewhat artificial assumption of a discrete frailty while still allowing an explicit cured fraction via the point mass at zero.

The marginal cure rates are $p_j = P(X_j = 1)$, and the joint distribution of (X_1, X_2) is parameterized via the odds ratio

$$R = \frac{P(X_1 = 1, X_2 = 1)P(X_1 = 0, X_2 = 0)}{P(X_1 = 1, X_2 = 0)P(X_1 = 0, X_2 = 1)}.$$

Given p_1, p_2, and R, the four joint probabilities $p_{11}, p_{10}, p_{01}, p_{00}$ are obtained explicitly. The parameter R controls dependence between cure indicators: $R = 1$ corresponds to independence, $R > 1$ to positive association, and $0 < R < 1$ to negative association, providing a simple scalar description of dependence at the cure-status level.

Conditional on the frailty, we adopt the joint frailty copula model (Wang and Emura [3]). Using the Laplace transform of the gamma distribution, we derive closed-form expressions for the unconditional bivariate survival

function when C_θ is chosen as a Gumbel copula, a Farlie-Gumbel-Morgenstern (FGM) copula, or the independence copula. For example, with the Gumbel copula, the bivariate survival function takes the form

$$S(t_1,t_2) = p_{11} + p_{01}\left(1+\gamma r_1 t_1^{\alpha_1}\right)^{-1/\gamma} + p_{10}\left(1+\gamma r_2 t_2^{\alpha_2}\right)^{-1/\gamma} + p_{00}\left[1+\gamma\left\{\left(r_1 t_1^{\alpha_1}\right)^{\theta+1} + \left(r_2 t_2^{\alpha_2}\right)^{\theta+1}\right\}^{\frac{1}{\theta+1}}\right]^{-1/\gamma}.$$

In all cases, the joint survival admits a mixture representation with mixing ratio equal to the joint cure probabilities. The marginal cure rates can also be modeled as functions of covariates via logistic regression,

$$p_{ij} = \frac{\exp\left(x_{ij}^\top \beta_j\right)}{1+\exp\left(x_{ij}^\top \beta_j\right)},$$

allowing direct interpretation of covariate effects on the probability of being cured.

A key contribution is the derivation of Kendall's tau for bivariate survival data with a cure fraction. Using Pimentel's framework for zero-inflated data (Pimentel [1]), we obtain closed-form expressions: for the cure zero-inflated gamma Gumbel model, the non-cured component corresponds to BB1 copula, and for the cure zero-inflated gamma independence model to Clayton copula, both with known Kendall's tau. This yields explicit formulas for Kendall's tau in terms of p_1, p_2, R, θ, and γ, enabling a familiar rank-based interpretation of dependence even with cured individuals.

For statistical inference, we derived the full likelihood for right-censored bivariate survival data, together with closed-form first and second derivatives with respect to all parameters. The likelihood accounts for four observation types per individual (event on margin 1 only, event on margin 2 only, events on both margins, and censoring on both margins). The parameter vector $(\theta, \gamma, p_1, p_2, R, \alpha_1, r_1, \alpha_2, r_2)$ is estimated by maximum likelihood using the `optim` function in R.

Finite-sample performance was evaluated by Monte Carlo simulation of the cure zero-inflated gamma Gumbel model under two dependence settings ($R = 3$ and $R = 0.5$) and sample sizes $n = 200, 400$ (200 replications each). The copula parameter θ, marginal cure rates p_1, p_2, and Weibull parameters $\alpha_1, r_1, \alpha_2, r_2$ were well estimated even at $n = 200$, while the frailty parameter γ and odds ratio R showed some bias that decreased at $n = 400$, with coverage probabilities approaching the nominal 95% level.

Finally, the model was applied to the Diabetic Retinopathy Study dataset from the `survival` package in R, comprising paired times to vision loss in 197 patients, with one eye randomized to laser treatment and the other left untreated. For the dependence between marginal cure rates, we considered four specifications $0 < R < 1, R = 1, 1 < R, R = \infty$. Models with and without covariates in the cure components (using age and a risk score) were compared via AIC and BIC, and the specification $R = 1$ was consistently favored, indicating no association between the marginal cure rates of the two eyes. The overall Kendall's tau for the joint survival times was estimated as $\tau_b \approx 0.455$, indicating moderate positive association attributable to shared frailty. When the fitted marginal survival functions from the cure zero-inflated gamma Gumbel model (without covariates) were overlaid on the Kaplan-Meier curves, the model-based curves closely matched the empirical ones and provided identifiable estimates of the cure fractions, which cannot be recovered from Kaplan-Meier alone.

In summary, the proposed model provides a unified cure frailty copula framework for bivariate survival data with long-term survivors, offering interpretable parameters for cure status, frailty, and event-time dependence, as well as closed-form Kendall's tau. Simulation results and the diabetic retinopathy application demonstrate accurate and stable parameter estimation, supporting the model's usefulness for analyzing paired time-to-event outcomes with cure fractions.

References

1. Ronald Silva Pimentel. *Kendall's Tau and Spearman's Rho for Zero-Inflated Data.* Disserations, Western Michigan University, 2009.
2. Marziye Rouzbahani, Mohammad Reza Akhoond, and Rahim Chinipardaz. A new bivariate survival model with a cured fraction: A mixed Poisson frailty–copula approach. *Japanese Journal of Statistics and Data Science*, 8(1):367–391, 2025.
3. Yin-Chen Wang and Takeshi Emura. Multivariate failure time distributions derived from shared frailty and copulas. *Japanese Journal of Statistics and Data Science*, 4(2):1105–1131, 2021.

微小撹乱パラメータを持つ線形放物型確率偏微分方程式モデルのパラメータ推定およびその応用

神戸大学大学院海事科学研究科 貝野友祐

1　空間 2 次元線形放物型確率偏微分方程式モデル

$D=[0,1]^2$ 上の確率偏微分方程式

$$\begin{cases} \mathrm{d}X_t(y,z)=\left\{\theta_2\left(\dfrac{\partial^2}{\partial y^2}+\dfrac{\partial^2}{\partial z^2}\right)+\theta_1\dfrac{\partial}{\partial y}+\eta_1\dfrac{\partial}{\partial z}+\theta_0\right\}X_t(y,z)\mathrm{d}t \\ \qquad\qquad +\epsilon\mathrm{d}W_t^Q(y,z), \quad (t,y,z)\in[0,1]\times D, \\ X_0(y,z)=\xi(y,z), \quad (y,z)\in D, \\ X_t(y,z)=0, \quad (t,y,z)\in[0,1]\times\partial D \end{cases} \tag{1}$$

を考える．ただし, $\{W_t^Q\}$ は D 上のソボレフ空間における Q-ウィーナー過程, パラメータ空間 $\Theta\subset\mathbb{R}^3\times(0,\infty)$ はコンパクト凸集合, $\theta:=(\theta_0,\theta_1,\eta_1,\theta_2)\in\Theta$ は未知パラメータ, $\theta^*=(\theta_0^*,\theta_1^*,\eta_1^*,\theta_2^*)$ は真値で $\theta^*\in\mathrm{Int}\,\Theta$ とする．既知の $\alpha\in(0,1)$ に対して Q-ウィーナー過程を

$$W_t^Q=\sum_{l_1,l_2\geq 1}\mu_{l_1,l_2}^{-\alpha/2}w_{l_1,l_2}(t)e_{l_1,l_2}$$

で定める．ただし $\{w_{l_1,l_2}\}_{l_1,l_2\geq1}$ は独立な 1 次元標準ブラウン運動であり, $\mu_0\in(-2\pi^2,\infty)$ に対して, $e_{l_1,l_2}(y,z)=2\sin(\pi l_1y)\sin(\pi l_2z)\mathrm{e}^{-(\kappa y+\eta z)/2}$, $\mu_{l_1,l_2}=\pi^2\left(l_1^2+l_2^2\right)+\mu_0, l_1,l_2\in\mathbb{N}, (y,z)\in D$, である．$\epsilon\in(0,1)$ は既知の微小撹乱パラメータとする．高頻度時空間データとして $\mathbb{X}_{N,M}=\{X_{t_i}(y_j,z_k)\}$, $0\leq i\leq N, 0\leq j\leq M_1, 0\leq k\leq M_2$ を考える．ただし, $t_i=i\Delta=i/N$, $y_j=j/M_1$, $z_k=k/M_2$, $M:=M_1M_2$ とする．

2　係数パラメータの推定

θ_1,η_1,θ_2 の最小コントラスト推定について考える．空間間引きデータ $\mathbb{X}_{N,m}^{(1)}=\{X_{t_i}(\widetilde{y}_j,\widetilde{z}_k)\}$, $0\leq i\leq N, 0\leq j\leq m_1, 0\leq k\leq m_2$ を考える．ただし, $m:=m_1m_2$, $m=O(N)$, $N=O(m)$, $\Delta=1/N$, $b\in(0,1/2)$ に対して,

$$b\leq\widetilde{y}_0<\widetilde{y}_1<\cdots<\widetilde{y}_{m_1}\leq 1-b, \quad b\leq\widetilde{z}_0<\widetilde{z}_1<\cdots<\widetilde{z}_{m_2}\leq 1-b$$

とする．また,

$$\bar{y}_j=\frac{\widetilde{y}_{j-1}+\widetilde{y}_j}{2}, \quad \bar{z}_k=\frac{\widetilde{z}_{k-1}+\widetilde{z}_k}{2}, \quad j=1,\cdots,m_1, k=1,\cdots,m_2$$

とする．

$$\begin{aligned} T_{i,j,k}X=&X_{t_i}(\widetilde{y}_j,\widetilde{z}_k)-X_{t_i}(\widetilde{y}_{j-1},\widetilde{z}_k)-\left(X_{t_i}(\widetilde{y}_j,\widetilde{z}_{k-1})-X_{t_i}(\widetilde{y}_{j-1},\widetilde{z}_{k-1})\right) \\ &-\Big(X_{t_{i-1}}(\widetilde{y}_j,\widetilde{z}_k)-X_{t_{i-1}}(\widetilde{y}_{j-1},\widetilde{z}_k)-\left(X_{t_{i-1}}(\widetilde{y}_j,\widetilde{z}_{k-1})-X_{t_{i-1}}(\widetilde{y}_{j-1},\widetilde{z}_{k-1})\right)\Big). \end{aligned}$$

とし, 次のコントラスト関数を考える

$$\mathcal{U}^{\epsilon}_{m,N}(\kappa,\eta,\theta_2)=\frac{1}{m}\sum_{k=1}^{m_2}\sum_{j=1}^{m_1}\left\{\frac{1}{\epsilon^2N\Delta^{\alpha}}\sum_{i=1}^{N}(T_{i,j,k}X)^2-\mathrm{e}^{-\kappa\overline{y}_j-\eta\overline{z}_k}\phi_{r,\alpha}(\theta_2)\right\}^2.$$

ただし, J_0 を次数 0 の第 1 種ベッセル関数とし, $r=\frac{(1-2b)\sqrt{N}}{\sqrt{m}}, \alpha>0$ に対して,

$$\phi_{r,\alpha}(\theta_2)=\frac{2}{\theta_2^{1-\alpha}\pi}\int_0^{\infty}\frac{1-\mathrm{e}^{-x^2}}{x^{1+2\alpha}}\left(J_0\Big(\frac{\sqrt{2}rx}{\sqrt{\theta_2}}\Big)-2J_0\Big(\frac{rx}{\sqrt{\theta_2}}\Big)+1\right)\mathrm{d}x \tag{2}$$

とする. Ξ を $\mathbb{R}^2\times(0,\infty)$ の部分集合とし, θ_1, η_1, θ_2 の最小コントラスト推定量を $(\widehat{\kappa},\widehat{\eta},\widehat{\theta}_2)=\underset{(\kappa,\eta,\theta_2)\in\Xi}{\operatorname{argmin}}\,\mathcal{U}^{\epsilon}_{m,N}(\kappa,\eta,\theta_2), \widehat{\theta}_1=\widehat{\kappa}\widehat{\theta}_2, \widehat{\eta}_1=\widehat{\eta}\widehat{\theta}_2$ とする. 正則条件の下, $N,m\to\infty$, $\epsilon\to0$ のとき $(\widehat{\theta}_1,\widehat{\eta}_1,\widehat{\theta}_2)$ は一致性をもつ.

θ_0 および μ_0 の適応的推定について考える. 時間間引きデータ $\mathbb{X}^{(2)}_{n,M}=\{X_{\widetilde{t}_i}(y_j,z_k)\}$, $0\le i\le n, 0\le j\le M_1, 0\le k\le M_2$ を考える. ただし, $n\le N$ に対して

$$\widetilde{t}_i=i\Delta_n=\left\lfloor\frac{N}{n}\right\rfloor\frac{i}{N},\quad i=0,1,\dots,n$$

とする. 座標過程 $x_{l_1,l_2}(t)$ を時間間引きデータ $\mathbb{X}^{(2)}_{n,M}$ を用いて

$$\widehat{x}_{l_1,l_2}(t)=\sum_{j=1}^{M_1}\sum_{k=1}^{M_2}X_t(y_{j-1},z_{k-1})\delta_j^{[y]}g_{l_1}(\widehat{\kappa})\delta_k^{[z]}g_{l_2}(\widehat{\eta})$$

で近似する. ここで,

$$g_l(x:a)=\frac{\sqrt{2}\mathrm{e}^{ax/2}}{(a/2)^2+(\pi l)^2}\left(\frac{a}{2}\sin(\pi lx)-\pi l\cos(\pi lx)\right),\quad a,x\in\mathbb{R},\ l\in\mathbb{N},$$
$$\delta_j^{[y]}g_l(a)=g_l(y_j:a)-g_l(y_{j-1}:a),\quad \delta_k^{[z]}g_l(a)=g_l(z_k:a)-g_l(z_{k-1}:a)$$

である. コントラスト関数を

$$\mathcal{V}^{\epsilon}_n(\lambda,\mu:\widehat{\mathbf{x}}_{l_1,l_2})=\sum_{i=1}^{n}\frac{(\widehat{x}_{l_1,l_2}(\widetilde{t}_i)-\mathrm{e}^{-\lambda\Delta_n}\widehat{x}_{l_1,l_2}(\widetilde{t}_{i-1}))^2}{\frac{\epsilon^2(1-\mathrm{e}^{-2\lambda\Delta_n})}{2\lambda\mu^{\alpha}}}+n\log\frac{1-\mathrm{e}^{-2\lambda\Delta_n}}{2\lambda\mu^{\alpha}\Delta_n}$$

と定義し, $(\widehat{\lambda}_{l_1,l_2},\widehat{\mu}_{l_1,l_2})=\underset{\lambda,\mu}{\operatorname{arginf}}\,\mathcal{V}^{\epsilon}_n(\lambda,\mu:\widehat{\mathbf{x}}_{l_1,l_2})$ とする. θ_0 および μ_0 の適応的推定量は $\widehat{\theta}_0=-\widehat{\lambda}_{l_1,l_2}+\widehat{\theta}_2\left(\frac{\widehat{\kappa}^2+\widehat{\eta}^2}{4}+\pi^2(l_1^2+l_2^2)\right), \widehat{\mu}_0=\widehat{\mu}_{l_1,l_2}-\pi^2(l_1^2+l_2^2)$ で定まる. 正則条件の下, $n\to\infty$, $\epsilon\to0$ のとき $(\widehat{\theta}_0,\widehat{\mu}_0)$ は一致性および漸近正規性を持つ.

3　数値シミュレーション

$N=10^3$, $M_1=M_2=200$ として SPDE (1) の数値解を生成し, 高頻度データに基づいた大規模数値シミュレーションを行う. 具体的には，最小コントラスト推定量 $(\widehat{\theta}_1,\widehat{\eta}_1,\widehat{\theta}_2)$ および適応的推定量 $(\widehat{\theta}_0,\widehat{\mu}_0)$ を計算し，その漸近挙動を検証する.

参考文献

[1] Y. Tonaki, Y. Kaino, and M. Uchida. (2026). Small dispersion asymptotics for an SPDE in two space dimensions using triple increments. *Journal of Statistical Planning and Inference*, 241, 106333.

形状制約下での関数パラメータの事後分析

入江 薫（東京大学経済学部）

要旨

モデルに含まれる未知の関数（関数パラメータ）の推測問題を考える。問題によっては、当該の関数は単調性や凸性などの形状制約 (shape constraints) を満たすことを要請される。形状制約下での関数パラメータの推測に関する研究の中でも、ベイズ統計学の立場を取り、関数パラメータの事後分布を計算することを目的とする研究は近年盛んに行われている。関数パラメータに関する不確実性を表現するためには、関数空間上の事前分布を設定しなければならないが、形状制約は「事前の知識」として事前分布の構成において表現される。そのような事前分布の構成方法として、関数値の差分をモデル化する方法と、基底関数展開による方法が知られる。本講演では、それぞれの方法について形状制約を検討した研究を三つ紹介する。

第一の研究では、関数値の差分に切断分布を適用することで、単調増加・単調減少する平均関数を表現する手法について検討する（Okano et al., 2024）。切断正規分布の尺度混合である切断 Horseshoe prior を用いることで、階段関数のような形状の関数を表現することができる。ナイル川の水量データに対して提案手法を適用したところ、ダム建設の年に急激な水量減が検出された。これは「縮小すべきでない時には縮小しない」という、点推定量の Tail-robustness という性質を反映したものであり、本研究でも数学的証明を与えている。

第二の研究では、同手法を分散関数の推定に応用する（Miyatake et al., 2025）。分散が単調に増加するという状況は、微分方程式モデルの数値解の誤差に現れる。誤差分散の対数差分をモデル化することになるが、これは金融時系列モデルのボラティリティ変動モデルに対応する。よって、第一の研究の成果と、ボラティリティモデルに関する計算方法を組み合わせることで、単調な分散関数の推定が可能になる。数値例として、提案手法を FitzHugh–Nagumo モデルの数値解の誤差分析に用いたところ、非ベイズ的な点推定の場合と比べて異常値に頑健な推定結果が得られた。これは Horseshoe prior が裾厚な分布によるものと思われる。

第三の研究では、基底関数展開に基づく方法について検討する（Hiraki et al., 2024）。具体的には、関心のある関数は有界閉区間上で定義され、単調増加であり、凸であり、かつ境界条件を満たす（端点で指定された点を通る）ことが要請される。そこで、それらの形状制約をすべて満たす関数（ベータ分布の分布関数など）を基底関数として採用し、関心のある関数を基底関数の凸結合で表すことで、形状制約を保証する。モデルは回帰分析に帰着するが、回帰係数が単体上に制約されることから、マルコフ連鎖モンテカルロ法には Pólya-gamma 拡大等の工夫が必要となる。上記の形状制約を必要とする好例として、所得の不平等を測るローレンツ曲線が挙げられる。提案手法を用いることで、ローレンツ曲線の時系列モデルを構成・推定できる。また、ローレンツ曲線の積分はジニ係数という経済指標となるが、その事後分析も簡易に実行可能である。

本講演は岡野遼（一橋大学）、羽村靖之（京都大学）、菅澤翔之助（慶應義塾大学）、宮武勇登（大阪大学）、松田孟留（東京大学・理化学研究所）、平木大智（東京大学）との共同研究に基づく。

報告

当日は要旨に基づき、参考文献にある三本の論文の内容を報告した。

質疑応答では Hiraki et al. (2024) で用いた、基底関数展開に基づく関数パラメータのモデリングに関して議論があった。当該の方法においては、関心のある関数 $f(x)$ を、基底関数 $h_k(x)$ を用いて、

$$f(x) = \sum_{k=1}^{K} \theta_k h_k(x),$$

とモデル化する。関数 f が単調増加、下に凸、かつ $f(0) = 0, f(1) = 1$ という境界条件を満たすように形状制約を課したい場合には、基底関数 h_k に同種の形状制約を課し、係数 $(\theta_1, \ldots, \theta_K)$ を単体上に制約（非負の値をとり和が 1 になるように）すればよい。そのような基底関数の例として、ベータ関数の分布関数（正規化された不完全ベータ関数）やパレート分布の分布関数が挙げられる。

- このうち、係数の制約の必要性について指摘があった。すなわち、単調性と凸性のみを達成したい場合には係数は非負であればよく、単体上に制約されている必要はない。係数の和が 1 であるという制約は、境界条件を満たす上で必要となる。
- 先行研究では単調性のみが議論されることが多い（単調回帰）。関数の正確な推定という点では、凸性は不要ではないかという考え方もある。しかし、本研究の関心であるローレンツ曲線はその積分値（ジニ係数）に重要な意味があり、凸性が保証されない場合には積分値が過小評価される傾向にある。講演では制約無しのモデルとの比較がなされたが、単調だが凸でない場合との厳密な比較は今後の研究課題である。
- 仮に単調性のみを仮定する場合に、差分アプローチ（Okano et al., 2024 ほか）との違いに関して質問があった。厳密な性能比較はなされていないが、特性の違いはあきらかである。積分値など、非観測点上での関数値も必要とする分析が可能なのは基底関数展開アプローチである。一方、差分アプローチは MCMC による事後分析が非常に簡易であり、MCMC の効率性も高いことが分かっている。目的によって使い分けるのがよいと思われる。

Okano, R., Hamura, Y., Irie, K., and Sugasawa, S. (2024), “Locally adaptive Bayesian isotonic regression with half shrinkage priors,” *Scandinavian Journal of Statistics*, 5(1), 109-141. arXiv:2208.05121

Miyatake, Y., Irie, K. & Matsuda, T. (2025). “Quantifying uncertainty in the numerical integration of evolution equations based on Bayesian isotonic regression,” *Japan Journal of Industrial and Applied Mathematics*, 42, 983–1001. arXiv:2411.08338

Hiraki, D., Hamura, Y., Irie, K., & Sugasawa, S. (2024). “State-space modeling of shape-constrained functional time series,” arXiv:2404.07586.

シンポジウム「データサイエンスの基盤を支える次世代統計理論・方法論の挑戦と革新」報告書

平木 大智（東京大学経済学研究科）

作成日：2025 年 12 月 14 日

1　基本情報

- シンポジウム名：データサイエンスの基盤を支える次世代統計理論・方法論の挑戦と革新
- 主催：科学研究費補助金 基盤研究（A）課題番号 25H01107「大規模複雑データの理論と方法論の深化と展開」研究代表者：青嶋誠（筑波大学）
- 開催日時：2025 年 12 月 1 日 - 12 月 3 日
- 会場：九州大学 西新プラザ

2　シンポジウム内容・所感

2.1　12 月 1 日 通常セッション

1 日目は 6 名の発表者が講演をなされた。私が Bayesian time-series を専攻していることもあり、もっとも印象的であったのは中北先生（理化学研究所 AIP）の「Algorithm fo Fast Gibbs Sampling in Hierarchical Bayesian Panel Modeling」であった。ベイズ時系列の世界ではもはや一般的となった ASIS という手法の効率性を理論的な側面から議論する内容であり、簡単なモデルに対してのみ結果を与えているが、非常に示唆に富んだご研究であった。

加えて岩重さん（広島大学）のクラスタリングにおけるクラスターの事前分布について、そして草野先生（熊本大学）の高頻度気金融データの疑似 BIC についてのご講演は、私の研究対象であるファイナンス分野とも共通点のあり、非常に興味深い内容であった。

2.2　12 月 2 日 通常セッション（報告者講演日）

2 日目のセッションは 3 つに分かれており、それぞれのセッションで 3 名の先生・学生がご講演された。どの内容もとても興味深いものであったが、特に赤間先生（東北大学）のご講演は DCC という動的な分散共分散行列の構造を課すことでファーマ・フレンチのファクターモデルについて示唆を与える結果をご提示になっていた。これは近年のマクロ経済分野にも関連する内容であり、深く考えさせられるものであった。

また、吉田先生（東京大学）や Ziyue さん（東北大学）のご講演内容も、私の興味に合致する内容であり非常に楽しむことができた。吉田先生は株価の取引の構造を反映したモデリングをなされており、Ziyue さんは

最適輸送写像（Wasserstain distance）を用いた分布回帰・分布予測を行っていた。これらは私の今後の研究にも一部反映が可能であり、非常に有意義であった。

2.3　12 月 3 日 通常セッション（報告者講演日）

3 日目のセッションでは私を含め 6 名の発表者による講演がなされた。

2 つ目のセッションにて報告者は講演を行った。演題は「Dynamic factor stochastic volatility in mean model」であり、内容は多変量マクロ経済時系列データの特徴を捉えるのに優れた動的因子モデルを拡張することで、経済が不安定な時期における変数の挙動を解析するというものである。本講演において、草野先生（熊本大学）をはじめとする、主に時系列解析を専門とする先生方から、非常に有益なご質問・関連情報のご提示をいただいた。加えて川野先生（九州大学）と江村先生（広島大学）には講演後に発表内容についてお話をさせていただいた。非常に収穫のある講演をさせていただき、今後の研究活動も一層邁進する所存である。

他のご講演においては、私の専攻するベイズ統計学に関連するものとして、入江先生（東京大学）、さらに高頻度データ（とみなせる）に関するものとして貝野先生（神戸大学）のご講演は非常に興味深いものであった。特に入江先生による「形状制約下での関数パラメータの事後分析」は prior information を適切にモデルに含まれる関数の形状に反映する興味深い内容であり、直接的な応用の可能性を感じさせる実りのあるご講演であった。その他では修士学生による素晴らしいご講演もあり、私自身の研究のモチベーションを保つのにも良い機会となった。

3　おわりに

開催責任者である川野先生（九州大学）、主催の青嶋先生（筑波大学）、また私の研究にアドバイスをいただいた先生方をはじめとする参加者の方々にあらためて感謝いたします。

ジョイントフレイルティコピュラモデルを用いた欠測を含む肺腺癌データの解析

梶谷文乃（発表者）[1] 江村剛志 [1]

1 広島大学 情報科学部：〒739-8511 東広島市鏡山一丁目 3 番 2 号

E-mail: b222334@hiroshima-u.ac.jp

抄録 生存時間データには、再発や死亡といった複数のイベントが存在する。しかし、古典的な Cox 比例ハザードモデルでは、これらのイベントを独立と仮定して解析してしまうという課題がある。また、複数の研究から統合されたデータに存在する未観測の異質性を十分に考慮できないという問題も指摘されている。

本研究では、Gene Expression Omnibus（GEO）に登録された肺腺癌患者 853 例のデータを用い、細胞外マトリックス（ECM: Extracellular Matrix）関連遺伝子である CD36、COL11A1、HMMR の発現量および、年齢・性別などの背景因子が予後に与える影響を評価した。まず、古典的な Cox 回帰モデルを用いて、再発と生存に対する影響を個別に評価した。さらに、ジョイントフレイルティコピュラモデルを適用し、再発と死亡に対する影響を同時に評価した。本モデルにより、未観測の異質性（フレイルティ）およびイベント間の依存関係を同時に考慮することが可能となった。

加えて、実データにおいて再発や死亡の有無が観測されているにもかかわらず、再発までの期間（TTP）が欠測している症例が存在するという課題に着目し、全生存期間（OS）の情報を用いて TTP を補完する方法をシミュレーションにより検討した。

キーワード エンドポイント、全生存時間、ジョイントモデル、コピュラ、無増悪期間、遺伝子発現量、比例ハザードモデル，欠測値

1. はじめに

生存時間データには再発や死亡といった複数のイベントが存在し、古典的な Cox 回帰モデルのような解析では、これらのイベントを独立と仮定してしまう問題がある。また，通常 Cox の回帰は，複数の研究から得られたデータの未観測異質性を考慮できない．しかしながら，がんゲノム研究の代表的なデータベースである，Gene Expression Omnibus (GEO)や Cancer Genome Atlas (TCGA)には，患者に発生した複数のイベントが記録されているものがある．また，これらのデータを利用した解析において，イベント間の相関や未観測異質性が考慮されることは殆ど無い．

本研究では、GEO データベースに登録された GSE30219(Rousseaux et al., 2013)、GSE31210(Yamauchi et al.,2012)、GSE68465(Shedden et al., 2008)、GSE50081(Der, S et al., 2014)、および GSE37745(Botling J et al., 2013)のデータセットを用い、細胞外マトリックス（ECM: Extracellular Matrix）関連遺伝子 *CD36*、*COL11A1*、*HMMR* の発現量および、年齢・性別などの背景因子が生存予後に与える影響を評価することを考える。ECM 関連遺伝子は腫瘍の進展や転移、免疫細胞浸潤に深く関与することが知られており、特定の ECM 関連遺伝子群が肺腺癌の予後予測に有用である可能性が指摘されている（Chai et al., 2024）。そのため、本研究でも ECM 関連遺伝子を解析対象とすることで、肺腺癌における予後評価に役立つと考えられる。しかしながら、古典的な生存時間解析の手法で肺腺癌患者のデータを解析する場合、未観測の異質性（フレイルティ(江村・古川 2024)）や、観測されるイベント間の相関構造を考慮することが難しい。本研究では、まず古典的な Cox 回帰により因子が再発と生存に与える影響を別々に評価した。つぎに、ジョイントフレイルティコピュラモデル(Emura et al. 2017, 2019)を用いて因子が再発と生存に与える影響を同時に評価した。このモデルにより、未観測の異質性（フレイルティ）とイベント間の従属関係（コピュラ）を同時に考慮することができた。ジョイントフレイルティコピュラモデルの結果を用いると、遺伝

子や背景因子の影響だけでなく、イベント間の関連も考察することができた。さらに、実際の生存時間データでは再発や死亡の有無が観測されているにもかかわらず、進行・再発までの期間（TTP）が欠測しているデータが存在する。このような欠測値を含むデータに対して、全生存期間（OS）の情報を用いて進行・再発までの期間を補完する方法をシミュレーションにより検討した。

参考文献

[1] Botling, J., Edlund, K., Lohr, M., Hellwig, B., et al. (2013). Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clinical Cancer Research, 19(1), 194-204.

[2] Chai, Y., Ma, Y., et al. (2024). Identification and validation of a 4-extracellular matrix gene signature associated with prognosis and immune infiltration in lung adenocarcinoma. *Heliyon*, 10(2).

[3] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden, K., Taylor, J. M., Enkemann, S. A., et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature Medicine, 14(8), 822-7

[4] Der, S. D., Sykes, J., Pintilie, M., Zhu, C. Q., et al. (2014). Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. Journal of Thoracic Oncology, 9(1), 59-64

[5] 江村剛志, & 古川恭治. (2024). フレイルティモデル― 生存分析におけるハザードのランダム効果―. *計量生物学*, 45(2), 215-245.

[6] 江村剛志 (2025). コピュラ理論の基礎. コロナ社.

[7] Rousseaux, S., Debernardi, A., Jacquiau, B., Vitte, A.-L., Vesin, A., Nagy-Mignotte, H., Moro-Sibilot, D., Brichon, P.-Y., Lantuejoul, S., Hainaut, P., Laffaire, J., de Reyniès, A., Beer, D. G., Timsit, J.-F., Brambilla, C., Brambilla, E., & Khochbin, S. (2013). Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Science Translational Medicine,

[8] Yamauchi, M., Yamaguchi, R., Nakata, A., Kohno, T., Nagasaki, M., Shimamura, T., Imoto, S., Saito, A., Ueno, K., Hatanaka, Y., Yoshida, R., Higuchi, T., Nomura, M., Beer, D. G., Yokota, J., Miyano, S., & Gotoh, N. (2012). Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma.

Modified AIC for Canonical-Link GLMs with Known Scale Parameter

大阪公立大学 数学研究所 栁原宏和

本発表では, Nelder & Wedderburn (1972) により提案された一般化線形モデル (Generalized Linear Model; GLM) における変数選択のための赤池情報量規準 (Akaike Information Criterion; AIC) の修正について取り扱う. GLM は, McCullagh & Nelder (1989) や Fahrmeir & Tutz (2001) などの標準的な統計学の教科書に示されているように, 重回帰モデル, ロジスティック回帰モデル, ポアソン回帰モデルなどを包含する汎用的な統計モデルであり, 確率分布やリンク関数を適切に設定することで, さまざまなデータ構造に対応することができる.

Akaike (1973;1974) により提案された AIC は, Kullback–Leibler (KL) 情報量 (Kullback & Leibler, 1951) に基づくリスク関数の漸近不偏推定量であり, 統計モデルの最大対数尤度の -2 倍に, "2× パラメータ数" というモデルの複雑さに対する罰則項を加えることで定義される. その簡便な定義式ゆえに応用分野で広く普及し, 提案から 50 年以上経た現在においても, 多様な分野で複数の統計モデルの比較や選択に利用されている. しかし, AIC はリスク関数に対して漸近的には不偏であるものの, 標本数が十分に大きくない場合にはバイアスが生じやすく, 結果として過剰に複雑なモデルを選択する傾向があることが知られている. この欠点を克服するため, 多くの研究者によってバイアス補正に関する研究が進められてきた.

正規分布を仮定した重回帰モデルにおいては, Sugiura (1978) により, 真のモデルを含む過大モデルの下でリスク関数の完全不偏推定量となる Corrected AIC (CAIC) が提案された. その後, Hurvich & Tsai (1989) により CAIC と同型の補正版 AIC, すなわち AIC_c が提案され, こちらが広く普及した結果, 現在では "修正 AIC" といえば AIC_c を指すのが一般的である. ただし, 完全な不偏性が成立するのは重回帰モデルかつ正規分布を仮定した場合に限られ, 他の分布やモデルでは完全不偏推定量を構成することは非常に難しい. ロジスティック回帰やポアソン回帰では, 過大モデルの下でバイアスを $O(n^{-1})$ の項まで補正した CAIC が提案されている (Yanagihara, Sekiguchi & Fujikoshi, 2003; Kamo, Yanagihara & Satoh, 2013). さらに, これらの結果は Imori, Yanagihara & Wakaki (2014) により, スケールパラメータが既知の場合の GLM へと拡張されている.

一方, 真のモデルを含まない過小モデルの下でのバイアス補正は, 過大モデルの場合に比べて格段に難しい. その理由は, 過大モデルでは回帰係数の最尤推定量 (MLE) が漸近不偏 (重回帰モデルでは完全不偏) であるのに対し, 過小モデルでは定数バイアスを持つためである. この定数バイアスの影響により, 補正の際に考慮すべき確率分布が非心度を持つ分布となる. 正規性を仮定した重回帰モデルでは, 過大モデルではカイ二乗分布を用いた補正が可能であるが, 過小モデルでは非心カイ二乗分布を用いる必要がある. 特に, バイアスは非心カイ二乗分布の逆数の期待値に依存しており, このため過大モデルのように完全な不偏推定量を構成することができない. Fujikoshi & Satoh (1997) は, 漸近展開を用いて過小モデルに対してもバイアス補正を行った AIC を提案した. 彼らは過大モデルの補正のみを行う "Corrected" 版との差異を明確にするため, この指標を "Modified AIC (MAIC)" と命名した. MAIC は CAIC とは違い過大モデルで不偏とはならないが, AIC や CAIC が持つ過小モデルでの定数バイアスを補正しており, また過大モデルにおいても $O(n^{-1})$ の項までバイアス補正がなされている. Fujikoshi & Satoh [**??**] の MAIC は多変量回帰モデルにおいて提案されたものであり, その特別な場合として重回帰モデルでの MAIC を導くことができる. 一方, GLM においては, 過小モデルでもバイアス補正を行う MAIC は未だ提案されていない.

重回帰モデルにおいては, リスク関数を最小にするモデルが真のモデルまたは過小モデルであることが理論的に証明されている (Yanagihara, Kamo, Imori & Yamamura, 2017 参照). GLM において同様の性質が成り立つかは未解明であるが, 少なくとも漸近的には, リスク関数を最小にするモデルが真のモデルまたは過小モデルであることが示唆される. したがって, GLM においても MAIC を導出することは, 応用上重要

である.

Fujikoshi & Satoh (1997) による MAIC では, 過小モデルにおいてバイアスの漸近展開の主要項の漸近不偏推定量を用いることでバイアスが補正されている. これをさらに漸近展開の第 2 項まで導出できれば, それらを用いて過小モデルにおいても $O(n^{-1})$ の項まで補正された AIC を提案することが可能である. しかし, 過小モデルの下では漸近展開の第 2 項まで導出することが困難である. さらに導出できたとしても, その展開式には多数の複雑な項で表現され, それらを打ち消すように各項の漸近不偏推定量に -1 をかけて加算するという退屈で煩わしい作業を行わなければならない. そこで本論文では, まず過小モデルにおけるバイアスの第 2 項までの導出を試みる. バイアスの展開式をできるだけ簡単な形で得るために, スケールパラメータが既知でリンク関数が正準リンク関数である GLM を考える. つまり, 以下のような対数尤度関数を持つ統計モデルを考えることになる.

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - b(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{a(\phi)} + c(y_i, \phi) \right\}.$$

ただし, $y_1, \ldots, y_n$ は互いに独立な目的変数, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ は k 次元説明変数ベクトルである. さらに, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ は未知の k 次元回帰係数ベクトル, ϕ は既知のスケールパラメータ, $a(\cdot), b(\cdot), c(\cdot, \cdot)$ は微分可能な既知の関数である. この場合, 対数尤度関数の導関数と目的変数の高次モーメントが一致するため, Yanagihara, Kamo, Imori & Satoh (2012) の付録 B の結果と同様に, 以下の 3 つの行列のみでバイアス項を表現することが可能となる.

$$\boldsymbol{G}_2(\boldsymbol{\theta}) = \frac{1}{na(\phi)^2} \sum_{i=1}^{n} \kappa_2(\theta_i) \boldsymbol{x}_i \boldsymbol{x}_i^\top, \quad \boldsymbol{G}_3(\boldsymbol{\theta}) = \frac{1}{na(\phi)^3} \sum_{i=1}^{n} \kappa_3(\theta_i) \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top \otimes \boldsymbol{x}_i \right),$$
$$\boldsymbol{G}_4(\boldsymbol{\theta}) = \frac{1}{na(\phi)^4} \sum_{i=1}^{n} \kappa_4(\theta_i) \left(\boldsymbol{x}_i \boldsymbol{x}_i^\top \otimes \boldsymbol{x}_i \boldsymbol{x}_i^\top \right).$$

ただし, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. ここで, $\kappa_2(\theta)$, $\kappa_3(\theta)$, $\kappa_4(\theta)$ は, y の分散, 3 次, 4 次キュムラントであり, 以下のような微分により求めることができる.

$$\kappa_r(\theta) = a(\phi)^{r-1} \frac{\partial^r}{\partial \theta^r} b(\theta), \quad (r = 2, 3, 4).$$

次に, 実際のバイアス補正においては, 従来のように補正項を逐次加算する方法は採らない. Bartlett 補正 (Bartlett, 1937) やその拡張 (Corderio & Ferrari, 1991; Fujikoshi, 2000), さらに歪度や尖度を打ち消す改良変換 (Konishi, 1981; Hall, 1992; Yanagihara & Yuan, 2005) に倣い, 統計量に変換を施すことで, 展開後に補正項が自然に現れるような方法によりバイアス補正を行う. この結果, 提案する新たな MAIC は, 簡潔な表現式を持ち, 過大モデルのみならず過小モデルにおいても $O(n^{-1})$ の項まで補正された情報量規準となっている.

発表当日には, 候補モデルと KL 情報量に基づくリスク関数を定義し, モデルが過大モデルであるという仮定を課さずに, $O(n^{-1})$ の項までバイアスを導出を行った. さらに, 得られたバイアスの展開式を基に, 補正項を明示的に追加することなく MAIC を提案した. 提案した MAIC は, モデルが過小・過大いずれの場合でも $O(n^{-1})$ の項までバイアスを補正した情報量規準となっている. また, 数値実験により, 提案した MAIC の性能を従来の AIC および CAIC と比較を行った.