科研費シンポジウム「統計学と機械学習における数理とモデリング」報告書

日時：２０１６年2月２１日（日）～2月２２日（月）
場所：東京工業大学大岡山キャンパス西八号館Ｗ８３３号室
開催責任者：鈴木大慈・野村俊一（東京工業大学）

本シンポジウムは下記の助成を受け開催された。
科学研究費 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）

【プログラム】
１日目　２０１６年2月２１日（日）

13:25～13:30：オープニング

13:30～15:30：機械学習と計算量セッション
1. Nonlinear Feature Selection for Large and Ultra High-dimensional data　山田 誠（京都大学）
2. Bayesian Network の構造学習における分枝限定法の適用に関する考察　岩本 俊亮（大阪大学）
3. 部分モニタリング問題における漸近最適アルゴリズム　小宮山 純平（東京大学）
4. マルコフ連鎖モンテカルロ法の高次元漸近論　鎌谷 研吾（大阪大学）※病欠により講演はキャンセル

16:00～18:00：機械学習理論セッション
5. 変分鞍点近似によるベイズ推論とモデル選択　林 浩平（国立情報学研究所）
6. グラフ表現を用いた異常変数同定における一致推定量　原 聡（IBM）
7. Non-convex Compressed Sensing with the Sum-of-Squares Method　相馬 輔（東京大学）
8. 教師付き学習における Barron and Cover 理論　川喜田 雅則（九州大学）

２日目　２０１６年2月２２日（月）

10:00～12:00：統計理論セッション
9. 状態空間モデルを用いた時系列の周期成分分解と位相推定　松田 孟留（東京大学）
10. 関数予測における漸近ミニマックス予測分布　矢野 恵佑（東京大学）
11. 低次元多様体上の Data Sharpening　工藤 雅紀（島根大学）
12. 確率的交互方向乗数法とマルチクラスグラフ型正則化学習への応用　鈴木 大慈（東京工業大学）

13:30〜15:30：ベイズセッション

13. 余震活動の確率予測：ベイズ統計を用いたアプローチ　近江 崇宏（東京大学）

14. 点過程モデルによる地震活動異常の解析と群発地震への応用　熊澤 貴雄（統計数理研究所）

15. 統計的言語モデルに基づく化学構造の特徴抽出と分子設計への応用　池端 久貴（総合研究大学院大学）

16. Bay-Bridge：local linear approximated bridge and its model building procedure via the Bayesian model　保科 架風（中央大学）

16:00〜17:30：生体ログデータ分析セッション

17. 潜在的歩数増加パターンと体組成指標との統計的関連性についての分析　大橋 洸太郎（立教大学）

18. 生体ログデータを使用した、歩行パターンと体組成指標との関連を検討する分析方法の提案　渡辺 真弓（慶應義塾大学）

19. 体型を示す新たな体組成指標の提案とそのダイナミクス　野村 俊一（東京工業大学）

17:30〜：クロージング

Nonlinear Feature Selection for Large and Ultra High-dimensional Data

Makoto Yamada
Kyoto University
myamada@kuicr.kyoto-u.ac.jp

January 18, 2016

Abstract

We propose *nonlinear non-negative least-angle regression* (N^3LARS), that can select non-redundant features from a large and *ultra* high-dimensional data in nonlinear way. A key advantage of N^3LARS is that it is a convex method and can find a globally optimal solution. Moreover, N^3LARS can easily incorporate with map-reduce frameworks such as Hadoop and Spark. Thus, with the help of distributed computing, a set of features can be efficiently selected from a large and ultra high-dimensional data. The effectiveness of the proposed method is demonstrated through an ultra high-dimensional biology dataset.

1 Problem Formulation

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]^\top \in \mathbb{R}^{d \times n}$ denotes the input data, $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ denotes the output data or labels and suppose that we are given n independent and identically distributed (i.i.d.) paired samples $\{(\boldsymbol{x}_i, y_i) \mid \boldsymbol{x}_i \in \mathcal{X},\ \ y_i \in \mathcal{Y},\ i = 1, \ldots, n\}$ drawn from a joint distribution with density $p(\boldsymbol{x}, y)$. Note, $\mathcal{Y}$ could be either continuous (i.e., regression) or categorical (i.e., classification). For the purposes of this paper consider $\boldsymbol{x}$ to be a dense vector.

The goal of supervised feature selection is to find m features ($m < d$) of input vector $\boldsymbol{x}$ that are responsible for predicting output y.

2 Nonlinear Non-Negative LARS (N^3LARS)

The optimization problem of N^3LARS is given as

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^d} \ \|\widetilde{\boldsymbol{L}} - \sum_{k=1}^{d} \alpha_k \widetilde{\boldsymbol{K}}^{(k)}\|_{\mathrm{Frob}}^2,\ \|\boldsymbol{\alpha}\|_1 \leq \eta,\ \text{s.t. } \alpha_1, \ldots, \alpha_d \geq 0, \tag{1}$$

where $\widetilde{\boldsymbol{K}} = \bar{\boldsymbol{K}}/\|\bar{\boldsymbol{K}}\|_{\mathrm{Fro}}$, $\widetilde{\boldsymbol{L}} = \bar{\boldsymbol{L}}/\|\bar{\boldsymbol{L}}\|_{\mathrm{Fro}}$ such that $\mathbf{1}_n^\top \widetilde{\boldsymbol{K}} \mathbf{1}_n = 0$, $\mathbf{1}_n^\top \widetilde{\boldsymbol{L}} \mathbf{1}_n = 0$, and $\|\widetilde{\boldsymbol{K}}\|_{\mathrm{Fro}}^2 = 1.$, $\bar{\boldsymbol{K}}^{(k)} = \boldsymbol{\Gamma} \boldsymbol{K}^{(k)} \boldsymbol{\Gamma}$ and $\bar{\boldsymbol{L}} = \boldsymbol{\Gamma} \boldsymbol{L} \boldsymbol{\Gamma}$ are centered Gram matrices, $\boldsymbol{K}_{i,j}^{(k)} = K(x_{k,i}, x_{k,j})$ and $\boldsymbol{L}_{i,j} = L(y_i, y_j)$ are Gram matrices, $K(x, x')$ and $L(y, y')$ are kernel functions, $\boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ is the centering matrix, $\boldsymbol{I}_n$ is the n-dimensional identity matrix, $\mathbf{1}_n$ is the n-dimensional vector with all ones, $\eta \geq 0$ is the regularization parameter, and $\|\cdot\|_{\mathrm{Frob}}$ is the Frobenius norm.

It is worth pointing out the objective function of N^3LARS can be reframed as

$$1 - 2 \sum_{k=1}^{d} \alpha_k \mathrm{NHSIC}(\boldsymbol{u}_k, \boldsymbol{y}) + \sum_{k,l=1}^{d} \alpha_k \alpha_l \mathrm{NHSIC}(\boldsymbol{u}_k, \boldsymbol{u}_l),$$

where $\mathrm{NHSIC}(\boldsymbol{u}, \boldsymbol{y}) = \mathrm{tr}(\widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{L}})$ is the normalized version of Hilbert-Schmidt Independence Criterion (HSIC) (Gretton *et al.*, 2005) and $\mathrm{NHSIC}(\boldsymbol{y}, \boldsymbol{y}) = 1$. Note that NHSIC always takes a non-negative value, and is zero if and only if two random variables are statistically independent when a *universal reproducing kernel* (Steinwart, 2001) such as the Gaussian kernel is used.

If the k-th feature $\boldsymbol{u}_k$ has high dependence on output $\boldsymbol{y}$, $\mathrm{NHSIC}(\boldsymbol{u}_k, \boldsymbol{y})$ becomes a large value and thus α_k should also be large. On the other hand, if $\boldsymbol{u}_k$ and $\boldsymbol{y}$ are independent, $\mathrm{NHSIC}(\boldsymbol{u}_k, \boldsymbol{y})$ is close to zero; $\boldsymbol{u}_k$ tends to be not selected by the ℓ_1-regularizer. Furthermore, if $\boldsymbol{u}_k$ and $\boldsymbol{u}_l$ are strongly dependent (i.e., redundant features), $\mathrm{NHSIC}(\boldsymbol{u}_k, \boldsymbol{u}_l)$ is large and thus either of α_k and α_l tends to be zero. That is, non-redundant features that have a strong dependence on output $\boldsymbol{y}$ tend to be selected by the N^{3}LARS. Thus, N^{3}LARS can be regarded as a widely used *minimum redundancy maximum relevance* (mRMR) feature selection method (Peng *et al.*, 2005).

2.1 Nyström Approximation for NHSIC

We start introducing kernels used in this paper.

For input x, we first normalize the input x to have unit standard deviation, and then use the Gaussian kernel, $K(x, x') = \exp\left(-\frac{(x-x')^2}{2\sigma_{\mathrm{x}}^2}\right)$, where σ_{x} is the Gaussian kernel width. In regression cases (i.e., $y \in \mathbb{R}$), we normalize an output y to have unit standard deviation, and then use the Gaussian kernel, $L(y, y') = \exp\left(-\frac{(y-y')^2}{2\sigma_{\mathrm{y}}^2}\right)$, where σ_{y} is the Gaussian kernel width. In classification cases (i.e., y is categorical), we use the delta kernel for y, $L(y, y') = \begin{cases} 1/n_y & \text{if } y = y', \\ 0 & \text{otherwise,} \end{cases}$ where n_y is the number of samples in class y.

Nyström approximation: To reduce the computational cost of Gram matrices, we use the Nyström approximation for NHSIC as

$$\begin{aligned} \mathrm{NHSIC}(\boldsymbol{u}, \boldsymbol{y}) &= \mathrm{tr}((\boldsymbol{F}^\top \boldsymbol{G})^2), \\ \boldsymbol{F} &= \boldsymbol{\Gamma} \boldsymbol{K}_{nb} \boldsymbol{K}_{bb}^{-1/2} / (\mathrm{tr}((\boldsymbol{K}_{bb}^{-1/2} \boldsymbol{K}_{nb}^\top \boldsymbol{K}_{nb} \boldsymbol{K}_{bb}^{-1/2})^2))^{1/4}, \\ \boldsymbol{G} &= \boldsymbol{\Gamma} \boldsymbol{L}_{nb} \boldsymbol{L}_{bb}^{-1/2} / (\mathrm{tr}((\boldsymbol{L}_{bb}^{-1/2} \boldsymbol{L}_{nb}^\top \boldsymbol{L}_{nb} \boldsymbol{L}_{bb}^{-1/2})^2))^{1/4}, \end{aligned}$$

where $\boldsymbol{K}_{nb} \in \mathbb{R}^{n \times b}$, $[\boldsymbol{K}_{nb}]_{ij} = K(u_i, u_{b,j})$, $\boldsymbol{K}_{bb} \in \mathbb{R}^{b \times b}$, $[\boldsymbol{K}_{bb}]_{ij} = K(u_{b,i}, u_{b,j})$, $\boldsymbol{L}_{nb} \in \mathbb{R}^{n \times b}$, $\boldsymbol{L}_{bb} \in \mathbb{R}^{b \times b}$, and b is the number of basis function.

2.2 Distributed computation with Map-Reduce

The Nyström approximation is useful for large n. However, for ultra high-dimensional cases (e.g., $d > 10^6$), the computation cost of N^{3}LARS is still large. To deal with this issue, we separately compute c_k in N^{3}LARS in parallel with distributed computing such as Spark[1] . Since we can independently compute c_k, we can easily speed up N^{3}LARS by simply increasing the number of servers.

References

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1237.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012). Feature selection via dependence maximization. *JMLR*, **13**, 1393–1434.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *JMLR*, **2**, 67–93.

[1] `https://spark.apache.org/`

BDスコアによるベイジアンネットワーク構造学習の効率化

岩本 俊亮　　鈴木 譲

大阪大学大学院理学研究科

本研究ではベイジアンネットワーク (BN) の構造学習における BD スコアの新たな上界を導き，構造学習にかかる計算量をこの上界を用いることで削減できることを示した．

離散型確率変数 $\boldsymbol{X} = (X_1, \ldots, X_n)$ について，各変数を頂点とする有向非巡回グラフ (Directed Acyclic Graph, DAG)B_s を考える．このとき BN の構造 B_s は同時分布 $P(X_1, \ldots, X_n)$ の条件付き分布による因数分解を与える．つまり，Π_i を X_i の親 (依存対象の変数，DAG における X_i を指す矢印の元にある変数) の集合とすると $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\Pi_i)$ となる．$B_p = \{P(X_i|\Pi_i)\}_{i=1}^{n}$ とすると，BN は (B_s, B_p) で定義することができる．ただし，条件付き分布による因数分解が同じものを表現する DAG は区別しない．

$\boldsymbol{X} = (X_1, \ldots, X_n)$ について，データ $D = \{C_1, \ldots, C_m\}$ を与える．このとき新たなデータ C が生じる予測分布は $P(C|D) = \sum_{all B_s} P(C|B_s) \cdot P(B_s|D)$ となる．ベイズ的アプローチによる BN の構造学習とは事後確率 $P(B_s|D)$ が最大となる B_s を探索すること．ある (B_s, B_p) について，$r_i = |X_i|$，$q_i = |\Pi_i| = \prod_{X_j \in \Pi_i} r_j$ とし，X_i の k 番目の値を y_k，Π_i の j 番目の組を w_j で表す．また，$\theta_{ijk} = P(X_i = y_k|\Pi_i = w_j)$，$\Theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijr_i})$，$\Theta_i = (\Theta_{i1}, \ldots, \Theta_{iq_i})$，$\Theta_{B_s} = (\Theta_1, \ldots, \Theta_n)$ とする．このとき，Θ_{ij} がディリクレ分布であること，つまり確率密度関数が $\rho(\Theta_{ij}|B_s) = \frac{\Gamma(N'_{ij})}{\prod_{k=1}^{r_i}\Gamma(N'_{ijk})} \cdot \prod_{k=1}^{r_i} {\theta_{ijk}}^{N'_{ijk}-1} (N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$，$N'_{ijk}$ はディリクレ分布のパラメータ) となることを仮定すると

$$\begin{aligned} P(D, B_s) &= P(B_s) \int P(D|\Theta_{B_s}, B_s)\rho(\Theta_{B_s}|B_s)d\Theta_{B_s} \\ &= P(B_s) \cdot \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \end{aligned} \tag{1}$$

を得る．ここで $N_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$，$N_{ijk}$ は $X_i = y_k$ かつ $\Pi_i = w_j$ となるデータの数である．この同時確率 $P(D, B_s)$ を BD(Bayesian Dirichlet) スコアという．また，式 (1) において $N_{ijk} = \frac{\epsilon}{q_i r_i}$ とするとき，同時確率 $P(D, B_s)$ を BDeu スコアという．ϵ は定数で，ESS(Equivalent Sample Size) という．事後確率 $P(B_s|D)$ が最大となる B_s については同時確率 $P(D, B_s)$ も最大となるので，実際の構造学習では同時確率が最大となるような B_s を探索する．これを BD スコア (BDeu スコア) による構造学習という．

式 (1) について，

$$Q(X_i, \Pi_i) = \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}$$

を局所 BD スコアという．局所 BD スコアについては次の定理が成り立つ．

定理 1 離散型確率変数 X について，Π_X, Π'_X を X の親集合で $\Pi_X \subset \Pi'_X$ なるものとする．

$$Q(X, \Pi_X) > Q(X, \Pi'_X) \tag{2}$$

となるとき，Π'_X を X の親集合とした DAG による BD スコアは最大とならない．

BD スコアによる構造学習のアルゴリズムとしては Dynamic Programming，Branch and Bound，AStar などがあるが，これらは

1. 局所 BD スコアの計算
2. BD スコアを最大にする DAG の探索

というステップからなる．1 にかかる計算量は $n \cdot 2^{n-1}$ であるが，定理 1 式 (2) 右辺の局所 BD スコアの代わりに上界を比べることで，特定の条件下では局所 BD スコアの計算量を減らすことができる．また，この上界は小さいほどよい．de Campos と Ji は局所 BDeu スコアに対して上界を導き [1]，鈴木はこの上界を局所 BD スコアに一般化した [2]．定理 2 はその上界である．

定理 2 X を与えられたデータの i 番目の確率変数，Π_{X_i} を X_i の親集合とする．このとき，

$$Q(X_i, \Pi_{X_i}) \leq \prod_{j=1}^{q_i} \frac{\max N'_{ijk}}{N'_{ij}}$$

が成り立つ．

本研究ではこれよりも小さい上界を導いた．

主定理 X を与えられたデータの i 番目の確率変数，Π_{X_i} を X_i の親集合とする．このとき，

$$Q(X_i, \Pi_{X_i}) \leq \prod_{j=1}^{q_i} \prod_{\alpha=0}^{N_{ij}-1} \frac{\alpha + \max N'_{ijk}}{\alpha + N'_{ij}}$$

が成り立つ．

この上界により BDeu スコアによる構造学習において計算実験を行い，定理 2 の上界を用いた場合と比較したところ，効率化できることが確認できた．

参考文献

[1] de Campos, C. P., & Ji, Q, "Properties of Bayesian Dirichlet scores to learn Bayesian network structures", Fox, M., & Poole, D. (Eds.), AAAI, pages 431-436, (2010).

[2] Suzuki, J., "Efficiently Learning Bayesian Network Structures Based on the B&B Strategy: A Theoretical Analysis", Advanced Methodologies for Bayesian Networks 2015, pages 1-14, (2015).

部分モニタリング問題における漸近最適アルゴリズム

小宮山純平（東京大学）

1　部分モニタリング問題

以下の定式化は，基本的に部分モニタリング問題の既存研究 [1] に準拠している。部分モニタリング問題は，繰り返しゲームとして定義される。各ラウンド $t = 1, \ldots, T$ に学習者 (アルゴリズム) はアクション $i(t) \in [N] := \{1, \ldots, N\}$ を選択し，また，結果 $j(t) \in [M] := \{1, \ldots, M\}$ が未知の確率分布 $p^* \in \mathcal{P}_M$($\mathcal{P}_M$ は M 次元確率単体）から引かれる。最初のラウンドの開始前に，損失行列 $L = (l_{i,j}) \in \mathbb{R}^{N \times M}$ とフィードバック行列 $H = (h_{i,j}) \in [A]^{N \times M}$（ここで $[A]$ は有限個のシグナル集合）が学習者に与えられる。各ラウンドごとに，学習者は損失 $l_{i(t),j(t)} \in \mathbb{R}$ を被り，シグナル $h_{i(t),j(t)} \in [A]$ を受け取る。ここで，結果や損失について学習者は知ることができず，学習者が知りえる情報はシグナルが $[A]$ のいずれであったかという情報のみである。つまり，結果に関する部分的な情報が（アクションに依存して）与えられる状況で，如何にして結果の分布 p^* を学習するかという問題である。

学習者の目的は，T ラウンドを通した損失の最小化である。結果の分布に対して，損失の期待値を最小化するようなアクションを多く選択したい。このアクションは，$i^* = \arg\min_{i \in [N]} L_i^\top p^*$ と書くことができる。ここで，L_i は L の i 行目である. ここで，i^* は unique だと仮定し，一般性を失わずに $i^* = 1$ とする. ここで，$\Delta_i = (L_i - L_1)^\top p^* \in \mathbb{R}^+$ と定義し，$N_i(t)$ を t ラウンド目までにアクション i を選択した数とすると，損失の最小化問題は，以下の Regret（最適なアクションを選択し続けた場合とアルゴリズムの期待損失の差分）の最小化問題と等価である。

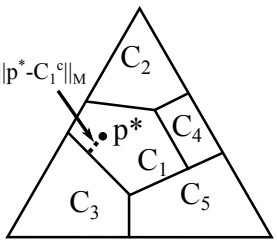


図 1: $N = 5, M = 3$ でのセル分解での一例.

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \Delta_{i(t)} = \sum_{i \in [N]} \Delta_i N_i(T+1),$$

この Regret のうち，本研究ではラウンド数 T を十分大きくとったときの挙動に興味がある。

部分モニタリング問題での結果分布に関し，セル分解が定義できる。アクション i が最適である結果分布の集合

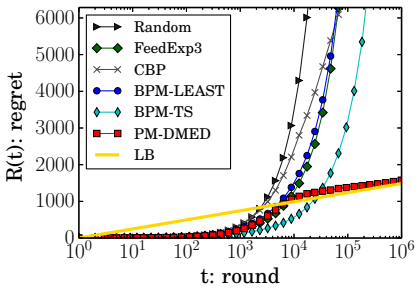
$$\mathcal{C}_i = \{q \in \mathcal{P}_M : \forall_{j \neq i} (L_i - L_j)^\top q \leq 0\}.$$

を i の最適セルと呼ぶ。すべてのアクションの最適セルは閉凸多面体であり，結果の分布の仮説空間である $\mathcal{P}_M$ の各セルへの分解（図 1）が定義できる。

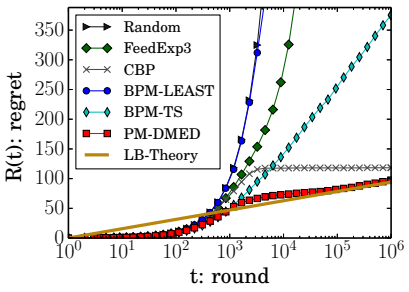
最適なアクションが発見できることの必要十分条件は，２つのアクションのどちらが期待損失が大きいか，すべてのアクションから得られるフィードバックを総合して判断可能であることである。つまり，任意のアクションのペア i, j に対し，$L_i - L_j \in \oplus_{k \in [N]} \mathrm{Im} S_k^\top$ である（大域観測可能（globally observable））ことである。

これまで，大域観測可能な問題のうち一部の問題に関しては Regret が $\Theta(\log T)$ であることが知られている [2]。本研究では，大域観測可能なすべての問題に関して Regret が $\Omega(\log T)$ であることを示す。さらに，その $\log T$ の定数を明らかにし，また漸近的に最適性能のアルゴリズムを提案する。全ての証明や一部の定理，実験詳細などは省略するが，これらに興味がある方は著者らの論文 [3] を参照されたい。

上記のように単純な問題設定であるが，オンライン広告のクリック確率の最適化や実験計画の最適化のモデルである確率的バンディット問題，検索エンジンのランキング最適化のモデルである効用ベース比較バンディット問題，繰り返しオークションにおける価格最適化問題など，実用上広いクラスの問題が部分モニタリング問題に属することが分かっており，この問題に関する効率的なアルゴリズムには大きな価値があると考える。



(a) 繰り返しオークション問題



(b) 確率的バンディット問題

図 2: Regret-round 片対数プロット．LB は Regret 漸近下界，Random は一様ランダムなアクション選択，FeedExp3, CBP, BPM-LEAST, BPM-TS は既存手法である．

2 Regret 下限

任意の結果分布 $p \in \mathcal{P}_M$ と $a > 0$ に関して $\mathbb{E}[\mathrm{Regret}(T)] = \mathrm{o}(T^a)$ を満たすアルゴリズムを強一致なアルゴリズムと定義する。強一致なアルゴリズムは，全ての真に最適ではないアームが実は最適であったという仮説を各ラウンド t に $1/t$ 程度の有意水準で棄却する必要がある。

Lemma 1. 強一致なアルゴリズムに関して以下が成立する。

$$\forall_{q \in \mathcal{C}_1^c} \sum_{i \in [N]} \mathbb{E}[N_i(T)] D(p_i^* \| S_i q) \geq \log T - \mathrm{o}(\log T),$$

ここで，p_i^*，$S_i q$ は，真の p^* および仮説 q における，アクション i を選択したときのシグナルの確率分布である。また，$D(p\|q) = \sum_i (p)_i \log\left((p)_i/(q)_i\right)$ は 2 つの分布間の KL ダイバージェンスである。

Regret の最小化解は本稿では省略するが，上記の制約を満たしつつ $\mathrm{Regret}(T)$ を最小化するようなアームの選択方法から，$C^* \log T$（C^* は最適係数）の形で導出できる．

3 PM-DMED アルゴリズム

アームを引く回数に関して，Lemma 1 の制約を満たしつつ $\mathrm{Regret}(T)$ を満たすという問題は，真の分布が分かっていれば半無限計画（LSIP）を解くことによって求められる．しかし，真の確率分布はもちろん学習者にとって未知である．本研究では，この問題を確率分布に関しての経験推定値をプラグインすることで解く partial monitoring deterministic minimum empirical divergence (PM-DMED) アルゴリズムを提案する．詳細は [3] を参照されたい．推定精度を上げるための項や，推定失敗からの復帰の項などいくつかの理論保証のための項を含めることによって，このアルゴリズムは T に関して下界と漸近一致する Regret を持つ．

繰り返しオークションおよび確率的バンディット問題におけるアルゴリズムの計算機シミュレーションでの経験的な性能を図 3 に示す．図から明らかであるように，既存手法の Regret が漸近下界よりかなり大きいのに対し，提案手法である PM-RMED は漸近に下界と一致する Regret の傾きを持つ．

参考文献

[1] Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014.

[2] Gábor Bartók, Navid Zolghadr, and Csaba Szepesvári. An adaptive algorithm for finite stochastic partial monitoring. In *ICML*, 2012.

[3] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *NIPS*, 2015.

マルコフ連鎖モンテカルロ法の高次元漸近論

大阪大学大学院基礎工学研究科，数理・データ科学教育研究センター　　鎌谷研吾

マルコフ連鎖モンテカルロ (MCMC) 法のより良いデザインのための高次元漸近論の枠組みと，高次元で有効な MCMC 法を紹介する．MCMC 法の高次元漸近論は 1980 年台後半から始まり，ベイズ統計学においても理論だけでなくその実用性が注目され，今も活発に研究がなされている．

$P_d(\mathrm{d}x) = p_d(x)\mathrm{d}x$ は $\mathbb{R}^d$ 上の確率分布とする．最も基本的な MCMC 法であるランダムウォーク型メトロポリス (RWM) 法は次の手続きで定義された．ある $x \in \mathbb{R}^d$ を出発点として以下を繰り返す：

$$\left\{ \begin{array}{lcl} x^* & \leftarrow & x + w,\ w \sim N_d(0, \sigma^2 I_d), \\ x & \leftarrow & x^* \text{ with probabilty } \alpha(x, x^*) \end{array} \right.$$

ただし，

$$\alpha(x, x^*) = \min\left\{1, \frac{p_d(x^*)}{p_d(x)}\right\}.$$

このステップにより d 次元のマルコフ連鎖 $\{X_m^d\}_m$ が定義される．このマルコフ連鎖は $P_d(\mathrm{d}x)$ を不変分布として持つように出来ていて，例えば $p_d(x) > 0$ $(x \in \mathbb{R}^d)$ であればエルゴード的になり，下記右辺の積分が存在する限り大数の法則

$$\frac{1}{M}\sum_{m=1}^{M} f(X_m^d) \to \int f(x) P_d(\mathrm{d}x) \text{ a.s. } (M \to \infty) \tag{1}$$

が成り立つ [7]．したがって右辺の積分の近似値として左辺の有限和を用いることが正当化されるわけである．

次元が高くなるに連れ，(1) の右辺の数値積分は極めて困難になり，MCMC 法の意義が出る．しかし，高次元での計算負荷の増大は MCMC 法でも不可避である．これは $d \to \infty$ としたときの $\{X_m^d\}_m$ の振る舞いを見ることでその一端を説明できる．このような次元の極限の議論の枠組みを与えるのが高次元漸近論である．高次元漸近論で様々なことがわかる．

1. P_d が正規分布に十分近い時は，採択率がおおよそ 23% 程度であるように $\sigma > 0$ を調整するのが良いと示唆される [5].
2. P_d の裾の形状に依存するが，RWM 法よりも収束レートの良い MCMC 法が存在する [6].
3. P_d の裾が軽ければ，MCMC の繰り返し回数は $O(d)$ 回必要であり，裾が重い場合は $O(d^2)$ 回必要になる [3].

本発表では高次元漸近論を使って，高次元でも有効な MCMC 法のデザインを考える．上述のように高次元では MCMC 法の繰り返し回数を長くしなければならないが，加えて一回の $p_d(x)$ の値の計算量も d に依存する．従って高次元で MCMC 法を適用する際は，計算量を抑える工夫が不可欠である．

RWM 法が困難なほどの高次元積分では，ラプラス変換を援用する方法や，並列化を用いる方法など様々な方法が存在する．本発表では少ない繰り返し回数で収束する MCMC 法を考える．近年，エルゴード的な提案カーネルを使った preconditioned Crank-Nicolson (pCN; [1]) が理論的に注目されている．本発表では pCN 法を改良した mixed preconditioned Crank-Nicolson (MpCN; [2]) 法を紹介する．MpCN 法は高次元漸近論に基づいてデザインされた有効な手法である：$\rho \in (0, 1)$ とする．

$$\left\{ \begin{array}{lcl} r & \sim & \mathrm{Gamma}(d/2, \|x\|^2/2), \\ x^* & \leftarrow & \rho^{1/2} x + (1-\rho)^{1/2} r^{-1/2} w, w \sim N_d(0, \sigma^2 I_d), \\ x & \leftarrow & x^* \text{ with probabilty } \alpha(x, x^*) \end{array} \right.$$

ただし，

$$\alpha(x,y) = \min\left\{1, \frac{p_d(y)\|x\|^{-d}}{p_d(x)\|y\|^{-d}}\right\}.$$

この手法がなぜ有効であるか，どれくらい従来手法を改善するかを数値実験と高次元漸近論を用いて考察する．高次元漸近論から示唆される，MCMC 法デザインの一般的な注意も述べる．MpCN 法を実際に使う際の注意についても述べたい．本発表は論文 [2, 4] にもとづいている．

本研究は JSPS 科研費 24740062 及び CREST, JST の助成を受けている．

参考文献

[1] Beskos, A., Roberts, G., Stuart, A., and Voss, J. MCMC methods for diffusion bridges. *Stoch. Dyn. 8*, 3 (2008), 319–350.

[2] Kamatani, K. Efficient strategy for the Markov chain Monte Carlo in high-dimension with heavy-tailed target probability distribution. *Arxiv* (2014).

[3] Kamatani, K. Rate optimality of Random walk Metropolis algorithm in high-dimension with heavy-tailed target distribution. *Arxiv* (2014).

[4] Kamatani, K., and Uchida, M. Hybrid multi-step estimators for stochastic differential equations based on sampled data. *Statistical Inference for Stochastic Processes 18*, 2 (2015), 177–204.

[5] Roberts, G. O., Gelman, A., and Gilks, W. R. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab. 7*, 1 (1997), 110–120.

[6] Roberts, G. O., and Rosenthal, J. S. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci. 16*, 4 (2001), 351–367.

[7] Tierney, L. Markov chains for exploring posterior distributions. *Ann. Statist. 22*, 4 (1994), 1701–1762. With discussion and a rejoinder by the author.

変分鞍点近似によるベイズ推論とモデル選択

国立情報学研究所，JST ERATO 河原林巨大グラフプロジェクト
林浩平

January 21, 2016

高次元データに対してなんらかの構造を仮定し，低次元の特徴を取り出すことはデータ解析においてしばしば行われている．例えば多数の文書から単語のクラスタリングを行うトピック抽出は，単語の生成確率がそれぞれの文書で異なる重み付けを持つ (しかし基底となる分布は全文書で共有される) 混合分布に従うという仮定のもとでモデリングし，パラメータ推定を行う．このような例は次元圧縮，ネットワーク分割，特徴抽出などさまざまな機械学習の問題に頻出する．

このような問題を統一的に扱う枠組みの 1 つとして，隠れ変数モデルを用いたアプローチが存在する．N 個のサンプルからなるデータを $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ とする．隠れ変数モデルは個々のサンプル $\mathbf{x}_n$ を隠れ変数 $\mathbf{z}_n$ とパラメータ $\mathbf{\Theta}$ で表現する．さきほどのトピック抽出の例を用いると $\mathbf{x}_n$ が文書 n の bag-of-words 表現，$\mathbf{\Theta}$ が基底となる単語分布，$\mathbf{z}_n$ が文書 n が従う単語分布を構成する混合分布の重みとなる．ベイズ統計では隠れ変数モデルは同時確率分布

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{\Theta} \mid K) = p(\mathbf{X} \mid \mathbf{Z}, \mathbf{\Theta}, K) p(\mathbf{Z} \mid \mathbf{\Theta}, K) p(\mathbf{\Theta} \mid K) \tag{1}$$

として表現され (K は $\mathbf{Z}$ の次元)，隠れ変数やパラメータはベイズ推論によって推定される．

隠れ変数モデルを扱う上で問題となるのが

- ベイズ推論の計算困難性
- K をどう決めるか

の 2 点である．まずベイズ推論に関しては同時確率分布 (1) の積分計算を解く必要があるが，この計算は一般的には解析解を持たないためなんらかの近似計算を行う必要がある．また K の決めかたとして，周辺尤度

$$p(\mathbf{X} \mid K) = \int \left\{ \int p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}) p(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta} \right\} \mathrm{d}\mathbf{Z} \tag{2}$$

が最大となる K を選択するベイズモデル選択が存在するが，ここの積分計算も一般に計算困難である．

この 2 つの問題点を解決する新しい方法として，因子化漸近ベイズ推論法が近年提案された [1]．因子化漸近ベイズ推論の基本的なアイデアは変分法と鞍点近似を組み合わせた点にある．まず変分法を使うと周辺尤度は

$$\log p(\mathbf{X} \mid K) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z} \mid K)] + H(q) + \mathrm{KL}(q \| p^*) \tag{3}$$

と変形できる．ここで $q(\mathbf{Z})$ は $\mathbf{Z}$ の任意の分布，$H(q)$ はエントロピー，$\mathrm{KL}(q\|p^*)$ は KL 距離，$p^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, K)$ は周辺事後分布である．さらに $p(\mathbf{X}, \mathbf{Z} \mid K) = \int p(\mathbf{X}, \mathbf{Z}, \mathbf{\Theta} \mid K)\mathrm{d}\mathbf{\Theta}$ に対して鞍点近似を用いることで，ある仮定を満たすとき，以下のように精度のよい近似が得られる：

$$\log p(\mathbf{X}, \mathbf{Z} \mid K) = \log p(\mathbf{X}, \mathbf{Z} \mid \hat{\mathbf{\Theta}}, K) + \log p(\hat{\mathbf{\Theta}}) - \frac{1}{2}\log\det|\mathbf{F}| - \frac{|\mathbf{\Theta}|}{2}\log\frac{N}{2\pi} + O(\frac{1}{N}). \quad (4)$$

ここで $\hat{\mathbf{\Theta}}$ は最尤推定量，$\mathbf{F}$ は $\hat{\mathbf{\Theta}}$ まわりでの $-\log p(\mathbf{X}, \mathbf{Z} \mid \hat{\mathbf{\Theta}}, K)$ のヘッセ行列，$|\mathbf{\Theta}|$ はパラメータの次元を表すとする．この結果 (4) を (3) に代入することで周辺尤度の近似が得られる．この近似はもとの周辺尤度と異なり計算困難な積分計算を持たないか，あるいは持っていたとしてもある程度タイトな下限を構築することができ，反復法による推定アルゴリズムで解くことができる，

因子化漸近ベイズ推論法はもともと混合分布 [1] に対して提案されたが，その後隠れマルコフモデル [2]，混合エキスパート [3, 4]，スパース特徴選択 [5]，関係データモデル [6, 7, 8] などさまざまなモデルに対して拡張された．本発表ではその基本的なアイデアと応用事例について紹介する．

References

[1] Ryohei Fujimaki and Satoshi Morinaga. Factorized asymptotic bayesian inference for mixture modeling. In *AISTATS*, 2012.

[2] Ryohei Fujimaki and Kohei Hayashi. Factorized asymptotic bayesian hidden markov model. In *International Conference on Machine Learning (ICML)*, 2012.

[3] Riki Eto, Ryohei Fujimaki, Satoshi Morinaga, and Hiroshi Tamano. Fully-automatic bayesian piecewise sparse linear models. In *AISTATS*, 2014.

[4] Jialei Wang, Ryohei Fujimaki, and Yosuke Motohashi. Trading interpretability for accuracy: Oblique treed sparse additive models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[5] Yohei Kondo, Kohei Hayashi, and Shin ichi Maeda. Bayesian masking: Sparse bayesian estimation with weaker shrinkage bias. In *7th Asian Conference on Machine Learning (ACML)*, 2015.

[6] Kohei Hayashi and Ryohei Fujimaki. Factorized asymptotic bayesian inference for latent feature models. In *27th Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.

[7] Kohei Hayashi, Shin-ishi Maeda, and Ryohei Fujimaki. Rebuilding factorized information criterion: Asymptotically accurate marginal likelihood. In *International Conference on Machine Learning (ICML)*, 2015.

[8] Chunchen Liu, Lu Feng, Ryohei Fujimaki, and Yusuke Muraoka. Scalable model selection for large-scale factorial relational models. In *International Conference on Machine Learning (ICML)*, 2015.

グラフ表現を用いた異常変数同定における一致推定量 *†

原 聡 ‡

IBM 東京基礎研究所

1 グラフ表現を用いた異常変数同定

D 次元の確率変数を $X = (X_1, X_2, \ldots, X_D) \in \mathbb{R}^D$ とする．異常変数同定とは，2 つのデータセット $\mathcal{P} := \{\boldsymbol{x}^{(n)}\}_{n=1}^{N} \overset{\text{i.i.d.}}{\sim} p(X)$ と $\mathcal{Q} := \{\boldsymbol{y}^{(m)}\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} q(X)$ とを比較し，分布に変化のある異常な確率変数の集合 $S \subseteq \{1, 2, \ldots, D\}$ を探す問題である．異常変数同定に対して複数の先行研究があるか，どれも実験的な性能評価に留まっており理論的な性能保証は行われていなかった [1, 2, 3, 4]．本研究では異常変数同定に対して解の一致性を保証する手法を提案する．なお，問題としては井手ら [1, 2] のグラフ表現を用いた定式化を対象とする．ここで，グラフ表現とは確率変数間の関係性をグラフで表現したものである．例えば，井手らの研究ではグラフ表現として確率変数間の相関や，ガウシアン・グラフィカルモデルが用いられている [1, 2].

問題 1（グラフ表現を用いた異常変数同定） $\Lambda^p, \Lambda^q \in \mathbb{R}^{D\times D}$ をそれぞれ分布 p, q のグラフ表現の重みつき隣接行列とする．このとき，以下の (i), (ii) を満たす集合 $S \subseteq \{1, 2, \ldots, D\}$ を見つける．ただし S^{c} は S の補集合である．

$$\text{(i) } \Lambda^p_{dd'} = \Lambda^q_{dd'}, \forall d, d' \in S^{\mathrm{c}}, \ \text{(ii) } \Lambda^p_{dd'} \neq \Lambda^q_{dd'}, \forall d \in S, \exists d' \in \{1, 2, \ldots, D\} \setminus \{d\}.$$

条件 (i) は異常変数集合 S の補集合，つまり正常変数においてグラフ表現に変化がないことを意味する．他方，条件 (ii) は異常変数においては隣接する少なくとも 1 つの辺においてその重み，つまり確率変数間の関係に変化があることを意味する．これらは近傍構造保存仮定として井手ら [1, 2] により提唱された条件である．

2 k-疎部分グラフ問題による定式化，解法及び一致性

データセット $\mathcal{P}$, $\mathcal{Q}$ から推定されたグラフ表現をそれぞれ $\hat{\Lambda}^p, \hat{\Lambda}^q$ とする．また，行列 $\hat{\Gamma} \in \mathbb{R}^{D\times D}$ を $\hat{\Gamma}_{dd'} := |\hat{\Lambda}^p_{dd'} - \hat{\Lambda}^q_{dd'}|$ により定義する．このとき，推定量 $\hat{\Lambda}^p, \hat{\Lambda}^q$ が真の隣接行列 Λ^p, Λ^q に十分近ければ，条件 (i), (ii) から以下が言える：(i') $\hat{\Gamma}^p_{dd'} \approx 0, \forall d, d' \in S^{\mathrm{c}}$, (ii') $\hat{\Gamma}_{dd'} > 0, \forall d \in S, \exists d' \in \{1, 2, \ldots, D\} \setminus \{d\}$. (i'), (ii') は集合 S^{c}（及びその補集合 S）が，行列 $\hat{\Gamma}$ の値の小さい部分行列を探すことで同定できることを示している．異常変数集合 S のサイズが k だとすると，これは以下の最適化問題として定式化できる．

$$\hat{S} = \underset{S\subseteq\{1,2,\ldots,D\}}{\operatorname{argmin}} \sum_{d,d'\in S^{\mathrm{c}}} \hat{\Gamma}_{dd'}, \ \text{s.t. } |S| = k. \tag{1}$$

* 本稿の執筆にあたり JST CREST の助成を受けた．

† 本研究は IBM 東京基礎研究所の森村哲郎氏，高橋俊博氏，柳沢弘輝氏，及び東工大/さきがけの鈴木大慈氏との共同研究である．

‡ satohara@jp.ibm.com

問題 (1) は $\hat{\Gamma}$ を隣接行列とするグラフの k-疎部分グラフ問題 [5] であり，一般には NP 困難である．k の値が既知であれば問題 (1) の解法としては，厳密解法や貪欲法による近似解の導出などが考えられる．特に，厳密解については以下の一致性定理が成り立つ．

定理 1（$\hat{S}$ の一致性） 推定量 $\hat{\Lambda}^p, \hat{\Lambda}^q$ が真の隣接行列 Λ^p, Λ^q の一致推定量ならば，問題 (1) の解 $\hat{S}$ は集合 S の一致推定量である．つまり，$\lim_{N,M\to\infty} P(\hat{S} \neq S) = 0$ が成り立つ．

集合 S のサイズ k が未知の場合は，問題 (1) の凸近似から導かれる凸二次計画を解くことで S の推定量を得ることができる．

$$\tilde{S} = \{d \mid \tilde{s}_d = 0\},\ \tilde{\boldsymbol{s}} = \operatorname*{argmin}_{\boldsymbol{s}\in\mathbb{R}^D} \boldsymbol{s}^\top A_\mu \boldsymbol{s},\ \text{s.t.}\ \sum_{d=1}^{D} s_d = 1,\ s_d \geq 0, \forall d.$$

ただし，行列 $A_\mu \in \mathbb{R}^{D\times D}$ は $A_\mu := \hat{\Gamma} + \mu I_D$（$I_D$ は D 次元単位行列）であり，定数 μ は A_μ が正定値になる十分大きな実数とする．このとき，推定量 $\tilde{S}$ について以下の一致性定理が成り立つ．

定理 2（$\tilde{S}$ の一致性） 推定量 $\hat{\Lambda}^p, \hat{\Lambda}^q$ が真の隣接行列 Λ^p, Λ^q の一致推定量であり，かつ以下を満たす $\delta > 0$ が存在するならば $\tilde{S}$ は集合 S の一致推定量である．

$$\min_{d\in S} \sum_{d'\in S^{\mathrm{c}}} |\Lambda^p_{dd'} - \Lambda^q_{dd'}| \geq (1+\delta)\mu.$$

定理 2 は k の値が未知でも一定条件下で集合 S の一致推定量が得られることを示している．また，$\tilde{S}$ は凸二次計画を解くことで得られるため，必要な計算時間を多項式時間で抑えることができる．これは NP 困難な問題 (1) を直接解くよりも計算時間の面で有利である．

参考文献

[1] T. Idé, S. Papadimitriou, and M. Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. *Proceedings of the 7th IEEE International Conference on Data Mining*, pages 523–528, 2007.

[2] T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 97–108, 2009.

[3] S. Hirose, K. Yamanishi, T. Nakata, and R. Fujimaki. Network anomaly detection based on eigen equation compression. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1185–1194, 2009.

[4] R. Jiang, H. Fei, and J. Huan. Anomaly localization for network data streams with graph joint sparse PCA. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 886–894, 2011.

[5] R. Watrigant, M. Bougeret, and R. Giroudeau. Approximating the sparsest k-subgraph in chordal graphs. *Approximation and Online Algorithms, Lecture Notes in Computer Science*, 8447:73–84, 2014.

Non-convex Compressed Sensing with the Sum-of-Squares Method

相馬 輔 (東京大学)
吉田 悠一 (国立情報学研究所・Preferred Infrastructure)

1 概要

本研究では，ℓ_q ノルム $(0 < q \leq 1)$ に対する **stable signal recovery** を考える．この問題では，未知の信号 $\boldsymbol{x} \in \mathbb{R}^n$ に対して，行列 $A \in \mathbb{R}^{m \times n}$ を用いたベクトル $\boldsymbol{y} = A\boldsymbol{x}$ が観測できるものとする．目標は，観測行列 A と 復元アルゴリズム $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ であって，

$$\|\boldsymbol{x} - \Delta(A\boldsymbol{x})\|_q \leq O(1) \cdot \min\{\|\boldsymbol{x} - \boldsymbol{x}^*\|_q : \|\boldsymbol{x}^*\|_0 \leq s\}$$

となるものを設計することである．既存の ℓ_q-stable signal recovery に対するアルゴリズム [1] は ℓ_q 最小化

$$\min \|\boldsymbol{z}\|_q^q \quad \text{subject to } A\boldsymbol{z} = \boldsymbol{y} \tag{1}$$

を解く必要があるが，これは非凸最適化問題であるため，多項式時間の ℓ_q-stable signal recovery アルゴリズムの存在は分かっていなかった．本研究では，**二乗和多項式法 (Sum-of-Squares Method)** を利用し，Rademacher 行列では，定数 $0 < q < 1$ に対して (ほぼ)ℓ_q-stable signal recovery が多項式時間で可能であることを示す．

定理 1. $q = 2^{-k}$ $(k \in \mathbb{Z}_+)$, $s, n \in \mathbb{Z}_+$ とする．$A \in \mathbb{R}^{m \times n}$ を $m = O(s^{2/q} \log n)$ を満たす正規化 Rademacher 行列とすると，以下が高確率で成り立つ: 任意の $\boldsymbol{x} \in [-1, 1]^n$ と $\epsilon > 0$ に対して

$$\|\boldsymbol{x} - \Delta(A\boldsymbol{x})\|_q \leq O(1) \cdot \min\{\|\boldsymbol{x} - \boldsymbol{x}^*\|_q : \|\boldsymbol{x}^*\|_0 \leq s\} + \epsilon$$

を満たす多項式時間アルゴリズム Δ が存在する．

表 1　ℓ_q-stable signal recovery の既存手法と提案手法

	q	観測行列 A	サンプル数 m	アルゴリズム Δ	Δ の計算量
[2, 3]	$q = 1$	Subgaussian	$O(s \log(n/s))$	ℓ_1 最小化 (LP)	多項式時間
[1]	$0 < q \leq 1$	Subgaussian	$C_1(q)s \log(n/s) + C_2(q)s$	ℓ_q 最小化	—
本研究	$0 < q \leq 1$(定数)	Rademacher	$O(s^{2/q} \log n)$	二乗和多項式	多項式時間

2 多項式最適化と二乗和多項式法

多項式最適化は，多項式 $f, g_1, \ldots, g_i \in \mathbb{R}[\boldsymbol{z}]$ $(\boldsymbol{z} = [z(1), \ldots, z(n)]^\top)$ に対して

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & f(\boldsymbol{z}) \\ \text{sub. to} \quad & g_i(\boldsymbol{z}) = 0 \quad (i = 1, \ldots, m) \end{aligned}$$

で定義される最適化問題である．多項式最適化は一般に非凸最適化問題であり，整数計画問題等を含むため NP 困難である．

多項式最適化に対する強力な凸緩和法として二乗和多項式法 [4, 5, 6, 7] がある．二乗和多項式法においては，次数と呼ばれるパラメータ d を指定する．そして，元の変数 $\boldsymbol{z}$ のかわりに **pseudoexpectation** と呼ばれる線形作用素 $\widetilde{\mathbf{E}} : \mathbb{R}[\boldsymbol{z}]_d \to \mathbb{R}$ を変数とした最適化問題を解く ($\mathbb{R}[\boldsymbol{z}]_d$ は次数 d 以下の多項式の集合)．

$$
\begin{aligned}
\min_{\widetilde{\mathbf{E}}} \quad & \widetilde{\mathbf{E}}[f(\boldsymbol{z})] \\
\text{sub. to} \quad & \widetilde{\mathbf{E}}[1] = 1 \\
& \widetilde{\mathbf{E}}[p(\boldsymbol{z})^2] \geq 0 \qquad (p \in \mathbb{R}[\boldsymbol{z}] : \deg(p) \leq d/2) \\
& \widetilde{\mathbf{E}}[g_i(\boldsymbol{z})p(\boldsymbol{z})] = 0 \quad (p \in \mathbb{R}[\boldsymbol{z}] : \deg(g_i p) \leq d,\, i = 1, \ldots, m)
\end{aligned}
$$

二乗和多項式法は $n^{O(d)}$ サイズの行列変数を持つ**半正定値計画 (SDP)** に帰着可能であることが知られている．特に，d が定数であれば多項式時間で (任意の精度で) 解くことが可能である．

3 提案アルゴリズム

ℓ_q 最小化は多項式最適化として定式化可能であるが，愚直な定式化のままでは緩和ギャップが大きすぎるため，ℓ_q 最小化の最適解の情報を得ることができない．そこで提案アルゴリズムでは，冗長な変数と制約を追加した多項式最適化に対して二乗和多項式緩和を適用することで，緩和ギャップを減らし ℓ_q 最小化の最適解の情報を得られるようにした．追加する制約は，ℓ_q 最小化の stability の証明に必要な ℓ_q^q 擬距離における三角不等式などである．また，緩和解 $\widetilde{\mathbf{E}}$ から元信号の推定量であるベクトルを抽出するアルゴリズムも構成した．詳細は [8] を参照されたい．

参考文献

[1] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, 2008.

[2] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[3] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[4] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 796–817, 2006.

[5] Y. Nesterov, "Squared functional systems and optimization problems," in *High Performance Optimization*. Springer, 2000, pp. 405–440.

[6] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, California Institute of Technology, 2000.

[7] N. Z. Shor, "An approach to obtaining global extremums in polynomial mathematical programming problems," *Cybernetics*, vol. 23, no. 5, pp. 695–700, 1987.

[8] T. Soma and Y. Yoshida, "Non-convex compressed sensing with the sum-of-squares method," in *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2016, pp. 570–579.

教師付き学習における Barron and Cover 理論

BARRON AND COVERS' THEORY IN SUPERVISED LEARNING

川喜田雅則
Masanori Kawakita

九州大学システム情報科学研究院
Information Science and Electrical Engineering, Kyushu University

1 まえがき

機械学習，統計学，情報理論において，様々な推定法の精度を理論的に評価，あるいは保証することは重要であるが，これまで従来の理論の多くは標本数，あるいは共変量の数を無限大と仮定したり，確率変数の有界性，あるいはモーメントに関する様々な技術的な条件を必要とする．これらの条件は現実に推定を行うときには成り立っていないものが少なくない．そこで，本稿ではこのような仮定をほとんど必要とせず性能評価を行う方法を考えてみたい．そのための土台となる理論が MDL(最小記述長) 原理 [1, 2, 3, 4] に関する代表的な成果の一つである Barron and Cover 理論 (BC 理論)[5] である．MDL 原理とはデータを可能な限り短く記述することが，データを生成した確率分布に関する推定の精度を高めることにつながるという原理である．MDL 原理に基づくモデル選択規準として最も有名なものの一つが Rissanen によって提案された二段階符号に基づく情報量基準 [1] である．ここでいう二段階符号とは，まずデータを記述するのに用いる確率モデルを記述し，次にそのモデルを用いてデータを記述するという符号化を指し，このモデル記述長とデータ記述長の和を最小にする推定量を MDL 推定量と呼ぶ．BC 理論は MDL 推定量に対して，次のような理想的な性質を満たすリスクの上界評価を与える理論である．まず BC 理論は確率変数の有界性やモーメント条件などの技術的な条件をほとんど必要としない．また有限の標本数で成立し，(経験的に) タイトなリスクの上界評価を与える．このような理論は著者の知る限り他にない．しかし BC 理論を機械学習に教師付き学習に応用しようとすると，これまで近似的かつ限定的にしか可能ではなかった．この問題の解決には BC 理論にとって本質的な困難さが伴う．最近我々はこの問題を一定程度解決し，BC 理論を教師付き学習に適用するための一つの方法を提案し，実際に lasso の乱数計画におけるリスク上界を導出することに成功した [6, 7, 8, 9] ので，それを本稿で紹介する．

2 教師付き学習における MDL 推定量

ある分布 $p_*(x,y) = p_*(x)p_*(y|x)$ から独立に生成された n 個の教師付きデータ $D := \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ が与えられたとする．以降では $x^n := x_1x_2\cdots x_n$，$y^n := y_1y_2\cdots y_n$ と書く．教師付き学習とは，このデータセット D を用いて，$p_*(y|x)$ を推定する問題である．ここでは $p_*(y|x)$ を推定するためにパラメトリックモデル $p_\theta(y|x)$ を用いる．ただし $\theta \in \Theta(=\Re^p)$ は p 次元のパラメータベクトルとする．MDL 推定量を定義するためにはパラメータ空間 Θ を量子化しなくてはならない．量子化近似したパラメータ空間を $\overline{\Theta}(x^n)(\subset \Theta)$ と書く．また $\overline{\Theta}(x^n)$ に含まれる各モデルの記述長を $L(\bar{\theta}|x^n)$ と定義する．モデル記述長は Kraft の不等式を満たすとする．モデル $p_\theta(y|x)$ に関する MDL 推定量 $\ddot{\theta}(x^n,y^n)$ は

$$\ddot{\theta}(x^n,y^n) = \arg\min_{\bar{\theta}\in\overline{\Theta}(x^n)} \left\{-\log p_{\bar{\theta}}(y^n|x^n) + \beta L(\bar{\theta}|x^n)\right\}$$

と定義される．ただし，β は 1 より大きい実数とする．この記述長は，共変量 x^n を与えられた下での，目的変数 y^n の条件付き記述長であることに注意する．最小化の目的関数の第 1 項をデータ記述長，第 2 項をモデル記述長と呼ぶ．また，それらの和を全記述長と呼ぶ．なお，本稿では log で自然対数を表し，符号長の単位に nat を用いる．この二段階符号化によって達成できる記述長の最小値を $L_{\beta\text{2-}p}(y^n|x^n)$ で表す．このとき，$L_{\beta\text{2-}p}(y^n|x^n)$ は y^n について Kraft の不等式を満たすため，語頭符号の符号長とみなせる．従って，$p_{\beta\text{2-}p}(y^n|x^n) := \exp(-L_{\beta\text{2-}p}(y^n|x^n)) = p_{\ddot{\theta}}(y^n|x^n)\exp(-\beta L(\ddot{\theta}|x^n))$ とすれば，$p_{\beta\text{2-}p}$ は条件付きの劣確率分布となる．

3 教師付き学習における BC 理論

オリジナルの BC 理論 [5] は，MDL 推定量についての理論であり，推定量は量子化近似された空間上で定義されていなければならなかった．それに対して本節では量子化近似をしない推定量

$$\hat{\theta}(x^n,y^n) := \arg\min_{\theta\in\Theta}\{-\log p_\theta(y^n|x^n) + \mathrm{Pen}(\theta|x^n)\}$$

を考える．BC 理論を教師付き学習について，このように量子化近似をしない推定量に適用することは本質的に難しかったが，[6, 10, 7, 8, 9] で提案された解決法について紹介する．まず BC 理論では推定量の精度を測るための損失関数に Rényi ダイバージェンス

$$d_\lambda(p,q) = -\frac{1}{1-\lambda}\log E_{p_*(x)p(y|x)}\Big(\frac{q(Y|X)}{p(Y|X)}\Big)^{1-\lambda}$$

を用いる．ここで $\lambda \in (0,1)$ は Rényi ダイバージェンスの次数と呼ばれる．また真の分布 $p_*(x)$ についての適当な典型集合 $A_\epsilon^{(n)}$ を導入し，$P_\epsilon^{(n)} = \Pr\{x^n \in A_\epsilon^{(n)}\}$ と書く．この典型集合を用いて次の定義を導入する．

Definition 1. ある $\beta > 1$ について，各 $p_*(x^n) \in \mathcal{P}_X$ に対して，量子化空間 $\overline{\Theta}(p_*) \subset \Theta$ と，Kraft の不等式を

満たすモデル記述長 $L_n^* : \overline{\Theta}(p_*) \to [0,\infty)$ が存在し，それらが共変量 x^n に依存せず，

$$\begin{aligned}
&\forall x^n \in A_\epsilon^{(n)}, \forall y^n, \\
&\max_{\bar\theta \in \overline{\Theta}(p_X^*)} \Big(n d_\lambda(p_*, p_{\bar\theta}) - \log \frac{p_*(y^n|x^n)}{p_{\bar\theta}(y^n|x^n)} - \beta L_n^*(\bar\theta) \Big) \\
&\geq \max_{\theta \in \Theta} \Big(n d_\lambda(p_*, p_\theta) - \log \frac{p_*(y^n|x^n)}{p_\theta(y^n|x^n)} - \mathrm{Pen}(\theta|x^n) \Big)
\end{aligned}$$

が成り立つとき，Pen は $(\mathcal{P}_X, \beta, A_\epsilon^{(n)})$ に対して ϵ リスク妥当であるという．

ϵ リスク妥当である罰則関数について，次の定理を示せる．

定理 2. $\mathrm{Pen}(\cdot|x^n)$ が $(\mathcal{P}_X, \beta, A_\epsilon^{(n)})$ について ϵ リスク妥当であるとき，任意の $\lambda \in (0, 1-\beta^{-1}]$ について

$$E_1 d_\lambda(p_*, p_{\hat\theta}) \leq E_1 \frac{1}{n} \log \frac{p_*(y^n|x^n)}{p_{\beta 2\text{-}p}(y^n|x^n)} + \frac{\beta}{n} \log \frac{1}{P_\epsilon^{(n)}}$$

が成り立つ．ただし，$p_{\beta 2\text{-}p}$ の定義の中で用いられる推定量は $\hat\theta$ とし，E_1 は $x^n \in A_\epsilon^{(n)}$ の下での条件付き期待値を表わす．

証明は例えば [9] を参照されたい．リスク上界の第二項は典型集合の性質から n に応じて小さくなる項である．従ってリスク上界の主要な項は第一項であり，これはオリジナルの BC 理論と同様に二段階符号の（条件付き）冗長度の形をしている．従って冗長度が小さくなるように二段階符号を設計することがリスクの良い保証（上界）を得ることにつながるという意味で，「BC 理論による MDL 原理の正当化」を教師付き学習の場合に拡張していることになる．ただし主要項が厳密に冗長度とみなせるためには，一般には罰則関数に別途条件が必要であることに注意する．

4 lasso のリスク上界

本節では lasso について，前節の ϵ リスク妥当性を用いて乱数計画における lasso の具体的なリスク上界を導出する．まず教師付きデータ $\{(x_i, y_i)|i = 1, 2, \cdots, n\}$ は回帰モデル $y_i = x_i^T \theta^* + \epsilon_i$ に従うとする．パラメータ θ の次元を p とし，n より大きくても良いとする．ノイズ ϵ_i は平均 0, 分散 σ^2 を持つ任意の確率変数とする．また $\mathcal{P}_X := \{p(x^n) = \Pi_{i=1}^n N(x_i; 0, \Sigma) |$ 任意の正則な$\Sigma\}$ とする．ただし $N(x; \mu, \Sigma)$ は平均ベクトル μ，分散共分散行列 Σ の正規分布を表すとする．正規性を仮定しても ϵ リスク妥当な罰則関数を求めるのは容易ではないが，著者らが文献 [6, 9] において導出した ϵ リスク妥当な重み付き ℓ_1 ノルムを紹介する．まず典型集合 $A_\epsilon^{(n)}$ には

$$A_\epsilon^{(n)} := \left\{ x^n \,\middle|\, \forall j, 1-\epsilon \leq \frac{(1/n)\sum_{i=1}^n x_{ij}^2}{\Sigma_{jj}} \leq 1+\epsilon \right\}$$

を採用する．ただし $\epsilon \in (0,1)$ とする．このとき以下の定理が成り立つ．

定理 3. 次数 λ の Rényi ダイバージェンスに対して，次の条件を満たす重み付き ℓ_1 ノルムは，$(\mathcal{P}_X, \beta, A_\epsilon^{(n)})$ についての ϵ リスク妥当性を満たす罰則関数となる．

$$\begin{aligned}
\mathrm{Pen}(\theta|x^n) &:= \mu_1 \|\theta\|_{w,1} + \mu_2, \\
\mu_1 &\geq \sqrt{\frac{2n \log 4p}{\sigma^2 \beta} \left(\frac{\lambda}{4(1-\epsilon)} + \sqrt{\frac{1+\epsilon}{1-\epsilon}} \right)}, \\
\mu_2 &\geq \frac{\log 2}{\beta}.
\end{aligned}$$

lasso 推定量は定理 2 に示されたリスク上界を持つ．また，文献 [6] の補題 2 に示されているように $P_\epsilon^{(n)} \geq 1 - 2p \exp\left(-\frac{n\epsilon^2}{7}\right)$ が成り立つ．従って定理 2 のリスク上界の第二項は指数的に小さくなる項であることがわかる．

参考文献

[1] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[2] A. R. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743–2760, 1998.

[3] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2005.

[4] P. Grünwald, I. Myung, and M. Pitt, *Advances in Minimum Description Length: Theory and Applications*, ser. Neural information processing series. Bradford Book, 2005.

[5] A. R. Barron and T. M. Cover, “Minimum complexity density estimation,” *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[6] 川喜田雅則, 豊暉原侑心, 竹内純一, “MDL 理論による lasso のリスク上界,” 信学技法 IBISML2015-16, vol. 115, no. 112, pp. 101–107, 2015.

[7] 竹内純一, 川喜田雅則, “教師付き学習における MDL 推定量の設計とその収束速度,” 信学技法 IT2015-26, vol. 115, no. 137, pp. 53–58, 2015.

[8] ——, “教師付き学習における MDL 推定,” 第 9 回シャノン理論ワークショップ予稿集, 2015.

[9] 川喜田雅則, 竹内純一, “Barron and cover 理論の教師付き学習への拡張及び lasso への応用.” 第 18 回情報論的学習理論ワークショップ（IBIS 2015), 11 2015.

[10] ——, “教師付き学習における MDL 推定のリスクバウンドについて,” 第 38 回情報理論とその応用シンポジウム (SITA 2015), pp. 78–83, 11 2015.

状態空間モデルを用いた時系列の周期成分分解と位相推定

東京大・情報理工　松田 孟留
東京大・情報理工，理研・脳センター　駒木 文保

多くの時系列は，さまざまな周波数の周期成分が重ね合わさったものと見なすことができる．たとえば脳波や Local Field Potential などの神経電位時系列は，その周期成分の位相が生理学的に重要な意味をもつことが近年明らかにされてきている [2]．従来はバンドパスフィルタリングとヒルベルト変換によって位相を推定している．すなわち，バンドパスフィルタをかけた時系列 $y(t)$ を

$$y_H(t) = \int_{-\infty}^{\infty} \frac{y(\tau)}{t-\tau} \mathrm{d}\tau, \quad z(t) = y(t) + iy_H(t)$$

によって解析信号 $z(t)$ に変換し，その偏角を位相としている．しかし，このような推定法は観測ノイズやフィルタリング方法によって結果が不安定になってしまう [4]．

本研究では，状態空間モデルに基づいた時系列の周期成分分解と位相推定の手法を開発した．Wiener は，脳波のスペクトルにおいてアルファ帯域（10Hz 付近）にピークができる原理について，確率微分方程式のモデルを用いて考察した [5]．このモデルは，振動子の周波数がランダムにゆらぐ（「時計の針が震える」）という発想から導入されたものである．提案手法では，このランダム周波数変調の発想をもとに線形ガウス型の状態空間モデル

$$\begin{pmatrix} x_{t+1,1}^{(k)} \\ x_{t+1,2}^{(k)} \end{pmatrix} \sim \mathrm{N}\left(a_k \begin{pmatrix} \cos(2\pi f_k \Delta t) & -\sin(2\pi f_k \Delta t) \\ \sin(2\pi f_k \Delta t) & \cos(2\pi f_k \Delta t) \end{pmatrix} \begin{pmatrix} x_{t,1}^{(k)} \\ x_{t,2}^{(k)} \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix} \right) \quad (k = 1, \cdots, K),$$

$$y_t \sim \mathrm{N}\left(\sum_{k=1}^{K} x_{t,1}^{(k)}, \tau^2 \right)$$

を用いる．ここで Δt は y_t のサンプリング間隔である．このモデルでは，時系列 y_t には K 個の（ランダム周波数変調あり）振動子の活動が重ね合わさっていると仮定し，k 番目の振動子の周波数を f_k，座標を $(x_{t,1}^{(k)}, x_{t,2}^{(k)})^\top$ としている．よって，y_t にカルマンフィルタ・スムーザを適用することで，各振動子の振幅と位相が推定できる．これは，時系列をトレンド成分や季節成分に分解する [3] の季節調整法と同じ発想である．状態空間モデルのパラメータ $a_1, \cdots, a_K$, $f_1, \cdots, f_K$, $\sigma_1^2, \cdots, \sigma_K^2, \tau^2$ は経験ベイズ法によって，また振動子の個数 K は情報量規準 ABIC[1] の最小化によってあらかじめ決定しておく．これにより，時系列を自然な形で（ランダム周波数変調ありの）周期成分に分解することができる．

参考文献

[1] H. Akaike. Likelihood and the Bayes procedure. In *Bayesian Statistics*, Eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 1–13. Valencia, Spain, 1980.

[2] G. Buzsáki. *Rhythms of the Brain*. Oxford University Press, Oxford, 2011.

[3] G. Kitagawa and W. Gersch. A smoothness priors-state space modeling of time series with trend and seasonality. *Journal of American Statistical Association*, Vol. 79, 378–389, 1984.

[4] A. G. Siapas, E. V. Lubenov and M. A. Wilson. Prefrontal phase locking to hippocampal theta oscillations. *Neuron*, Vol. 46, pp. 141–151, 2005.

[5] N. Wiener. *Nonlinear problems in random theory*. MIT Press, Cambridge, 1966.

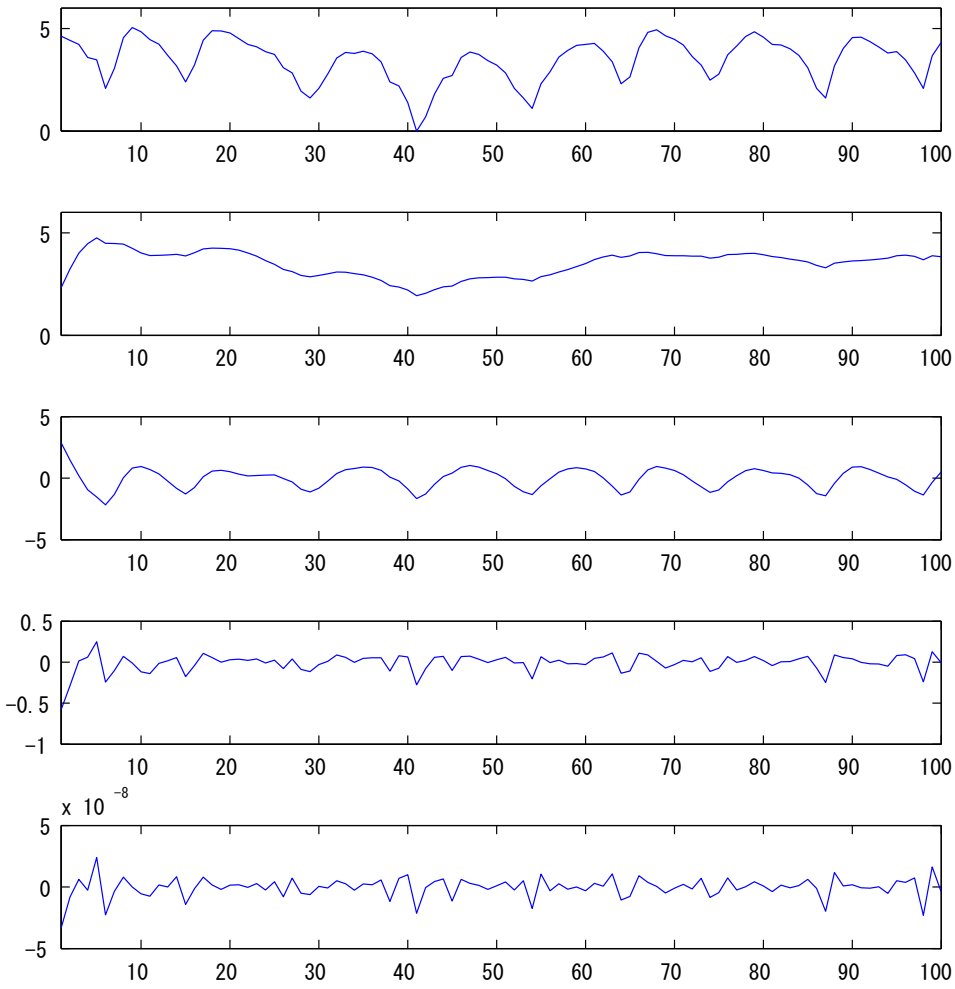


図 1　提案手法による Wolfer sunspot data の分解結果．1 行目：元データ（対数変換した黒点時系列），2〜4 行目：周期成分，5 行目：観測ノイズ

関数予測における漸近ミニマックス予測分布

東京大・情報理工　矢野 恵佑
東京大・情報理工, 理研・脳科学総合研究センター　駒木 文保

1 概要

本発表では，ノンパラメトリックな関数予測における漸近ミニマックス予測分布の構成を議論する．まず，真の関数が楕円型の制約をもち我々がその制約値を知っている場合の漸近ミニマックス予測分布を構成する．次に，真の関数が楕円型の制約をもち我々がその制約値を知らない場合の漸近ミニマックス予測分布を構成する．最後に，数値実験を通して我々の構成した漸近ミニマックス予測分布の性質を議論する．

2 関数予測における漸近ミニマックス予測分布

未知の関数 $\theta(\cdot)$ にガウス過程 $W(\cdot)$ が加法された曲線 $X(\cdot)$ を観測し，$\theta(\cdot)$ に $W(\cdot)$ と独立なガウス過程 $\widetilde{W}(\cdot)$ が加法された曲線 $Y(\cdot)$ を予測する状況を考える．$X(\cdot)$ は既知の $\varepsilon > 0$ および $[0,1]$ 上のガウス過程 $W(\cdot)$ を利用して

$$X(t) = \theta(t) + \varepsilon W(t) \text{ for } t \in [0,1] \tag{1}$$

と与えられ，$Y(\cdot)$ は既知の $\tilde{\varepsilon} > 0$ および $[0,1]$ 上のガウス過程 $\widetilde{W}(\cdot)$ を利用して

$$Y(t) = \theta(t) + \tilde{\varepsilon}\widetilde{W}(t) \text{ for } t \in [0,1] \tag{2}$$

と与えられるとする．

我々は上記の予測を次のようなガウス列モデルでの予測に帰着して考える．すなわち，W と $\widetilde{W}$ を $\mathbb{R}^\infty$ 上のガウス分布 $\otimes_{i=1}^\infty \mathcal{N}(0,1)$ に従う独立な確率変数とし，ε と $\tilde{\varepsilon}$ を既知の正数とした時に，$X = \theta + \varepsilon W$ に従う X を観測し，$Y = \theta + \tilde{\varepsilon}\widetilde{W}$ に従う Y の分布を予測する．X の従う分布を P_θ と書き，Y の従う分布を Q_θ と書く．

本発表では，パラメータ空間が楕円型である場合の Kullback–Leibler リスクのもとでの漸近ミニマックスな予測分布を構成する．予測分布とは観測 X をもとに構成した $\mathbb{R}^\infty$ 上の確率分布のことである．楕円型のパラメータ空間は非零な増加列 $a = (a_1, a_2, \cdots)$ および正数 C を利用して

$$\Theta(a, C) = \{\theta \in l_2 : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \leq C\} \tag{3}$$

と与えられる．Kullback–Leibler リスクのもとで漸近ミニマックスな予測分布とは

$$\begin{aligned}
&\inf_{\hat{Q}(\cdot;X)\in\mathcal{A}} \sup_{\theta\in\Theta(a,C)} \int \log \frac{\mathrm{d}Q_\theta}{\mathrm{d}\hat{Q}(\cdot; X=x)}(y)\mathrm{d}Q_\theta(y)\mathrm{d}P_\theta(x) \\
&\qquad = (1+\mathrm{o}(1)) \sup_{\theta\in\Theta(a,C)} \int \log \frac{\mathrm{d}Q_\theta}{\mathrm{d}\hat{Q}_{\mathrm{minimax}}(\cdot; X=x)}(y)\mathrm{d}Q_\theta(y)\mathrm{d}P_\theta(x) \text{ as } \varepsilon \to 0
\end{aligned} \tag{4}$$

を満たす予測分布 $\hat{Q}_{\mathrm{minimax}}(\cdot; X)$ である．ここで，$\mathcal{A}$ は $\mathbb{R}^\infty$ 上の確率分布全体であり，$\tilde{\varepsilon}$ は $1/\tilde{\varepsilon} = \mathrm{O}(1/\varepsilon)$ と与えられているとする．

まず，増加列 a および C が既知である場合にある $\tau^*(\varepsilon, a, C) \in l_2$ によるガウス分布 $G_{\tau^*(\varepsilon,a,C)} = \otimes_{i=1}^{\infty} \mathcal{N}(0, \tau_i^{*2}(\varepsilon, a, C))$ を事前分布に用いたベイズ予測分布

$$Q_{G_{\tau^*(\varepsilon,a,C)}}(\cdot|X = x) := \int Q_\theta(\cdot)\mathrm{d}G_{\tau^*(\varepsilon,a,C)}(\theta|X = x) \tag{5}$$

が漸近ミニマックスであることを示す．ここで，$G_{\tau^*(\varepsilon,a,C)}(\cdot|X = x)$ は $G_{\tau^*(\varepsilon,a,C)}$ の事後分布である．[1] はパラメータの次元数が ε に応じて増加するという漸近論を利用して楕円型のパラメータ制約下での Kullback–Leibler リスクのもとでの漸近ミニマックス予測分布を求めている．本結果は [1] の結果の無限次元のパラメータ空間への拡張である．

次に，増加列 a および C が未知の場合での漸近ミニマックスを達成する予測分布の構成法について紹介する．我々の構成法は有限次元の Stein の事前分布を利用する．d 次元ユークリッド空間における Stein の事前分布とは密度が $h(\theta^{(d)}) := (\sum_{i=1}^{d} \theta_i^2)^{(2-d)/2}$ と表される事前分布である．

最後に，数値実験を通して有限の ε に対する漸近ミニマックス予測分布の性質を議論する．

参考文献

[1] Xu, X. and Liang, F. (2010), Asymptotic minimax risk of predictive density estimation for non-parametric regression. Bernoulli 16, pp. 543–560.

低次元多様体上の Data Sharpening

島根大学総合理工学研究科　工藤　雅紀　　島根大学総合理工学研究科　内藤　貫太

1. はじめに ノンパラメトリック核型回帰において，バイアス縮小を導く方法として Data Sharpening が知られている．Data Sharpening という手法は [2] で提案され，局所線形 (以下，LL) 推定量に残差の LL 推定量を加えることで Data Sharpening(以下，DS) 推定量が得られる．そして，p 変量の説明変数に対する DS 推定量の漸近的振る舞いは，バンド幅 $h \to 0$ と標本数 $n \to \infty, nh^p \to \infty$ のもとで，漸近バイアスが $O_p(h^4)$ であり，漸近分散は $O_p(1/(nh^p))$ である．他方，LL 推定量の漸近的振る舞いは [4] で議論され，その漸近バイアスは $O_p(h^2)$ であり，その漸近分散は $O_p(1/(nh^p))$ であることが知られている．したがって，LL 推定量よりも，DS 推定量の方が漸近的には小さいバイアスを持つことがわかる．

一方，説明変数がある低次元の多様体に埋め込まれているという仮定を含む回帰モデルが議論されてきた．そこでのモデルは 2 種類あり，説明変数が埋め込まれている多様体を既知とするモデルと，未知とするモデルに大別される．多様体を既知とするモデルの代表例は方向統計学であり，多様体として円周や球面を想定しており，これまで盛んに研究されている．しかし，多様体が未知であるとする研究は，既知である場合の研究に比べるとあまり多くないように思われ，例えば [1] では，LL 推定量を未知の d 次元多様体上に構成し漸近的振る舞いが調べられ，漸近バイアスが $O_p(h^2)$ であり，その漸近分散は $O_p(1/(nh^d))$ であることを導いている．

本発表では，[1] と [2] の結果を伴わせて，説明変数が未知の低次元多様体に埋め込まれている設定で DS 推定量を構成し，その理論的・数値的検証について報告する．

2. Data Sharpening 推定量 p 変量の説明変数を X，1 変量の目的変数を Y とし，その独立な n 個の観測値を $\{(X_i, Y_i)|1 \le i \le n\}$ とする．$\varepsilon_1, \cdots, \varepsilon_n$ を i.i.d の誤差とし，回帰モデル

$$Y_i = m(X_i) + v(X_i)^{1/2}\varepsilon_i,\ E[\varepsilon_i] = 0,\ Var[\varepsilon_i] = 1,\ i = 1, \cdots, n$$

を想定する．推定のターゲットは m である．[1] での設定を踏まえ，説明変数が d 次元多様体 $\mathcal{X}(d \le p,\ \mathcal{X} \subset \mathbb{R}^p)$ に埋め込まれているとする．すなわち，$X_i \in \mathcal{X}$ を仮定する．このとき，$x \in \mathcal{X}$ に対して，Local chart と言われる写像 $\varphi : \mathcal{B}^d_{0,\mu} \to \mathcal{X} \cap \mathcal{B}^p_{x,\nu}$ が存在する．ここで，$\mathcal{B}^q_{y,r} = \{z \in \mathbb{R}^q | \sqrt{(z-y)^T(z-y)} < r\}$ は中心が y で半径が r である q 次元の球．さらに非退化な d 次元の密度関数 $f(z)$ により，$f(\varphi^{-1}(x))$ を X の密度関数とする．$\mathbb{R}^p \ni x$ における $m(x)$ の初期推定量 $\widehat{m}_{LL}(x,h)$ は，局所線形推定量で構成する．ここで，$\widehat{m}_{LL}(x,h) = e_1^T(X_x^T W_x X_x)^T X_x^T W_x \mathbf{Y}$ であり，

$$X_x = \begin{bmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{bmatrix},\ \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ W_x = \mathrm{diag}(K_h(X_1 - x), \cdots, K_h(X_n - x)),$$

K はカーネル関数，$h > 0$ はバンド幅，$K_h(U) = K\left(h^{-p}U\right), e_1 = [1, 0, \cdots, 0]^T$ である．

残差 $r_i := Y_i - \widehat{m}_{LL}(X_i, h)$ から，$\{(X_i, r_i)|1 \le i \le n\}$ による局所線形推定量として $\widehat{r}_{LL}(x,h) = \mathrm{e}_1^T(X_x^T W_x X_x)^T X_x^T W_x [r_1, \cdots, r_n]^T$ を構成する．Data Sharpening 推定量は

$$\widehat{m}_{DS}(x,h) := \widehat{m}_{LL}(x,h) + \widehat{r}_{LL}(x,h) = e_1^T(X_x^T W_x X_x)^T X_x^T W_x (2Y - \widetilde{M})$$

として得られる．ここで，$\widetilde{M} = [\widehat{m}_{LL}(X_1, h) \cdots \widehat{m}_{LL}(X_n, h)]^T$ である．

3. 記号 関数 $g : \mathbb{R}^q \to \mathbb{R}$ $(q = d \text{ or } q = p)$ に対して，$\nabla^2 g(x)$ はヘッセ行列を表す．$\varphi(z) = [\varphi_1(z) \cdots \varphi_p(z)]^T$

と $z=(z_1,\cdots,z_d)$ により，a における φ のヤコビ行列を

$$\mathcal{J}(a|\varphi)=\left[\frac{\partial}{\partial z_j}\varphi_i(z)\right]_{1\le i\le p,1\le j\le d}\Bigg|_{z=a}$$

とし，また $G:\mathbb{R}^p\to\mathbb{R}$ に対して，

$$B(x|g,\varphi)=C(x|\varphi)^{-1}\int_{\mathbb{R}^d}u^T\mathcal{J}(\varphi^{-1}(x)|\varphi)^T\nabla^2 g(x)\mathcal{J}(\varphi^{-1}(x)|\varphi)uK(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)du,$$

$$R(x|G,\varphi)=\int_{\mathbb{R}^d}G(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)^2du$$

とおく．ここで

$$C(x|\varphi)=\int_{\mathbb{R}^d}K(\mathcal{J}(\varphi^{-1}(x)|\varphi)u)du$$

であり，さらに，

$$G^*(y)=2G(y)-C(x|\varphi)^{-1}G*G(y),\ G*G(y)=\int_{\mathbb{R}^d}G(\mathcal{J}(\varphi^{-1}(x)|\varphi)v)G(y-\mathcal{J}(\varphi^{-1}(x)|\varphi)v))dv.$$

以上の記号を用いると，[1] の Theorem 2.1 における $\widehat{m}_{LL}(x,h)$ のバイアスと分散，それぞれのリーディングタームは $J_1(x)=B(x|m,\varphi)$ と $J_2(x)=v(x)R(x|K,\varphi)/\{f(\varphi^{-1}(x))C(x|\varphi)^2\}$ である．

4. 漸近挙動 適当な条件のもと，$x\in\mathcal{X},\ n\to\infty,\ h\to 0$ のとき

$$Bias[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]=-\frac{h^4}{4}B(x|J_1,\varphi)+o_p(h^4),$$

$$Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]=\frac{1}{nh^d}\frac{R(x|K^*,\varphi)}{R(x|K,\varphi)}J_2(x)(1+o_p(1)).$$

[1] で議論された LL 推定量は漸近バイアスが $O_p(h^2)$，漸近分散は $O_p(1/(nh^d))$ であった．一方，DS 推定量の漸近バイアスは $O_p(h^4)$，漸近分散は $O_p(1/(nh^d))$ であり，バイアス縮小が生じていることがわかる．

5. DS 推定量の漸近正規性 3 節の条件のもと，$E[\varepsilon^4]<\infty$ かつ $x\in\mathcal{X}$ で $h=\kappa n^{-1/(d+8)}$ のとき

$$\left(\widehat{m}_{DS}(x,h)-m(x)-\frac{h^4}{4}B(x|J_1,\varphi)\right)\Big/\sqrt{Var[\widehat{m}_{DS}(x,h)|X_1,\cdots,X_n]}\xrightarrow{D}N(0,1)$$

が成り立つ．ここで，κ はある正の定数で，“$\xrightarrow{D}$”は分布収束を表し，$N(0,1)$ は標準正規分布である．

6. バイアス縮小の検証 $\widehat{m}_{DS}(x,h)$ のバイアス縮小が実際に生じていることを，シミュレーションと実データへの適用を通して確認する．結果は本発表にて報告する．

参考文献

[1] Bickel, P. and Li, B. (2007). Local polynomial regression on unknown manifolds, *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Institute of Mathematical Statistics Lecture Notes-Monograph Series*, **54**, 177-186.

[2] Choi, E., Hall, P. and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression, *The Annals of Statistics*, **28**, 1339-1355.

[3] Naito, K. and Yoshizaki, M. (2009). Bandwidth selection for a data sharpening estimator in nonparametric regression, *Journal of Multivariate Analysis*, **100**, 1465-1486.

[4] Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression, *The Annals of Statistics*, **22**, 1346-1370.

[5] Yoshizaki, M. and Naito, K. (2004). Practical aspects of bias reducing estimators in nonparametric regression, *Japanese Journal of Applied Statistics*, **33**, 131-155 (in Japanese).

確率的交互方向乗数法と
マルチクラスグラフ型正則化学習への応用

鈴木大慈 †, 吉田悠一 ‡
† 東京工業大学，JST さきがけ (s-taiji@is.titech.ac.jp)
‡ 国立情報学研究所，Preferred Infrastructure, Inc., (yyoshida@nii.ac.jp)

1 はじめに

本発表では，「多クラス連結正則化」と呼ばれる新しい正則化を導入し，その効率的確率的最適化手法を提案する．グラフに沿った正則化は半教師あり学習や画像のデノイジング，セグメンテーション，遺伝子データ解析など，多くの場面で現れる．本研究では，そのような正則化の中でもスパース性を有する正則化に焦点を当てる．既存の手法はグラフの各頂点に当てられたラベルが一つであるか，複数であってもそれらが独立に扱われていたり，スパース性を有していないかった．本研究では，k-劣モジュラ関数に発想を得て，新しい正則化を提案する．これまでの研究で一般化 Fused Lasso 正則化 (Tibshirani et al., 2005) はグラフカットの Lovász 拡張となることが知られていた (Bach, 2013)．グラフカットは劣モジュラ関数であるが，劣モジュラ関数は $\{0,1\}^N$ 変数を入力として受け取るため，複数ラベルは扱えてない．そのため，我々はグラフカットは異なる $\{0,1,2,\dots,k\}^N$ 上の離散関数の線形拡張を正則化として導入する．提案する離散関数はある特殊な状況で k-劣モジュラ関数と呼ばれる関数になる (Huber & Kolmogorov, 2012; Kolmogorov, 2011)．また，提案した正則化学習問題を効率的に解く確率的最適化手法を提案する．提案する最適化手法は，双対問題を確率的交互方向乗数法 (Suzuki, 2014) を用いて解くものである．双対問題の実効定義域の多面体的性質を利用することで，提案手法は線形収束することが示される．

2 多クラス連結正則化

$G=(V,E,\boldsymbol{\ell})$ を部分的にラベルの付けられたグラフとする．すなわち，V は頂点の集合，E は辺の集合であり，$\boldsymbol{\ell}\in\{0,1\}^S$ は頂点の部分集合 $S\subseteq V$ の上に与えられたラベルを表している．

我々の目的は部分的に与えられたラベル $\boldsymbol{\ell}$ から V 全体の本当のラベルを当てることである．なお，$\boldsymbol{\ell}$ は必ずしも正しい値であるとは限らず，ノイズが乗っていることも想定する．この問題を解くために，辺で連結された頂点のラベルは “近い” と仮定して，隣接頂点のラベルを近づけるような正則化をかける．そのような正則化の代表的な手法として一般化 Fused Lasso 正則化がある．これは次のような問題で記述される:

$$\min_{\boldsymbol{x}\in\mathbb{R}^V}\frac{1}{2}\|\boldsymbol{x}_S-\boldsymbol{\ell}_S\|^2+\gamma\sum_{(u,v)\in E}|\boldsymbol{x}_u-\boldsymbol{x}_v|, \tag{1}$$

ただし，γ は正則化パラメータである．上の問題は連続最適化問題であるが，$\boldsymbol{x}$ を丸めることで離散的な解を得る．

上の問題はラベルが $\{0,1\}$ に二値だったが，$k\geq 3$ 値の問題も解けるよう，正則化を拡張する．再び $G=(V,E,\boldsymbol{\ell})$ を部分的にラベルの与えられたグラフとする．ただし，$\boldsymbol{\ell}\in[k]^S$ とする．この時，次の正則化学習問題を考え，これを多クラス連結正則化と呼ぶ:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{V\times[k]}}\frac{1}{2}\|\boldsymbol{x}_S-\boldsymbol{\ell}_S\|^2+\frac{\gamma}{2}\sum_{(u,v)\in E}\sum_{i=1}^{k}|\boldsymbol{x}_{u,i}-\boldsymbol{x}_{v,i}|+\frac{\gamma(a-1)}{2}\Big|\sum_{i=1}^{k}\boldsymbol{x}_{u,i}-\sum_{i=1}^{k}\boldsymbol{x}_{v,i}\Big|. \tag{2}$$

$k = 1, a = 1$ のとき，問題 (2) は問題 (1) と一致する．第三項のため，あるラベルに対応する座標が大きな値を取れば，他のラベルに対応する座標の値が小さくならなくてはいけない．このため，複数のラベルのうち，ただひとつのラベルが大きな値を取るような作用が働く．

3　双対問題とアルゴリズム

問題 (2) の目的関数は微分できず，単純な勾配法を当てはめることはできない．そこで，我々は双対問題を考え，その緩和問題を確率的交互方向乗数法で解く方法を提案する．$\boldsymbol{F} \in \mathbb{R}^{E \times V}$ を隣接行列，すなわち $(\boldsymbol{F}\boldsymbol{x})_e = \boldsymbol{x}_{u,i} - \boldsymbol{x}_{v,i}$ が $\boldsymbol{x} \in \mathbb{R}^{V \times [k]}$ に対して成り立つ行列とする．すると，問題 (2) の双対問題は

$$\min_{\boldsymbol{y} \in \mathbb{R}^{E \times [k]}} \quad \gamma \psi^*(\boldsymbol{y}/\gamma) + \frac{1}{2} \|\boldsymbol{F}_S^\top \boldsymbol{y} - \boldsymbol{\ell}_S\|^2 \tag{3a}$$

$$\text{subject to} \quad \boldsymbol{F}_{:,v}^\top \boldsymbol{y} = \boldsymbol{0} \quad (v \notin S), \tag{3b}$$

で与えられる．ここで，$\boldsymbol{F}_C = [\boldsymbol{F}_{:,v}]_{v \in C}$ $(C \subseteq V)$ で，ψ^* はある関数 f^* を用いて $\psi^*(\boldsymbol{y}) = \sum_{e \in E} f^*(\boldsymbol{y}_{e,:})$ と表される．ここで，$f^*(\boldsymbol{y}_{e,:})$ は集合 C_1 と C_2 を用いて次のように与えられる:

$$f^*(\boldsymbol{y}) = \begin{cases} 0 & (\boldsymbol{y} \in C_1 + C_2), \\ \infty & (\text{otherwise}). \end{cases}$$

ただし，$C_1 = \{\boldsymbol{y} \in \mathbb{R}^k \mid \|\boldsymbol{y}\|_\infty \leq 1/2\}$, $C_2 = \{\boldsymbol{y} \in \mathbb{R}^k \mid \boldsymbol{y} = \alpha \boldsymbol{1},\ |\alpha| \leq (a-1)/2\}$ である．この双対問題を直接勾配法で解くことは難しいが，f^* が二つの単純な凸集合 C_1 と C_2 のミンコフスキー和であることを利用し，$\boldsymbol{y}_{e,:} = \boldsymbol{y}_{1,e} + \boldsymbol{y}_{2,e}$ $(\boldsymbol{y}_{1,e} \in C_1,\ \boldsymbol{y}_{2,e} \in C_2)$ と分解することで，各変数 $\boldsymbol{y}_{1,e}, \boldsymbol{y}_{2,e}$ に関しては f^* への射影が容易に計算できるようになる．このことを用いることで，確率的交互方向乗数法に基づいた線形収束する方法を構築することができる．

Theorem 3.1. $\boldsymbol{x}^{(t)}$ を確率的交互方向乗数法の t ステップ目の解，$\boldsymbol{x}^*$ を最適解とする．すると，ある $\nu > 0$ と $C > 0$ が存在して，$\boldsymbol{x}^{(t)}$ は次のように *R*-線形収束する*:*

$$\mathrm{E}[\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|^2] \leq C \exp\left(-\frac{\nu}{|E|} t\right),$$

収束レートの定数 ν は最適解を含むアフィン空間と実効定義域の角度で決まる．

参考文献

Bach, Francis. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

Huber, Anna and Kolmogorov, Vladimir. Towards minimizing k-submodular functions. In *Combinatorial Optimization*, pp. 451–462. Springer, 2012.

Kolmogorov, Vladimir. Submodularity on a tree: Unifying $L^\natural$-convex and bisubmodular functions. In *MFCS*, pp. 400–411. Springer, 2011.

Suzuki, Taiji. Stochastic dual coordinate ascent with alternating direction method of multipliers. In *ICML*, pp. 736–744, 2014.

Tibshirani, Robert, Saunders, Michael, Rosset, Saharon, Zhu, Ji, and Knight, Keith. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

余震活動の確率予測：ベイズ統計を用いたアプローチ

東京大学生産技術研究所
近江崇宏

概要：大きな地震（本震）が起こると、その後に多くの地震（余震）が引き起こされる。大きな余震は被災地に追加の被害を及ぼしうることから、日本では気象庁が余震活動の確率予報を業務として行っている。その一方で、本震後に迅速に余震活動の予測を行うには、様々な技術的な問題が存在する。例えば、本震の直後では、多くの特に小さな余震が観測から抜け落ちてしまっている。そのため、本震直後ではこのような不完全な観測データに基づいて予測をしなければならない。また通常、観測データに基づき予測モデルのパラメータ値を一つだけ決めて、それに基づき予測を行っているが、短期間のデータからの推定には大きな不確定性が伴うため、そのような予測方式では予測が実際の観測から大きく外れてしまうことがある。我々はこのような問題を統計的な手法を用いて解決し、予測をより有用なものとするための研究をこれまで行ってきた。本講演では、私たちのこれまでの研究を紹介し、余震活動の予測において、統計モデリングがどのように用いられ、いかに重要な役割を果たしているかを議論する。

余震活動の統計モデルと確率予測：本震後の時刻 t におけるマグニチュード M の余震の発生率$\lambda(t,M)$は次のようなモデルで与えられる。

$$\lambda(t,M)=\frac{K}{(t+c)^p}e^{-\beta M}.$$

ここで$K/(t+c)^p$ は余震活動のべき的な時間減衰（大森−宇津則）を、$e^{-\beta M}$ は地震頻度のマグニチュードに対する指数減衰（グーテンベルク・リヒター則）をそれぞれ表している。

原理的にはこのモデルのパラメータ$\{K,p,c,\beta\}$を決めてやれば将来の余震活動の予測を立てることができる。その一方で、パラメータ値を決定するためには幾つかの難点がある。まず、余震活動はそれぞれの本震に応じて非常に個性が強く、たとえ本震のマグニチュードが同一だとしても、活動強度が全く異なることもある。そのため事前にパラメータ値を決めることが難しく、その時点で得られている対象となる余震活動の観測データからパラメータ値を直接推定する必要がある。しかしながら、特に本震直後の観測データは多くの観測漏れが含まれていることが知られており、不完全な観測データからパラメータ値の推定を行わなくてはならない。このことがこれまで、余震活動の迅速

を難しくしていた。

このような問題に対して、私たちはデータの欠損を統計的に特徴づけるために、次のような時刻 t におけるマグニチュード M の余震の検出率関数 $\Phi(t,M)$

$$\Phi(t,M)=\int_{-\infty}^{M} dx \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{(x-\mu(t))^2}{2\sigma^2}]$$

を導入した。ここで $\mu(t)$ は検出率が 50%に対応するマグニチュードを表し、時間変動すると仮定する。ここではベイズ平滑化により $\mu(t)$ を推定する[1]。この検出率関数と予測モデルを組み合わせることにより、不完全なデータからのパラメータの推定が可能になり、本震直後からより精度の高い予測が行えるようになる[1,2]。

さらに予測行う際にはベイズ型の予測を導入した。これまでは最適なパラメータを一つ選び予測を行う（プラグイン予測）というようなことが行われてきたが、短期の余震観測データを用いた場合には、大きな推定の不確定性を伴う。そのためプラグイン予測はしばしば観測値から大きく外れることがある。そこで、よりロバストな予測を行うため、MCMC によって事後分布からサンプルされた多数のパラメータ値からの予測を組み合わせるベイズ型予測を用いた。その結果、ベイズ型予測によって予測精度が優位に改善することが明らかになった[3]。

参考文献：

[1] T. Omi, Y. Ogata, Y. Hirata and K. Aihara, "Forecasting large aftershocks within one day after the main shock", Scientific Reports 3, 2218 (2013).

[2] T. Omi, Y. Ogata, Y. Hirata and K. Aihara, "Estimating the ETAS model from an early aftershock sequence", Geophysical Research Letters 41, 850 (2014).

[3] T. Omi, Y. Ogata, Y. Hirata and K. Aihara, "Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches", Journal of Geophysical Research: Solid Earth 120, 2561 (2015).

点過程モデルによる地震活動異常の解析と群発地震への応用

統計数理研究所　熊澤貴雄
統計数理研究所、東京大学　尾形良彦
気象研究所　木村一洋
気象研究所　前田憲二
気象研究所　小林昭夫

地震活動の通常パターンからの異常はモデルのパラメータの時間的な変化として表れ、その変化は地震断層の状態が変化したことを反映する。断層の状態が変化する原因として近隣の地震性・非地震性の断層すべりに伴う応力場変化の他に、間隙流体圧の増加に伴う断層強度の低下等が報告されている。非地震性の断層すべりや間隙流体圧の増減は直接的な観測が困難だが、一定の震源域において地震が多発する群発性地震を引き起こし、通常の地震活動パターンから統計的に外れる活動を見せる。これらの異常は短期的或いは断続的であり、時々刻々と変化する状態を記述するためには変化点的解析では不十分な場合が多く、より連続的にパラメータの時間変化を扱うモデルが必要となる。

報告者はETASモデル(Ogata, 1988)のパラメータの時間変化をベイズ平滑化により推定するベイズ型非定常ETASモデル（１）を導入し、

(1)
$$\lambda_\theta(t \mid H_t) = \mu q_\mu(t) + \sum_{\{i:\, S<t_i<t\}} \frac{K_0 q_K(t_i) e^{\alpha\left(M_i - M_z\right)}}{\left(t - t_i + c\right)^p}$$

ただし
$$q_*(t) = \sum_{i=1}^{N} I_{(t_i, t_{i+1})}(t) \left\{ \frac{q_{*,i+1} - q_{*,i}}{t_{i+1} - t_i}(t - t_i) + q_{\mu,i} \right\} = \sum_{i=1}^{N} q_{*,i} F_i(t)$$、

2011年東北地方太平洋沖地震に誘発された内陸地震群を解析した。誘発地震の一部は本震の断層活動からは説明できないものであったが、誘発地震系列のパラメータの時間変化を推定することでそれらが間隙流耐圧の増加による断層強度減少に起因することを示すことがでた。

また、火山活動と地震活動の相関をモデル化する為に、非定常ETASモデルで推定したパラメータの時間変化と体積歪み計測量の関係を用いた予測モデルを

提案した。伊豆地方東部でのマグマ貫入に伴う地殻の体積歪みは群発地震発生数と高い相関を示すことが報告されているが、モデルの或るパラメータの時間変化がより高い相関を示すことが分かってきた。この関係をETASモデルに組み込むことで群発地震の短期的予測が可能となる。

より具体的には、ベイズ型非定常 ETAS モデル（１)でこれらの地震活動からインバージョンで求めた背景地震活動度$\mu(t)$が体積歪の変動に約半日ほど遅れた相互相関を示すことが分かった。そこで先行時間の体積歪の変動から常時地震活動への線形回帰で予測すると、回帰係数は遡る先行時間に関して指数関数的に減少することが分かった。これより、先行する体積歪の変動時系列と指数関数の線形畳み込み（linear convolution）から背景地震活動を与える非定常 ETAS モデルの最尤推定値を求めると、どの群発地震でも殆ど同一な減衰係数となることが解った。しかし他方で、指数関数の切片パラメータは群発地震の開始位置から歪計までの距離に依存していることが分かった。すなわち群発開始位置を与えれば、歪データから正確な背景地震発生率を得ることができる。結局二つの独立した解析結果が正当であることを裏付けられた。すなわち（1）ベイズ型非定常ETASモデルを使用した群発地震の背景活動度のインバージョン結果の精度が高いこと（2）群発地震の背景率とマグマ貫入の場所に歪観測点からの距離に依存する体積応力の時間変化との間に物理的に自然な因果関係がみられることである。マグマ貫入などの流体の作用によるストレス変化の定量化のメカニズムを理解する有力なヒントになりうると考える。

統計的言語モデルを用いた化学構造の特徴抽出と分子設計への応用

総合研究大学院大学　池端 久貴　北陸先端科学技術大学院大学　本郷 研太
(株)地球快適化インスティテュート　磯村 哲
北陸先端科学技術大学院大学　前園 涼　統計数理研究所　吉田 亮

1　背景

医薬品や材料を含む多く製品の開発現場において、計算機を利用した分子設計は非常に有用な手法である．データベースに含まれる既存の化学構造からの特性予測はQSAR(Quantitative Structure Activity Relationship)モデルと呼ばれ，これまでに多くの特徴抽出方法の提案，回帰モデルを始めとしたさまざまな統計的手法が適用されている[A. Varnek , and I. Baskin]. しかしながら，特性から化学構造の予測であるInverse-QSAR問題は，QSAR問題に比べると研究成果の報告件数はかなり少ない．その理由としては化合物空間の広大さが考えられ，例えば，元素をC，N，O，Sに限定し，非水素原子の総数を30までとした場合でも，可能な有機分子の組み合わせ総数は10^{60}個と膨れ上がる．

これまでInverse-QSAR問題に対して提案されている手法は２種類に大別できる．前者は特徴空間上のQSARモデルから得られた領域の内点に対応する化学構造を全列挙する方法で，後者は環構造を維持したまま化学構造を断片化して得られたフラグメントを構成単位とし，遺伝的アルゴリズムを用いて目的関数を最小化する化学構造を探索するといった手法である．

本研究では，統計的言語モデルと逐次モンテカルロ法を組み合わせた分子設計法を提案する．既存化合物から学習されたパターンを基に単語単位の生成モデルを構成できるため，これまでのフラグメントを用いた手法でカバーしきれなかった多様な化学構造を得ることができる．また，逐次モンテカルロ法を用いることで，遺伝的アルゴリズムや全列挙による手法ではできなかった確率的な取り扱いが可能となる．

2　手法

2.1　n-gramモデルによる化学構造のモデリング

SMILES形式は文字列で化学構造を表現するために広く用いられている形式の一つである. 各元素は標準的な略記形が用いられ，有機分子骨格を形成するB，C，N，O，P，S，F，Cl，Br，I以外は[Fe]のように角括弧を用いる. 環構造は始点と終点を数字で, 側鎖部分は丸括弧で囲むことで表現される．このSMILES形式で表現された化学構造をn-gramモデルを仮定し，学習することで，化学構造に特徴的なパターンを抽出し，化学構造のランダムサンプリングが可能となる．ただし，化合物の環構造，側鎖などの分岐構造を文字列で表現する際は離れた文字間の依存関係が現れるため，以下のような工夫を施す.

- 開始している環（ペアが現れていない数字）の個数，開始している側鎖の有無の条件毎に学習を行う
- 閉じた側鎖の情報を削減するオペレータΨの導入　$\Psi(\mathrm{CCC(CCCC)O}) = \mathrm{CCC(C)O}$

以上を考慮した上で，SMILES文字列を$\boldsymbol{s} = (s_1, \cdots, s_{|s|})$，$a$文字目から$b$文字目までの部分文字列を$\boldsymbol{s}_{[a:b]}$とするとき，改良n-gramモデルは以下のように表すことができる.

$$p_n(\boldsymbol{s}) = \prod_{i=1}^{|s|} \sum_{k=1}^{|A|} {p_n}^{(k)} \left(s_i \mid \Psi\left(\boldsymbol{s}_{[i-n+1:i-1]}\right) \right) I\left(\boldsymbol{s}_{[1:i-1]} \in A_k\right), \tag{1}$$

ここで${p_n}^{(k)}$はグループA_kにおける文字生成確率を表し，$|A|$個あるグループへの割り当ては生成文字以前の部分文字列の非ペアの数字の個数，閉じていない側鎖の有無から非確率的に決定される．パラメータ推定は既存化学構造からKneser-Ney smoothing [S. F. Chen *et.al*] などの推定方法を用いる.

2.2　Inverse-QSARモデリング

ある目標の特性y^*が与えられたときに，その構造$\boldsymbol{s}$についての条件付き分布$p(\boldsymbol{s} \mid y^*)$を得たい. ベイズの定理を用いると$p(\boldsymbol{s} \mid y^*) \propto p(y^* \mid \boldsymbol{s})p(\boldsymbol{s})$となり，右辺第一項はQSARモデルとなり, 第二項は構造に関する事前知識となる．これは式(1)で表される構造モデルに対応する． QSARモデルを構築する際には，化学構造を記述子と呼ばれるベクトルで表現されたものを入力とすることが多い．本稿では適当な記述子 $\boldsymbol{\phi}: \mathbf{s} \to x$ を用いてp次元のベクトル$\boldsymbol{x}$が得られたとする．目的の特性値$\boldsymbol{y}$は1次元実数値とし，未知の係数ベクトル$\boldsymbol{w}$を用いて，$y = \boldsymbol{w}^T\boldsymbol{x} + \epsilon^2$のように$\boldsymbol{x}$との間に線形の関係を仮定する．ここで$\epsilon^2 \sim^{iid} N(0, \sigma^2)$とする．更に事前分布として，$\boldsymbol{w} \sim N\left(\mathbf{0}, {V_0}^{-1}\right)$，$\sigma^2 \sim IG(a_0, b_0)$を仮定する．訓練データ$D = \{X = (\boldsymbol{x^T}_1, \cdots, \boldsymbol{x^T}_N)^T, \boldsymbol{y} = (y_1, \cdots, y_N)\}$が得られたとき, 新しい入力$\boldsymbol{x}^*$に対する特性値$y^*$は次のようなnon-standardized Student's t-distributionに従う．

(2) $$p(Y = y^* | \boldsymbol{x}^*, D) = T\left({\boldsymbol{w}_N}^T\boldsymbol{x}^*, \frac{b_N}{a_N}\left(1 + \boldsymbol{x}^{*T}V_N\boldsymbol{x}^*\right), 2a_N\right),$$

ここで，$\boldsymbol{w}_N = V_N X^T\boldsymbol{y}$，$V_N = \left({V_0}^{-1} + X^TX\right)^{-1}, a_N = a_0 + \frac{N}{2}, b_N = b_0 + \frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y} - {\boldsymbol{w}_N}^T{V_N}^{-1}\boldsymbol{w}_N\right)$である．

通常、目標の特性は$[y_1, y_2]$ のように区間を取ることが多い．このとき$\boldsymbol{s}$の事後分布は，以下のように表される．

(3) $$p(\boldsymbol{s} \mid y_1 \le Y \le y_2) \propto \int_{y_1}^{y_2} p(Y = y^* | \boldsymbol{s}, D)dy^*\, p_n(\boldsymbol{s})$$

2.3 逐次モンテカルロ法

得られた$\boldsymbol{s}$の事後分布を目標分布とし，逐次モンテカルロ法［P. Del Moral *et.al*］を適用することで式(3)に対応した化学構造の事後分布を得る．ステップ t における目標分布γ_tを

(4) $$\gamma_t = \{q(\boldsymbol{s})\}^{\kappa(t)}p_n(\boldsymbol{s}).$$

とする．ここで，$q(\boldsymbol{s}) = \int_{y_1}^{y_2} p(Y = y^* | \boldsymbol{s}, D)dy^*$で，$\kappa(t)$は正の単調関数で少なくとも最終ステップにおいては$\kappa(t) = 1$ となる関数である．逐次モンテカルロ法の手続きとしては以下のようになる．

1. initialize
2. at time t
3. Local move from ${\boldsymbol{s}_{t-1}}^{(i)}$ to ${\boldsymbol{s}_t}^{(i)}$ with the following procedure ($i = 1, \cdots N$)
   - obtain $u \sim Binom(L, 0.5)$ for prefixed L.
   - reduce u letters from the right edge and generate $L - u$ letters from the right edge with probability in the learned probability from the equation(1).
4. update weights as ${w_t}^{(i)} := {w_{t-1}}^{(i)} \frac{\{q({s_t}^{(i)})\}^{\kappa(t)}}{\{q({s_{t-1}}^{(i)})\}^{\kappa(t-1)}}$, for $i = 1, \cdots N$, and normalize them to obtain ${W_t}^{(i)}$
5. if $\left(\left(\frac{1}{N}\sum_{i=1}^{N} {W_t}^{(i)^2}\right)^{-1} \le {N}/{2}\right)$, then resample with weights as the probability, and ${W_t}^{(i)} = \frac{1}{N}$
6. Reordering the randomly chosen atom being at the first position by some software such as OpenBabel.
7. n:=n+1, then go back to 2.

3 検証

材料開発への適用例を発表にて報告する．

参考文献

[1] A. Varnek , and I. Baskin, Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis, *Journal of chemical information and modeling*, 52, 1413-1437, 2012.

[2] S. F. Chen and J. Goodman, An Empirical Study of Smoothing Techniques for Language Modeling, *Computer Speech and Language*, **13**, 359-394, 1999.

[3] P. Del Moral et.al, Sequential Monte Carlo samplers, *J. R. Statist. Soc. B*, 68, 411-436, 2006.

Bay-Bridge : local linear approximated bridge and its model building procedure via the Bayesian model

中央大学大学院理工学研究科　　保科　架風

1　はじめに

近年の様々な技術の発展に伴い, 多くの場面で大規模かつ複雑な構造を有するデータが獲得・蓄積されている. そのようなデータから線形回帰モデリングによって現象のメカニズムの特定や予測を行う際, 最尤・最小 2 乗推定やモデル選択基準に基づく変数選択が適用されてきた. しかし, これらの手法では推定の精度や計算コストの面で問題が生じることがある. これに対し, モデルの推定と変数選択を同時に行うスパース回帰モデリングでは, このようなデータを効率的に取り扱うことが可能であり, この 20 年間盛んに研究が進められ, 飛躍的な発展を遂げている.

スパース回帰モデリング手法の 1 つである bridge 回帰 (Frank and Friedman, 1993) は 2 つの調整パラメータの値を適切に選ぶことで回帰係数ベクトルの幾つかの成分を厳密に 0 と推定することが可能であり, 様々なデータから有益な情報を抽出することができる. これに対し, Kawano (2014) では GBIC 基準 (Konishi *et al.*, 2004) によって bridge 回帰の調整パラメータを選択する手法を提案している. しかし, 複数の調整パラメータの存在はモデル選択のコストを増大させる. また, bridge 回帰の推定では非凸最適化問題を扱う必要がある.

これに対し本研究では, 1. bridge 回帰の回帰係数ベクトル, 誤差分散, 1 つの調整パラメータをベイズモデルに基づく MAP 推定を行うことでモデル選択の次元を下げ, 2. 残りの調整パラメータの値を選ぶためのベイズモデルに基づくモデル選択基準の導出を行った. 一般的にベイズモデルに基づくモデル選択基準はサンプルサイズに依存するラプラス近似を使用しており, 高次元データに対してしばしば有効に機能しなくなる. これに対し本研究におけるモデル選択基準の導出では近似にモンテカルロ積分を適用し, また, 高次元モデルでの有効性が知られている EBIC (Chen and Chen, 2008) 基準の考え方を取り入れた. これにより, 提案手法は高次元データのモデリングにおいても有効に機能することが期待できる.

2　提案手法

目的変数 y, p 次元説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ に対するサンプルサイズ n の線形回帰モデル

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

を考える. ただし, n 次元目的変数ベクトル $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, 切片項 β_0, すべての成分が 1 の n 次元ベクトル $\mathbf{1}_n$, $n \times p$ 計画行列 $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ $(\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T,\ i = 1, \ldots, n)$, p 次元回帰係数ベクトル $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ であり, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ を n 次元誤差ベクトルとし, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, σ^2 を未知の誤差分散とする. また, 一般性を失うことなく $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{ij} = 0$, $\sum_{i=1}^n x_{ij}^2 = n$, $(j = 1, \ldots, p)$ と目的変数と説明変数を基準化する.

このとき, bridge 回帰 (Frank and Friedman, 1993) は次の L_q 正則化法として定義が与えられる:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right]. \tag{2}$$

$q \leq 1$ のとき, bridge 回帰は lasso と同様にスパース性を持つが, $q < 1$ のときの推定では非凸最適化問題を解く必要がある. また, bridge のモデル選択では λ と q の最適な組み合わせを選ぶ必要があり, 候補のモデルの数が増大する.

これに対し, 本研究では bridge にベイズモデルの視点から $\boldsymbol{\beta}$, $\delta = 1/\sigma^2$, λ^2 を MAP 推定する以下の Bay-Bridge (Bayesian MAP estimation with local linear approximated bridge regression) を提案した:

$$(\hat{\beta}, \hat{\delta}, \hat{\lambda}^2) := \underset{\boldsymbol{\beta}, \delta, \lambda^2}{\operatorname{argmin}} \left[-\frac{n-2}{2} \log \delta + \frac{\delta}{2} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda q \sqrt{\delta} \sum_{j=1}^{p} w_j |\beta_j| - \frac{p + \nu_\lambda - 2}{2} \log \lambda^2 + \frac{\eta_\lambda}{2} \lambda^2 \right]. \quad (3)$$

ただし, $w_j = |\beta_j^{(0)}|^q (j = 1, \ldots, p)$, $(\nu_\lambda/2, \eta_\lambda/2)$: λ^2 のガンマ事前分布のパラメータである. また, この MAP 推定では bridge に local linear approximation (Zou and Li, 2008) を適用したものを基にしている.

さらに本研究では Bay-Bridge に対するモデル選択基準として EBIC 基準の周辺尤度の近似にモンテカルロ積分を適用し, 以下のモデル選択基準を導出した:

$$\begin{aligned} \mathrm{EBIC}(q) =& n \log 2\pi + 2p^* \log 2 - \log \Gamma\left(\frac{n-p^*}{2}\right) - \nu_\lambda \log \frac{\eta_\lambda}{2} + 2 \log \Gamma\left(\frac{\nu_\lambda}{2}\right) \\ & - 2 \log \Gamma\left(p^* + \frac{\nu_\lambda}{2}\right) - \log |W| + \log \gamma + \Gamma\left(\frac{1}{2}\right) + 2 \log M \\ & - 2 \log \sum_{m=1}^{M} |A_{(m)}|^{-1/2} \left\{ \frac{1}{2} \boldsymbol{y}^T (I_n - X_{\mathcal{A}_q} A_{(m)}^{-1} X_{\mathcal{A}_q}^T) \boldsymbol{y} \right\}^{-(n-p^*)/2} \\ & \left\{ \frac{1}{2} \left(\sum_{\hat{\beta}_j \neq 0} \tau_{j(m)}^2 + \eta_\lambda \right) \right\}^{-p^* + (\nu_\lambda/2)} \exp\left(\sum_{\hat{\beta}_j \neq 0} \gamma \tau_{j(m)}^2 \right) + 2\alpha \binom{p}{p^*}. \end{aligned} \quad (4)$$

ただし, $p^* = |\mathcal{A}_q|$, $\mathcal{A}_q = \{j \mid \hat{\beta}_j \neq 0, j = 1, \ldots, p\}$, $X_{\mathcal{A}_q}$ は $\mathcal{A}_q$ に対応する説明変数で構成される計画行列, $A_{(m)} = X_{\mathcal{A}_q}^T X_{\mathcal{A}_q} + W D_{(m)}^{-1}$, $W = \mathrm{diag}(\xi_j^2 | j \in \mathcal{A}_q)$, $\xi_j = q w_j$ $(j = 1, \ldots, p)$, $D = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$, $D_{(m)}$: 対角成分に $\{\tau_{j(m)}^2 \mid j \in \mathcal{A}_q\}$ を持つ対角行列, $\{\tau_{j(m)}^2 \mid j \in \mathcal{A}_q, m = 1, \ldots, M\}$: $\prod_{j \in \mathcal{A}_q} \mathrm{Gamma}(\tau_j^2 | 1/2, \gamma)$ からのサイズ M のランダムサンプルである. これにより, 高次元データに対するモデリングにおいて有効に機能するモデル選択基準を導出することができた.

参考文献

[1] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

[2] Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35, 109–135.

[3] Kawano, S (2014). Selection of tuning parameters in bridge regression models via Bayesian information criterion. *Statistical Paper*, 55, 1207–1223.

[4] Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91, 27–43.

[5] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36, 1509–1553.

潜在的歩数増加パターンと体組成指標との統計的関連性についての分析

大橋洸太郎（立教大学社会情報教育研究センター・慶應義塾大学システムデザインマネジメント研究所）　小熊祐子（慶應義塾大学大学院健康マネジメント研究科・スポーツ医学研究センター）渡辺美智子（慶應義塾大学大学院健康マネジメント研究科）

1. 目的

本研究は，生体ログデータの計量化のために混合成長曲線モデルを適用する一連の統計的分析処理技法を提案し，その適用例として分析対象を年代や性別によって層別した上で，「歩数」特に「累積歩数」に関する潜在的なパターンの相違が体組成に影響するプロセスモデルを例示することを目的としている。厚生労働省(2013)では，がん，循環器疾患，糖尿病や慢性閉塞性肺疾患を生活習慣病と位置づけ，日本国民のこれらの疾患への対策として，食生活の改善や運動習慣の定着に根差した一次予防と呼ばれる分野に重点を置いた対策が望まれていることを述べている。このような中で今回我々は，混合成長曲線モデル(Growth Mixture Model，GMM)の応用と，その結果に共分散分析を組み合わせることが，運動に伴う生体ログデータの解析に有効ではないかと考えた。また適用例として，独立行政法人情報通信研究機構(National Institute of Information and Communications Technology，NICT)の採択課題「課題 A ソーシャル・ビッグデータ利活用アプリケーションの研究開発」における「ヘルスリテラシー向上のための生体ログデータ分析に基づく健康情報フィードバック」プロジェクトで収集が行われている「歩行」と「体組成指標」のデータを用いた。

2. 方法

プロジェクトへの協力者の内，性別と年齢の情報と，4 週間分 28 日以上の歩行記録が揃っていた 2830 名(男性 1757 名，平均 48。38 歳(SD=11。86)，女性 1073 名，平均 41。59 歳(SD=12。29))を分析対象とした。また協力者を人数の多かった層を中心に「30 代男性」「30 代女性」「40 代男性」「40 代女性」「50 代男性」「50 代女性」の 6 層に分けた。続いてこれらの各層の協力者の 1 週目から 4 週目までの「累積歩数」に関して GMM を適用し，各層が何群に分けられるのかを検討した。そして分類による効果を吟味するために，分類された各「群」と「経過週数」を説明変数とする共分散分析を行った。検定の目的変数となる体組成指標には，株式会社タニタの測定器による「内臓脂肪レベル」，野村・渡辺・小熊(2015)で提案された「体型指標」「隠れ肥満指標」の 3 つを用いた。「経過週数」は 1 週目の累積歩行数と 2 週目の体組成の指標の平均値の組合せを「1 週経過」，2 週目の累積歩行数と 3 週目の体組成の指標の平均値を「2 週経過」，3 週目の累積歩行数と 4 週目の体組成の組合せを「3 週経過」として 3 時点のデータを作成し，前の週までの累積歩行数が次の週の体組成にどれほど影響しているのかを検討した。

3. 結果

GMM の結果については，50 代男性と 50 代女性について Table1 に掲載した。AIC による群数の比較の結果, 50 代男性, 50 代女性共に 5 つの群に分類することが適切であることが分かった。

Table 1 : GMMによる群数

	50代男性	50代女性
群数	AIC	AIC
4	30208.51	11458.79
5	29901.77	11306.53
6	29907.77	11312.53

Table 2 : 50代男性の「体型指標」の群毎の平均値

	50代男性		
群番号	体型指標	隠れ肥満指標	内蔵脂肪レベル
5	1.403	0.743	9.73
4	1.912	0.160	11.111
3	3.101	−0.045	12.789
2	3.298	−0.271	12.823
1	4.193	−0.825	13.991

Table2 に 50 代男性の「体型指標」「隠れ肥満指標」「内臓脂肪レベル」の群毎の平均値を示した。群番号が大きい程多く歩いていた群を指す。これらの指標は値が小さい程体組成が良いと判断されるものであり，この結果から多くの歩数を記録した群程体組成指標もよいとわかった。

共分散分析の結果については，50 代男性の「体型指標」の結果を Table 3 に掲載した。Table 3 における身長，体重，基礎代謝，筋肉量は共変量である。これらの変数を統制した上で 50 代男性は，「群」「経過週数」による体型指標の変化に有意な差がみられた。またこの他 50 代男性には，「隠れ肥満指標」「内臓脂肪レベル」に「群」による有意差がみられていた。

Table 3 : 50代男性「体型指標」

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
群	4	421.0	105	1.579E+03	2.000E−16 ***
経過週数	2	1.0	1	8.121E+00	3.330E−04 ***
身長	1	0.0	0	2.750E−01	5.990E−01
体重	1	6993.0	6993	1.049E+05	2.000E−16 ***
基礎代謝	1	1.0	1	2.027E+01	8.170E−06 ***
筋肉量	1	1.0	1	1.704E+01	4.210E−05 ***
群×経過週数	8	0.0	0	4.500E−02	9.990E−01
残差	570	38.0	0		

4. 考察

4 週間，1 ヶ月程度の歩行記録をつけ，歩いていくことで「体型指標」は有意に変化していくことが分かった。また，群による違いは「体型指標」「隠れ肥満指標」「内臓脂肪レベル」にみられ，より多く歩いた記録を持つ群の方が，そうでない群よりも体組成指標が改善する様子が統計的に確認された。ただし今回分析対象としたデータは, 記録のあった協力者の測定結果であり, この結果の中には測定機器の非装着による歩行数の記録の未計上分が含まれていると考えられる．例えば日に 10 時間以上の装着時間のあるものを有効データとするといった，データの峻別を視野に入れ，今後の分析を行いたい

5. 文献

・厚生労働省，2013，二十一世紀における第二次国民健康づくり運動 (健康 21(第二次)) ，厚生労働省 HP<http://www1.mhlw.go.jp/topics/kenko21_11/b2.html#A22>(最終閲覧日, 2016 年 1 月 7 日).

・野村俊一・渡辺美智子・小熊裕子，2015，体組成指標の推移パターンの解析と適正な減量の判別，第 6 回横幹連合コンファレンス論文集，pp223-226.

・タニタからだカルテ HP <http://www.karadakarute.jp/tanita/column/columndetail.do?columnId=172> (最終閲覧日 2016 年 1 月 7 日)

生体ログデータを使用した歩行パターンと体組成指標との関連を検討する分析方法の提案

渡辺　真弓[1]　渡辺　美智子[1]　小熊　祐子[1,2]

[1]慶應義塾大学健康マネジメント研究科　[2]スポーツ医学研究センター

E-mail: yamam00@sfc.keio.ac.jp

1.　はじめに

身体活動の増加は，将来的な疾病の予防という効果に留まらずメンタルヘルス・自己効力感の向上，腰痛・膝痛軽減，上気道感染予防等に寄与することが知られている．一方，身体活動の不足は生活習慣病発生の危険因子である[1)]．近年，スマートフォン等の普及等によって，日常生活中の情報をデジタルデータとして蓄積できる環境が整ってきた[2)]．本研究の目的は，歩数計や体重計において蓄積された生体ログデータの活用方法の提案である．身体活動に関する情報を用いて歩行パターンを導出し，健康状態との関連を検証する一連の過程を提示する．

2.　方法

歩数データには，活動量計で記録された 232300 件，3812 人分の 2013 年 4 月 1 日から 2014 年 3 月 31 日までの一年間の歩行データが時間単位で保存されている．まずは各データの ID 平均からの偏差を求め，この偏差を第 1～3 四分位点で 4 分割した．ただし，0 時から 5 時の間は 2 分割した．このデータを用いて，歩行パターン把握のために潜在クラス分析を行った．クラス数は AIC，BIC の情報量基準を参考にして決定した．分析には，MplusVersion7.11 を使用した．

体組成データは，歩行データと同じく 2013 年 4 月～2014 年 3 月の 1 年間に体組成計にて測定された，男性 361772 件（5771 人），女性 213102 件（5354 人）のデータである．体重，体脂肪率，筋肉量，基礎代謝量，体内年齢のデータを使用し男女別に主成分分析を行った．分析には，JMP12.0.1 を使用した。

3.　結果

3.1　歩行パターン

AIC, BIC の値と解釈妥当性を考慮して, 4 クラスを選択した. 各クラスの構成割合は, クラス 1 が 22.1%, クラス 2 が 21.6%，クラス 3 が 28.4%，クラス 4 が 27.6%だった．　応答確率から，クラス 1 を「休日型アクティブ日」，クラス 2 を「休日型非アクティブ日」，クラス 3 は「通勤型アクティブ日」，クラス 4 は「通勤型非アクティブ日」とした．各クラスへの所属確率の ID 平均を算出し，最も高い所属確率を示すクラスをその者のクラス（歩行パターン）とした．また，平均歩数を 3 分割し，歩行パターンと平均歩数を掛け合わせて，12 類型を求めた．　12 類型毎に，4 つの歩行パターンに対する所属確率割合の月次推移を Fig. 1 に示した.

3.2　体組成パターン

体組成データの主成分分析の結果, 主成分 1 は総合的な身体の大きさ，中でも太さに強く関連する指標であると考えられた．主成分 2 は，身体がどの程度引き締まっているかを示す指標である考えられた.

3.3　歩行パターンと主成分パターンの結合

主成分1と主成分2の主成分得点を算出し，歩行データと結合させた. この結果, 男性 690 人，女性 290 人のデータを得た．歩行パターン 12 類型毎に主成分 1 と主成分 2 の月次推移を示したのが Fig. 2 である．男性においては, 平均歩行数が「少」で「休日型アクティブ日多」「通勤型非アクティブ日多」の者は 5 人以下だったため，分析からは除外した．総合的な身体の大きさ（太さ）は，平均歩数が増加するほど下がり，痩せる傾向にあると言える．しかし，平均歩数が同程度であって

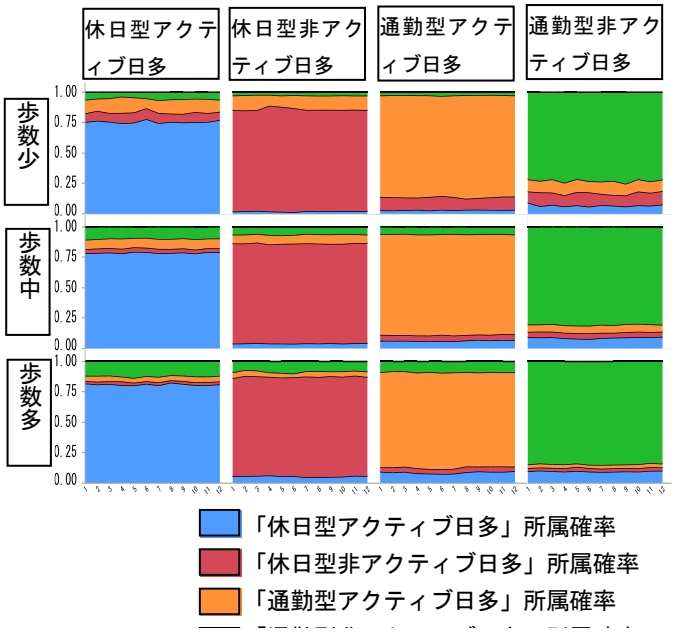


Fig. 1　Posterior probabilities for each walking pattern by average walk steps

も，どの歩行パターンが多いかということで主成分 1 の得点は異なり，通勤型非アクティブ日が多い者の得点が低く，痩せている傾向にあると言える．身体の引き締まり度は，やはり平均的な歩数が多くなるほど低くなる傾向にあり，身体が引き締まっていることが分かる．同程度の平均歩数であっても，特に「休日型アクティブ日」が多い者と「通勤型アクティブ日」が多い者で低く，身体が引き締まっている傾向にある．主成分 1 の得点が低く，総合的には痩せている傾向にある平均歩数「多」で「通勤型非アクティブ日多」の者は，引き締まり度合いはあまり高くないという結果であった．

次に女性においては，平均歩数「少」で「休日型アクティブ日多」「通勤型非アクティブ日多」の者，平均歩数が「多」で「通勤型アクティブ日多」の者は 1-2 人と非常に少数だったため，分析からは除外した．総合的な身体の大きさは，平均歩数が多くなるほど低くなり，痩せる傾向にある．同程度の平均歩数で比較すると，歩行パターンによる大きな違いは見られない．身体の引き締まり度は，平均歩数が少ない者で低く，身体が引き締まっている傾向にある．これは，一般的な認識とは逆の傾向であるが，データ数の少なさに起因することも考えられる．

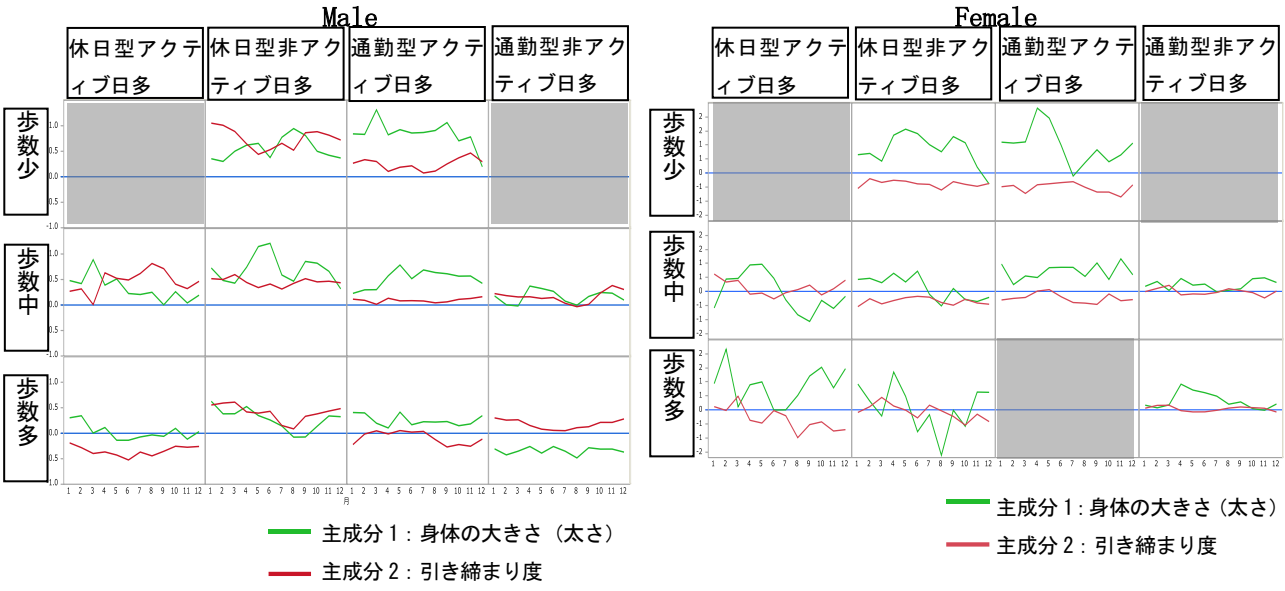


Fig.2 principal component scores by each walking patterns

4. 考察

平均的な歩数が多くなるほど，身体は痩せ，引き締まる傾向にあった．これは，歩行を行うほど筋肉が増加し健康に良いと言った一般的な認識を裏付けるものであると言える．ただし，男性においては同程度の平均歩数でも，歩行パターンによって体組成は異なる傾向を見せた．「休日型」「通勤型」の双方で，アクティブな日を多く過ごす者では，そうでない者と比べてより引き締まり度合いが高まり，健康的な身体状態であると考えられた．平均的に同程度の歩数であっても，歩行パターンが異なることで身体への影響も異なることが示唆される．また，外見的にはスリムでも，筋肉量が少なく不健全な痩せ方をしている者の存在が示唆され，今後の検討に値する．ただし，女性においてはこのような傾向は見られなかった．むしろ平均歩数が少ない程身体が引き締まるという，解釈が困難な結果となったが，この点に関してはデータ数の少なさが大きく影響している可能性を否定できず，今後データ数を増加した上で再検討する必要がある．

謝辞

本研究は，独立行政法人情報通信研究機構（NICT）の委託研究「ソーシャル・ビッグデータ利活用・基盤研究の研究開発」における採択課題「ヘルスリテラシー向上のための生体ログデータ分析に基づく健康情報フィードバック」の支援を受けたものである．

参考文献

1) 厚生労働省，健康づくりのための身体活動基準 2013（2013）
2) 西山勇毅他，ライフログデータを用いたチームの行動変容促進，情報処理学会論文誌，56 巻 1 号，349/361(2015)

体型を示す新たな体組成指標の提案とそのダイナミクス

野村 俊一[1,2]　渡辺 美智子[3]　小熊 祐子[3,4]

[1] 東京工業大学情報理工学研究科 [2] 慶應義塾大学システムデザイン・マネジメント研究科

[3] 慶應義塾大学健康マネジメント研究科 [4] 慶應義塾大学スポーツ医学研究センター

E-mail: nomura@is.titech.ac.jp

1　はじめに

本稿では，成人男女の体重および体脂肪率の分布に基づいて構成した，体型および隠れ肥満度を判定するための２つの指標を提案する．通常の肥満には該当しない体重であっても，提案指標により体脂肪量の多寡を基に隠れ肥満の傾向にあるかどうかを判断することができる．さらに，１年間にわたる個々人の体重および体脂肪率の計測記録を分析し，２つの提案指標のダイナミクスの違いを明らかにすることで，両指標を改善する減量ほど長期的に継続することが示された．また，提案指標について都道府県別に集計した結果を示し，メタボリックシンドロームの該当率と比較しながら地域的な体型の傾向を考察する． なお, 本稿における解析に用いるデータは，2013 年 4 月 1 日から 2014 年 3 月 31 日までの 1 年間にわたって継続的に計測された，個人別の体重および体脂肪率のログデータである．記録蓄積された体組成計測定データを分析して見出された体組成推移パターンに共通する性質を活用することで，個人ごとの健康状態や目的に応じた精度の高い健康指導情報の提供を目指す．

2　体型判定への新たな指標の提案

現在，肥満度の判定基準として最も利用されている指標に，BMI (Body Mass Index) ＝体重÷身長^2 (kg/m^2) がある．また，BMI と同様に，体脂肪率も肥満の判定に広く用いられている．BMI が健常な範囲内であっても体脂肪率が高いと，内臓脂肪が過剰な隠れ肥満の兆候があり，逆に体脂肪率が低いとアスリートのような筋肉質な体型といえる． BMI と体脂肪率は、両者の物理的単位が異なることから相関関係が非線形で捉えづらい．そこで，BMI を次式のように体脂肪にかかる部分（Fat Mass Index; FMI＝BMI × 体脂肪率）と体脂肪以外にかかる部分（Lean Mass Index; LMI＝BMI × (1－体脂肪率)）に分解する．FMI と LMI は同じ物理的単位 (kg/m^2) をもっており，体重に占める体脂肪量とその他の重量を，BMI のように身長に関して標準化した数値といえる．

ここで，FMI と LMI の変数の組に対して，男女ごとに主成分分析を行うと，それぞれ Fig. 2 の矢印のようにばらつきの大きい第一主成分と，それに直交するばらつきの小さい第二主成分とに分けることができる．この第一主成分および第二主成分をスケール変換および平行移動することにより得られる下式の体型指標（Body Size Index）および隠れ肥満指標（Hidden Obesity Index）を，体型分類のための新たな指標として提案する．

$$\text{体型指標（男性）} = 1.10 \times \text{FMI} + 0.85 \times \text{LMI} - 19.65$$
$$\text{隠れ肥満指標（男性）} = 0.85 \times \text{FMI} - 1.10 \times \text{LMI} + 16.70$$
$$\text{体型指標（女性）} = 1.20 \times \text{FMI} + 0.35 \times \text{LMI} - 13.10$$
$$\text{隠れ肥満指標（女性）} = 0.35 \times \text{FMI} - 1.20 \times \text{LMI} + 16.50$$

Fig. 1 の矢印は上式で提案した指標の座標軸を表しており，座標軸の原点，すなわち両指標が 0 である状態は，BMI の標準値 22 に対応している．また，隠れ肥満指標が一定の下，体型指標が 1 単位増えることは BMI が 1 単位増えることに相当している．従って，体型指標の符号は体重が標準に比べて過剰であるか否かを示しており，体型指標の正負から「痩せ型」か「肥満型」に体型を大きく分類することができる．一方，隠れ肥満指標は，体重の内訳を重視した指標であり，体重に占める体脂肪の比率が高くなるほど隠れ肥満指標は大きくなる．隠れ肥満指標は測定者の集団で平均 0 となるよう中心化されており，隠れ肥満指標の正負に

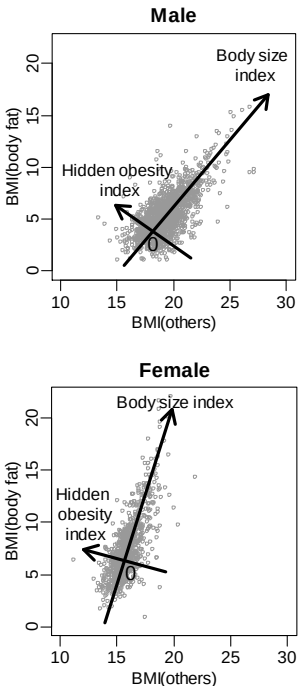


Fig. 1: Scatter plots of FMI and LMI with axes of proposed indices.

より「筋肉質型」と「隠れ肥満型」への分類が可能となる．隠れ肥満は外観からわかりづらいため，本人が自覚しないまま進行することが多く，それゆえ，この隠れ肥満指標は，隠れ肥満の進行度を表すバロメータとして有用であると考えられる．

3　提案指標の地域的傾向

上で提案した 2 つの指標について，計測者の居住都道府県別・男女別・年代別に集計を行い，提案指標の地域的な傾向を探る．Fig. 2 では 40 歳以上 59 歳以下男性の体型指標および隠れ肥満指標の平均値に加え，さらに，比較対象として厚生労働省の資料より集計したメタボリックシンドローム該当者の受診者に対する割合および該当者＋予備群該当者の割合を示している．全体的に見て，東北地方～北海道と，四国地方～九州地方にかけて，各指標の平均値およびメタボリックシンドローム該当割合が高い傾向にあることが分かる．男性の両指標の大きさとメタボリックシンドローム該当割合の高低は概ね同じ傾向を示している以上から，提案指標はメタボリックシンドロームと一定の整合性が取れた指標と見ることができるが，2 つの指標の地域性には違いも見られ，体型の類型により細かい地域性があるように考えられる．

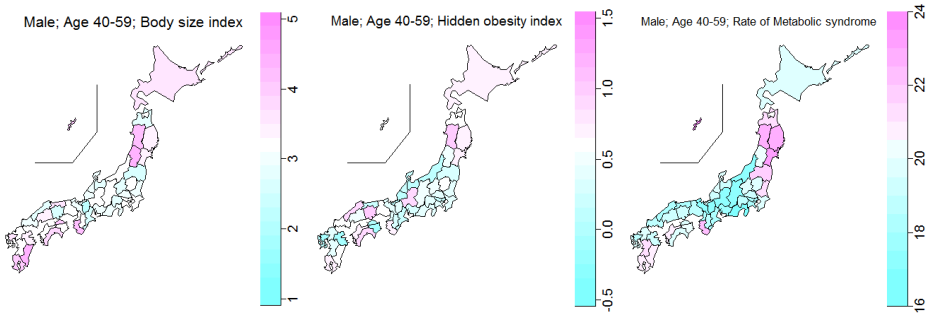


Fig. 2: Cartograms of proposed indices and rates of Metabolic syndrome (%).

4　提案指標の月別推移

ここでは，上で提案した 2 つの指標について，年間を通した各月における指標の推移を分析し，指標ごとの推移の仕方に現れる特徴を探る．分析するデータはここまでと同じ 2013 年 4 月から 2014 年 3 月の 1 年間の体重および体脂肪率の計測記録であるが，データの精度を確保するため，毎月 10 日以上にわたり計測が行われた男性 273 名，女性 97 名のデータに対象を絞った．体型指標と隠れ肥満指標それぞれの月次推移に有意水準 1%の単位根検定（Harris-Tsavalis test [1]）を行うと，男女ともに隠れ肥満指標の推移は定常性をもつという結論が得られる．つまり，隠れ肥満指標は，一時的に変化しても，長期的には各人の生活習慣に見合った水準へと落ち着くという平均回帰性をもっているといえる．一方，体型指標について同じ単位根検定を行っても定常性をもつという結論は得られず，一度増減した体型指標は自然と元には戻るものではなく，戻すための働きかけが必要になることを示唆している．

5　体重・体脂肪率の計測記録の平滑化

不定期かつ誤差を含む体組成計測記録から，精度の高い体重・体脂肪率の状態推移を得る手法として，ここではカルマンフィルターによる計測値の平滑化を提案する．Fig. 3 のように，体重および体脂肪率の実計測値は，大きく上下に振れているのに対して，カルマンフィルターによる平滑化推定値は安定した推移を示しており，計測ミスの影響も抑えられている．

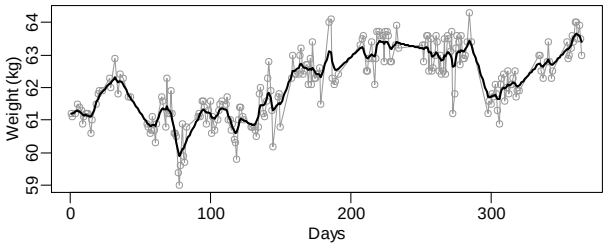


Fig. 3: Time series of weight with Kalman smoother.

謝辞

本研究は，独立行政法人情報通信研究機構（NICT）の委託研究「ソーシャル・ビッグデータ利活用・基盤研究の研究開発」における採択課題「ヘルスリテラシー向上のための生体ログデータ分析に基づく健康情報フィードバック」の支援を受けたものである．

参考文献

[1] Harris, R. D. F., and Tsavalis, E. Inference for unit roots in dynamic panels where the time dimension is fixed, *Journal of Econometrics*, 91, 201--226 (1999).