科研費シンポジウム「統計的モデリングと計算アルゴリズムの数理と展開」

基盤研究（A）15H01678「大規模複雑データの理論と方法論の総合的研究」

研究代表者： 青嶋 誠 (筑波大学)

開催責任者： 金森 敬文 (名古屋大学)

日時： 2017 年 2 月 18 日 (土)〜19 日 (日)

場所： 名古屋大学 情報科学研究科棟

プログラム2/18 (土)

09:55 - 10:00 オープニング

青嶋誠 (筑波大)

10:00 - 11:40 (25 分 ×4)

竹之内高志 (はこだて未来大/理研 **AIP**) 変形ブレグマン擬距離とその周辺

矢田和善 (筑波大) 高次元固有値・固有ベクトルの一致推定量について

呉偉 (名大) 組合せ問題に対する min-max regret 基準のロバスト最適化

今泉允聡 (東大) 作用素推定：関数データ回帰の統計的性質

13:10-14:50 (25 分 ×4)

吉澤真太郎 (御殿場基礎研) リッジ回帰の双対表現

小崎敏寛 (ステラリンク (株)) 回帰分析と錐計画

狩野修平 (東大) Affine 形式によるホロノミック勾配法の誤差評価

木村圭児 (九大) 変数選択に対する混合整数非線形計画法

15:10-16:20 (25 分 ×1 & 45 分 (特別講演)×1)

黒河天 (東大) 多角形充填構造を持つ画像データへの最適グラフ埋め込み手法の開発

林正人 (名大/**NUS**) 特別講演：Information Geometry Approach to Parameter Estimation in Hidden Markov Models

16:40-18:20 (25 分 ×4)

中山優吾 (筑波大) 高次元小標本におけるバイアス補正 SVM

江口真透 (統数研) Generalized Boltzmann machines and activation functions

熊谷亘 (神奈川大) パラメータ転移学習の理論解析

松島慎 (東大) 大規模な線形予測器のための 非同期特徴抽出スキーム

2/19 (日)

10:00 - 11:40 (25 分 ×4)

- 平尾将剛 (愛知県立大) 行列式点過程の準モンテカルロ積分への応用
- 汪金芳 (千葉大) Cell Regression and Reference Priors
- 澤正憲 (神戸大) Euclidean Design Theory
- 赤間陽二 (東北大) TBA

13:10-14:50 (25 分 ×4)

- 松井孝太 (名大) 入れ子型混合モデルに基づく cancer outlier profile の推定とがん診断への応用について
- 伊藤直紀 (東大) 実用的な加速近接勾配法の実装と 2 値判別モデルへの応用
- 本谷秀堅 (名工大) 空間の低ランク性と平滑性を考慮した フーリエ係数最適化による MR 超解像
- 伊藤伸一 (東大) 大自由度系のデータ同化のための 2nd-order adjoint 法を用いた高速不確実性評価法

15:10-16:50 (25 分 ×4)

- 清智也 (東大) 座標ごとの変数変換によって得られる Stein 型の分布
- 只木孝太郎 (中部大) An operational characterization of the notion of probability by algorithmic randomness and its applications
- 中川健治 (長岡技科大) 通信路容量を達成する出力分布の射影アルゴリズムによる探索について
- 渡辺一帆 (豊橋技科大) レート歪み理論と一般化事後分布

16:50 - 17:00 クロージング

変形ブレグマン擬距離とその周辺

竹之内高志

はこだて未来大学, 理化学研究所 革新知能統合研究センター

概要

離散空間上の確率モデルは様々な現象・データを柔軟に表現可能な有用なツールであり, データから適切にパラメーターを決定することで精度の高い推論が可能となる. その一方で, 離散空間上の確率モデルは正規化項の計算に指数オーダーの計算量を必要とすることがあるため, パラメーターの推定量を構成することが難しいことが多い. 本研究では, 変形ブレグマンダイバージェンスと e-混合モデルを用いることで正規化項の計算をすることなく構成可能な推定量を提案し, その性質について議論する.

1 導入

X を d 次元の離散空間 $\mathcal{X}$ 上の確率変数ベクトルとし, $\langle f \rangle = \sum_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ とする. 本稿では確率モデル

$$\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}},\ q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\psi_{\boldsymbol{\theta}}(\boldsymbol{x})),\ Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}} \rangle \tag{1}$$

に着目し, パラメーター $\boldsymbol{\theta}$ を推定することを目的とする. ただし, $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ は非正規化モデル, $Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}} \rangle$ はモデル $\bar{q}_{\boldsymbol{\theta}}$ が確率測度であることを要請する正規化項である. データセット $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ に対応する経験分布を $\tilde{p}(\boldsymbol{x})$ とする.

$$\tilde{p}(\boldsymbol{x}) = \begin{cases} \frac{n_{\boldsymbol{x}}}{n} & \boldsymbol{x} \in \mathcal{Z}, \\ 0 & \text{otherwise}, \end{cases}$$

ただし $\mathcal{Z} = \{\boldsymbol{x} | \boldsymbol{x} \in \mathcal{D}\}$, $n_{\boldsymbol{x}}$ はデータセット $\mathcal{D}$ に含まれる $\boldsymbol{x}$ の個数とする. 最尤推定量は経験分布 $\tilde{p}(\boldsymbol{x})$ と確率モデル $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ 間の KL ダイバージェンス最小化として定式化することができるが, 正規化項の計算に由来する計算量の問題から推定量の構成が困難であることが多い. 本稿では変形ブレグマンダイバージェンスと e-混合モデルを用いて, 正規化項の計算を行わずに構築可能な推定量を提案する.

2 提案法

2 つの正値測度 f, g に対して, 以下の変形ブレグマンダイバージェンスを考える. ただし U は凸関数, f は単調関数とする.

$$D(p, q; U, f) = \left\langle U(f(q)) - U(f(p)) - U'(f(p))\left\{f(q) - f(p)\right\} \right\rangle. \tag{2}$$

$D(p,q;U,f) \geq 0$ が常に成立し, $D(p,q;U,f) = 0$ が成立するのは $p = q$ のときのみである (Murata et al., 2004). 経験分布 $\tilde{p}(\boldsymbol{x})$ と確率モデル $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ の e-混合モデルを以下のように定義する (Amari and Nagaoka, 2000) .

$$\tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\tilde{p}^{\alpha}(\boldsymbol{x})\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \tilde{p}^{\alpha}\bar{q}_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle} = \frac{\tilde{p}^{\alpha}(\boldsymbol{x})q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \tilde{p}^{\alpha}q_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle} \propto \begin{cases} \left(\frac{n_{\boldsymbol{x}}}{n}\right)^{\alpha} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} & \boldsymbol{x} \in \mathcal{Z}, \\ 0 & \text{otherwise}, \end{cases}.$$

ただし $\alpha(\neq 0, 1)$ を定数とし, $\tilde{r}_{0,\boldsymbol{\theta}}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$, $\tilde{r}_{1,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})$ が成立する. e-混合モデルは正規化項を計算することなく構成可能であることに注意する.

本稿では, これらを用いて, 以下のような推定量を考える.

$$\hat{\boldsymbol{\theta}}_{U,f} = \operatorname*{argmin}_{\boldsymbol{\theta}} D(\tilde{r}_{\alpha,\boldsymbol{\theta}}, \tilde{r}_{\alpha',\boldsymbol{\theta}}; U, f). \tag{3}$$

e-混合モデルが正規化項の計算をすることなく構成可能であるため, (3) も正規化項の計算を行うことなく構成可能な推定量となる. この推定量に対して, 以下が成立する.

命題 1 推定量 $\hat{\boldsymbol{\theta}}_{U,f}$ はフィッシャー一致性を持つ. すなわち, データを生成する真の分布が $p(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ であり, $r_{\alpha,\boldsymbol{\theta}} = \frac{\bar{q}^{\alpha}_{\boldsymbol{\theta}_0}(\boldsymbol{x})\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \bar{q}^{\alpha}_{\boldsymbol{\theta}_0}\bar{q}^{1-\alpha}_{\boldsymbol{\theta}} \right\rangle}$ とするとき, 以下が成立する.

$$\boldsymbol{\theta}_0 = \operatorname*{argmin}_{\boldsymbol{\theta}} D(r_{\alpha,\boldsymbol{\theta}}, r_{\alpha',\boldsymbol{\theta}}; U, f). \tag{4}$$

補題 1 推定量 $\hat{\boldsymbol{\theta}}_{U,f}$ の漸近分布は漸近的に以下の多変量正規分布に従う.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{U,f} - \boldsymbol{\theta}_0) \sim N(0, H^{-1}_{U,f,\boldsymbol{\theta}_0} J_{U,f,\boldsymbol{\theta}_0} H^{-1}_{U,f,\boldsymbol{\theta}_0}). \tag{5}$$

ただし $H_{U,f,\boldsymbol{\theta}_0} = \left\langle \xi_{U,f}(\bar{q}_{\boldsymbol{\theta}_0})\bar{q}_{\boldsymbol{\theta}_0}(\psi'_{\boldsymbol{\theta}_0} - \mu_0)(\psi'_{\boldsymbol{\theta}_0} - \mu_0)^T \right\rangle$, $J_{U,f,\boldsymbol{\theta}_0} = \left\langle \bar{q}_{\boldsymbol{\theta}_0}\zeta_{U,f,\boldsymbol{\theta}_0}\zeta^T_{U,f,\boldsymbol{\theta}_0} \right\rangle$,
$\mu_0 = \left\langle \bar{q}_{\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0} \right\rangle$, $\xi_{U,f}(z) = U''(f(z))f'(z)^2 z$,
$\zeta_{U,f,\theta}(\boldsymbol{x}) = \xi_{U,f}(\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}))(\psi'_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mu_0) - \langle \bar{q}_{\boldsymbol{\theta}}\xi_{U,f}(\bar{q}_{\boldsymbol{\theta}})(\psi'_{\boldsymbol{\theta}} - \mu_0)\rangle$ である.

定理 1 関数 f が

$$U''(f(z))f'(z)^2 z = 1 \tag{6}$$

を満たす時, 推定量 $\hat{\boldsymbol{\theta}}_{U,f}$ の漸近分散 $H^{-1}_{U,f,\boldsymbol{\theta}_0} J_{U,f,\boldsymbol{\theta}_0} H^{-1}_{U,f,\boldsymbol{\theta}_0}$ はフィッシャー情報量行列の逆行列となり, 漸近有効となる.

References

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. *MIT Press, Cambridge, Mass*, 1:282–317, 1986.

N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-boost and bregman divergence. *Neural Computation*, 16(7): 1437–1481, 2004.

T. J. Sejnowski. Higher-order Boltzmann machines. In *American Institute of Physics Conference Series*, 151(7):398–403, 1986.

高次元固有値・固有ベクトルの一致推定量について

矢田　和善　(筑波大数理物質)
青嶋　誠　　(筑波大数理物質)

1. はじめに

情報化の進展に伴い，高次元データの統計解析が，ますます重要になってきている．2000年以降，確率論と理論物理の方面から，ランダム行列の理論に基づく幾つかの重要な結果がもたらされた．Johnstone (2001, AS) 等は，標本固有値の漸近分布を導出した．しかしながら，そこでは，データの次元数 d と標本数 n が $n/d \to c > 0$ を満たす場合を考え，高次元において標本数は次元数と同程度を仮定した．例えば，次元数は優に 10,000 を超えるが標本数は高々100 程度といった高次元小標本においては，標本数を次元数と同程度には仮定できない．それゆえ，n が d に依存しないような設定で，もしくは，$n = n(d)$ であっても $n/d \to 0$ となる設定で，高次元漸近理論を展開する必要がある．Yata and Aoshima [2] は，高次元小標本における PCA の性質を研究し，PCA が一致性をもつための標本数 n の d に関するオーダー条件を導き，高次元小標本において PCA が不適解を起こすことを示した．この問題を解決する策として，Yata and Aoshima [3] は，高次元小標本データ空間の幾何学的表現を研究し，それに基づいて “ノイズ掃き出し法” とよばれる方法論を考案した．一方で，Yata and Aoshima [4] は，高次元大標本も含む一般的な高次元データに対して，power spiked モデルと呼ばれる固有値モデルを考案し，高次元データに対する新しい PCA を構築した．最近，Aoshima and Yata [1] は，ノイズ掃き出し法による固有ベクトルの推定量を用いることで新たな高次元二標本検定法を考案した．

本講演では，高次元固有ベクトルの一致性について論じ，閾値を用いてノイズ掃き出し法による固有ベクトルの推定量を補正することで，緩い仮定のもと高次元固有ベクトルの一致性を与える新たな方法論を提案した．

2. 高次元固有ベクトルの一致性

平均に d 次のベクトル $\boldsymbol{\mu}$，共分散行列に d 次の半正定値行列 $\boldsymbol{\Sigma}$ をもつ母集団を考える．母集団から n (≥ 3) 個の d 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出する．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_d]$ で $\boldsymbol{\Sigma}$ を $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ と分解する．標本共分散行列 $\boldsymbol{S}$ のスペクトル分解を $\boldsymbol{S} = \sum_{i=1}^{d} \hat{\lambda}_i \hat{\boldsymbol{h}}_i \hat{\boldsymbol{h}}_i^T$ とする．最近，Yata and Aoshima [4] は，power spiked モデルとよばれる固有値モデルを考案し，高次元データに対する新しい PCA を研究した．いま，$\boldsymbol{\Sigma}_{(1)} = \sum_{j=1}^{m} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$，$\boldsymbol{\Sigma}_{(2)} = \sum_{j=m+1}^{p} \lambda_j \boldsymbol{h}_j \boldsymbol{h}_j^T$ とおき，$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{(1)} + \boldsymbol{\Sigma}_{(2)}$ という分解を考える．そのとき，次の条件を満たすような $\lambda_1 \geq \cdots \geq \lambda_d$ を power spiked モデルと定義する．

λ_m に対して，$\lim_{d\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_{(2)}^{k_m})/\lambda_m^{k_m} = 0$ なる (有界な) ある自然数 k_m が存在する．(1)

いま，$\delta_j = \lambda_j^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_{(2)})/(n-1)$，$j = 1, ..., m$ とおく．power spiked モデル (1) のもと，次の定理を得る．

定理 1 ([4]). 各 $j = 1, ..., m$ について，適当な正則条件と条件

(C-i) $\mathrm{tr}(\Sigma_{(2)}^2)^2/(n\lambda_j^4) = o(1)$

のもと，$d, n \to \infty$ のとき次が成り立つ．

$$\frac{\hat{\lambda}_j}{\lambda_j} = 1 + \delta_j + o_p(1) \ \text{ and } \ \boldsymbol{h}_j^T \hat{\boldsymbol{h}}_j = (1+\delta_j)^{-1/2} + o_p(1).$$

定理1より，適当な正則条件と (C-i) のもと次を得る．

$$||\hat{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = 2\{1 - (1+\delta_j)^{-1/2}\} + o_p(1). \tag{2}$$

ここで，$||\cdot||$ はユークリッドノルムを表す．一方で，Yata and Aoshima [3] は，高次元小標本データ空間の幾何学的表現を研究し，それに基づいて “ノイズ掃き出し法” とよばれる方法論を考案し，次のような固有値の推定量を提案した．

$$\tilde{\lambda}_j = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}) - \sum_{i=1}^{j} \hat{\lambda}_i}{n-1-j} \quad (j = 1, ..., n-2). \tag{3}$$

さらに，$\boldsymbol{\Sigma}$の固有ベクトルについて，ノイズ掃き出し法による推定を考える．推定量 (3) に基づいて，$\boldsymbol{\Sigma}$ の固有ベクトル $\boldsymbol{h}_j$ を $\tilde{\boldsymbol{h}}_j = (\hat{\lambda}_j/\tilde{\lambda}_j)^{1/2}\hat{\boldsymbol{h}}_j$ で推定する．そのとき，power spiked モデル (1) のもと，次の定理を得る．

定理2 ([4]). 各 $j = 1, ..., m$ について，適当な正則条件と (C-i) のもと，$d, n \to \infty$ のとき次が成り立つ．

$$\frac{\tilde{\lambda}_j}{\lambda_j} = 1 + o_p(1) \ \text{ and } \ \boldsymbol{h}_j^T \tilde{\boldsymbol{h}}_j = 1 + o_p(1).$$

それゆえ，$\boldsymbol{h}_j$ の内積に関する一致性をもつ．

ここで，$||\tilde{\boldsymbol{h}}_j||^2 = \hat{\lambda}_j/\tilde{\lambda}_j \geq 1$ であることに注意し，定理1と2より，適当な正則条件と (C-i) のもと次を得る．

$$||\tilde{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = \delta_j + o_p(1). \tag{4}$$

すなわち，(2) と (4) より，$\liminf_{d,n\to\infty} \delta_j > 0$ のとき，$\hat{\boldsymbol{h}}_j$ と $\tilde{\boldsymbol{h}}_j$ はノルムに関する一致性をもたない．

本講演では，閾値を用いて $\tilde{\boldsymbol{h}}_j$ を補正することで，$\delta_j \to \infty$ のもとでも，

$$||\acute{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = o_p(1)$$

が成り立つような新たな固有ベクトルの推定量 $\acute{\boldsymbol{h}}_j$ を提案し，理論的かつ数値的に既存の推定量と比較した．

参考文献

[1] Aoshima, M., Yata, K. (2016). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statist. Sinica*, in press (arXiv:1602.02491).

[2] Yata, K., Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. *Commun. Statist. Theory Methods* **38** 2634-2652.

[3] Yata, K., Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* **105** 193-215.

[4] Yata, K., Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.* **122** 334-354.

組合せ問題に対する min-max regret 基準のロバスト最適化

Nagoya University Wei WU
Nagoya University Mutsunori YAGIURA

Several optimization problems arising in real world applications do not have accurate estimates of the problem parameters when the optimization decision is taken. Stochastic programming and robust optimization are two common approaches for the solution of optimization problems under uncertainty. The min-max and min-max regret criteria are two of the typical approaches for robust optimization. The min-max criterion aims at obtaining a solution with the best worst-case value across all scenarios. The *regret* is defined as the difference between the actual cost and the optimal cost that would have been obtained if a different solution had been chosen. The min-max regret approach is to minimize the worst-case regret. This criterion is not as pessimistic as the min-max approach.

For the max-min regret criterion, we consider the *knapsack problem* (KP) under discrete profits. For the discrete scenario case, we assume that the scenario set is described explicitly. The discrete *max-min knapsack problem* (MM-KP) is known to be strongly NP-hard for unbounded scenario set [3]. However, the MM-KP is solvable by a pseudo-polynomial time algorithm when the size of scenario set is bounded by a constant. We examine this pseudo-polynomial time method based on dynamic programming. We also examine a branch-and-cut algorithm based a *mixed integer programming* (MIP) model. We propose a heuristic algorithm for the MM-KP that solves the underlying KP to optimality under a fixed scenario. For the average-profit scenario, we show that the optimal value of this fixed-scenario KP under the average-profit scenario is a valid upper bound. We further propose an iterative method to improve the performance of the fixed scenario heuristic. In each iteration, we generate a new scenario based on solutions obtained by then.

For the min-max criterion, we consider the *generalized assignment problem* (GAP) and the *multidimensional knapsack problem* (MKP) under interval costs. The classical GAP is a strongly NP-hard combinatorial optimization problem [6] having many applications (see [1, 4], and [5]). The classical MKP is a strongly NP-hard combinatorial optimization problem [2] and has been widely studied over many decades due to both theoretical interests and its broad applications in several engineering fields, such as cargo loading, cutting stock, bin-packing, financial and other management issues [7].

The interval *min-max regret generalized assignment problem* (MMR-GAP) is a generalization of the GAP to the case in which the cost coefficients are uncertain. In real life applications, the costs are often affected by many factors, and they can be unknown at the optimization stage. We assume that every cost coefficient can take any value in a corresponding given interval, regardless of the values taken by the other cost coefficients. The problem requires to find a robust solution that minimizes the maximum regret. We prove that the decision version of MMR-GAP is Σ_2^p-complete. We propose a heuristic algorithm for the MMR-GAP that solves the underlying GAP to optimality under a fixed scenario. We consider three scenarios (lowest cost, highest cost, and median cost), and we show that the median cost scenario leads to a solution of the MMR-GAP whose objective function value is within twice the optimal value. We also propose

a dual substitution heuristic based on a MIP model obtained by replacing some constraints with the dual of their continuous relaxation. We also propose exact algorithmic approaches that iteratively solve the problem by only including a subset of scenarios. The first approach is based on logic-based Benders decomposition: it solves a MIP with incomplete scenarios and iteratively supplements the scenarios corresponding to violated constraints. We then introduce a basic branch-and-cut algorithm and enhance it through: (i) Lagrangian relaxations, to provide tighter lower bounds than those produced by the linear programming relaxation; (ii) an efficient variable fixing technique; (iii) a two-direction dynamic programming approach to efficiently solve the Lagrangian subproblems. We compare the introduced algorithms through computational experiments on different benchmarks.

For the interval *min-max regret multidimensional knapsack problem* (MMR-MKP), we propose a new heuristic framework, which we call the *iterated dual substitution* (IDS) algorithm. The IDS iteratively generates linear constraints (rows) based on a mixed integer programming model. Computational experiments on a wide set of benchmark instances are carried out, and the proposed iterated dual substitution algorithm performs best on all of the tested instances.

References

[1] M.L. Fisher, R. Jaikumar, "A generalized assignment heuristic for vehicle routing," Networks, 11 (1981) 109–124.

[2] M. Garey, D. Johnson, Computers and Intractability: A Guide to the Theory of NP-completeness, W.H. Freeman, San Francisco, 1979.

[3] P. Kouvelis and G. Yu, Robust Discrete Optimization and Its Applications, Kluwer, Academic Publishers, Dordrecht, 1997.

[4] S. Martello, P. Toth, Knapsack Problems: Algorithms and Computer Implementations, John Wiley & Sons, Chichester, New York, 1990.

[5] K.S. Ruland, "A model for aeromedical routing and scheduling," International Transactions in Operational Research, 6 (1999) 57–73.

[6] S. Sahni, T. Gonzalez, "P-complete approximation problems," Journal of the Association for Computing Machinery, 23 (1976) 555–565.

[7] M. Varnamkhasti, "Overview of the algorithms for solving the multidimensional knapsack problem," Advanced Studies in Biology, 4 (2012) 37–47.

Operator Estimation : Analysis for Functional Regression with Functional Input and Output

Masaaki Imaizumi (UT)
(joint work with Kengo Kato (UT))

In this presentation, we investigate a regression problem where both covariate and response variables are random functions. Let the covariate X and the response Y be $L^2(I)$-valued random variables with $I = [0, 1]$, and $T : L^2(I) \to L^2(I)$ be a regression operator. Then, we consider that the random functions are generated from the following regression model

$$Y(t) = T(X)(t) + \epsilon(t), \quad t \in I, \tag{1}$$

where ϵ is a Gaussian noise process in $L^2(I)$.

Data representation by functions is used for analyzing data which are observed on a large number of grids in its domain. The fields of statistical methods for analyzing such data is called *functional data analysis*, and it is summarized in Ramsay and Silverman (2005).

The functional regression with functional covariates and functional responses is investigated by many studies, especially the linear regression case. Cuevas *et al.* (2002), Yao *et al.* (2005), Crambes and Mas (2013), Hörmann and Kidziński (2015) and more studies propose their estimators and clarify their theoretical properties. Since it is possible to represent the linear regression operator by a form $\int b(\cdot, t)x(t)dt$ using a bivariate function $b(s, t)$, some of the methods for the linear regression are conducted by estimating $b(s, t)$.

The nonlinear regression with functional covariate and functional response is a developing problem, since there is no standard way to represent the nonlinear operator T. Such the problem is considered by Bosq and Delecroix (1985), and various methods are proposed. The Nadaraya-Watson method with the kernel function is investigated by Ferraty *et al.* (2011), Lian (2011), Ferraty *et al.* (2012), and others. Another method for the problem is a functional reproducing kernel Hilbert space (fRKHS) method which is studied by Preda (2007), Lian (2007), and Kadri *et al.* (2015).

In this presentation, we propose an estimator for the linear and nonlinear functional regression problem, and clarify some regularity conditions and properties of the estimators. Then, we derive the convergence rate of the estimator which is characterized by smoothness of the variables and the operator, where the smoothness represents a speed of delay of coefficients measured by the basis function decomposition. About the convergence rate of the estimator for the linear regression, we

find that the rate is independent from the smoothness of the operator in the direction of responses, and the rate is identical to the convergence rate with the linear regression with a scalar output case provided by Hall and Horowitz (2007). For the nonlinear case, we also find the similar properties of the convergence rate with some assumptions.

References

Bosq, D. and Delecroix, M. (1985) Nonparametric prediction of a hilbert space valued random variable, *Stochastic processes and their applications*, **19**, 271–280.

Crambes, C. and Mas, A. (2013) Asymptotics of prediction in functional linear regression with functional outputs, *Bernoulli*, **19**, 2627–2651.

Cuevas, A., Febrero, M. and Fraiman, R. (2002) Linear functional regression: the case of fixed design and functional response, *Canadian Journal of Statistics*, **30**, 285–300.

Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011) Kernel regression with functional response, *Electronic Journal of Statistics*, **5**, 159–171.

Ferraty, F., Van Keilegom, I. and Vieu, P. (2012) Regression when both response and predictor are functions, *Journal of Multivariate Analysis*, **109**, 10–28.

Hall, P. and Horowitz, J. L. (2007) Methodology and convergence rates for functional linear regression, *The Annals of Statistics*, **35**, 70–91.

Hörmann, S. and Kidziński, Ł. (2015) A note on estimation in hilbertian linear models, *Scandinavian journal of statistics*, **42**, 43–62.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A. and Audiffren, J. (2015) Operator-valued kernels for learning from functional response data, *Journal of Machine Learning Research*, **16**, 1–54.

Lian, H. (2007) Nonlinear functional models for functional responses in reproducing kernel hilbert spaces, *Canadian Journal of Statistics*, **35**, 597–606.

Lian, H. (2011) Convergence of functional k-nearest neighbor regression estimate with functional responses, *Electronic Journal of Statistics*, **5**, 31–40.

Preda, C. (2007) Regression models for functional data by reproducing kernel hilbert spaces methods, *Journal of Statistical Planning and Inference*, **137**, 829–840.

Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis*, 2ne Edition, Springer.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional linear regression analysis for longitudinal data, *The Annals of Statistics*, **33**, 2873–2903.

リッジ回帰の双対表現

吉澤 真太郎　(御殿場基礎科学研究会)*

概　　要

リッジ回帰の 1 つに Hodrick-Prescott フィルタ (1997) と呼ばれるものがある（以下、HP フィルタと呼ぶ)。このフィルタは Leser(1961) が起源とされている。HP フィルタは、観測時系列ベクトルとトレンド成分ベクトルの差の 2 乗ノルムに、差分行列がトレンド成分ベクトルに作用したベクトルの 2 乗ノルムをスカラ倍して加えたものとなっている。Yamada(2015) は、HP フィルタのトレンド成分ベクトルに、差分行列から構成される変換が作用した族に対して、具体的な 2 種類の作用について考察している。この 2 種類の作用は、トレンド成分ベクトルに作用する差分行列の核空間と像空間から具体的に構成されたものである。本来、トレンド成分ベクトルへの変換作用すべてに対し、リッジ回帰の表現を検討すべきであると考える。そこで、本報告では差分行列の核空間と像空間とを陽に表示することで、HP フィルタの一般表現及びその双対表現を考察する。これら 2 つの一般表現のペアから通常の双対幾何構造（情報幾何）を考察した。ただし、通常のリッジ回帰の正則化パラメータは非負を前提としているが、一般表現とその双対表現においては、正則化パラメータが非負とは限らないところが通常の正則化と相違している。正則化パラメータが負の値も取りうることは、幾何学的には不自然なことではない。正則化パラメータと差分行列の特異値の大小関係が（一般表現及びその双対表現のヘッシアンが正定、不定、負定の 3 つの場合)、HP フィルタの性能に与える影響を数理的に考察することは、今後の課題とする。

参考文献

[1] C.E.V.Leser, A simple method of trend construction, *Journal of the Royal Statistical Society, Series B (Methodological)*, 23, (1961), 91-107.

[2] R.J.Hodrick and E.C. Prescott, Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 29(1),(1997), 1-16

[3] H.Yamada, Ridge regression representations of the generalized Hodrick-Prescott filter, *J. Japan Statist.Soc.*, Vol.45, No.2 (2015), 121-128.

2016 年度 科研費シンポジウム：統計的モデリングと計算アルゴリズムの数理と展開（於：名古屋大学情報学研究科, 2016 年 2 月 18 日～19 日）

* e-mail: yzw2003@mail.goo.ne.jp

報告書 (回帰分析と錐計画)

小崎敏寛 (Toshihiro Kosaki)*
ステラリンク株式会社 (Stera Link, Co., Ltd.)

1　課題

回帰分析は，一つの被説明変数の変動をいくつかの説明変数によって説明する手法である．その中で，あてはめからの誤差からなるベクトルの 2 ノルムの 2 乗を最小にするようにパラメータを決める手法が最小二乗法である．最小化する基準として，1 ノルムや ∞ ノルムを考えることも古くから行われてきた [2, 8].

2　解決手法

近年，錐計画問題に注目が集まっている．錐計画問題の一つのクラスとして，二次錐計画問題 [1, 6] がある．最近，この問題の 2 ノルムを一般のノルムにしたノルム錐計画問題 [4, 5] を提案した．本研究では，ノルムをさらに一般化したゲージ関数 [3, 7] を考える．ゲージ関数からなる錐を持つ問題をゲージ錐計画問題と呼ぶ．最小化する基準として，2-ノルムでなく，ゲージ関数を考える．すると問題はゲージ錐計画問題になる．

3　結果

ゲージ錐計画問題の双対問題を考えることで，弱双対定理がなりたつことを示した．

4　まとめと今後の課題

回帰分析の計算法にゲージ錐計画が適用できることを示した．この問題に対して，弱双対定理がなりたつことを示した．

今後の課題として次のようなものがある．ゲージ錐計画を解くアルゴリズムを考え，実際に数値実験を行う．そして得られた結果の解釈を行う．また，双対ギャップが存在しない条件を調べる．

参考文献

[1] F. Alizadeh and D. Goldfarb, Second-Order Cone Programming, Mathematical Programming, 95, 3-51, 2003.

*toshihirokosaki@gmail.com

[2] T. S. Arthanari and Y. Dodge, Mathematical Programming in Statistics, John Wiley & Sons, Inc., 1981.

[3] H-H Chao and L. Vandenberghe, Semidefinite Representations of Gauge Functions for Structured Low-Rank Matrix Decomposition, arXiv, 2016.

[4] 小崎敏寛, ノルム錐計画問題の双対性, 京都大学数理解析研究所講究録 1931 : 最適化アルゴリズムの進展 : 理論・応用・実装, 89-93, 2015.

[5] 小崎敏寛, 一般のノルム錐計画問題の弱双対定理, 統計数理研究所共同研究リポート 369, 最適化 : モデリングとアルゴリズム 28, 75-78, 2016.

[6] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, Applications of Second-Order Cone Programming, Linear Algebra and its Applications, 284, 193-228, 1998.

[7] D. G. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons, Inc., 1969.

[8] G. A. Watson, Approximation in Normed Linear Spaces, Journal of Computational and Applied Mathematics, 121, 1-36, 2000.

Affine 形式によるホロノミック勾配法の誤差評価

東京大・情報理工　狩野　修平
東京大・情報理工　清　智也

Nakayama et al. (2011) によって提案されたホロノミック勾配法は，計算対象の関数をそれが満たす微分方程式系に変換して数値計算を行うという手法である．ホロノミック勾配法を適用する場合，微分方程式を解くために適当な初期値を与える必要がある．既存研究の多くではパラメータ付き積分があるパラメータにおいて陽に書ける場合はその値を用い，それが不可能な場合は級数展開を用いて初期値を設定している (Sei and Kume (2015) 等)．しかし，級数展開による初期値設定は真の初期値を計算できているわけではなく，実際に計算した積分値がどれぐらい信頼できるのかということはあまり考察されてこなかった．その問題に対処するために，本発表では初期値が区間で与えられたもとでの Pfaffian 系の数値計算について考える．

数値計算したいホロノミック関数を $f(\theta) = f(\theta_1, \ldots, \theta_p)$, $\theta \in \Theta \subset \mathbb{R}^p$ とし，$f(\theta)$ を含む q 次のベクトルを $m(\theta)$ と書く．$i = 1, \ldots, p$ について，$m(\theta)$ に

$$\partial_i m(\theta) = P_i(\theta) m(\theta), \quad \left(\partial_i = \frac{\partial}{\partial \theta_i}\right) \tag{1}$$

を満たすような Pfaffian 系 $P_i : \Theta \to \mathbb{R}^{q \times q}$ が存在しているものとする．この時，興味のあるパラメータを $\theta(1)$，それと異なるパラメータを $\theta(0)$ と書き，それらをつなぐ curve を $\theta(t)$, $t \in [0, 1]$ と表す．この時式 (1) は

$$\partial_t m(\theta(t)) = P_t^* m(\theta(t)) = g(t, m), \quad P_t^* := \sum_{i=1}^{p} (\partial_t \theta_i(t)) P_i(\theta(t)), \quad \partial_t = \frac{\partial}{\partial t} \tag{2}$$

のような常微分方程式系に書き換えられる．式 (2) で $m(\theta(0))$ が与えられた時に $m(\theta(1))$ を計算することを考える．ここで，与えられた初期値 $\hat{m}(\theta(0))$ については，

$$|\hat{m}(\theta(0)) - m(\theta(0))| \leq C, \; C = (c_1, \ldots, c_q)^\top \in \mathbb{R}^q \tag{3}$$

のような評価が成立していると仮定する．式 (2)(3) で与えられた微分方程式に対するアプローチとしては，Taylor model による古典的な方法 (Lohner (1987) 等) と，既存の Runge-Kutta 法のスキームを利用して解の包含を生成する手法が提案されている (Dit Sandretto and Chapoutot (2016) 等)．本研究では，ホロノミック勾配法の既存研究において幅広く Runge-Kutta 法が用いられていることを考慮し，後者の枠組みを用いて微分方程式を数値的に解き，既存研究と比較することを考える．

式 (2)(3) で表される微分方程式の数値解を計算するには，区間演算と呼ばれる変数の上界と下界を与えて区間を表す演算法が用いることができる．しかし，区間演算は変数間の相関性を反映できないという欠点があるため，本研究では Affine 形式 (Stolfi and De Figueiredo (2003)) と呼ばれる変数の存在範囲を多項式の形で表す手法を使って計算を行う．区間演算と Affine 形式は，演算の前後に対して包含性が保存されるという性質があるため，初期値を区間の形で与え，その中に真の値が含まれている限り，計算結果として出力される区間に真の解が含まれているとみなすことができる．具体的には，$t = t_j$ における Affine 形式の $\tilde{m}_j$ とあるステップ幅 h について，次のように $t_{j+1} = t_j + h$ における Affine 形式 $\tilde{m}_{j+1}$ を計算する．

$$\tilde{m}_{j+1} = \tilde{y}_{n+1} + \tilde{\mathrm{LTE}}([t_j, t_{j+1}], m([t_j, t_{j+1}]))$$

ただし，$\tilde{y}_{n+1}$ は

$$\tilde{y}_{n+1} = \tilde{m}_n + h\sum_{i=1}^{s} b_i k_i, \quad k_i = g(t_n + c_i h, \tilde{m}_n + h\sum_{j=1}^{s} a_{ij} k_j)$$

のように，通常の Runge-Kutta 法で計算を行い，$\mathrm{LTE}([t_j, t_{j+1}], m)$ は各ステップにおける打ち切り誤差を示す．発表では実際に用いる計算アルゴリズムの詳細と数値実験の結果について説明する．

参考文献

Dit Sandretto, J. A. and Chapoutot, A. (2016). Validated Explicit and Implicit Runge-Kutta Methods, *Reliable Computing*, Vol. 22, pp. 78–103.

Lohner, R. J. (1987). Enclosing the Solutions of Ordinary Initial and Boundary Value Problems, in *Computer Arithmetic: Scientific Computation and Programming Languages*, Stuttgart: Wiley-Teubner, pp. 255–286.

Nakayama, H., Nishiyama, K., Noro, M., Ohara, K., Sei, T., Takayama, N., and Takemura, A. (2011). Holonomic gradient descent and its application to the Fisher-Bingham integral, *Advances in Applied Mathematics*, Vol. 47, No. 3, pp. 639–658.

Sei, T. and Kume, A. (2015). Calculating the normalising constant of the Bingham distribution on the sphere using the holonomic gradient method, *Statistics and Computing*, Vol. 25, No. 2, pp. 321–332.

Stolfi, J. and De Figueiredo, L. (2003). An Introduction to Affine Arithmetic, *TEMA - Tendências em Matemática Aplicada e Computacional*, Vol. 4, No. 3, pp. 297–312.

変数選択に対する混合整数非線形計画法

九州大学大学院数理学府数理学専攻　木村 圭児
九州大学マス・フォア・インダストリ研究所　脇　隼人

1. 変数選択と AIC 最小化

統計学におけるモデル推定では, 情報量規準を用いてモデルに採用するパラメータを選択することがあり, これは変数選択と呼ばれる. 赤池情報量規準 (Akaike's information criterion: AIC) などの情報量規準が用いられ, AIC が最小であるようなパラメータの組合せを求めることで, データとの当てはまりの良さを損なわないように予測精度の良いモデルを推定することができる. AIC ができるだけ小さい値をとる変数選択の一般的な手法の一つとして, ステップワイズ法が知られている. ステップワイズ法は, R などの既存の統計ソフトウェアに実装されていて, 計算が高速で広く用いられている. しかし, ステップワイズ法は局所探索をしているとみなせるので, 推定されたモデルの AIC が最小であるとは限らない.

2. 既存手法と提案する手法

ロジスティック回帰モデルに対して, 線形近似を用いた混合整数線形計画法による変数選択が提案されている [3]. この手法によって得られる解は, AIC が最小であると保証されないが, ステップワイズ法に比べ質の良い解である. 一方, 線形回帰モデルに対する変数選択の手法として, 混合整数二次錐計画法を用いた手法 [2] や混合整数二次計画法を用いた手法 [1] が提案されている.

本研究は, 最適化を基にした変数選択に対して, 混合整数非線形計画問題として定式化し, 定式化された問題を効率良く解くための様々な工夫を提案している. 提案する手法を実装するために, 分枝限定法のフレームワークを提供している SCIP (Solving Constraint Integer Program [4]) を用いる. SCIP は自由度が高いソフトウェアであり, 解法を細かく制御するプログラムを実装することができる. そのため, 本研究は定式化した問題をソルバーのみで解くというわけでは無く, 効率良く工夫を提案し, 実装まで実現している. 線形回帰における AIC 最小化に対して, 提案する手法がどの程度の規模の問題まで求解できるか紹介する.

3. MINLP として定式化

最適化を用いる変数選択では, 一般に, 目的関数は "与えられたデータとモデルの誤差" と "説明変数の数" の二つの項から形成される. モデルに現れるパラメータを $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ とする. 変数選択では, 説明変数の候補が選択されなかった場合, 対応するパラメータ β_j は 0 である. 変数選択のための最適化問題は, 次の混合整数非線形計画問題として定式化できる.

$$\begin{aligned} \min_{\boldsymbol{\beta}, \boldsymbol{z}} \ & f(\boldsymbol{\beta}) + \lambda \sum_{j \in I_p} z_j \\ \text{s.t. } & z_j = 0 \Rightarrow \beta_j = 0 \ (j \in I_p) \\ & \beta_j \in \mathbb{R}, \ z_j \in \{0, 1\} \ (j \in I_p) \end{aligned} \tag{1}$$

ただし, λ は正の定数で, $I_p = \{1, \ldots, p\}$ である. $f(\boldsymbol{\beta})$ が与えられたデータとモデルの誤差を表す関数ならば, 問題 (1) の目的関数はモデルを評価することができる. 目的関数の第二項 $\lambda \sum_{j=1}^{p} z_j$ は説明変数の数のペナルティとして機能する.

4. 下界値と上界値の計算

分枝限定法を用いた問題 (1) を効率良く手法を提案する. 分枝限定法は, 問題 (1) の部分問題の最小値の下界値と問題 (1) の最小値の上界値が必要である. 分枝限定法のフレームワークを提供している SCIP[4] に下界値と上界値の計算を実装することで, 問題 (1) を解くことができる.

分枝限定法における分枝操作によって, 部分問題が生成される際, z_j $(j \in I_p)$ は 1 あるいは 0 に固定される. 生成された部分問題に対して, z_j $(j \in I_p)$ の添字に関して次の集合を定義する.

$$\begin{aligned} Z_1 &= \{j \in I_p : z_j \text{ は } 1 \text{ に固定されている }\}, \\ Z_0 &= \{j \in I_p : z_j \text{ は } 0 \text{ に固定されている }\}, \\ Z &= \{j \in I_p : z_j \text{ はまだ固定されていない }\}. \end{aligned}$$

部分集合 $Z_1, Z_0, Z \subseteq I_p$ に対して, 問題 (1) の部

分問題 $Q(Z_1, Z_0, Z)$ は次のように表される.

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{z}} \ f(\boldsymbol{\beta}) + \lambda \sum_{j \in I_p} z_j \\ &\text{s.t. } z_j = 0 \Rightarrow \beta_j = 0, \ z_j \in \{0,1\} \ (j \in Z) \\ &\qquad \beta_j = 0, \ z_j = 0 \ (j \in Z_0) \\ &\qquad z_j = 1 \ (j \in Z_1), \ \beta_j \in \mathbb{R} \ (j \in I_p) \end{aligned} \tag{2}$$

部分問題 $Q(Z_1, Z_0, Z)$ の $z_j \in \{0,1\}$ $(j \in Z)$ の整数性を緩和した緩和問題は次のように表される.

$$\begin{aligned} &\min_{\boldsymbol{\beta},\boldsymbol{z}} f(\boldsymbol{\beta}) + \lambda \sum_{j \in I_p} z_j \\ &\text{s.t. } z_j = 0 \Rightarrow \beta_j = 0, \ 0 \le z_j \le 1 \ (j \in Z) \\ &\qquad \beta_j = 0, \ z_j = 0 \ (j \in Z_0) \\ &\qquad z_j = 1 \ (j \in Z_1), \ \beta_j \in \mathbb{R} \ (j \in I_p) \end{aligned} \tag{3}$$

問題 (3) の最小値は, 部分問題 $Q(Z_1, Z_0, Z)$ の最小値の下界値である. 次に問題 (3) から制約 $z_j = 0 \Rightarrow \beta_j = 0, \ 0 \le z_j \le 1 \ (j \in Z)$ と $z_j \ (j \in \{1, \ldots, p\})$ を取り除いた以下の問題について考える.

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ f(\boldsymbol{\beta}) + \lambda \#(Z_1) \\ &\text{s.t. } \beta_j = 0 \ (j \in Z_0), \ \beta_j \in \mathbb{R} \ (j \in I_p) \end{aligned} \tag{4}$$

問題 (3) の任意の実行可能解 $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$ に対して, $\sum_{j \in I_p} \bar{z}_j = \sum_{j \in Z} \bar{z}_j + \lambda \#(Z_1)$ より, 次の不等式が成立する.

$$f(\bar{\boldsymbol{\beta}}) + \lambda \sum_{j \in I_p} \bar{z}_j \ge f(\bar{\boldsymbol{\beta}}) + \lambda \#(Z_1).$$

$\bar{\boldsymbol{\beta}}$ は (4) の実行可能解である. よって, 問題 (4) の最小値は部分問題 $Q(Z_1, Z_0, Z)$ の最小値の下界値である. したがって, 本研究では, 問題 (4) を部分問題 $Q(Z_1, Z_0, Z)$ の下界値を求めるための緩和問題として扱う.

線形回帰における AIC 最小化の場合, 問題 (4) は制約無し凸二次計画問題となる. よって, 線形方程式を解くことで問題 (4) の最小解を求めることができる. ロジスティック回帰における AIC 最小化の場合, 問題 (4) は制約無し凸計画問題となる. よって, ニュートン法などの既存のアルゴリズムを適用することで問題 (4) の最小解を求めることができる.

分枝限定法では, 問題 (1) の最小値の上界値も必要である. 問題 (4) の最小解から問題 (1) の実行可能解を生成することができる. 生成した解を用いて, 問題 (1) の最小値の上界値を得る.

5. 効率良く解くための工夫

本研究では, 問題 (1) を効率良く解くために以下の工夫を提案している.

(I) 親問題の緩和問題を用いた下界値の更新
(II) 解の傾向を利用した計算コストの削減
(III) 分枝変数の決め方
(IV) ステップワイズ法を基にした上界値の更新

分枝限定法のフレームワークを提供している SCIP[4] にこれらを実装している.

6. 数値実験

線形回帰における AIC 最小化に対して, 既存の手法 (MISOCP[2] と MIQP[1]) と提案する手法 (MINLP) の実験結果を紹介する. 数値実験には, [5] で公開されているデータを使用する. 計算時間は制限時間を 5000 秒とする. 5000 秒で解けない場合は, >5000 と記し, AIC は得られた上界値を記す.

name	p	手法	AIC	time(sec)
sfC	26	MINLP	2816.3	**10.5**
		MISOCP[2]	2816.3	489.8
		MIQP[1]	2816.3	26.49
fires	63	MINLP	**1429.6**	>5000
		MISOCP[2]	1431.6	>5000
		MIQP[1]	1435.1	>5000
crime	100	MINLP	**3410.3**	>5000
		MISOCP[2]	3690.7	>5000
		MIQP[1]	3646.4	>5000

参考文献

[1] D. Bertsimas, A. King and R. Mazumder, "Best Subset Selection via a Modern Optimization Lenz" , Ann. Stat., 44, 2, 813 – 852, 2016.

[2] R. Miyashiro and Y. Takano, "Mixed Integer second-order cone programming formulations for variable selection," Eur. J. Oper. Res., 247, 721–731, 2015.

[3] T. Sato, Y. Takano, R. Miyashiro and A. Yoshise, "Feature Subset Selection for Logistic Regression via Mixed Integer Optimization," Computational Optimization and Applications, 2015.

[4] SCIP: Soving Constraint Integer Programs, `http://scip.zib.de/`

[5] UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/`

多角形充填構造を持つ画像データへの最適グラフ埋め込み手法の開発

黒河天 * 伊藤伸一 † 長尾大道 *† 糟谷正 ‡ 井上純哉 ‡

1 はじめに

細胞や組織の画像から，それらの境界線を抽出する問題は，多くの分野で重要である．例えば，材料科学分野では，材料内部における結晶粒成長の写真から粒の境界を抽出し，データ同化を用いたパラメータ推定や状態予測に生かすことができる．結晶粒の成長モデルに適用可能なデータ同化手法として [1] がある．領域分割 [2] やクラスタリング [3] といった，各ピクセルにラベルを与えるアプローチは，境界線そのものを求めることができない．一方，エッジ検出や直線検出は，物理的に解釈可能な線群を与えるとは限らない．

頂点集合 V，辺集合 E の定める無向グラフ $G = (V, E)$ と 2 次元標本点上のスカラー場 $f = (f_{m,n})_{m=1,\dots,M,n=1,\dots,N}$（例えば，写真の画素値など）に対して，$f$ を最もよく説明する G の平面埋め込みを求める問題をグラフ当てはめ (graph fitting)，略して graphit と呼ぶことにする．我々は graphit の観点から，多角形充填構造を持つ画像の境界線を求める手法を提案する．

2 提案手法

2次元標本点上のスカラー場 f を連続領域上に拡張したスカラー場 Ψ を定義する．頂点 $V = \{v_1, \dots, v_I\}$ の座標 $(x_1, y_1), \dots, (x_I, y_I)$ が定める G の平面埋め込みを考え，埋め込みに沿った線積分

$$J(x_1, y_1, \dots, x_I, y_I) = \sum_{\{v_i, v_j\} \in E} \int_{\Gamma_{ij}} \Psi \, \mathrm{d}l \tag{1}$$

を導入する．ただし，Γ_{ij} は (x_i, y_i) から (x_j, y_j) への線分を表す．このとき，最適化問題

$$\min J(x_1, y_1, \dots, x_I, y_I) \text{ s.t. } (x_1, y_1, \dots, x_I, y_I) \in S \tag{2}$$

の解を，求める graphit とする．ただし，S は適当な実行可能領域を表す．また，Ψ ではなく Ψ^2 の線積分を最小化する方法を，2乗 graphit という．

線分 Γ_{ij} を関数

$$\Phi_{ij}(\theta) = (1 - \theta) \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \theta \begin{pmatrix} x_j \\ y_j \end{pmatrix}, \theta \in [0, 1] \tag{3}$$

* 東京大学大学院情報理工学系研究科
† 東京大学地震研究所
‡ 東京大学大学院工学系研究科

でパラメータ付けるとき，graphit における J の勾配は陽に書けて

$$\frac{\partial J}{\partial x_i} = \sum_{j:\{v_i,v_j\}\in E} \left(\frac{x_i - x_j}{L_{ij}^2} \int_{\Gamma_{ij}} \Psi \,\mathrm{d}l + L_{ij} \int_0^1 (1-\theta) \Big(\frac{\partial \Psi}{\partial x} \circ \Phi_{ij} \Big)(\theta) \,\mathrm{d}\theta \right), \tag{4}$$

$$\frac{\partial J}{\partial y_i} = \sum_{j:\{v_i,v_j\}\in E} \left(\frac{y_i - y_j}{L_{ij}^2} \int_{\Gamma_{ij}} \Psi \,\mathrm{d}l + L_{ij} \int_0^1 (1-\theta) \Big(\frac{\partial \Psi}{\partial y} \circ \Phi_{ij} \Big)(\theta) \,\mathrm{d}\theta \right) \tag{5}$$

となる．L_{ij} は Γ_{ij} の長さとする．したがって graphit は逐次２次計画法 [4] や L-BFGS-B [5] で計算できる．Ψ の定義には幾つかの方法が考えられる．

1. 以下の $\Psi : \big[0, M\big) \times \big[0, N\big) \to \mathbf{R}$ の定める graphit を，素朴な graphit という：

$$\Psi|_{[m-1,m)\times[n-1,n)}(x, y) \coloneqq f_{m,n} \text{ for } m = 1, \ldots, M, n = 1, \ldots, N. \tag{6}$$

2. 以下の $\Psi : \big[0, M-1\big) \times \big[0, N-1\big) \to \mathbf{R}$ の定める graphit を，１次スプラインに基づく graphit という：

$$\begin{aligned} \Psi|_{[m-1,m)\times[n-1,n)}(x, y) \coloneqq & (m-x)(n-y) f_{m,n} + (m-x)(y-n+1) f_{m,n+1} \\ & + (x-m+1)(n-y) f_{m+1,n} + (x-m+1)(y-n+1) f_{m+1,n+1} \\ & \text{for } m = 1, \ldots, M-1, n = 1, \ldots, N-1. \end{aligned} \tag{7}$$

それぞれの定める graphit 及び２乗 graphit を用いて，多角的充填構造を持つ人工画像の境界を抽出する数値実験を行った．この結果，１次スプラインに基づく２乗 graphit が，最も遠い初期頂点配置から真の頂点配置を探索する傾向にあった．

3 おわりに

本研究では graphit に基づく画像からの境界抽出の手法を提案した．提案手法は必ず多角形充填構造を持つ解を与えるが，解は頂点の初期配置に強く依存するため，手動で引いた境界線を補正するに止まる．今後の課題として，良い初期配置を自動的に構成する手法の開発が挙げられる．

参考文献

[1] S. Ito, H. Nagao, A. Yamanaka, Y. Tsukada, T. Koyama, M. Kano, and J. Inoue. Data assimilation for massive autonomous systems based on a second-order adjoint method. *Physical Review E*, Vol. 94, No. 043307, 2016.

[2] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, Vol. 26, pp. 1277–1294, 1993.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, Vol. 31, pp. 264–323, 1999.

[4] P. T. Boggs and J. W. Tolle. Sequential quadratic programming. *Acta Numerica*, Vol. 4, pp. 1–51, 1995.

[5] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, Vol. 16, pp. 1190–1208, 1995.

Information Geometry Approach to Parameter Estimation in Hidden Markov Models

Masahito Hayashi[1]
[1]Graduate School of Mathematics, Nagoya University, Japan,
and Centre for Quantum Technologies, National University of Singapore,
Singapore.

Abstract: We consider the estimation of hidden Markovian process by using information geometry with respect to transition matrices. We consider the case when we use only the histogram of k-memory data. Firstly, we focus on a partial observation model with Markovian process and we show that the asymptotic estimation error of this model is given as the inverse of projective Fisher information of transition matrices. Next, we apply this result to the estimation of hidden Markovian process. For this purpose, we define an exponential family of $\mathcal{Y}$-valued transition matrices. We carefully equivalence problem for hidden Markovian process on the tangent space. Then, we propose an novel method to estimate hidden Markovian process.

Information geometry established by Amari and Nagaoka [1] is a very powerful method for statistical inference. Recently, the paper [4] applied this approach to estimation of Markovian process. In the paper [4], they employed information geometry of transition matrices given by Nakagawa and Kanaya [2] and Nagaoka [3]. Since this geometric structure depends only on the transition matrices, it does not change as the number n of observation increases while the geometry based on the probability distribution changes according to the increase of the number n. In particular, the paper [4] introduced the curved exponential family of transition matrices, and derived the Cramér-Rao inequality for the family, which shows the optimality of the inverse of the transition matrix version of Fisher information matrix.

On the other hand, some of preceding studies [5, 6] of hidden Markov process employed information geometry. However, they studied em-algorithm based on the geometry of probability distributions, which changes according to the increase of the number n. So, the estimation process becomes complicated when n is large. Hence, they could not evaluate the asymptotic behavior of the estimation error.

In this paper, we apply the information geometry of transition matrices to estimation of hidden Markovian process. Since we need to estimate the hidden structure from the observed value, we apply the em algorithm based on the geometry of transition matrices. That is, for this purpose, we formulate a partial observation model of Markovian process and the em algorithm based on the

geometry of transition matrices for this model. Then, using the transition matrix version of the projective Fisher information, we evaluate the asymptotic error in the Markovian case under a certain regularity condition.

However, we have another difficulty for the hidden Markovian process. There is ambiguity for the transition matrix to express the hidden Markovian process. That is, there is a possibility that two different transition matrices express the same hidden Markovian process. This problem is called the equivalence problem and is solved by Ito, Kobayashi, and Amari [7]. However, to discuss the estimation of the hidden Markovian process, we need to consider this problem in the tangent space because the asymptotic error is characterized by the local geometrical structure.

In this paper, for this purpose, we establish the local equivalence relation for the hidden Markovian process. When we apply the above em algorithm to the hidden Markovian process, for the regularity condition, we need to guarantee that the tangent space is non-degenerate with respect to the local equivalence condition.

Acknowledgment

The authors are very grateful to Professor Takafumi Kanamori and Professor Vincent Y. F. Tan for helpful discussions and comments. The works reported here were supported in part by the JSPS Grant-in-Aid for Scientific Research (B) No. 16KT0017, the Okawa Research Grant and Kayamori Foundation of Informational Science Advancement.

References

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press (2000).

[2] K. Nakagawa and F. Kanaya, "On the converse theorem in statistical hypothesis testing for markov chains," *IEEE Trans. Inform. Theory*, Vol. 39, No. 2, 629-633 (1993).

[3] H. Nagaoka, "The exponential family of Markov chains and its information geometry" Proceedings of The 28th Symposium on Information Theory and Its Applications (SITA2005), Okinawa, Japan, Nov. 20-23, (2005).

[4] M. Hayashi and S. Watanabe, "Information Geometry Approach to Parameter Estimation in Markov Chains," *Annals of Statistics*, Volume 44, Number 4, 1495-1535 (2016).

[5] S. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, Vol. 8, No. 9, 13791408 (1995).

[6] Y. Fujimoto and N. Murata, "A modified EM algorithm for mixture models based on Bregman divergence," *Annals of the Institute of Statistical Mathematics*, Vol. 59, No. 1, 325 (2007).

[7] H. Ito, S. -I. Amari, and K. Kobayash, "Identifiability of Hidden Markov Information Sources and Their Minimum Degrees of Freedom," *IEEE Trans. Inform. Theory*, Vol. 38, No. 2, 324-333, (1992).

高次元小標本におけるバイアス補正 SVM

筑波大学・数理物質科学　　中山　優吾
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, 高次元小標本データにおける判別分析を考えた．母集団が 2 個あると想定し, 各母集団 π_i $(i = 1, 2)$ は平均に p 次のベクトル $\boldsymbol{\mu}_i$, 共分散行列に p 次の正定値対称行列 $\boldsymbol{\Sigma}_i$ $(> \boldsymbol{O})$ をもつと仮定する．ここで, 高次元データに対して $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ を想定することは現実的ではないので, 共分散行列の共通性は仮定しない．ただし, $\liminf_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} > 0$, $\limsup_{p\to\infty}\{\mathrm{tr}(\boldsymbol{\Sigma}_1)/\mathrm{tr}(\boldsymbol{\Sigma}_2)\} < \infty$ を仮定する．各母集団 π_i から，n_i (≥ 2) 個のトレーニングデータ $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i}$ を無作為に抽出する．判別対象のデータを $\boldsymbol{x}_0$ $(\in \pi_1$ もしくは $\in \pi_2)$ とし，$\boldsymbol{x}_0 \in \pi_i$ を誤判別する確率を $e(i)$ と表記する．$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とおく．

高次元の 2 群における判別分析について，Aoshima and Yata (2015) は，$\Delta \to \infty$, $p \to \infty$ なる非スパース性と適当な正則条件のもと，一般的な高次元判別方式に

$$e(i) \to 0,\ p \to \infty,\ i = 1, 2 \tag{1}$$

なる一致性が得られることを証明した．

一方で，サポートベクターマシン (SVM) は高次元データ解析において疎な解が得られ，汎化性能が良いことも知られているが，Nakayama et al. (2016) は SVM の漸近的性質を高次元小標本の枠組みで研究した．本講演では，トレーニングデータが線形分離可能であることに着目し，線形 SVM (LSVM) の高次元小標本における漸近的性質を導出し，一致性を与えるための正則条件も導出した．LSVM はある正則条件のもとで (1) を示すが，高次元小標本においてはバイアス項によって不一致性を起こす恐れがある．LSVM はある正則条件のもとで (1) を示すが，高次元小標本においてはバイアス項によって不一致性を起こす恐れがある．そこで，そのバイアス項を補正したバイアス補正 LSVM (BC-LSVM) を提案した．さらに，バイアス補正非線形 SVM も提案し，その判別性能を理論的かつ数値的に検証した．

2　高次元データにおける LSVM の漸近的性質とバイアス補正

線形 SVM (LSVM) の判別関数を $y(\boldsymbol{x}_0)$ とする．いま，$p \to \infty$ で次を仮定する．

(A-i)　$\mathrm{Var}(||\boldsymbol{x}_{ik} - \boldsymbol{\mu}_i||^2) = O\{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)\}$，$i = 1, 2$;

(A-ii)　$\dfrac{\max_{i=1,2}\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}{\Delta^2} = o(1)$.

$\Delta_* = \Delta + \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$ とおく．ここで，$\kappa = \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 - \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$ とし，次を仮定する．

(A-iii) $\quad \limsup_{p\to\infty} \dfrac{|\kappa|}{\Delta} < 1.$

定理 1 (Nakayama et al., 2016)**.** (A-i)-(A-iii) を仮定する．$p \to \infty$ のとき判別関数 $y(\boldsymbol{x}_0)$ について (1) が成り立つ．

(A-iii) は LSVM のバイアス項に関する仮定であることに注意する．しかしながら，通常，高次元の枠組みで n_1 と n_2，もしくは $\mathrm{tr}(\boldsymbol{\Sigma}_1)$ と $\mathrm{tr}(\boldsymbol{\Sigma}_2)$ が不均等の場合，(A-iii) は仮定できない．もし，(A-iii) が仮定できない場合は以下の不一致性が成り立つ．

系 1 (Nakayama et al., 2016)**.** (A-i) と (A-ii) を仮定する．$p \to \infty$ のとき判別関数 $y(\boldsymbol{x}_0)$ について次が成り立つ．

$$e(1) \to 1 \quad \textit{and} \quad e(2) \to 0 \quad \textit{as } p \to \infty \quad \textit{if} \quad \liminf_{p\to\infty} \frac{\kappa}{\Delta} > 1; \quad \textit{and}$$
$$e(1) \to 0 \quad \textit{and} \quad e(2) \to 1 \quad \textit{as } p \to \infty \quad \textit{if} \quad \limsup_{p\to\infty} \frac{\kappa}{\Delta} < -1.$$

系 1 からもし (A-iii) が成り立たない場合は，もはや LSVM は高次元データ解析において用いるべきではない．そこで，バイアス補正 LSVM (BC-LSVM) を次のように定義する．

$$y_{BC}(\boldsymbol{x}_0) = y(\boldsymbol{x}_0) - \frac{\hat{\kappa}}{\hat{\Delta}_*}.$$

ここで, $\hat{\Delta}_* = \|\overline{\boldsymbol{x}}_{1n_1} - \overline{\boldsymbol{x}}_{2n_2}\|^2$, $\hat{\kappa} = \mathrm{tr}(\boldsymbol{S}_{1n_1})/n_1 - \mathrm{tr}(\boldsymbol{S}_{2n_2})/n_2$ である．ただし, $\overline{\boldsymbol{x}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ と $\boldsymbol{S}_{in_i} = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})(\boldsymbol{x}_{ij} - \overline{\boldsymbol{x}}_{in_i})^T/(n_i - 1)$.

定理 2 (Nakayama et al., 2016)**.** (A-i) と (A-ii) を仮定する．$p \to \infty$ のとき判別関数 $y_{BC}(\boldsymbol{x}_0)$ について (1) が成り立つ．

BC-LSVM は (A-iii) を仮定せずに一致性を示すことができる．

さらに，バイアス補正非線形 SVM も提案し，その判別性能を数値実験と実データ解析を用いて検証した．

参考文献

[1] Aoshima, M. and Yata, K. (2015). High-dimensional quadratic classifiers in non-sparse settings. arXiv:1503.04549.

[2] Nakayama, Y., Yata, K. and Aoshima, M. (2016). Support vector machine and its bias correction in high-dimension, low-sample-size settings. Revised in J. Stat. Plan. Infer.

Generalized Boltzmann machines and activation functions

Shinto Eguchi
Institute of Statistical Mathematics, Japan

1 Introduction

Recently the neural network has been revived and attracted lots of interests toward a direction to deep multilayer networks in machine learning, see LeCun et al. (2015). Boltzmann machines are highly explored and focused on the general architecture in a paradigm of deep learning. In particular, restricted Boltzmann machines present a simple understanding for the conditional distributions with logistic sigmoidal functions. On the other hand, several activation functions including the softplus, leaky rectified linear and max functions rather than the logistic sigmoidal function contribute to efficient learning for network parameters. In principle, Boltzmann machine is based on Boltzmann distributions which is extended to an exponential family, cf. Welling (2004) for exponential family harmoniums.This is closely connected with the classical statistics that is formalized by the notion of sufficiency, likelihood, invariance, unbiasedness, efficiency and so forth under the assumption of the exponential family.

We present a general framework for Boltzmann machines associated with an activation function, called the generalized Boltzmann machine (GBM). For this a generalized mean rather than the linear form as in the arithmetic mean is introduced employing the activation function. We will discuss to combine energy functions E_ℓ for $\ell, 1 \le \ell \le L$ by the generalized mean

$$\frac{1}{\tau}\phi^{-1}\Big(\frac{1}{L}\sum_{\ell=1}^{L}\phi(\tau E_\ell)\Big) \tag{1}$$

via the activation function ϕ with the inverse temperature τ. We note that if ϕ is an identity function on $\mathbb{R}$, then (1) is nothing but the arithmetic mean; if $\phi = \exp$, then (1) equals the log-sum-exponential function. We show that a generator function defines a generalized exponential family. As a result, we give a probabilistic understanding for the GBM including the maxout neural network (Goodfellow et al. 2013) if we view the generator function as an activation function in a neural network.

2 Generalized Boltzmann machine

We generalize both the Boltzmann distribuiona nd the log-likelihood hunction as follows. Let ϕ be a strictly increasing on $\mathbb{R}$ and E an energy function of an input vector x. Then the generalized Boltzmann distribution is given by

$$f^{(\phi)}(x,\theta) = \frac{1}{z_\theta}\phi\big(-\tau E(x,\theta)\big) \tag{2}$$

with a scale parameter τ and $z_\theta = \sum_x \phi(-\tau E(x,\theta))$. On the other hand, the generalized loss function for a given data set $\mathcal{D}$ is

$$L^{(\phi)}(\theta,\mathcal{D}) = -\sum_{\mathcal{D}}\frac{1}{\tau}\phi^{-1}\big(f^{(\phi)}(x,\theta)\big) + n\Psi_\tau(\theta), \tag{3}$$

where $\Psi_\tau(\theta) = \frac{1}{\tau}\sum_x \Phi\big(\phi^{-1}(f^{(\phi)}(x,\theta))\big)$ with $\Phi(s)$ being $\int_{-\infty}^{s}\phi(t)dt$. We see that the generalization is just to replace the pair $(\exp, \log)$ with (ϕ, ϕ^{-1}), or if $\phi = \exp$, then $f^{(\phi)}(x,\theta)$ is reduced to a Boltzman distribution and $L^{(\phi)}(\theta, \mathcal{D})$ equals the minus log-likelihood function up to a constant.

We next consider an energy function of a visible variable x and hidden variable h such that $E(x,h,\theta) = -b^\top x - c^\top h - x^\top W h$, where $\theta = (b, c, W)$. In fact $E(x,h,\theta)$ is a restricted version of energy function that is connected only between the visible and hidden components. The generalized Boltzmann distribution is given by $f^{(\phi)}(x,h,\theta) = \phi(-\tau E(x,h,\theta))/z_\theta$, which has the marginal distribution for the visible variable,

$$f^{(\phi)}(x,\theta) = \frac{1}{z_\theta}\sum_h \phi(-\tau E(x,h,\theta)). \tag{4}$$

Basically we can apply the maximum likelihood for this model (4), however we adopt the generalized estimation method with the loss function

$$L^{(\phi)}(\theta,\mathcal{D}) = -\sum_{x\in\mathcal{D}}\frac{1}{\tau}\phi^{-1}\Big(f^{(\phi)}(x,\theta)\Big) + n\Psi_\tau(\theta). \tag{5}$$

The gradient is given by a weighted sum of estimating functions as follows.

Theorem 1 *Let $L^{(\phi)}(\theta,\mathcal{D})$ be the objective function defined in* (5)*. Then, the gradient is written by*

$$\begin{aligned}\frac{\partial}{\partial\theta}L^{(\phi)}(\theta,\mathcal{D}) = &\sum_{x\in\mathcal{D}}\sum_h w^{(\phi)}(x,h,\theta)\Big\{\frac{\partial}{\partial\theta}\frac{\phi(-\tau E(x,h,\theta))}{z_\theta}\Big\}\\ &-n\mathbb{E}_{f^{(\phi)}(x,\theta)}\Big[\sum_h w^{(\phi)}(x,h,\theta)\Big\{\frac{\partial}{\partial\theta}\frac{\phi(-\tau E(x,h,\theta))}{z_\theta}\Big\}\Big]\end{aligned} \tag{6}$$

where

$$w^{(\phi)}(x,h,\theta) = \frac{1}{\phi'\left(\phi^{-1}\left(\sum_h \phi(-\tau E(x,h,\theta))/z_\theta\right)\right)}. \tag{7}$$

We can consider a deep architecture of restricted Boltzmann machines connecting among the input vector x and hidden vectors $h_1, \cdots, h_L$ with energy functions.

3 Discussion

The recent developments in deep learning will open a new paradigm for data sciences, however, the understanding to allow uncertainty is not fully elucidated. We discuss to extend Boltzmann machine to permit various activation functions in a probabilistic argument. This implies to escape from the standard idea by the exponential family and log-likelihood function. However any extension satisfy a kind of minimax principle. In effect it is supported if we consider the generalized entropy and divergence. It is necessary to proceed further investigations in maximum entropy principle.

References

[1] Goodfellow, D. Warde-Farley, M. Mirza, A. Courville & Y. Bengio (2013). "Maxout Networks" (http://arxiv.org/abs/1302.4389). 30th International Conference on Machine Learning.

[2] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[3] Welling, Max, Michal Rosen-Zvi, and Geoffrey E. Hinton. "Exponential family harmoniums with an application to information retrieval." Advances in neural information processing systems. 2004

パラメータ転移学習の理論解析

熊谷亘

神奈川大学 工学部

本研究では, パラメータ転移アプローチを用いた転移学習アルゴリズムについて考察する．特に, パラメトリックな特徴写像に対し，局所安定性とパラメータ転移学習性という新たな概念を導入することで，パラメータ転移アルゴリズムの学習バウンドを導出できることを示す．

従来の機械学習では, データは単一の分布から独立同一に発生すると仮定されている．しかし, この仮定は実際の応用では必ずしも成立しない．そのため, 異なる分布から発生したサンプルを扱うことができる方法を発展させることが重要であると考えられる．このとき，転移学習はこれらの状況に対処するための一般的な方法を提供する．転移学習では典型的に, 目的のタスクに関連する少数のサンプルに加えて, 他のドメインから発生した豊富なサンプルが利用可能であることが想定されている．そのとき, 転移学習の目的は，他のドメインのデータから有用な知識を抽出し，それを用いて目的のタスクに対するアルゴリズムの性能を向上させることである．転移される知識の種類に応じて, 転移学習の問題を解決するためのアプローチは, インスタンス転移, 特徴転移, パラメータ転移などに分類することができる，本研究では, ある種のパラメトリックモデルが想定され，転移された知識はパラメータにエンコードされるようなパラメータ転移アプローチについて考察する．本研究の目的は, パラメータ転移アプローチに基づくアルゴリズムに対して，理論的解析を行うことである．

以下では，パラメータ転移アプローチについて問題設定を述べる．はじめに，幾つかの記法を簡単に導入する．まず，$\mathcal{X}$ と $\mathcal{Y}$ はそれぞれ，サンプル空間とラベル空間であるとする．ラベル付きサンプルの空間 $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ と，その上の結合分布 $P(\mathbf{x}, y)$ の組みを領域と呼ぶ．そのとき，サンプル空間 $\mathcal{X}$ と $\mathcal{X}$ 上の周辺分布 $P(\mathbf{x})$ の組をドメインと呼び，また，ラベル集合 $\mathcal{Y}$ と 条件付き分布 $P(y|\mathbf{x})$ の組みをタスクと呼ぶ．さらに， $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ を仮説空間とし，$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ を損失関数とする．そのとき, 期待リスクと経験リスクを $\mathcal{R}(h) := \mathbb{E}_{(\mathbf{x},y)\sim P}\left[\ell(y, h(\mathbf{x}))\right]$ と $\widehat{\mathcal{R}}_n(h) := \frac{1}{n}\sum_{j=1}^{n} \ell(y_j, h(\mathbf{x}_j))$ によって定める．転移学習の設定において，目標ドメインと呼ばれる，興味のあるドメインから生成されるサンプルに加え，元ドメインと呼ばれる，別のドメインから生成されるサンプルが利用可能であると仮定される．本研究では，目標ドメインと元ドメインを区別するために，$P_{\mathcal{T}}$ や $\mathcal{R}_{\mathcal{S}}$ のように，$\mathcal{T}$ もしくは $\mathcal{S}$ で表される添字を付すこととする．

以下，目標ドメインでは $\mathcal{Y}_{\mathcal{T}} \subset \mathbb{R}$ と仮定し, 目標ドメインのパラメトリックな特徴写像 $\psi_{\boldsymbol{\theta}}$:

$\mathcal{X}_{\mathcal{T}} \to \mathbb{R}^m$ を用いて，仮説 $h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}} : \mathcal{X}_{\mathcal{T}} \to \mathcal{Y}_{\mathcal{T}}$ が次のように表されると仮定する:

$$h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}}(\mathbf{x}) := \langle \mathbf{w}, \psi_{\boldsymbol{\theta}}(\mathbf{x}) \rangle. \tag{1}$$

ここで，パラメータは $\boldsymbol{\theta} \in \Theta$ と $\mathbf{w} \in \mathcal{W}_{\mathcal{T}}$ とし，Θ はノルム $\|\cdot\|$ が付随したノルム空間の部分集合，$\mathcal{W}_{\mathcal{T}}$ は $\mathbb{R}^m$ の部分集合とする．以下では, 単純に $\mathcal{R}_{\mathcal{T}}(h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}})$ および $\widehat{\mathcal{R}}_{\mathcal{T}}(h_{\mathcal{T},\boldsymbol{\theta},\mathbf{w}})$ を $\mathcal{R}_{\mathcal{T}}(\boldsymbol{\theta},\mathbf{w})$ および $\widehat{\mathcal{R}}_{\mathcal{T}}(\boldsymbol{\theta},\mathbf{w})$ のように記述する．元ドメインには, サンプル分布 $P_{\mathcal{S},\boldsymbol{\theta},\mathbf{w}}$ や仮説 $h_{\mathcal{S},\boldsymbol{\theta},\mathbf{w}}$ などのパラメトリックモデルが存在するとし，パラメータ空間の一部 Θ は元ドメインと目標ドメインと共有されていると仮定する．そのとき, $\boldsymbol{\theta}^*_{\mathcal{S}} \in \Theta$ と $\mathbf{w}^*_{\mathcal{S}} \in \mathcal{W}_{\mathcal{S}}$ は元領域において何らかの指標に関して有効なパラメータであるする．しかしながら, 本研究では $\boldsymbol{\theta}^*_{\mathcal{S}} \in \Theta$ と $w^*_{\mathcal{S}} \in \mathcal{W}_{\mathcal{S}}$ に対して明示的な仮定は課さない．

次に，本研究で扱うパラメータ転移アルゴリズムについて説明する．元ドメインと目標ドメインでそれぞれ N 個と n 個のサンプルを使用できるとする．パラメータ転移アルゴリズムは，まず N 個のサンプルを使用して，$\boldsymbol{\theta}^*_{\mathcal{S}}$ の推定値 $\widehat{\boldsymbol{\theta}}_N \in \Theta$ を出力する．次にアルゴリズムは目標ドメインのパラメータ

$$\mathbf{w}^*_{\mathcal{T}} \quad := \quad \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_{\mathcal{T}}} \mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}\right)$$

に対し，n 個のサンプルを用いて推定値

$$\widehat{\mathbf{w}}_{N,n} \quad := \quad \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}_{\mathcal{T}}} \widehat{R}_{\mathcal{T},n}(\widehat{\boldsymbol{\theta}}_N, \mathbf{w}) + \rho r(\mathbf{w})$$

を出力する．ここで $r(\mathbf{w})$ は $\|\cdot\|_2$ に関して 1-強凸な正則化項とし，ρ は正の実数とする. 元ドメインが何らかの意味で目標ドメインに関係している場合，元ドメインでの有効なパラメータ $\boldsymbol{\theta}^*_{\mathcal{S}}$ も目標タスクにとって有用であると期待される．$\mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}\right)$ を予測性能の基準値として採用し，学習バウンドを導き出す．局所安定性と転移学習可能性の他に，幾つかの技術的仮定をおくことで，以下の学習バウンドが成り立つ．

定理 1 (学習バウンド)**.** パラメトリック特徴写像 $\psi_{\boldsymbol{\theta}}$ が局所安定であるとする．また，元ドメインで学習された $\boldsymbol{\theta}^*_{\mathcal{S}} \in \Theta$ の推定量 $\widehat{\boldsymbol{\theta}}_N$ は確率 $1 - \bar{\delta}$ でパラメータ転移学習可能性を満たすとする．ここで， ρ を適切に設定するとき, 以下の不等式が確率 $1 - (\delta + 2\bar{\delta})$ で成り立つ :

$$\mathcal{R}_{\mathcal{T}}\left(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}\right) - \mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}\right) \quad \leq \quad C_1 \frac{1}{\sqrt{n}} + C_2 \left\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\right\| + C_3 n^{\frac{1}{4}} \sqrt{\left\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\right\|}. \tag{2}$$

ここで C_1, C_2, C_3 は定数である．

サンプル数 N を用いて，推定誤差 $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\|$ を評価することができるならば, 定理 1により, どの項が支配的であるかがわかる．特に, 目標ドメインのサンプルと比較して，元ドメインのサンプルがどの程度あれば十分かを評価することができる．

大規模な線形予測器のための非同期特徴抽出スキーム

東京大学　情報理工学系研究科　松島 慎

平成 29 年 1 月 27 日

本発表では $\mathbf{w} \in \mathbb{R}^p$ をパラメータとする以下の関数を最小化することを考える：

$$P(\mathbf{w}) \triangleq \|\mathbf{w}\|_1 + C \sum_{i=1}^{n} \ell(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle, y_i). \tag{1}$$

ここで $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ は微分可能な凸関数とし、$\mathbf{x}_i \in \mathcal{X}$ と $y_i \in \mathcal{Y}$ は学習に用いる入力と出力とする。また、$\phi : \mathcal{X} \to \mathbb{R}^p$ はデータの特徴を記述する関数である。この最小化問題は機械学習の分野で L1 正則化項付きの経験誤差最小化問題と呼ばれる一般的な枠組みで現れ、分類問題や回帰問題だけでなく、構造予測など複雑な出力を入力から予測する場合にも多く用いられる [4]。

一般にサンプル数 n が増大するにつれ、ϕ として複雑かつ高次元な関数を用いても過学習が起こらず、結果として性能の高い予測器を得ることが可能になる。近年ニューラルネットを用いて表現されるような複雑な予測器を用いて、性能が高い予測が可能になったのも、使用可能なサンプル数が増大したことが一つの要因になっている。このことからも線形予測器においても高い予測器を得るためには、ϕ として複雑かつ高次元な関数を用いる必要があると考えられる。

標準的に最適化問題 (1) を解く場合、$\Phi \triangleq \{\phi_j(\mathbf{x}_i)\}_{1 \le i \le n, 1 \le j \le p}$ の計算と実際の求解の段階は分離されている。そのため非常に大きな数の特徴を考えることで Φ のデータ量が計算機のメモリ容量を超す場合に、最適化を効率的に行うことができない問題がある。すなわち、典型的には勾配を求める場合などに Φ の値を読み込む際に、低次の記憶領域にアクセスしなければならず、実際の計算時間が大きくなってしまう。

L1 正則化を用いる上記の問題の場合、最適化問題に必要ない特徴が多く存在し、最適解として疎な解を得られることが期待される。すなわち、予測器が予測に必要な特徴は特徴関数のうち一部の関数のみであり、L1 正則化の影響でこれらに対応するパラメータが零化されることが期待される。そのような場合、結果的に零化される特徴を考えない問題を解くことで実質的に同じ予測器を得ることができる。実際に不必要と推測される特徴を無視して経験的に学習の効率を上げる方法が提案されている。このような方法を利用すれば、必要な特徴のみを展開しメモリに載せることで、上記の問題を回避することができると考えられる。

メモリ容量より大きなデータを用いて正則化付き経験誤差最小化問題を解く最初のスキームとして、Yu et al. のブロック最小化スキーム [5] が挙げられる。ブロック最小化スキームや Matsushima et al. の Dual Cached Loops[2] は確率的勾配法を用いることによって (1) の最小化問題を解くことが可能である。しかし、確率的勾配法を用いる場合、最適化問題に必要ない特徴を除外するなど、利用することができない。

本発表では、Φ がメモリ容量より大きいような場合に最適化問題 (1) を効率的に解くためのスキームを提案する。本スキームでは二種類の異なるアルゴリズムを非同期的に動作させることで特徴の選択的抽出とパラメータの最適化を同時に行う。本スキームは任意の基底関数について学習することができるスキームであるが、基底関数が持つ構造を用いて効率的な学習を行うことも可能である。本スキームでは書き込みスレッドと訓練スレッドと呼ばれる二つのスレッドがパラメータ $\mathbf{w}$、補助パラメータ $\mathbf{u}$ と特徴キャッシュΦ_J を共有し、非同期的に動作する。書き込みスレッドはデータから特徴 ϕ_j を読み込み、これを特徴キャッシュに追加する。ただし、抽出された特徴が現在の解を改良することができなければ、特徴キャッシュに追加し訓練スレッドと共有することなく、展開された特徴は破棄される。すなわち特徴を破棄するための条件は

$$-1 < C\nabla_j L(\mathbf{w}^t) < 1. \tag{2}$$

と記述することができる。この値は前述のように補助パラメータ $\mathbf{u}^t$ を用いて簡単に計算することができる。もし特徴キャッシュの容量を超過するようであれば、無作為にキャッシュ内の特徴を破棄し、解放された領域に特徴を追加する。

一方で、訓練スレッドは特徴キャッシュ内に格納されている列に対応する座標を無作為に一つ選び、座標降下法を用いてパラメータ更新を行う。選んだ座標に関して (2) かつ $w_j = 0$ が成り立つ場合、対応する列は特徴キャッシュから破棄することで座標降下法の効率を上げることを目指す。この条件は前節で説明したヒューリスティクスに比べより多くの列を不必要とみなすため、実際には必要だった列を破棄する機会がより多くなると考えられる。しかし書き込みスレッドが同時に各列の必要性を確認しているため、より積極的に列を除外することが全体の効率を上げると考えられる。

一般に、書き込みスレッドは特徴を展開することに大部分の時間を費やす一方で、訓練スレッドはパラメータの更新に大部分の時間を費やすため、特徴キャッシュへのアクセスにおける衝突やそれに伴うスレッドの待機はほとんど生じないと考えられる。訓練スレッドに関しては非同期的な座標降下法 [1, 3] を用いて複数のスレッドを利用することができる。また書き込みスレッドは自然に同時に複数のスレッドを利用することができる。さらに特徴空間の構造を利用することにより、より効率的な特徴抽出が可能であると考えられる。

参考文献

[1] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16(285-322):1–5, 2015.

[2] S. Matsushima, S.V.N. Vishwanathan, and A. J. Smola. Linear support vector machines via dual cached loops. In *Proceedings of Knowledge Discovery and Data Mining*, pages 177–185, 2012.

[3] Z. Peng, Y. Xu, M. Yan, and W. Yin. Arock: an algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing*, 38(5):A2851–A2879, 2016.

[4] I. Rish and G. Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC Press, Inc., 2014.

[5] H.-F. Yu, C.-J. Hsieh, K.-W. Chang, and C.-J. Lin. Large linear classification when data cannot fit in memory. In *Proceedings of Knowledge Discovery and Data Mining*, pages 833–842, 2010.

行列式点過程の準モンテカルロ積分への応用

平尾 将剛（愛知県立大学）*

1 はじめに

行列点過程 (determinantal point process, DPP) は，1970 年代中頃に Macchi [4] によって量子力学的粒子のひとつであるフェルミ粒子をモデル化するために提案された確率点過程である．この点過程は応用上よく用いられる点過程のひとつであるポアソン点過程とは異なり，反発力がある粒子系を記述することができ，近年盛んに研究されている対象である．

本講演では，近年の行列式点過程を用いた数値積分法への応用について，自身の研究 [3] における結果内容を踏まえ報告を行なった．特に**球面アンサンブル**，**ハーモニックアンサンブル**と呼ばれる球面上の行列式点過程について着目し，後述するソボレフ空間に対する**最悪誤差**の評価を与えた．またその結果，球面アンサンブルは漸近的に最適な収束レートを達成する点列を与えること (準モンテカルロデザイン系列を漸近的に与えること)，及びハーモニックアンサンブルは通常のモンテカルロ法より速い収束レートを達成する点列を与えることを報告した．これらの主結果については次節で必要事項について準備したのちに，3 節においてその詳細を述べる．

2 準備

主結果を紹介するため，ここでは幾つかの準備を与える．通常，準モンテカルロ法では，積分領域が d 次元立方体 $[0,1]^d$ のものを考えるが，ここでは d 次元球面 $\mathbb{S}^d$ のものを考える．Brauchart et al. [2] により，近年，球面上の積分に対して**準モンテカルロデザイン系列** (QMC design sequence) の概念が導入された．これは球面上の積分に関して，より高速な誤差の収束を実現する球面上の点列を構成する問題である．[2] では，特に球面上に制限された滑らかさ s の Sobolev 空間 $\mathbb{H}^s(\mathbb{S}^d)$ において準モンテカルロデザイン系列を定義し，その考察を行っている．

次節で与える主結果の証明において重要なのは，Sobolev 空間 $\mathbb{H}^s(\mathbb{S}^d)$ は再生核ヒルベルト空間であることである．Brauchart et al. [2] では，内積を適切に選んだ場合，例えば，$d/2 < s < d/2+1$ に対して，その再生核は

$$K^{(s)}(\boldsymbol{x},\boldsymbol{y}) := 2V_{d-2s}(\mathbb{S}^2) - |\boldsymbol{x}-\boldsymbol{y}|^{2s-d}, \quad \boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^d.$$

と表されることが知られている．また，$s > d/2+1$ においても似た様な表現が得られることが知られている (Brauchart-Womersly [1] や [2] を参照)．

Sobolev 空間 $\mathbb{H}^s(\mathbb{S}^d)$ 上の準モンテカルロデザイン系列を定義するために，球面上の積分 $I(f) := \int_{\mathbb{S}^d} f(\boldsymbol{x})\, d\sigma_d(\boldsymbol{x})$ と N 点集合 $X_N \subset \mathbb{S}^d$ での近似 $Q[X_N](f) := \frac{1}{N}\sum_{\boldsymbol{x}\in X_N} f(\boldsymbol{x})$ とにおける最悪誤差 (worst-case error) を次で定義する．

$$\mathrm{wce}(Q[X_N]; \mathbb{H}^s(\mathbb{S}^d)) := \sup_{\substack{f\in\mathbb{H}^s(\mathbb{S}^d) \\ \|f\|_{\mathbb{H}^s}\le 1}} |Q[X_N](f) - I(f)|. \tag{1}$$

このとき，$\mathbb{H}^s(\mathbb{S}^d)$ 上の準モンテカルロデザイン系列は次で定義される．

*博士 (情報科学) 愛知県立大学 情報科学部（〒 480-1198 愛知県長久手市茨ケ廻間 1522-3, E-mail: hirao@ist.aichi-pu.ac.jp）

Definition 2.1 ($\mathbb{H}^s(\mathbb{S}^d)$ 上の準モンテカルロデザイン系列, [2])**.** $s > d/2$ とし，$\{X_N\}$ を $\mathbb{S}^d$ 上の点集合の増大列とする．このとき，N に依存しない定数 $c(s,d) > 0$ が存在し，

$$\mathrm{wce}(Q[X_N]; \mathbb{H}^s(\mathbb{S}^d)) \leq \frac{c(s,d)}{N^{s/d}}. \tag{2}$$

が成り立つとき，増大列 (X_N) を $\mathbb{H}^s(\mathbb{S}^d)$ 上の準モンテカルロデザイン系列であると言う．

特に上の再生核を用いれば，$d/2 < s < d/2+1$ における最悪誤差は，

$$\{\mathrm{wce}(Q[X_N]; \mathbb{H}^s(\mathbb{S}^d))\}^2 = V_{d-2s}(\mathbb{S}^d) - \frac{1}{N^2}\sum_{i\neq j}|\boldsymbol{x}_j - \boldsymbol{x}_i|^{2s-d}$$

で与えられることは重要である．主定理ではこの最悪誤差の評価を行なう．

3　主結果

ここで主定理について述べる．我々は前述した 2 種類の球面上の行列式点過程を用いた際の最悪誤差の評価を与える．最初に $\mathbb{S}^2$ 上の球面アンサンブルを用いた場合である．

Theorem 3.1 ([3])**.** $1 < s < 2$ とする．このとき，N 点球面アンサンブル $\mathcal{X}_N$ に対して，

$$E\left[\mathrm{wce}(Q[\mathcal{X}_N]; \mathbb{H}^s(\mathbb{S}^2))^2\right] = 2^{2s-2}B(s,N) \tag{3}$$

が成り立つ．ここで $B(s,N)$ はベータ関数である．

ここでスターリングの公式を用いると，固定した s に対して，$B(s,N) \sim \Gamma(s)N^{-s}$ $(N \to \infty)$ であることが分かる．したがって，十分大きな N に対して (3) は平均的に (2) を満たしていることが分かる．すなわち，驚くべきことに球面アンサンブルは漸近的に $\mathbb{H}^s(\mathbb{S}^2)$ に対する準モンテカルロデザイン系列を与えるのである．

さらに一般次元の球面 $\mathbb{S}^d$ に関しては，ハーモニックアンサンブルを用いると次の評価式を得ることができる．この定理は通常のモンテカルロ法より，速いオーダーで最悪誤差が収束していることを述べている．

Theorem 3.2 ([3])**.** $\mathcal{X}_N$ を $N = \dim(\mathcal{P}_L(\mathbb{S}^d))$ 点のハーモニックアンサンブルであるとする．このとき，$d/2+1/2 < s < d/2+1$ に対して，N とは独立な定数 $C(s,d) > 0$ が存在し，

$$\mathbb{E}[\{\mathrm{wce}(Q[\mathcal{X}_N]; \mathbb{H}^s(\mathbb{S}^d))\}^2] \leq \frac{C(s,d)}{N^{1+1/(2d)}}$$

を満たす．

謝辞　本研究は，科学研究費補助金（若手研究 (B)，課題番号 16K17645）の助成を受けている．

References

[1] J.S. Brauchart, R.S. Womersley. Numerical integration over the unit sphere, $\mathbb{L}_2$-discrepancy and sum of distances. In preparation.

[2] J.S. Brauchart, E.B. Saff, I.H. Sloan, R.S. Womersley. QMC Designs: optimal order quasi-Monte Carlo integration schemes on the sphere. *Math. Comp.*, 83(290): 2821–2851, 2014.

[3] M. Hirao. QMC designs and determinantal point processes. Submitted to MCQMC2016 conference proceedings.

[4] O. Macchi. The coincidence approach to stochastic point processes, *Adv. Appl. Prob.*, 7: 83–122, 1975.

Cell Regression and Reference Priors

Jinfang Wang and Shigetoshi Hosaka

Chiba University and Hosaka Clinic of Internal Medicine

1. Introduction Many data collected by survey agencies, such as government offices, are published in table forms. A large portion of these tables show cell frequencies for categorized continuous variables. In this paper, we shall refer to such coarse tables as *reference tables*. *Cell regression* methods proposed in this paper only explore the information on these cell frequencies and the intervals defining the cells. We use parametric models to predict the cell probabilities, and estimate the regression parameters by minimizing the Kullback-Leibler divergence from the reference distribution to the predictive distribution. Bayesian extensions are also considered. We propose an approximate Markov chain Monte Carlo algorithm to compute the posterior distribution of the parameters. The posterior distribution so obtained may be used as prior distribution in a second stage analysis based on more detailed data. For this reason, we call the posterior distribution the *reference prior*. Indeed, the primary motivation of cell regression is to *mining* prior information embedded in often large but coarse tables available with low costs.

2. Predictive Distributions We assume that both x_i and y_i are continuous, and are interval-censored. The exact values of x_i and y_i are not known. We only know the cell membership for each pair (x_i, y_i), for $i = 1, \cdots, R$.

Definition 1 (Predictive Cell Distributions). *Assume that the following cell regression model holds.*

$$y_i \,|(x_i, \boldsymbol{\theta}) \sim f(y_i|x_i, \boldsymbol{\theta}) \tag{1}$$

$$x_i \sim f(x_i) \tag{2}$$

where both $f(y_i|x_i, \boldsymbol{\theta})$ and $f(x_i)$ are known up to an unknown parameter $\boldsymbol{\theta}$. A predictive cell distribution $p_{ij}(\boldsymbol{\theta})$ is defined by

$$p_{ij}(\boldsymbol{\theta}) = \int_{c_{j-1}}^{c_j} \left\{ \int_{d_{i-1}}^{d_i} f(x, y|\boldsymbol{\theta})\, dy \right\} dx \qquad (\boldsymbol{\theta} \in \Theta)\,. \tag{3}$$

Let R_{ij} be cell frequencies, $r_{ij} = R_{ij}/R$ and $R = \sum R_{ij}$. Let $\mathcal{P}$ be the predictive distribution defined by (3). The goodness of $\mathcal{P}$ is measured through the *Kullback-Leibler divergence* from the reference distribution $\mathcal{R}$ to $\mathcal{P}$,

$$\mathrm{KL}(\mathcal{R}||\mathcal{P}) = \sum_{i,j} r_{ij} \log \frac{r_{ij}}{p_{ij}} \tag{4}$$

We define the minimum contrast estimator $\widehat{\boldsymbol{\theta}}$ to be the solution to the following constrained minimization problem,

$$\min_{\boldsymbol{\theta}} \quad \mathrm{KL}(\mathcal{R}||\mathcal{P}) \qquad \text{subject to} \quad \sum_{i,j} p_{ij}(\boldsymbol{\theta}) = 1 \tag{5}$$

3. Reference Priors Bayesian random effects models are often used in deriving the posterior predictive distribution for a particular subject based on repeated measurements for each subject. The cell regression analyses come into play when we want to replace the assumption on the prior distribution of the fixed parameter. We propose to use the posterior distribution of the corresponding parameters derived from a Bayesian cell regression analyses as the prior distribution for the fixed parameter. We consider the following Bayesian cell regression model.

$$\begin{aligned} y_i\,|(x_i, \boldsymbol{\theta}) &\sim f(y_i|x_i, \boldsymbol{\theta}) && (6)\\ x_i &\sim f(x_i) && (7)\\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) && (8) \end{aligned}$$

We define the Bayesian predictive cell distribution $p^b_{ij}(\boldsymbol{\theta})$ by

$$p^b_{ij}(\boldsymbol{\theta}) = \int_{c_{j-1}}^{c_j} \left\{ \int_{d_{i-1}}^{d_i} f(x,y)\, dy \right\} dx \qquad (\boldsymbol{\theta} \in \Theta) \tag{9}$$

It is easily seen that

$$p^b_{ij}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) p_{ij}(\boldsymbol{\theta}) \tag{10}$$

Denote by $\mathcal{P}^b$ the Bayesian predictive distribution. Let

$$\widehat{\boldsymbol{\theta}}^b = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\mathcal{R}||\mathcal{P}^b) = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\mathcal{R}||\mathcal{P}) - log\ \pi(\boldsymbol{\theta}) \tag{11}$$

The following algorithm computes the posterior distribution for $\boldsymbol{\theta}$.

(i) Compute $\widehat{\boldsymbol{\theta}}^b$ of (11).

(ii) For the (i,j) cell, generate R_{ij} independent pairs of (x,y) according to

$$\begin{aligned} x &\sim f(x) && (12)\\ y\,|(x, \boldsymbol{\theta}) &\sim f(y|x, \widehat{\boldsymbol{\theta}}^b) && (13) \end{aligned}$$

And do this for all cells to generate R independent samples

$$(x_1, y_1), \cdots, (x_R, y_R) \tag{14}$$

(iii) Apply a typical Markov chain Monte Carlo method to compute the posterior distribution of $\boldsymbol{\theta}$ based on the approximate sample (14) and the Bayesian cell regression model

$$\begin{aligned} y_i\,|(x_i, \boldsymbol{\theta}) &\sim f(y_i|x_i, \boldsymbol{\theta}) && (15)\\ x_i &\sim f(x_i) && (16)\\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) && (17) \end{aligned}$$

Euclidean Design Theory

澤正憲
神戸大学大学院システム情報学研究科
sawa@people.kobe-u.ac.jp

関数の積分値を，定義域の有限個の点 (ノード) での関数値の重み付き平均で近似する公式を，cubature 公式 (cubature formula) という．特に，次数 t 以下の任意の多項式 f について，積分値と f の重み付き平均値の誤差が 0 になるとき，次数 t の公式 (cubature of degree t) という．

Simpson や Gauss の名を冠する公式があるように，少ないノードからなり，かつ次数の大きな公式の存在および構成問題は，直交多項式の零点の解析と並行して，数値解析，特殊関数論などの分野を中心に古くから調べられてきた．実は，そのような cubature 公式は

- 実・複素幾何における球面デザイン (spherical design),
- 微分幾何学における大対蹠集合 (great antipodal set),
- 離散幾何学における距離集合 (distance set),
- 関数解析学におけるバナッハ空間のノルム不変線形作用素 (isometric embedding),
- 高次形式論におけるヒルベルト恒等式 (Hilbert identity),
- 整数論のテータ級数に関する Lehmer 予想 (Lehmer conjecture),
- 量子情報理論における mutually unbiased basis (MUB),
- 統計的実験計画法における D 最適計画 (D-optimal design),

などの異分野の対象と深く結び付いているのだが，その多くは個々の分野で独自に調べられてきた経緯があり，最近でも，双方にとって有益な研究成果が互いに認識されていないというケースも少なくない．

そこで，本研究では，上述の諸研究対象を統一的に扱うべく，被積分関数のクラスを一般の関数空間に押し上げて，測度空間上の cubature 公式 (cubature on measure space) の概念を導入した．そして，以下の主成果を得た：

(1) 多項式型 cubature 公式のノードの個数に関する Stroud 不等式 (Stroud bound) の一般化．特に BIB デザインのブロック数に関する Fisher 不等式の解析的別証明の提示．

(2) (1) の不等式においてタイト (tight) な公式の再生核による特徴付け．

(3) 多項式型 cubature 公式の漸近存在定理として知られている Tchakaloff の定理 (Tchakaloff Theorem) の一般化．特に位相幾何学的な諸条件下での Chebyshev 型の cubature 公式，すなわち「重み付き平均」が「算術平均」であるような公式の漸近存在証明．

(4) 多項式型 cubature 公式に対する Sobolev の不変式論の一般化．

(5) 多項式型 cubature 公式の逐次的構成法の一般化．

このように,「測度空間上の cubature 公式論」では，多項式型の cubature 公式に関する諸概念・諸定理を一般化して，異分野の研究対象を網羅的に扱うための理論的な枠組みを提供する．また，統計的には，BIB デザインの Fisher 不等式の解析的別証明 (主結果 (1))，直交配列の「直交性」の解析的な解釈，そして D 最適計画の構成の提示など，主に実験計画法への応用に期待感がある．

本研究は拙論文 [2] に基づいている．なお，多項式型の cubature 公式の理論と応用については，Dunkl-Xu [1], Sobolev-Vaskevich [3], Stroud [4] などに詳しく書かれているので，そちらを参照されるとよい．そのうち，測度空間上の cubature 公式論の統計学者向けの書「Euclidean Design Theory」が Springer Briefs から出版される予定なので，そちらも手にとっていただけると嬉しく思います．

参考文献

[1] C.F. Dunkl, Y. Xu. *Orthogonal Polynomials of Several Variables.* Encyclopedia of Mathematics and its Applications, 81. Cambridge University Press, 2001.

[2] M. Sawa. Cubature 公式の理論．数学 **62** (2016), 24-53.

[3] S.L. Sobolev, V.L. Vaskevich. *The Theory of Cubature Formulas.* Mathematics and its Applications, 415. Kluwer Academic Publishers Group, Dordrecht, 1997.

[4] A.H. Stroud. *Approximate Calculation of Multiple Integrals.* Prentice-Hall, N.J., 1971.

EXACT VC DIMENSION OF ELLIPSOIDS AND CONSISTENCY OF MAXIMUM LOG-LIKELIHOOD ESTIMATOR FOR MULTIVARIATE GAUSSIAN MIXTURES

YOHJI AKAMA

1. Introduction

The log-likelihood function $L_n(\theta)$ to estimate the parameter θ of a d-dimensional ($d > 1$) K-component Gaussian mixture is unbounded, where n is the size of a sample. This is overcome in Chen and Tan [4], by adding a term $p_n(\theta)$ to $L_n(\theta)$. This penalizes small variances or ratios of variances to $L_n(\theta)$. Chen and Tan attempted to prove the consistency of their penalized maximum likelihood estimator (MLE), by reducing the argument to a univariate argument combined with a Bernstein inequality and Borel-Cantelli Lemma. But, at some point, they did not uniformly handle the unbounded subsets of $\mathbb{R}^d$, although they should. See [3], for detail. To fix the flaw in Chen and Tan's consistency proof, Alexandrovich slightly modified Chen and Tan's sufficient condition for the penalized MLE to be consistent, and used (1) a *uniform* law of iterated logarithm for classes having finite VC dimensions and (2) that the set of the d-dimensional ellipsoids has finite VC dimension. VC dimension of a class is a combinatorial measure of the complexity of the class. It is difficult to give the exact value, in general. In this note, we give the exact VC-dimension of the d-dimensional ellipsoids (Akama-Irie [1]). Then we review Alexandrovich's approach to consistency proof [3] of Chen and Tan's penalized MLE to Gaussian mixture.

For sets $X \subset \mathbb{R}^d$ and $Y \subset X$, we say that a set $B \subset \mathbb{R}^d$ *cuts* Y *out of* X if $Y = X \cap B$. A class $\mathcal{C}$ of subsets of $\mathbb{R}^d$ is said to *shatter* a set $X \subset \mathbb{R}^d$, if for any $Y \subset X$ there exists $B \in \mathcal{C}$ such that B cuts Y out of X.

The *Vapnik-Chervonenkis dimension* (VC *dimension* for short) of a class $\mathcal{C}$ is

$$\mathrm{VCdim}(\mathcal{C}) := \sup\{\ \sharp S \ \mid\ S \subset \mathbb{R}^d, \mathcal{C} \text{ shatters } S\ \}.$$

By a d-dimensional ellipsoid, we mean a region in $\mathbb{R}^d$ defined as $\{x \in \mathbb{R}^d \mid Q(x-\mu) < 1\}$, where $\mu \in \mathbb{R}^d$ and Q is a positive definite quadratic form. Our main result is:

Theorem 1. *The class of d-dimensional ellipsoids has* VC *dimension* $(d^2+3d)/2$.

2. Alexandrovich's consistency proof of Chen and Tan's penalized MLE

In order to uniformly handle unbounded subsets of $\mathbb{R}^d$ missed in Chen and Tan's consistency proof, Alexandrovich slightly modified the condition C_3, so that he can employ uniform law of iterated logarithm developed in empirical process theory [5].

Condition 1.

$\tilde{C}_3$: $\tilde{p}_n(\Sigma) \le a(n)\log|\Sigma|$ *for* $|\Sigma| \le cn^{-2d}$. *Here c is a constant and* $a(n) = o(n)$.

Here n is the size of a sample, $p_n(\theta) = \sum_{j=1}^K \tilde{p}_n(\Sigma_j)$ is the penalty function for a K-component Gaussian mixture parameter $\theta = (\mu_1, \Sigma_1, p_1, \ldots, \mu_K, \Sigma_K, p_K)$, $\mu_j \in \mathbb{R}^d$, a positive definite $\Sigma_j \in \mathbb{R}^{d\times d}$, and $\sum_{j=1}^K p_j = 1$.

From Chen and Tan's argument [4], Alexandrovich observed that the following condition implies the consistency of Chen and Tan's penalized MLE, if the penalty function satisfies $\tilde{C}_3$ instead of C_3. Condition 2 gives an upper bound of the number of the sample random variables lying in a ellipsoid, asymptotic in the size n of the sample and the "volume" of the ellipsoid.

Condition 2. *Y_n $(n \in \mathbb{N})$ is an i.i.d. process such that*

$$\sum_{i=1}^n 1_{\{(Y_i-\mu)^\top \Sigma^{-1}(Y_i-\mu)\le(\log|\Sigma|)^2\}} \le a(n) + b(n,\, |\Sigma|)$$

where $a(n) = o(n), b(n,s) = O(n), \lim_{s\to 0+} \frac{b(n,s)}{\sqrt{\log s}} = 0$.

Proposition 1 (Alexander [2])**.** *Suppose that $\mathcal{C}$ is a class of Borel subsets of $\mathbb{R}^d$ and $\mathcal{C}$ has a finite VC dimension. Let Y_n $(n \in \mathbb{N})$ be a d-dimensional i.i.d. process. Then a.s.,*

$$\limsup_{n\to\infty} \sup_{C\in\mathcal{C}} \frac{|\sum_{i=1}^n 1_C(Y_i) - nP_{Y_1}(C)|}{\sqrt{2n\log\log n}} = \sup_{C\in\mathcal{C}} \sqrt{P_{Y_1}(C)(1-P_{Y_1}(C))}.$$

By Theorem 1 and Proposition 1 with $\mathcal{C}$ being the set of d-dimensional ellipsoids, Alexandrovich derived:

Corollary 1. *Let $(Y_n)_{n\in\mathbb{N}}$ be a d-dimensional i.i.d. process with a bounded Lebesgue density f, $M := \sup_y f(y)$. Then a.s. there exists a positive integer N such that*

$$\sum_{i=1}^n 1_{\{(Y_i-\mu)^\top \Sigma^{-1}(Y_i-\mu)\le(\log|\Sigma|)^2\}} \le \frac{3}{4}\sqrt{n\log\log n} + nMv_d|\Sigma|^{1/2}(\log|\Sigma|)^d$$

for every $\mu \in \mathbb{R}^d$, every positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, and every $n \ge N$. Here v_d is the volume of the d-dimensional unit ball.

Then this satisfies Condition 2. Alexandrovich's observation explained above implies the consistency of Chen and Tan's penalized MLE under the modified condition $\tilde{C}_3$.

References

[1] Akama, Y. and Irie, K., *VC dimension of ellipsoids*, arXiv:1109.4347 [math.CO]., (2011)

[2] Alexander, K. S., *Probability inequalities for empirical processes and a law of the iterated logarithm*, Annals of Probability **12**, 1041—1067 (1984).

[3] Alexandrovich, G., *A Note on the Article 'Inference for multivariate normal mixtures' by J. Chen and X . Tan.*, J. Multivariate Analysis **129**, 245 — 248, 2014.

[4] Chen, J. and Tan, X., *Inference for multivariate normal mixtures*, Journal of Multivariate Analysis, **100** (2009), 1367–1383.

[5] Dudley, R.M., *Uniform central limit theorems*, Cambridge Studies in Advanced Mathematics, vol. 63, Cambridge University Press, 1999. MR MR1720712 (2000k:60040)

Mathematical Institute, Tohoku University, Sendai, 980-8578, Japan

E-mail address: akama@m.tohoku.ac.jp

入れ子型混合モデルに基づく cancer outlier profile の推定とがん診断への応用 について

松井 孝太 [1], 大浦 智紀 [2], and 松井 茂之 [1]

[1] 名古屋大学大学院医学系研究科 / JST CREST
[2]Eli Lilly Japan

1 はじめに

がん患者の集団における発がん遺伝子の異質性は, 疾患生理学の理解, リスクグループの特定および患者の治療の最適化に重要な意味を有すると考えられている. Tomlin et al. [4] は, 前立腺がんを対象とした研究で , 従来の 2 標本 t-検定のような全がん患者で共通な発現プロファイルを持つ遺伝子を特定しようとする手法では捉えられないプロファイルが存在することを示した. そのような, 一部のがん患者において一部のがん遺伝子が示す特異的な発現プロファイルを outlier expression profile (OEP) と呼び, そのようながん遺伝子のことを cancer outlier と呼ぶ. cancer outlier は, 上述した異質性の代表的な例と考えられ, [4] では OEP を持つ遺伝子の特定する方法として cancer outlier profile analysis (COPA) が開発され, これに基づいて前立腺がん患者のサブタイプが明らかとなった.

Tomlin et al. の研究に触発された COPA の研究が, その後複数のグループによって実施されている. Wu [5] は, 健常者の発現プロファイルを基準として cancer outlier を定義し, これを特定するための新しい検定統計量として ORT 統計量を提案した. Tibshirani and Hastie [3] は, Wu とは異なる cancer outlier の定義に基づいて outlier-sum statistic を用いた検定による cancer outlier のスクリーニングを提案した. Tomlin et al. を含めたこれらの COPA 研究は, それぞれ独自に定義した cancer outlier に対して行われている. すなわち, 健常者のプロファイルからどの程度乖離していれば outlier と見做すのか, というしきい値が研究ごとに異なる. Lian [1] は, これらの問題設定を統一的に評価するため, すべてのしきい値に関して考察を行い, MOST statistic を用いた outlier 遺伝子の検定手法を開発した.

ここまでに紹介した COPA 研究は, すべて cancer outlier として振舞っている遺伝子を仮説検定に基づいて特定するための手法の研究である. しかし, これらの多重検定を用いた outlier 遺伝子のスクリーニングは, cancer outlier analysis の第一歩に過ぎない. 例えば, がん関連遺伝子のあるグループは, 発がん経路で共調節されたり, また増幅された染色体領域や遺伝子融合に関連する染色体領域に共局在する特性を有する. したがって, よく似た OEP を持つがん関連遺伝子をクラスタリングすることで, このような生物学的にも意味のある遺伝子群の特定に繋がる. さらに, そうして特定された遺伝子群を用いた, cancer outlier に基づくがん患者の識別法 (すなわち, がんの診断法) の開発も期待される.

以上のような背景の下で, 本研究ではパラメトリック入れ子型混合モデルに基づいた遺伝子とサンプル両方向のクラスタリング (biclustering) 手法を提案する (Figure 1). 提案法は, cancer outlier を陽にモデル化し, さらに表現型 (がん or 健常) の情報も取り入れた教師付きのクラスタリング手法となっており, EM アルゴリズムを用いてパラメータを推定する. また, 適切な OEP 遺伝子のコンポーネント数を決定するために, 情報量基準に基づくモデル選択を行う. 一度提案法によって Figure 1 のような構造 (outlier 遺伝子の従う分布) が推定されれば, 新たな患者に対して, がんか健常かを確率的に判別するアルゴリズムを構成することができる. 本研究では, outlier 構造の推定からがん患者の判別までの一連のプロシージャを提案法として, シミュレーション及び実データによ

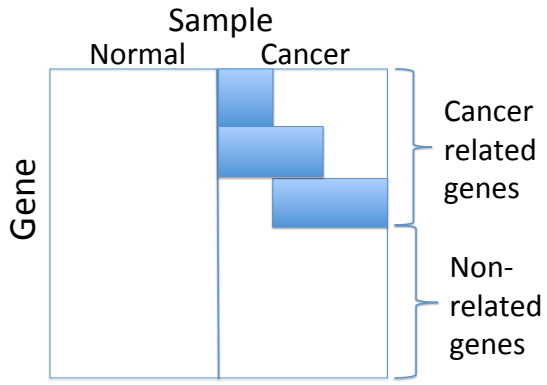


Figure 1: 本研究で提案する biclustering の結果の概念図. 行方向 (遺伝子方向) に見ると, がん関連遺伝子 (青の部分) と関連なし遺伝子 (白の部分) に分かれる. 列方向 (患者方向) に見ると, がん患者の群において, 関連遺伝子の発現プロファイルが患者毎に異なる.

る評価実験を行う. 実データとして, Mils et al. [2] で解析された骨髄異形成症候群の表現型データ及びマイクロアレー遺伝子発現量データを用いる.

References

[1] Heng Lian. Most: detecting cancer differential gene expression. *Biostatistics*, 9(3):411–418, 2008.

[2] Ken I Mills, Alexander Kohlmann, P Mickey Williams, Lothar Wieczorek, Wei-min Liu, Rachel Li, Wen Wei, David T Bowen, Helmut Loeffler, Jesus M Hernandez, et al. Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of aml transformation of myelodysplastic syndrome. *Blood*, 114(5):1063–1072, 2009.

[3] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1):2–8, 2007.

[4] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *science*, 310(5748):644–648, 2005.

[5] Baolin Wu. Cancer outlier differential gene expression detection. *Biostatistics*, 8(3):566–575, 2007.

実用的な加速近接勾配法の実装と 2 値判別モデルの応用

*伊藤直紀 (東京大学), 武田朗子 (統計数理研究所), TOH Kim-Chuan (シンガポール国立大学)

2017 年 2 月 19 日 科研費シンポジウム「統計的モデリングと計算アルゴリズムの数理と展開」

1 はじめに

2 値判別問題は機械学習における重要な問題の 1 つであり, サポートベクターマシン (SVM) をはじめとして様々な 2 値判別モデルが提案されている. 高い判別精度を達成するためには, データごとに様々なモデルを解き, 適切なモデルを選択する必要がある. そのため, 自由なモデル選択を行う上では, 個々のモデルに特化した解法ではなく, 汎用的かつ高速な最適化手法を開発することが重要である. そこで本稿では以下の最適化問題に対する解法を考える.

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} F(\boldsymbol{\alpha}) := f(\boldsymbol{\alpha}) + g(\boldsymbol{\alpha}). \tag{1}$$

ただし, 以下の仮定をおく.

- $f : \mathbb{R}^d \to \mathbb{R}$ は連続的微分可能な真閉凸関数で, その勾配 $\nabla f(\cdot)$ はリプシッツ連続である. すなわち, あるリプシッツ定数 $L_f > 0$ が存在して, 以下の不等式が成り立つ.

  $\|\nabla f(\boldsymbol{\alpha}) - \nabla f(\boldsymbol{\beta})\|_2 \le L_f\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2 \quad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d.$
- $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ は真閉凸関数で, 有効領域 $\mathrm{dom}(g) = \{\boldsymbol{\alpha} \in \mathbb{R}^d \mid g(\boldsymbol{\alpha}) < +\infty\}$ は閉凸である.

機械学習や信号処理で表れる様々なモデルが, f を損失項, g を正則化項と置くことで問題 (1) に帰着できる. 問題 (1) に対する効率的な解法として, 加速近接勾配法 (APG) [1] が知られており, APG を改良して実用上の性能を高める研究もなされている. 一部の改良法に対しては大域的収束性が示されていなかったが, 我々 [2] は複数の改良法を組み合わせることで, 理論的に収束レートの保証のある実用的に高速な APG (FAPG) を提案した. 本稿では FAPG の適用例を紹介する.

2 提案アルゴリズム

関数 g の近接写像は以下の様に定義される.

$$\mathrm{prox}_{g,L}(\bar{\boldsymbol{\alpha}}) = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^d}\Big\{g(\boldsymbol{\alpha}) + \tfrac{L}{2}\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_2^2\Big\}.$$

近接勾配法 (PG) は, 関数 f の勾配と, 関数 g の近接写像を用いて, 以下の様に点列を更新する.

$$\boldsymbol{\alpha}^{k+1} \leftarrow \mathrm{prox}_{g,L_k}\Big(\boldsymbol{\alpha}^k - \tfrac{1}{L_k}\nabla f(\boldsymbol{\alpha}^k)\Big).$$

ただし, $L_{k+1} = L_k \ge L_f$ $(k = 1, 2, \ldots)$ とおく. 実は, $F(\boldsymbol{\alpha})$ を $\boldsymbol{\beta}$ のまわりで近似した関数:

$$Q_L(\boldsymbol{\alpha};\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \langle\nabla f(\boldsymbol{\beta}), \boldsymbol{\alpha} - \boldsymbol{\beta}\rangle + g(\boldsymbol{\alpha}) + \tfrac{L}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2$$

を用いると,

$$\mathrm{prox}_{g,L_k}\Big(\boldsymbol{\alpha}^k - \tfrac{1}{L}\nabla f(\boldsymbol{\alpha}^k)\Big) = \mathrm{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^d}\, Q_{L_k}(\boldsymbol{\alpha};\boldsymbol{\alpha}^k)$$

と表すことができる. すなわち, PG は, 関数 F の近似関数 $Q_{L_k}(\boldsymbol{\alpha};\boldsymbol{\alpha}^k)$ を繰り返し最小化するアルゴリズムだと見なせる. この PG で生成される点列に “慣性” のような動きを加えたものが, 加速近接勾配法 (APG) [1] である. APG では以下の様に点列を更新する.

Step 1: $\boldsymbol{\alpha}^k \leftarrow \mathrm{prox}_{g,L_k}\Big(\boldsymbol{\beta}^k - \tfrac{1}{L_k}\nabla f(\boldsymbol{\beta}^k)\Big)$.

Step 2: $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$. (ただし $t_1 = 1$.)

Step 3: $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\alpha}^k + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})$.

PG の収束レートは $F(\boldsymbol{\alpha}^k) - F(\boldsymbol{\alpha}^*) \le O(1/k)$ であるのに対して, APG はより速い収束レート $F(\boldsymbol{\alpha}^k) - F(\boldsymbol{\alpha}^*) \le O(1/k^2)$ をもつ. FAPG [2] は, APG と様々な実用上の改良法を組み合わせた手法であり, Algorithm 1 のように点列を更新する. FAPG の収束レートは $F(\boldsymbol{\alpha}^k) - F(\boldsymbol{\alpha}^*) \le O\big((\log k/k)^2\big)$ だが, 実用的には APG よりも高速である.

3 線形判別モデルへの応用

2 値判別問題は, 特徴量 $\boldsymbol{x}$ とラベル y の m 個の組: $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}$, $i \in M := \{1, \ldots, m\}$ で与えられた訓練データをもとに, 未知のデータ $\hat{\boldsymbol{x}}$ がどちらのクラスに属しているかを判別する問題である. いま, 線形関数 $h(\hat{\boldsymbol{x}}) = \boldsymbol{w}^\top\hat{\boldsymbol{x}}$ の符号によって データ $\hat{\boldsymbol{x}}$ のクラスを予測することを考える. このとき適切な係数 $\boldsymbol{w}$ の定め方を最適化問題として定式化したものが線形判別モデルである. 様々な線形判別モデルが問題 (1) に帰着され, FAPG によって効率的に解くことができる.

3.1 ℓ_1 正則化モデル

$g(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ とおく. このとき,

- $f(\boldsymbol{w}) = C\sum_{i=1}^m \log(1 + \exp(-y_i\boldsymbol{w}^\top\boldsymbol{x}_i))$ とおくと, 問題 (1) は ℓ_1 正則化ロジスティック回帰となる.
- $f(\boldsymbol{w}) = C\sum_{i=1}^m (\max\{0, 1 - y_i\boldsymbol{w}^\top\boldsymbol{x}_i\})^2$ とおくと, 問題 (1) は ℓ_1 正則化 ℓ_2-SVM となる.

表 1　$\mathcal{U}'$ と $\boldsymbol{x}(\boldsymbol{\alpha})$ の定め方と, 2 値判別モデルの関連. ただし, $M_+ = \{i \in M \mid y_i = +1\}$, $M_- = \{i \in M \mid y_i = -1\}$ である. また, サンプル集合 $\{\boldsymbol{x}_i \mid i \in M_o\}$ に対して, $\bar{\boldsymbol{x}}_o$ は平均ベクトル, Σ_o は共分散行列を表す $(o \in \{+,-\})$.

モデル	$\mathcal{U}'$, $\boldsymbol{x}(\boldsymbol{\alpha})$
ミニマックス確率マシン (MPM)	$\mathcal{U}' := \{(\boldsymbol{\alpha}_+, \boldsymbol{\alpha}_-) \in \mathbb{R}^{2n} \mid \|\boldsymbol{\alpha}_o\|_2 \le \kappa,\ o \in \{+,-\}\}$ $\boldsymbol{x}(\boldsymbol{\alpha}) = (\bar{\boldsymbol{x}}_+ + \Sigma_+^{1/2}\boldsymbol{\alpha}_+) - (\bar{\boldsymbol{x}}_- + \Sigma_-^{1/2}\boldsymbol{\alpha}_-)$
フィッシャーの線形判別器 (FDA)	$\mathcal{U}' = \{\boldsymbol{\alpha} \in \mathbb{R}^n \mid \|\boldsymbol{\alpha}\|_2 \le \kappa\}$ $\boldsymbol{x}(\boldsymbol{\alpha}) = (\bar{\boldsymbol{x}}_+ - \bar{\boldsymbol{x}}_-) + (\Sigma_+ + \Sigma_-)^{1/2}\boldsymbol{\alpha}$
ν-サポートベクターマシン (ν-SVM)	$\mathcal{U}' = \{\boldsymbol{\alpha} \in \mathbb{R}^m \mid \sum_{i\in M_+}\alpha_i = \sum_{i\in M_-}\alpha_i = \frac{1}{2},\ \mathbf{0} \le \boldsymbol{\alpha} \le \frac{1}{m\nu}\boldsymbol{e}\}$ $\boldsymbol{x}(\boldsymbol{\alpha}) = \sum_{i\in M_+}\alpha_i\boldsymbol{x}_i - \sum_{i\in M_-}\alpha_i\boldsymbol{x}_i$

Algorithm 1 FAPG Method

Input: $L_0 > 0$, $\eta_u, \eta_d > 1$, $\boldsymbol{\alpha}^0$, $K_1 \ge 2$, $\delta \in (0,1)$
Init.: $t_1 \leftarrow 1$, $t_0 \leftarrow 0$, $\boldsymbol{\beta}^1 \leftarrow \boldsymbol{\alpha}^0$, $\boldsymbol{\alpha}^{-1} \leftarrow \boldsymbol{\alpha}^0$
$L_1 \leftarrow L_0$, $i \leftarrow 1$, $k_{re} \leftarrow 0$
for $k = 1, 2, \ldots$ **do**
  $\boldsymbol{\alpha}^k \leftarrow \mathrm{prox}_{g,L_k}\left(\boldsymbol{\beta}^k - \frac{1}{L_k}\nabla f(\boldsymbol{\beta}^k)\right)$ # Step 1
  while $F(\boldsymbol{\alpha}^k) > Q_{L_k}(\boldsymbol{\alpha}^k; \boldsymbol{\beta}^k)$ **do**
    $L_k \leftarrow \eta_u L_k$ # 'bt'
    $t_k \leftarrow \frac{1+\sqrt{1+4(L_k/L_{k-1})t_{k-1}^2}}{2}$ # 'dec'
    $\boldsymbol{\beta}^k \leftarrow \boldsymbol{\alpha}^{k-1} + \frac{t_{k-1}-1}{t_k}(\boldsymbol{\alpha}^{k-1} - \boldsymbol{\alpha}^{k-2})$ # 'dec'
    $\boldsymbol{\alpha}^k \leftarrow \mathrm{prox}_{g,L_k}\left(\boldsymbol{\beta}^k - \frac{1}{L_k}\nabla f(\boldsymbol{\beta}^k)\right)$ # 'bt'
  end while
  $L_{k+1} \leftarrow L_k/\eta_d$ # 'dec'
  $t_{k+1} \leftarrow \frac{1+\sqrt{1+4(L_{k+1}/L_k)t_k^2}}{2}$ # Step 2'
  $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\alpha}^k + \frac{t_k-1}{t_{k+1}}(\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1})$ # Step 3
  if $k > k_{re} + K_i$ and $F(\boldsymbol{\alpha}^k) > F(\boldsymbol{\alpha}^{k-1})$ **then**
    $k_{re} \leftarrow k$, $K_{i+1} \leftarrow 2K_i$, $i \leftarrow i+1$ # 'mt'
    $\eta_d \leftarrow \delta \cdot \eta_d + (1-\delta)\cdot 1$ # 'st'
    $\boldsymbol{\beta}^{k+1} \leftarrow \boldsymbol{\alpha}^{k-1}$, $\boldsymbol{\alpha}^k \leftarrow \boldsymbol{\alpha}^{k-1}$ # 're'
    $t_{k+1} \leftarrow 1$, $t_k \leftarrow 0$ # 're'
  end if
end for

ただし $C > 0$ はパラメータである. いずれの場合も FAPG を適用することができる. この $g(\boldsymbol{w})$ のもとでの近接写像はソフト閾値関数と呼ばれ, $(\mathrm{prox}_{g,L}(\boldsymbol{w}))_i = \mathrm{sign}(w_i)\max\{0, |w_i| - L\}$ と計算できる.

3.2　統一的判別モデル

ℓ_2 正則化モデルを含む様々な線形判別モデルが, 以下の形式で統一的に記述することができる.

$$\min_{\|\boldsymbol{w}\|_2 \le 1} \phi_{\mathcal{U}}(\boldsymbol{w}) \qquad (2)$$

ただし, $\phi_{\mathcal{U}}(\boldsymbol{w}) = \max_{\boldsymbol{x}\in\mathcal{U}}\{-\boldsymbol{w}^\top\boldsymbol{x}\}$ であり, $\mathcal{U} \subseteq \mathbb{R}^m$ は閉凸集合である. 集合 $\mathcal{U}$ の大きさや形を適切に定めることにより, 様々な 2 値判別モデルが (2) に帰着できる. いま, 媒介変数 $\boldsymbol{\alpha} \in \mathbb{R}^d$, 線型写像 $\boldsymbol{x}(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$, 閉凸集合 $\mathcal{U}' \subseteq \mathbb{R}^d$ を用いて, $\mathcal{U} = \{\boldsymbol{x}(\boldsymbol{\alpha}) \mid \boldsymbol{\alpha} \in \mathcal{U}'\}$ と表すと, 問題 (2) は以下の最小ノルム問題に帰着される.

$$\max_{\boldsymbol{\alpha}\in\mathcal{U}'} \min_{\|\boldsymbol{w}\|_2\le 1} -\boldsymbol{w}^\top\boldsymbol{x}(\boldsymbol{\alpha}) = -\min_{\boldsymbol{\alpha}\in\mathcal{U}'} \|\boldsymbol{x}(\boldsymbol{\alpha})\|_2, \qquad (3)$$

ただし, $\boldsymbol{w} = \boldsymbol{x}(\boldsymbol{\alpha})/\|\boldsymbol{x}(\boldsymbol{\alpha})\|_2$ である. 表 1 は $\mathcal{U}'$ と $\boldsymbol{x}(\cdot)$ の定め方と, 2 値判別モデルの関連を表している. いま, $f(\boldsymbol{\alpha}) = \|\boldsymbol{x}(\boldsymbol{\alpha})\|_2$, $g(\boldsymbol{\alpha}) = \delta_{\mathcal{U}'}(\boldsymbol{\alpha})$ とおけば, FAPG によって (3) を解くことができる. いずれの例においても, 近接写像 $\mathrm{prox}_{g,L}(\boldsymbol{w})$ は Breakpoint 探索法 [3] などを用いて線形時間 $O(m)$ で計算することができる.

4　数値実験

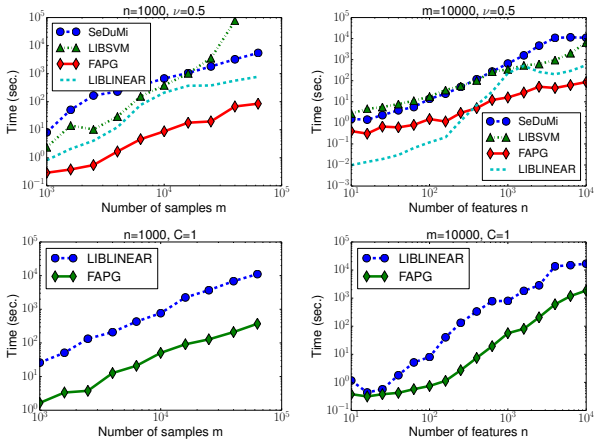


図 1　サンプル数 m および次元 n の大きさと計算時間の関係. (上段: ν-SVM. 下段: ℓ_1 正則化 ℓ_2-SVM.)

ν-SVM と ℓ_1 正則化 ℓ_2-SVM に対し, 人工データを用いて既存のソフトウェアと FAPG の計算時間を比較した. 特に大規模なデータを用いたときに, FAPG の優位性が見られた.

参考文献

[1] A. Beck and M. Teboulle: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[2] N. Ito, A. Takeda, and K.-C. Toh: A unified formulation and fast accelerated gradient method for classification. Journal of Machine Learning Research, to appear.

[3] K.C. Kiwiel: Breakpoint searching algorithms for the continuous quadratic knapsack problem. Mathematical Programming, 112:473–491, 2008

空間の低ランク性と平滑性を考慮したフーリエ係数最適化によるMR超解像

本谷　秀堅†　　河村　直輝†　　横田　達也†

† 名古屋工業大学 〒466-8555 愛知県名古屋市昭和区御器所町

あらまし　MR画像は撮影時に空間分解能を高く設定すると，S/N比が著しく低下する．本稿では同一試料を複数方向から撮影したMR画像を超解像化することで，S/N比を保持したまま高分解能なMR画像を構成する．提案法では，対象の輪郭情報を用いる超解像法にTV正則化と低ランク化を導入する．

キーワード　画像処理，超解像，Total Variation, 低ランクテンソル補完，ADMM

Hidekata HONTANI†, Naoki KAWAMURA†, and Tatsuya YOKOTA†

† Nagoya Institute of Technology

1. はじめに

磁気共鳴画像法 (MRI) は生体内部の3次元空間情報を高コントラストで観察するために有用であり，重要なモダリティの一つとして広く利用されている．MRIでは体内の水素原子濃度を磁気を用いて観測し，体内断面をスライス画像として撮影する．スライス厚を小さくすれば空間分解能を高くすることができるが，その場合はスライス毎の水素原子含有量が減少し，S/N比が大きく低下する．そのためMRIでは，十分なS/N比を確保するためにスライス厚を大きくとって撮影する．ただしスライス厚を大きくすると，MR画像は磁場方向の空間分解能が低く，非等方的な3次元画像となる．すなわち，MRIの磁場方向の高周波成分が欠損した状態で画像が得られる．

空間分解能が非等方的なMR画像を画像処理する場合，既存のアルゴリズムをそのまま適用できないケースが多々ある．そのため，MR画像を等方的に高解像度化することは非常に重要である．

超解像は画像信号における未知の高周波成分を復元する技術である．Gerchbergは対象の輪郭情報が既知の場合に，効果的に解くことができる超解像法を提案した [2]．Gerchbergの手法では信号領域と周波数領域の双方から誤差エネルギーを反復的に除去することで高解像度化を行う．一方で近年，Total Variation(TV) 正則化による超解像法が注目を浴びている．TV正則化は画像のエッジを保持しつつ滑らかさを評価できるため，超解像に有効である [3]．また，最近では行列の低ランク化によるテンソル補完技術も，その有用性が注目されている [4]．

提案法では，図1のように異なる方向から撮影した複数枚のMR画像を用いる．互いに異なる磁場方向から撮影した数枚のMR画像は，低分解能な方向を相互に補い合うため，S/N比を保ちつつ高解像なMR画像を構築できることが期待できる．しかし，軸から離れた高周波成分は観測できない．そこで，未知の高周波成分を補完するために新たに超解像法を提案し，効果的な高解像度化を目指す．

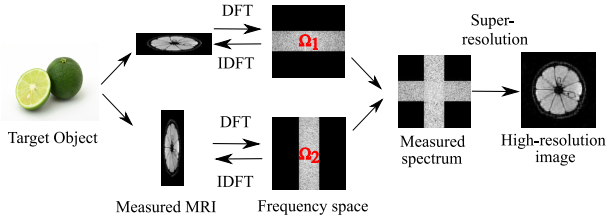


図1　提案法の流れ．MRIにより対象物体は異なる複数方向からそれぞれ非等方的に観測される．各画像を統合しても，観測されない周波数領域があり，これら未知の高周波成分を超解像で復元する．

2. 問題設定

本節では，提案手法について説明する．同一対象を異なる3つの方向から撮影したMR画像を $I_1 \in \mathbb{R}^{N\times N\times n}$, $I_2 \in \mathbb{R}^{n\times N\times N}$, $I_3 \in \mathbb{R}^{N\times n\times N}$ とする．これらの磁場方向は互いに直交し，$n < N$ である．すなわち，各スライス画像サイズは $N \times N$ であり，スライス枚数は n である．$I_1 \sim I_3$ から，等方的で高分解能な画像 $I \in R^{N\times N\times N}$ を復元したい．ここで，I_i のフーリエ変換を F_i とすると，目標の F 周波数空間のうち F_i の成分は図1に示すように，磁場方向に向かって狭帯域に計測されている．この観測領域を Ω_i とする．3つの観測領域を統合した $\Omega := \Omega_1 \cup \Omega_2 \cup \Omega_3$ は F 周波数空間全てを表現できず，未知の高周波成分が存在する．この未知の高周波成分を，次節で提案する超解像により復元する．

3. 提 案 法

我々の提案する超解像法では，対象の輪郭を用いる Gerchberg 法に TV 正則化とテンソルの低ランク補完を組み合わせる．提案法では，次のような凸最適化問題を解くことで，画像 $\boldsymbol{x} = \mathrm{vec}(I) \in \mathbb{R}^{N^3 \times 1}$ を復元する．

$$\min_{\boldsymbol{f}} \sum_{j=1}^{3} (\boldsymbol{x}_j - D_j H_j \boldsymbol{x}) + \lambda_{TV} ||\boldsymbol{x}||_{TV} + \lambda_{LR} \sum_{j=1}^{3} \frac{||M_{(j)}||_{tr}}{3} + \sum_{j=1}^{3} \frac{\rho}{2} ||\boldsymbol{x} - \mathrm{vec}(M_j)||_2^2 + \frac{\mu}{2} ||\boldsymbol{f}_0 - R_\Omega \boldsymbol{f}||_2^2,$$

$$\text{s.t. } \boldsymbol{x} = G\boldsymbol{f}, \quad \boldsymbol{0} = R_\Gamma \boldsymbol{x}, \quad M_{(j)} = \mathrm{unfold}_j(M_j). \qquad (1)$$

ここで，$\mathrm{vec}(\cdot)$ はテンソルからベクトルへの変換，$\mathrm{unfold}_j(\cdot)$ は 3 次元テンソルを方向 j に沿って行列に展開する．M_j はテンソル I の方向 j に対応する低ランク化のスラック変数であり，$\boldsymbol{f}_0$ は $\boldsymbol{f}$ の観測された初期値である．$\boldsymbol{x}_j = \mathrm{vec}(I_j) \in \mathbb{R}^{N^2 n \times 1}$ であり，H_j, D_j はそれぞれ磁場方向に向かって平滑化，サンプルを行う行列である．$||\cdot||_{\mathrm{TV}}$ は Total Variation，$||\cdot||_{\mathrm{tr}}$ は行列のトレースノルムであり，G は逆フーリエ基底行列である．また $R_\Gamma \in \{0,1\}^{N^3 \times N^3}$ は画像中の対象領域外部を指定し，$R_\Omega \in [0,1]^{N^3 \times N^3}$ は周波数空間の領域 Ω を指定する行列である．$\lambda_{TV}, \lambda_{LR}, \rho, \mu > 0$ はそれぞれ TV 正則化，低ランク化，スラック変数のフィッティング，周波数フィッティングの 4 つの項のバランスを調整するパラメータである．式 (1) は PDS や ADMM，MM アルゴリズム等で解く．ADMM では，式 (1) を $\boldsymbol{x}, M_j, \boldsymbol{f}$ と 2 つのラグランジュ係数について最適化する．

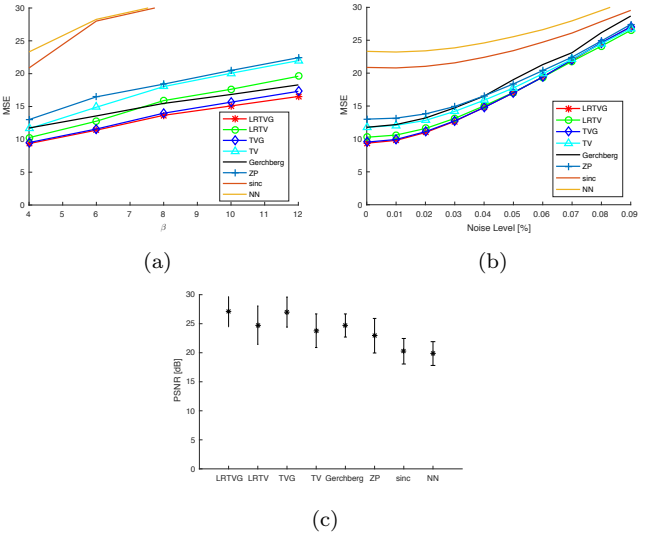
4. 実　　験

WEIZMANN データセット [1] に対してシミュレーション実験を行った．$N \times N$，$N = 120$ の元画像を目標画像とし，これを 1 方向に向かって間隔 β でダウンサンプルした画像 I_1，I_2，$(n = N/\beta)$ を得る．狭帯域な観測画像 I_1，I_2 から元画像を復元し，対象領域内部の誤差で従来法と比較評価を行った結果を図 2 に示す．比較する手法は最近傍法 (NN)，Gerchberg 法 [2]，TV 正則化超解像法 (TV) [3]，低ランク補間を導入した TV 超解像 (LRTV) [4]，提案法 (LRTVG)，提案法から低ランク補間項を除いたもの (TVG) である．提案法は従来法と比較して，β を変化させても精度の高い結果が得られた．しかし，強いノイズを付加すると従来法より性能が減少した．次にスダチの MR 実画像を用いて実験を行った結果を図 3 に示す．観測時には欠損していたエッジ成分が復元され，高解像度化された．

5. おわりに

本稿では，同一試料を複数方向から撮影した MR 画像を超解像化することで，等方的に高分解能な MR 画像を構成した．Gerchberg の超解像法を TV 正則化と低ランク化することで，効果的に超解像化できることを確認した．

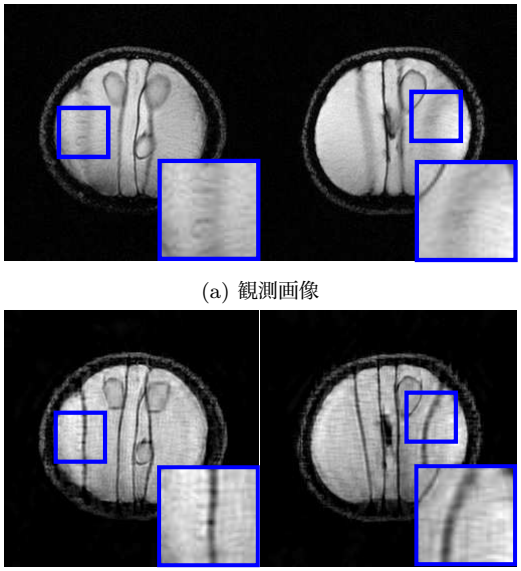
謝辞　本研究は JSPS 科研費 26108003，及び 15K16067 の助成を受けたものである．



(a)　(b)

(c)

図 2　WEIZMANN データセットによるシミュレーション結果．(a)*sharp image* における MSE w.r.t. β，(b)*sharp image* のノイズ強度変化における MSE($\beta = 4$)，(c)$\beta = 4$，データセット集合で計測した PSNR.



(a) 観測画像

(b) 復元結果

図 3　T2-MRI で撮影したスダチの実験結果．

文　　献

[1] S. Alpert, M. Galun, R. Basri, and A. Brandt, “Image segmentation by probabilistic bottom-up aggregation and cue integration,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1-8, 2007.

[2] R. W. Gerchberg. “Super-resolution through error energy reduction.” *Journal of Modern Optics*, vol.21, no.9, pp709-720, 1974.

[3] A. Marquina, and S. J. Osher. “Image super-resolution by TV-regularization and Bregman iteration.” *Journal of Scientific Computing*, vol.37, no.3, pp.367-382, 2008.

[4] F. Shi, J. Cheng, L. Wang, P-T. Yap, and D. Shen. “LRTV: MR image super-resolution with low-rank and total variation regularizations.” *IEEE transactions on Medical Imaging*, vol.34, no.12, pp.2459-2466, 2015.

大自由度系のデータ同化のための 2nd-order adjoint 法を用いた高速不確実性評価法

伊藤伸一 [1]、長尾大道 [1,2]、加納将行 [1]、黒河天 [2]

[1] 東京大学地震研究所
[2] 東京大学大学院情報理工学系研究科

データ同化 [1] はシミュレーションモデルと観測データをベイズ統計学を用いて融合する計算基盤技術であり、事後分布の形状を評価することにより、観測不可能な系の内部状態の推定やモデルパラメータの推定を行なうだけでなく、その推定値の不確実性を併せて評価することが可能である。事後分布の構成に用いられるデータ同化は逐次型と非逐次型に大別され、アンサンブルカルマンフィルタ [2] や粒子フィルタ [3] 等に基づく逐次データ同化は、さまざまな科学分野で用いられている。しかし、これらの手法はモンテカルロベースの手法であり一般に推定および不確実性評価に必要な計算量・メモリがシミュレーションモデルの自由度の指数オーダーとなるため、シミュレーションモデルの規模が大きくなると評価が難しくなる。一方で、非逐次データ同化法である 4 次元変分法 [4] は、推定に必要な計算量・メモリをモデルの自由度の線形オーダーに抑えることができるため、大規模シミュレーションモデルに適している。しかし、従来の 4 次元変分法の枠組みでは勾配法による事後確率最大解のみが分かるだけであり、その不確実性を評価することは原理的に不可能であるため、逐次データ同化法と組み合わせる等のさまざまなアドホックな工夫を凝らすことにより、不確実性の評価を行なっていた。

そこで我々は、その問題を解決するために、2nd-order adjoint 法 [5] という計算法を利用して、大規模なシミュレーションモデルに対しても高速かつ高精度な不確実性評価を可能にするアルゴリズムを開発した [6]。不確実性は事後分布を近似する正規分布の共分散行列の対角成分、つまりヘッセ逆行列の対角成分で近似されるが、本研究では、実用上不確実性を評価すべき変数の数がシミュレーションモデルの自由度に比べて圧倒的に少ないことに注目し、2nd-order adjoint 法をうまく利用することで、注目する要素の不確実性のみを直接求めるアルゴリズムを構築した。2nd-order adjoint 法を用いない従来までの不確実性の計算法では必要とするメモリがシミュレーションモデルの自由度の 2 乗、計算量がシミュレーションモデルの自由度の 3 乗以上必要とするのに対して、提案手法はメモリがシミュレーションモデルの自由度の 1 乗、計算量がシミュレーションモデルの自由度に依らない (シミュレーションモデルの時系列を計算する計算量と同程度) になるので、大規模なシミュレーションモデルに対して非常に有利である。

提案手法の検証のため、固体-液体相界面の移動ダイナミクスを記述するフェーズフィールドモ

デルの時間発展から得られる擬似データに本提案手法を適用し、モデルの初期状態およびモデルパラメータの推定、モデルパラメータの不確実性評価を行なった。検証の結果、本提案手法は仮定された初期状態およびモデルパラメータを正しく推定することができ、さらに、データの量・質に応じた不確実性を定量的かつ高速に評価することに成功した。本提案手法によって得られる不確実性は、実験等へのフィードバックに対する重要な知見を与えることができ、例えば、必要な精度の推定結果を得るにはどの程度データが必要かなどを教えてくれる。小規模自由度のシミュレーションモデルに対しては逐次データ同化法等を使うことでそれは可能だったが、大規模自由度のシミュレーションモデルに対しても、本提案手法によりそれが可能になった。

参考文献

[1] S. Reich and C. Cotter, Probabilistic Forecasting and Bayesian Data Assimilation (Cambridge University Press, Cambridge, 2015).

[2] G. Evensen, The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics **53**, 343 (2003).

[3] G. Kitagawa, Introduction to Time Series Modeling, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (CRC Press, Boca Raton, FL, 2010).

[4] F.-X. Le Dimet and O. Talagrand, Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects, Tellus A **38A**, 97 (1986).

[5] F.-X. Le Dimet, I. M. Navon, and D. N. Daescu, Second-order information in data assimilation, Monthly Weather Review **130**, 629 (2002).

[6] S. Ito, H. Nagao, A. Yamanaka, Y. Tsukada, T. Koyama, M. Kano, and J. Inoue, Data assimilation for massive autonomous systems based on a second-order adjoint method, Physical Review E **94**, 043307 (2016).

座標ごとの変数変換によって得られる Stein 型の分布

清 智也 （東大 情報理工）*

概要

与えられた多次元の確率分布に対し，座標ごとの確率積分変換によってコピュラが得られることはよく知られている（Sklar の定理）。本研究では，座標ごとの変数変換によって，Stein 型の等式を満たす分布が一意的に得られることを示す。ただしいくつかの正則条件を要する。本結果は Marshall and Olkin (1968, Numer. Math.) による対角スケーリング定理の非線形版と解釈できる。

1 Stein 型の等式

d を正の整数とする。$\mathbb{R}^d$ 上の確率分布 μ のうち，各周辺分布 μ_i $(i = 1, \dots, d)$ が絶対連続で，$\int_{\mathbb{R}} x_i \mathrm{d}\mu_i = 0$ かつ $\int_{\mathbb{R}} x_i^2 \mathrm{d}\mu_i < \infty$ となるようなものの全体を $\mathcal{P}^2$ とおく。μ_i は絶対連続だが μ 自身は絶対連続とは限らないことに注意する。また，局所絶対連続な関数 $f : \mathbb{R} \to \mathbb{R}$ のうち，その導関数 f' が本質的に有界であるようなものの全体を $\mathcal{B}$ とおく。

定義 1. 分布 $\mu \in \mathcal{P}^2$ が **Stein 型分布**であるとは，全ての $f \in \mathcal{B}$ に対して

$$\int f(x_i) \left(\sum_{j=1}^{d} x_j \right) \mathrm{d}\mu = \int f'(x_i) \mathrm{d}\mu, \quad i = 1, \dots, d, \tag{1}$$

を満たすこととする。また式 (1) を **Stein 型の等式**と呼ぶ。

$d = 1$ の場合，式 (1) はいわゆる Stein の等式 $\int f(x_1) x_1 \mathrm{d}\mu = \int f'(x_1) \mathrm{d}\mu$ に帰着され，μ は標準正規分布に限られる（例えば [1]）。同様に，独立な分布 $\mu = \prod_{i=1}^{d} \mu_i$ で Stein 型となるのは各 μ_i が標準正規分布のときのみである。多変量正規分布 $\mu = \mathrm{N}(0, S)$ の場合，μ が Stein 型となるのは各 i に対して $\sum_{j=1}^{d} S_{ij} = 1$ のときであることが示される。

もし確率変数ベクトル $(X_1, \dots, X_d)$ が Stein 型分布に従うならば，式 (1) より，それらの和 $\sum_j X_j$ は各 X_i の任意の単調増加変換 $f(X_i)$ と正の相関を持つことが分かる。この性質の応用として，多変量データを総合指数にまとめる方法が議論されている [4]。

座標ごとの変換

$$T(x) = (T_1(x_1), \dots, T_d(x_d)), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

* 本研究は科研費 (課題番号: 26108003, 26540013) の助成を受けたものである。

のうち，各 $T_i : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ が非減少関数で，かつ像測度 $T_\sharp\mu$ が $\mathcal{P}^2$ に属すようなものの全体を $\mathcal{T}_{\mathrm{cw}}(\mu)$ とおく。ここで cw は coordinate-wise の略である。

座標ごとの変換で得られる分布としてコピュラがよく知られている [3]。本研究では，与えられた $\mu \in \mathcal{P}^2$ に対し，$T_\sharp\mu$ が Stein 型となるような座標ごとの変換 $T \in \mathcal{T}_{\mathrm{cw}}(\mu)$ を見つける問題を考える。そのような T を μ の **Stein 型変換**と呼ぶことにしよう。

2 主結果

定理 1. $\mu \in \mathcal{P}^2$ とし，$T \in \mathcal{T}_{\mathrm{cw}}(\mu)$ とする。このとき次の 2 つの条件は同値である：

(i) T は μ の Stein 型変換である。
(ii) $T_\sharp\mu$ は集合 $\{S_\sharp\mu \mid S \in \mathcal{T}_{\mathrm{cw}}(\mu)\}$ において $\mathcal{E}$ を最小化する。

定理 2 (一意性)**.** $\mu \in \mathcal{P}^2$ のサポートが周辺分布 $\mu_1, \dots, \mu_d$ のサポートの直積に一致すると仮定する。このとき μ の Stein 型変換は高々一つである。

分布 $\mu \in \mathcal{P}^2$ が**共正値** (copositive) であるとは,

$$\inf_{T \in \mathcal{T}_{\mathrm{cw}}(\mu)} \frac{\int (\sum_i T_i)^2 \mathrm{d}\mu}{\sum_i \int T_i^2 \mathrm{d}\mu} > 0.$$

が成り立つこととする。

定理 3 (存在性)**.** μ が共正値ならば，μ の Stein 型変換が存在する。

定理 2，定理 3 は，次の対角スケーリング定理の一般化とみなすことができる。

定理 4 (Marshall and Olkin [2])**.** S は d 次の半正定値行列で, かつ真に共正値 (strictly copositive)

$$\inf_{w_1, \dots, w_d > 0} \frac{\sum_i \sum_j w_i S_{ij} w_j}{\sum_i w_i^2} > 0$$

であるとする。このとき，ある正の数 $w_1, \dots, w_d$ で，各 i について $\sum_{j=1}^d w_i S_{ij} w_j = 1$ となるものが一意的に存在する。

参考文献

[1] Chen, L. H. Y., Goldstein, L., and Shao, Q. (2011). *Normal Approximation by Stein's Method*, Springer.

[2] Marshall, A. W., Olkin, I., (1968). Scaling of matrices to achieve specified row and column sums. *Numer. Math.*, 12, 83–90.

[3] Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer.

[4] Sei, T. (2016). An objective general index for multivariate ordered data, *J. Multivariate Anal.*, 147, 247–264.

An operational characterization of the notion of probability by algorithmic randomness and its applications*

Kohtaro Tadaki

Department of Computer Science, College of Engineering, Chubu University
1200 Matsumoto-cho, Kasugai-shi, Aichi 487-8501, Japan
E-mail: tadaki@cs.chubu.ac.jp
http://www2.odn.ne.jp/tadaki/

The notion of probability plays an important role in almost all areas of science and technology. In modern mathematics, however, probability theory means nothing other than *measure theory* [10], and the *operational characterization* of the notion of probability is not established yet. In this talk, based on the toolkit of *algorithmic randomness* we present an operational characterization of the notion of probability, called an *ensemble*.

Algorithmic randomness, also known as *algorithmic information theory*, is a field of mathematics which enables us to consider the randomness of an *individual* infinite sequence [14, 11, 2, 12, 3, 4, 13, 7]. We use the notion of *Martin-Löf randomness with respect to Bernoulli measure* [12] to present the operational characterization. As the first step of the research of this line, in this talk we consider the case of finite probability space, i.e., the case where the sample space of the underlying probability space is finite, for simplicity.

We give a natural operational characterization of the notion of *conditional probability* in terms of ensemble, and give equivalent characterizations of the notion of *independence* between two events based on it. Furthermore, we give equivalent characterizations of the notion of *independence* of an arbitrary number of events/random variables in terms of ensembles. In particular, we show that the independence between events/random variables is *equivalent* to the independence in the sense of van Lambalgen's Theorem [17], in the case where the underlying finite probability space is computable.

In the talk we make applications of our framework to quantum mechanics, information theory, and cryptography in order to demonstrate the *wide applicability* of our framework to the general areas of science and technology.

For the detail of this talk, see Tadaki [15].

Key words: probability, algorithmic randomness, operational characterization, Martin-Löf randomness, Bernoulli measure, conditional probability, independence, van Lambalgen's Theorem, quantum mechanics, information theory, cryptography

References

[1] R. B. Ash, *Information Theory*. Dover Publications, Inc., New York, 1990.

[2] G. J. Chaitin, "On the length of programs for computing finite binary sequences," *J. Assoc. Comput. Mach.*, vol. 13, pp. 547–569, 1966.

*This work was partially supported by JSPS KAKENHI Grant Numbers 24540142, 15K04981. This work was partially done while the author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2014.

[3] G. J. Chaitin, "A theory of program size formally identical to information theory," *J. Assoc. Comput. Mach.*, vol. 22, pp. 329–340, 1975.

[4] G. J. Chaitin, *Algorithmic Information Theory*. Cambridge University Press, Cambridge, 1987.

[5] B. S. DeWitt and N. Graham (eds.), *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton University Press, Prenceton, 1973.

[6] P. A. M. Dirac, *The Principles of Quantum Mechanics*, 4th ed. Oxford University Press, London, 1958.

[7] R. G. Downey and D. R. Hirschfeldt, *Algorithmic Randomness and Complexity*. Springer-Verlag, New York, 2010.

[8] H. Everett, III, ""Relative State" formulation of quantum mechanics," *Rev. Mod. Phys.*, vol. 29, no. 3, pp. 454–462, 1957.

[9] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*. Chapman & Hall/CRC Press, 2007.

[10] A. N. Kolmogorov, *Foundations of the theory of probability*, Chelsea Publishing Company, New York, 1950.

[11] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Inform. Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[12] P. Martin-Löf, "The definition of random sequences," *Inform. and Control*, vol. 9, pp. 602–619, 1966.

[13] A. Nies, *Computability and Randomness*. Oxford University Press, Inc., New York, 2009.

[14] R. J. Solomonoff, "A formal theory of inductive inference. Part I and Part II," *Inform. and Control*, vol. 7, pp. 1–22, 1964; vol. 7, pp. 224–254, 1964.

[15] K. Tadaki, "An operational characterization of the notion of probability by algorithmic randomness and its applications," arXiv:1611.06201, November 2016. Available at URL: `https://arxiv.org/abs/1611.06201`

[16] K. Tadaki, "A refinement of quantum mechanics by algorithmic randomness and its application to quantum cryptography," Proceedings of the 2017 Symposium on Cryptography and Information Security (SCIS2017), 4A1-2, January 24-27, 2017, Naha, Japan.

[17] M. van Lambalgen, *Random Sequences*. Ph.D. dissertation, University of Amsterdam, 1987.

[18] R. von Mises, *Probability, Statistics and Truth*. Dover Publications, Inc., New York, 1957.

[19] R. von Mises, *Mathematical Theory of Probability and Statistics*. Academic Press Inc., New York, 1964.

［報告書］
通信路容量を達成する出力分布の射影アルゴリズムによる探索ついて

On the Search by the Projection Algorithm for the Output Distribution that Achieves the Channel Capacity

中川 健治，渡部 康平，佐武 拓斗
長岡技術科学大学

1．概要

本論文では離散的無記憶通信路の通信路容量を達成する出力分布を探索するアルゴリズムについて考察した。このアルゴリズムは，出力分布全体の空間における 1 つの出力分布からアフィン部分空間への射影の反復である。通信路容量の問題は，ユークリッド空間における有限個の点に対する最小包含円問題と同様の幾何学的構造を有する。ユークリッド空間内の尺度はユークリッド距離であり，出力分布全体の空間での尺度は Kullback-Leibler 情報量である。本論文では，まず，ユークリッド空間における最小包含円問題を考察し，最小包含円の中心を計算するアルゴリズムを開発した。そして，そこで得られたアルゴリズムを通信路容量の問題に適用し，通信路容量を達成する出力分布を計算するアルゴリズムを得た。

離散的無記憶通信路の通信路容量 C は相互情報量の最大値として定義される。C はまた，Kullback-Leibler (KL) 情報量に関する min max 問題の解としても定式化される。通信路容量に関する min max 問題における KL 情報量をユークリッド距離に置き換えると，ユークリッド空間における類似の問題が得られる。それは，有限点集合に対する最小包含円の問題である。

通信路容量の問題は出力分布に関する最適化問題であり，考える空間はユークリッド空間ではなく確率分布全体がなす集合である。その集合上の幾何学は情報幾何である。ユークリッド幾何と情報幾何は甘利の α 幾何の立場から統一的に考えることができる。したがって，最小包含円問題の解法アルゴリズムを，幾何的な類似性を通して，通信路容量問題に適用できる。その際に重要なことは，最小包含円問題を考察するのにユークリッド幾何のどんな性質を使ってもよいということではなく，通信路容量問題に適用可能な性質だけを使う必要がある，ということである。その結果，最小包含円問題で得られたアルゴリズムをほぼ自動的に通信路容量問題に移植することができる。本論文では，実際に両方の幾何に共通の性質として，距離，重心座標，内積，ピタゴラスの定理，アファイン部分空間への射影のみを使ってアルゴリズムを開発した。このアルゴリズムを本論文では「射影アルゴリズム」と呼んだ。さらに，従来扱いが困難とされていた入力アルファベット数が出力アルファベット数より多い場合において，その困難さを緩和して，通信路容量を達成する出力分布を計算する方法を提案した。

通信路容量の問題の基盤となる情報幾何がユークリッド幾何と異なる点は，ユークリッド幾何では 1 つの座標系を用いるのに対して，情報幾何では互いに双対な 2 つの座標系を用いることである。ユークリッド幾何と情報幾何は，実数 α によって α 幾何という統合的な幾何構造の族の一部として考えられる。ユークリッド幾何は $\alpha = 0$, 情報幾何は $\alpha = \pm 1$ という特別な値に対応する幾何として捉えることができる。α 幾何では，α divergence, 内積，Pythagoras の定

理，α 射影が使える。最小包含円問題に関する定理の証明において利用したユークリッド幾何の性質が距離，内積，Pythagoras の定理，射影だけなので，そこで得られた定理やアルゴリズムは容易に通信路容量の問題に適用することができた。

ユークリッド幾何は我々になじみの幾何なので，アルゴリズムの開発とその証明がしやすい。それらはある程度強引な方法でもできてしまうが，それはある意味でいいことである。しかし，それと同じ方法を情報幾何に移植しようとすると，うまくいかない場合があった。そこで，それは強引で自然な方法ではなかったと気がつく。情報幾何に適用できるように考え直すと，逆にユークリッド幾何に還元できて，ユークリッド幾何で別の方法を発見することができた。つまり，ユークリッド幾何と情報幾何に共通の性質だけを使ったアルゴリズムが最も自然であり，実は最も簡単な方法である。これは非常に興味深い結果である。たぶん，初めから通信路容量の問題を情報幾何だけで考えていたのでは本論文の射影アルゴリズムは得られなかったと思う。ユークリッド幾何と情報幾何が有機的に結びついた結果だと思う。

2. 本論文で得られた結果

本論文で得られた結果は下記の通りである。

- 通信路容量の問題はユークリッド空間 $\mathbb{R}^n$ における最小包含円問題と類似した幾何構造を持つことを明らかにした。その類似性に基づいて通信路容量を計算するための射影アルゴリズムを開発した。

- 通信路容量を達成する出力分布を射影アルゴリズムによって探索するための定理を示し，その定理において重心座標が重要な役割を持つことを明らかにした。

- 通信路行列 Φ の行ベクトルが必ずしも一般の位置にあるとは限らない場合にも，新たに Φ に近い通信路行列 $\widetilde{\Phi}$ を定義して，その行ベクトルが一般の位置にあるようにして，射影アルゴリズムを適用した。$\widetilde{\Phi}$ の各行ベクトルは，もとの Φ の各行ベクトルの次元を上げたものである。しばしば，ある問題において次元を上げることによって問題の困難さが緩和される場合がある。本論文で提案している方法は次元を上げる方法のひとつの例である。

なお，本論文の成果をまとめて論文として IEEE Transactions on Information Theory に投稿して採録となり，掲載される予定である。

3. 今後の課題

今後の課題として下記のことが挙げられる。

- 任意の入力アルファベット数 m に対して通信路容量を達成する出力分布を計算するアルゴリズムを作る。

- レート歪み関数や capacity constraint function を計算する射影アルゴリズムを作る。

- 通信路容量を計算する逐次アルゴリズムである有本アルゴリズムを最小包含円の問題に移植する。

レート 歪み理論と一般化事後分布

渡辺一帆 *

The rate-distortion (RD) theory studies and characterizes the performance of lossy compression systems by the RD function [1]. There are studies interpreting learning and prediction problems by rate-distortion theoretic views. In particular, clustering methods can be considered as vector quantizers and naturally related to the RD theory. RD theoretic interpretations have been obtained also for problems such as classification and sequence prediction.

In this study, we provide a RD theoretic view of Bayesian learning. We formulate a rate-distortion problem by the distortion measure defined by the pointwise regret of the model parameter θ,

$$d(x^n, \theta) = \log \frac{p(x^n|\hat{\theta}(x^n))}{p(x^n|\theta)},$$

where x^n is the data set and $\hat{\theta}$ is the maximum likelihood estimator. We show that the generalized Bayesian posterior distribution appears as the optimal solution to the problem. The generalized posterior distribution has been used for purposes such as the posterior consistency [2] and model selection [3].

We also discuss the connection of this view to the asymptotic theory of Bayesian learning, which characterizes the generalization error of a learning machine by a constant called learning coefficient [4]. It is demonstrated that the learning coefficient provides an upper bound of the RD dimension [5], which is defined by the asymptotic behavior of the rate-distortion function. Let λ be the learning coefficient. It is known that λ is the leading term of the negative log-marginal likelihood, and if the model is regular, $2\lambda = m$, the dimension of the parameter, whereas if it is non-regular, $2\lambda \leq m$ holds in general [4]. The asymptotic behavior of the RD function of the above problem shows that twice the learning coefficient provides an upper bound of the RD dimension,

$$\dim_{RD} \leq 2\lambda$$

which is defined by the RD function $R(D)$ as $\dim_{RD} = \lim_{D\to 0} \frac{R(D)}{-\frac{1}{2}\log D}$.

*豊橋技術科学大学: wkazuho@cs.tut.ac.jp

We also provide a RD theoretic interpretation of the Dirichlet process (DP) means clustering, which is a simple extension of the K-means clustering and estimates the number of clusters from data [6]. More specifically, DP-means runs the usual K-means with $K = 1$, and generates a new cluster when the distance from a data point to its nearest cluster center is larger than a constant η, which is called the penalty parameter and prespecified by the user.

From an RD theoretic point of view, the penalty parameter can be considered as controlling the maximum distortion in the training data set $\{x_1, \cdots, x_N\} (x_i \in R^L)$ [7],

$$\max_i d(x_i, \theta_{c(i)}),$$

where $c(i)$ denotes the cluster label of the ith data point and $\theta_1, \cdots, \theta_{\hat{K}}$ are the cluster centers obtained by the algorithm. The logarithm of the estimated number of clusters per dimension of data, $\log \hat{K}/L$, can be considered as the rate. The RD theory of the maximum distortion criterion [8] implies that

$$\frac{\log \hat{K}}{L} \to R(\eta)$$

as $N \to \infty$ and $L \to \infty$, where R is the RD function of the distortion measure used in the algorithm [7].

References

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] T. Zhang, "Information theoretical upper and lower bounds for statistical estimation," *IEEE Trans. on Inform. Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[3] M. Aitkin, "Posterior Bayes factors," *Journal of the Royal Statistical Society. Series B*, vol. 53, no. 1, pp. 111–142, 1991.

[4] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

[5] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. on Inform. Theory*, vol. 40, no. 5, pp. 1564–1572, 1994.

[6] B. Kulis and M. I. Jordan, "Revisiting k-means: new algorithms via Bayesian nonparametrics," in *Proc. of ICML*, 2012, pp. 513–520.

[7] M. Kobayashi and K. Watanabe, "A rate-distortion theoretic view of Dirichlet process means clustering," *IEICE Trans. on Fundamentals*, submitted.

[8] T. S. Han, *Information-Spectrum Methods in Information Theory*, Springer Berlin Heidelberg, 2003.