科研費シンポジウム
統計学，機械学習の数理とその応用
講演予稿集

日時: 2017 年 9 月 21 日(木)〜22 日(金)

場所: はこだて未来大学

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: 竹之内高志(はこだて未来大学)

公立はこだて未来大学
FUTURE UNIVERSITY HAKODATE

統計学，機械学習の数理とその応用

会場: はこだて未来大学

プログラム

9/21(木)

14:00-14:05 オープニング

14:05-15:35 (30 分×3)

・矢田 和善・青嶋 誠 (筑波大学)

スパース性に基づくノイズ掃き出し法について

・広津千尋 (明星大学、連携研究センター)

A unifying approach to the shape and change-point hypotheses

- All in one algorithm of p-value, power and confidence region-

・張 元宗, 篠崎信雄 (目白大学, 慶應大学)

事前情報に基づく P 次元ポアソン分布の母平均の同時推定とその応用

15:55-16:55　(30 分×2)

・江口真透 (統計数理研究所)

一般化平均によるモデルと推定

・林 賢一 (慶応義塾大学)

擬似線形関数を用いたクラスタワイズ回帰モデル

17:15-18:15　(30 分×2)

・小森　理 (福井大学)

準線形モデルによるポアソン点過程の拡張とその応用

・Stephen Wu (統計数理研究所)

From Small Data to Big Data: Integrating Machine Learning, Physical Model and Uncertainty Quantification for Efficient Polymer Design

19:30- 懇親会

9/22(金)

10:00-11:00　(30 分×2)

・川島 孝行,　藤澤洋徳 (総合研究大学院大学,　統計数理研究所)

一般化線形回帰のロバスト化およびスパース化

・富田 裕章 (総合研究大学院大学)

多重代入法を用いたバイアス補正推定量に関する考察

11:20-11:50　(30 分×2)

・川喜田雅則・藤澤洋徳 (九州大学, 統計数理研究所)

最適な半教師付き学習

・熊谷 亘, 金森敬文 (理化学研究所 革新知能統合研究センター, 名古屋大学)

パラメータ転移学習における汎化誤差の評価

12:20-12:25 クロージング

スパース性に基づくノイズ掃き出し法について

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

ゲノム科学，情報工学，金融工学などの現代科学の1つの特徴は，データがもつ次元数の膨大さにある．こういった高次元データの第一の特徴は，次元数が標本数を遥かに超えることである．第二の特徴は，高次元データは豊富な情報を有するものの，それが巨大なノイズに埋もれ見つけ難いことである．これらの理由から，通常の多変量解析法では高次元データの推測に精度を保証することができず，間違った解析結果を導くことさえある．そのため，高次元データの解析には，新しい理論と方法論が必要になる．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の次元数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，スパース性を基づくノイズ掃き出し法ついて論じた．閾値を用いて，固有ベクトルの推定量を補正することで，緩い仮定のもとそ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

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$n\ (\geq 3)$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，$\delta_j = \lambda_j^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_{(2)})/(n-1)$，$j = 1, ..., k$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．本講演では，$\tilde{\boldsymbol{h}}_j$ を補正した．いま，$\tilde{\boldsymbol{h}}_1 = (\tilde{h}_1,...,\tilde{h}_d)^T$ とおき，$\tilde{h}_1,...,\tilde{h}_d$ を絶対値の大きい順に並べ替えたものを $\tilde{h}_{(1)},...,\tilde{h}_{(d)}$ とおく．すなわち，$|\tilde{h}_{(1)}| \geq \cdots \geq |\tilde{h}_{(d)}|$ となる．$||\tilde{\boldsymbol{h}}_1||^2 > 1$ であることに注意すれば，$\sum_{s=1}^{k-1}\tilde{h}_{(s)}^2 < 1$，$\sum_{s=1}^{k}\tilde{h}_{(s)}^2 \geq 1$ となる k が一意に定まる．そのとき，$\tilde{\boldsymbol{h}}_1$ を次のようにスパース化する．

$$\acute{\boldsymbol{h}}_1 = (\acute{h}_1,...,\acute{h}_d)^T.$$

ただし，

$$\acute{h}_s = \begin{cases} \tilde{h}_s & (|\tilde{h}_s| \geq |\tilde{h}_{(k)}|) \\ 0 & (|\tilde{h}_s| < |\tilde{h}_{(k)}|) \end{cases}, \quad (s=1,...,d)$$

とする．そのとき，

$$||\acute{\boldsymbol{h}}_1 - \boldsymbol{h}_1||^2 = o_p(1)$$

なる一致性を高次元のもと示した．同様に，$\boldsymbol{h}_j$ $(j \geq 2)$ についてもスパース化し，その精度を理論的かつ数値的に検証した．

参考文献

[1] Aoshima, M. and Yata, K. (2017). Two-sample tests for high-dimension, strongly spiked eigenvalue models, *Statistica Sinica*, in press (arXiv:1602.02491).

[2] Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context, *Communications in Statistics. Theory and Methods, Special Issue: Honoring Zacks, S. (ed. Mukhopadhyay, N.)*, **38**, 2634-2652.

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

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

A unifying approach to the shape and change-point hypotheses -All in one program of *p*-value, power and confidence region-

広津千尋（明星大学連携研究センター）

1．序論

用量反応解析においては，通常，厳密な反応曲線を想定することが難しいために単調性，凸性，Ｓ字性のような形状制約がよく想定される．それらのうち正規分布モデルに対する単調仮説については，単調回帰(isotonic regression)がよく知られている．しかしながら，それは Bartholomew (1959, a, b)によってやや直観的に導入され，とくにこのような制約のある母数空間に対する最適性は自明ではない．さらに制約付き最小二乗法は，計算及び分布論が複雑で，凸性，Ｓ字性問題，一般の確率モデル，さらに 2 元配置交互作用問題への拡張には困難が伴う．一方，著者等のアプローチは Hirotsu (1982)で導かれた一般制約仮説に対する検定の完全類を基にしており，その意味での最適性を持っている．それによると，単調性，凸性，Ｓ字性仮説それぞれに対し，単純，2 重，3 重累積和に基づく単調増大な統計量が示唆される．そのうち本稿では規準化最大対比を用いる方法について論ずる．これらは理論的に一貫した方法で扱えるので，標題の'unifying'はその意味を込めている．

一方，これらの形状制約は変化点モデルと密接な関係がある．すなわち，単調性，凸性，Ｓ字性仮説はそれぞれ，段差変化点，スロープ変化点，変曲点モデルと対応する(Hirotsu and Marumo, 2002)．例えば，段差変化点モデルを表す対比は単調対比の典型であり，逆にすべての単調対比は段差変化点対比の一意正係数線形結合で表される．同様のことがスロープ変化点，変曲点モデルについても示される．最大対比統計量は，これら変化点モデルに対する efficient score 検定も与える．すなわち，従来統計学の二つの異なる流れの中で研究されてきた制約仮説と変化点問題を総合的に扱う事が出来，'unifying' にはこの意味も込められている．応用例として，医薬品医療機器総合機構では副作用自発報告が収集され，その経年変化が解析されている．その場合，単調増加傾向をいち早く検出すると同時に，増加傾向の生じた時点を推測することは応用上大変有意義である．さらに，正規分布に限らず，指数分布族が一貫した方法で扱えるという意味でも，本方法は 'unifying' である．

2．数理モデル

平均Λi の独立なポアソン系列 $yi\sim e-\Lambda i\Lambda iyi/yi!,\ \ i=1,\ \cdots,\ a,$ を想定する．帰無仮説 $Hm0:\ \Lambda 1=\cdots=\Lambda a$ を単調仮説 $Hm:\ \Lambda 1\leq\cdots\leq\Lambda a,$ に対して検定する最大対比検定は次で与えられる，

$$\max \text{acc.}\ t1=\max t1,\cdots,ta-1, \qquad (1)$$

$tk=tkYk=a-kak\Lambda-1/2\Lambda-Ykk,\ k=1,\ \cdots,\ a-1,\ Yk=y1+\cdots+yk,\ \ \Lambda=Ya/a.$

一方，それは次の段差変化点モデル

$$\begin{cases} \theta i=\log \Lambda i=\theta,\ i=1,\cdots,k, \\ \theta i=\log \Lambda i=\theta+\Delta,\ i=k+1,\cdots,a. \end{cases} \qquad Mmk: \tag{2}$$

における仮説

$$H\Delta:\ \Delta=0,\ \text{for all unknown } k. \tag{3}$$

に対する efficient score 検定も与える．

さらに，段差変化点モデル(2)に対しては，$K+1$ が変化点であるとする帰無仮説 $H0K+1:k+1=K+1$ を，そうではないとする対立仮説 $H1K+1:k+1\neq K+1$ に対して検定する問題にも興味があり，やはり最大対比検定統計量（1）が有用である．変化点の信頼領域がこの検定の反転で得られる．これらの検定とその検出力，および信頼領域計算を 1 プログラムで行うのが標題の 'all in one program' の意味である (Hirotsu and Tsuruta, 2017)．これらのうち，帰無仮説（3）に対する p-value 計算は Worsley (1986)と一致するが，他の計算は新しい提案である．これらの計算では，$t1,\cdots,tk$のtkを与えた同時条件付き分布に関する漸化式の更新を行う．その際，$tk+1$ を与えたtkの条件付き分布が必要になるが，それを既知とするか，プログラム内で計算するかがその違いとなる．既知とする Worsley 方式は 'all in one program' に纏められず，また，スロープ変化点，変曲点モデルには拡張出来ない．本法は一貫した方式でスロープ変化点，変曲点モデルに拡張出来る．最後に，累積和に基づく方法は数理的構造が極めて単純であるがゆえに，２元配置交互作用問題にも自然に拡張される．その場合は行，列の両方，あるいは一方に自然な順序を想定する等，様々な状況があり，理論上，応用上興味ある問題の宝庫である(Hirotsu, 2017)．

参考文献

Bartholomew, D. J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika* 46, 36-48.

Bartholomew, D. J. (1959b). A test of homogeneity for ordered alternatives II. *Biometrika* 46, 328-335.

Hirotsu C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* 69, 567-577.

Hirotsu, C. and Marumo, K. (2002). Change point analysis as a method for isotonic inference. *Scand. J. Statist.* 29, 125–138.

Hirotsu, C. (2017). *Advanced analysis of variance*. Wiley Series in Probability and Statistics.

Hirotsu, C, and Tsuruta, H. (2017). An algorithm for a new method of change-point analysis in the independent Poisson sequence. *Biometrical Letters* 54, 1-24, 2017.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family of random variables. *Biometrika* 73, 91–104.

事前情報に基づく p 個のポアソン母平均の縮小推定量とその応用

目白大学　張　元宗,　　慶応義塾大学　篠崎　信雄

1　はじめに

$X_1, \cdots, X_p$ を互いに独立にポアソン分布 $P_o(\lambda_i), i = 1, \cdots, p(\geq 2)$, にしたがう確率変数とする。標準化 2 乗誤差損失関数（normalized squared error loss）

$$L(\hat{\lambda}, \lambda) = \sum_{i=1}^{p} \lambda_i^{-1}(\hat{\lambda}_i - \lambda_i)^2 \tag{1.1}$$

を基準としたとき、母平均 $\lambda = (\lambda_{\mathbf{1}}, \cdots, \lambda_{\mathbf{p}})'$ の同時推定問題に対して、Clevension & Zidek (1975) は原点に縮小する推定量

$$\hat{\lambda}_i^{CZ}(\mathbf{X}) = \left(1 - \frac{\varphi(Z)}{Z + p - 1}\right) X_i, \qquad i = 1, \ldots, p,$$

を提案した。ここで、$Z = \sum_{i=1}^{p} X_i$ である。$\varphi(Z)$ が非減少関数、$0 \leq \varphi(Z) \leq 2(p-1)$ ならばこの推定量が不偏推定量 $\mathbf{X} = (X_1, \cdots, X_p)'$ を改良することを示した。しかし、いくつかの λ_i が大きな値である場合、原点に縮小する推定量は大きな改良を与えるとは言えない。Ghosh, Hwang & Tsui(1983) および Tsui (1984,1986) は、指定した非負の整数点あるいは順序統計量に縮小する推定量を提案した。しかし、両論文で提案された推定量は複雑で、改善の余地がある。ここで、事前情報に基づくポアソン母平均の縮小推定理論を統合し、ある正な値また観測値の最小値に縮小するような推定量のクラスを再構築する。さらに、ポアソン母平均に simple tree order 制約がある場合に、 isotonic regression 推定量を改良する方法を提案する。また、multiplicative Poisson models での母平均の同時推定問題も取り上げ、順序統計量への縮小推定量を提案する。

2　事前情報に基づく縮小

この節では、指定された非負な値及び順序統計量への縮小を論じる。次節で 1 つの応用例として、母平均に simple tree order 制約条件がある場合、isotonic regression 推定量を縮小する同時推定量を提案する。

2.1　非負な値への縮小

$a_i \geq 0, i = 1, \cdots, p$ とし、部分集合 $\mathcal{C} = \{(x_1, \ldots, x_p) | x_i \geq a_i, i = 1, \ldots, p\}$ とそのインジケータ関数を $I_{\mathcal{C}}$ とする。a_i に縮小する推定量をつぎのように考える。

$$\hat{\lambda}_i(\mathbf{X}) = X_i - \varphi(Z_{\mathcal{C}}) \frac{(X_i - a_i)}{Z_{\mathcal{C}} + d} I_{\mathcal{C}}, \qquad i = 1, \ldots, p.$$

ここで、$Z_{\mathcal{C}} = \sum_{i=1}^{p} (X_i - a_i)$ であり、$d > 0$ である。

部分集合 $\mathcal{C}$ で (1.1) の損失関数の下で、$\mathbf{X}$ と $\hat{\lambda}(\mathbf{X}) = (\hat{\lambda}_1(\mathbf{X}), \ldots, \hat{\lambda}_p(\mathbf{X}))$ との平均損失の差を評価することで、$\hat{\lambda}(\mathbf{X})$ が $\mathbf{X}$ を改良するための十分条件をつぎの定理で与える。

定理 2.1. $p \geq 2$ とする。損失関数 (1.1) の下で、$\hat{\lambda}(\mathbf{X})$ が $\mathbf{X}$ を改良するための十分条件は $\varphi(\cdot)$ は非減少関数で、$0 \leq \varphi(\cdot) \leq 2(p-1),\ d \geq \sup \varphi(\cdot)/2$ である。

次に、ある $k \geq 2$ に対して、部分集合 $\mathcal{C}_k = \{(x_1, \ldots, x_p) | x_i \geq a_i, i = 1, \ldots, k, x_j < a_j, j = k+1, \ldots, p\}$ とする。このような $2^p - p - 1$ 個の互いに排反な部分集合のそれぞれで a_i に縮小するような推定量を考える。つまり、$\mathbf{X} \in \mathcal{C}_k$ のとき、

$$\hat{\lambda}_i(\mathbf{X}) = \begin{cases} X_i - \varphi_k(Z_{\mathcal{C}_k}) \dfrac{(X_i - a_i)}{Z_{\mathcal{C}_k} + d_k}, & i = 1, \ldots, k, \\ X_i, & i = k+1, \ldots p, \end{cases}$$

を考える。ここで、$Z_{\mathcal{C}_k} = \sum_{i=1}^{k} (X_i - a_i)$ であり、 $d_k > 0$ である。各部分集合で $\mathbf{X}$ と $\hat{\lambda}(\mathbf{X})$ との平均損失の差を評価することで、$\hat{\lambda}(\mathbf{X})$ が $\mathbf{X}$ を改良するための十分条件をつぎの定理で与える。

定理 2.2　損失関数 (1.1) の下で、$\hat{\lambda}(\mathbf{X})$ が $\mathbf{X}$ を改良するための十分条件は $\varphi_k(\cdot)$ は非減少関数で、$0 \leq \varphi_k(\cdot) \leq 2(k-1),\ d_k \geq \sup \varphi_k(\cdot)/2$ である。

2.2　順序統計量への縮小

$p \geq 3$ とし、最小値 $X_{(1)} = min\{X_1, \ldots, X_p\}$ に縮小するような推定量を

$$\hat{\lambda}_i(\mathbf{X}) = X_i - \varphi(W) \frac{X_i - X_{(1)}}{W + d}, \qquad i = 1, \ldots, p$$

を考える。ここで、$W = \sum_{k=1}^{p} (X_k - X_{(1)})$ である。

定理 2.3 損失関数 (1.1) の下で、$\hat{\lambda}(\mathbf{X})$ が $\mathbf{X}$ を改良するための十分条件は $\varphi(\cdot)$ は非減少関数で、$0 \le \varphi(\cdot) \le 2(p-2), d \ge \sup \varphi(\cdot)/2$ である。

この十分条件は標本空間を p 個の互いに排反な部分集合に切り分け、各集合での平均損失の差を評価することで示される。

3 応用例－母平均に制約条件がある場合の isotonic regression 推定量の改良:

例 3.1. $X_i \sim P_o(\lambda_i), i = 0, 1, \ldots, p$ に従い、母数に simple tree order 制約条件、$\lambda_0 \le \lambda_i, i = 1, \ldots, p$ がある場合、$\boldsymbol{\lambda}$ の isotonic regression 推定量は次のように与えられる。

$$\hat{\lambda}_i^{st}(\mathbf{X}) = \begin{cases} X_i, & \text{for } i \in S^c \\ A_X(S), & \text{for } i \in S, \end{cases}$$

ここで、一般性を失うことなく、$S = \{0, 1, \ldots, k\}$, $S^c = \{k+1, \ldots, p\}$ であるとし、$A_X(S) = \sum_{i \in S} X_i/(k+1)$, $X_i \ge A_X(S)$, $i \in S^c$ である。$p - k \ge 2$ のとき、$\hat{\lambda}_i^{st}(\mathbf{X})$ を次のように縮小する。

$$\hat{\lambda}_i^{m}(\mathbf{X}) = \begin{cases} X_i - \varphi_{p-k}(W_{S^c}) \dfrac{X_i - A_X(S)}{W_{S^c} + d_{p-k}}, & \text{for } i \in S^c \\ A_X(S), & \text{for } i \in S, \end{cases}$$

ここで、$W_{S^c} = \sum_{i=k+1}^{p} (X_i - A_X(S))$ である。

定理 3.1 損失関数 (1.1) の下で、$\hat{\lambda}^m(\mathbf{X})$ が $\hat{\lambda}^{st}(\mathbf{X})$ を改良するための十分条件は $\varphi_{p-k}(\cdot)$ は非減少関数で,$d_{p-k} \ge \sup \varphi_{p-k}(\cdot)/2$ である。

また、母数に次のような制約条件が与えられる場合にも応用できる。

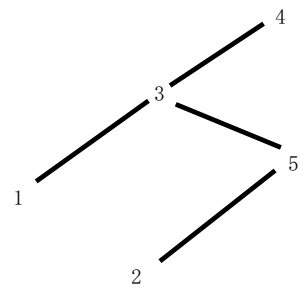

図 1

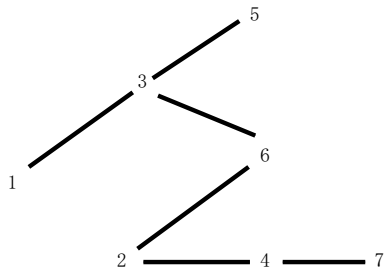

図 2

例 3.2. 図 1 に示されるように、母数に下記のような制約条件

$$\lambda_1 \le \lambda_3, \quad \lambda_3 \le \lambda_4, \quad \lambda_3 \le \lambda_5, \quad \lambda_2 \le \lambda_5$$

が与えられるとする。λ_i の isotonic regression 推定量を $X_i^*, i = 1, \ldots, 5$ とすると $X_4^* = X_4$、$X_5^* = X_5$ になるための必要十分条件は $X_4 \ge X_3^*$、$X_5 \ge \max(X_2^*, X_3^*)$ である。この場合に限って、(X_4, X_5) を $(X_3^*, \max(X_2^*, X_3^*))$ に縮小する。

例 3.3. 図 2 に示されるように、母数に下記のような制約条件

$$\lambda_1 \le \lambda_3, \quad \lambda_3 \le \lambda_5, \quad \lambda_3 \le \lambda_6, \quad \lambda_2 \le \lambda_4, \quad \lambda_2 \le \lambda_6, \quad \lambda_4 \le \lambda_7$$

が与えられるとし、λ_i の isotonic regression 推定量を $X_i^*, i = 1, \ldots, 7$ とする。そのとき、$X_5^* = X_5,\ X_6^* = X_6$ かつ $X_7^* = X_7$, の場合 (X_5, X_6, X_7) を $(X_3^*, \max(X_2^*, X_3^*), X_4^*)$ へ縮小する。$X_5^* = X_5,\ X_6^* = X_6$ かつ $X_7^* > X_7$ の場合。(X_5, X_6) を $(X_3^*, \max(X_2^*, X_3^*))$ に縮小する。同様に、$(X_5^* = X_5,\ X_6^* > X_6,\ X_7^* = X_7)$ 及び $(X_5^* > X_5,\ X_6^* = X_6,\ X_7^* = X_7)$ の場合にも縮小することができる。

4 multiplicative Poisson models での母平均の同時推定問題への応用

multiplicative Poisson models での母平均の同時推定問題を考えるとき、上記で述べた事前情報に基づく縮小に関する理論は最尤推定量の順序統計量への縮小にも応用することが出来る。

一般化平均によるモデルと推定

江口　真透
統計数理研究所

Let ϕ be a nonnegative, strictly increasing, convex function defined on $\mathbb{R}$. We discuss generalized geodesic connecting probability density functions $p(x)$ and $q(x)$ as

$$p_\phi(x,\pi) = \phi\big(\pi\phi^{-1}(q(x)) + (1-\pi)\phi^{-1}(p(x)) - \kappa_\phi(\pi)\big) \tag{1}$$

for π of $(0,1)$. A typical example of ϕ is given by an exponential function. Thus,

$$p_{\exp}(x,\pi) = \exp\big(\pi\log q(x) + (1-\pi)\log p(x) - \kappa_{\exp}(\pi)\big).$$

is nothing but the e-geodesic, and $\kappa_{\exp}(\pi)$ is called the cumulant function. We observe that

$$\frac{\partial}{\partial\pi}\kappa_{\exp}(\pi)\Big|_{\pi=0} = \int\{\log p(x) - \log q(x)\}p(x)dx,$$

which is the KL divergence between p and q. We find for a general ϕ that

$$\frac{\partial}{\partial\pi}\kappa_\phi(\pi)\Big|_{\pi=0} = \int\{\phi^{-1}(p(x)) - \phi^{-1}(q(x))\}\Xi(p(x))dx, \text{ say } \ D_\phi(p,q),$$

where

$$\Xi(p(x)) = \frac{\phi'(\phi(p(x)))}{\int\phi'(\phi(p(y)))dy}$$

Hence, we call $D_\phi(p,q)$ the generalized KL divergence confirming to satisfy the first axiom of a distance function, that is,

$$D_\phi(p,q) \geq 0 \text{ with equality if and only if } p = q \tag{2}$$

In summary, we the geodesic $\{p_\phi(x,\pi) : \pi \in [0,1]\}$ connecting density function p and q as given in (1) naturally associates with the divergence $D_\phi(p,q)$. In this discussion we can consider the dual geodesic defined by

$$p^*_\phi(x,\pi) = (1-\pi)\Xi(p(x)) + \pi\Xi(q(x)).$$

If $\phi = \exp$, then

$$p^*_{\exp}(x,\pi) = (1-\pi)p(x) + \pi q(x)$$

which is nothing but the m-geodesic. In accordance, $\{p_\phi(x,\pi) : \pi \in [0,1]\}$ is a generalization of e-geodesic; $\{p^*_\phi(x,\pi) : \pi \in [0,1]\}$ is that of m-geodesic. In this line, we observe a generalization of Pythagoras identity associated with orthogonal e-geodesic and m-geodesic.

If we adopt a parametric mode $p(x, \theta)$ for given data $\{x_i\}_{i=1}^n$, then we discuss a loss function given by

$$L_\phi(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \phi^{-1}(\Xi^{-1}(p(x_i, \theta))).$$

and the estimator for the parameter θ is proposed by $\hat{\theta}_\phi = \mathrm{argmin}_\theta\, L_\phi(\theta)$. In accordance with the formalation we confirm the asymptotic consistency for $\hat{\theta}_\phi$ as follows. Let $p(x, \theta_0)$ be the underlying density function of the data. Then the expected loss function is given by

$$\mathbb{L}_\phi(\theta, \theta_0) = -\int \phi^{-1}(\Xi^{-1}(p(x, \theta)))p(x, \theta_0).$$

Then we observe that

$$\mathbb{L}_\phi(\theta_0, \theta_0) - \mathbb{L}_\phi(\theta, \theta_0) = D_\phi(\Xi^{-1}(p(\cdot, \theta_0)), \Xi^{-1}(p(\cdot, \theta))),$$

which is nonnegative with equality if and only if $\theta = \theta_0$ because of (2). Hence we conclude the asymptotic consistency of $\hat{\theta}_\phi$ for θ_0 noting $L_\phi(\theta)$ is almost surely converges to $\mathbb{L}(\theta, \theta_0)$. The proof is essentially equivalent to that for the asymptotic consistency of the MLE, cf. Wald (1949). In fact, if $\phi = \exp$, then Ξ equals the identity function and $L_{\exp}(\theta)$ is the negative log-likelihood function, so that $\hat{\theta}_\phi$ is the MLE of θ_0.

We discuss to explore this view for the model and estimation methods in a context of generalized regression analysis and clustering analysis, cf. Rose (1998), Notsu et al. (2016) and Omae et al. (2017). A typical example of ϕ is considered in a class of cumulative distribution functions including exponential and Parate distributions.

References

[1] Rose, Kenneth. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems." Proceedings of the IEEE 86.11 (1998): 2210-2239.

[2] Notsu, Akifumi, and Shinto Eguchi. "Robust clustering method in the presence of scattered observations." Neural computation 28.6 (2016): 1141-1162.

[3] Omae, Katsuhiro, Osamu Komori, and Shinto Eguchi. "Quasi-linear score for capturing heterogeneous structure in biomarkers." BMC bioinformatics 18.1 (2017): 308.

[4] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Annals of Mathematical Statistics 20, 595-601.

擬似線形関数を用いたクラスタワイズ回帰モデル

慶應義塾大学 理工学部　林 賢一

1　背景

ビッグデータの時代といわれる昨今，3つのVとして挙げられる「量（volume），速度（velocity），多様性（variety）」は統計科学に多くの課題を突き付けている（Laney, 2001）．この流れの中で，データの規模や複雑性は増し，異質な集団を複数個含むようなデータが得られることが多くなった．じっさい，異質な集団を含むデータは生命科学や認知科学，マーケティング・サイエンスなど，諸種の分野において観察されている（O'Driscoll et al., 2012; Suk et al., 2014 など）．このような場合，比較的小規模な統制されたデータにおいて仮定するような，単一の確率分布によってデータの背景に潜む構造を解き明かす試みには限界がある．

いま，データ $\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ が観測されているとする．$\boldsymbol{x}_i$ は d 個の説明変数を含むベクトル，y_i は被説明変数である．このとき，y_i を関数 $\mu(\boldsymbol{x}_i)$ により予測・説明する問題を考える．また，データには K 個の異質な部分集団が含まれ，各個体はそれらのうちのいずれかに属するものとする．

このような設定下での回帰問題について，Desarbo and Cron (1988) は以下のような確率モデルを導入した．

$$y_i \sim \sum_{k=1}^{K} p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i - \boldsymbol{\beta}_k^\top \widetilde{\boldsymbol{x}}_i)^2}{2\sigma_k^2}\right). \tag{1}$$

ここで，$\widetilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top$，$p_1, \ldots, p_K$ は $\sum_{k=1}^{K} p_k = 1$，$p_k \geq 0$ $(k = 1, \ldots, K)$ をみたす未知パラメータである．これと類似のモデルとして，混合エキスパートモデル（MoE; mixture of experts）が挙げられる（Jacobs et al., 1991）．MoE モデルは，条件付確率（密度）関数 $f(y_i|\boldsymbol{x}_i)$ を次のように表現する．

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i; \boldsymbol{\gamma}) f_k(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_k). \tag{2}$$

ここで，$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)^\top$，$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)^\top$，$\pi_k(\boldsymbol{x}_i; \boldsymbol{\gamma}) = \dfrac{\exp(\boldsymbol{\gamma}_k^\top \widetilde{\boldsymbol{x}}_i)}{\sum_{\ell=1}^{K} \exp(\boldsymbol{\gamma}_\ell^\top \widetilde{\boldsymbol{x}}_i)}$，$f_k(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}_k)$ はパラメータ $\boldsymbol{\beta}_k$ をもつ，第 k 部分集団における条件付確率（密度）関数である．

2　擬似線形関数を用いた回帰関数

本研究では，一般化線形モデル $\mathrm{E}[Y|\boldsymbol{x}] = g^{-1}(\mu(\boldsymbol{x}; \boldsymbol{\theta}))$ に対し（g はリンク関数，$\boldsymbol{x} \in \mathbb{R}^d$，$\boldsymbol{\theta}$ はパラメータベクトル），以下のような回帰関数を提案する．

$$\mu(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{\tau} \log\left(\sum_{k=1}^{K} p_k(\boldsymbol{x}; \boldsymbol{\mu}) \exp\left(\tau \mu_k(\boldsymbol{x}; \boldsymbol{\beta}_k)\right)\right). \tag{3}$$

ここで，τ は実数値をとるハイパーパラメータ，$p_k(\boldsymbol{x};\boldsymbol{\mu})$ はパラメータ $\boldsymbol{\mu}$ をもち，任意の $\boldsymbol{x} \in \mathbb{R}^d$ に対し $\sum_{k=1}^K p_k(\boldsymbol{x};\boldsymbol{\mu}) = 1$ をみたす非負の関数，$\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top, \boldsymbol{\mu}^\top)^\top$ である．回帰関数 (3) は，K 個の回帰関数を狭義単調増加な関数 $\phi(\cdot)$ を用いた結合 $\phi^{-1}\left(\sum_{k=1}^K p_k(\boldsymbol{x};\boldsymbol{\mu})\phi(\mu(\boldsymbol{x};\boldsymbol{\beta}_k))\right)$ と見ることができる．このような ϕ による結合を擬似線形結合とよび（Omae et al., 2017），本稿では擬似線形結合を用いた回帰モデル (3) をクラスタワイズ擬似線形（cluster-wise quasi-linear; CWQL）回帰モデルとよぶことにする．関数 (3) は，$\phi(z) = \exp(\tau z)$ の場合の擬似線形結合である．実数 τ は，各回帰関数 μ_k の重みを調整するハイパーパラメータであり，$\tau \to 0$ のときに単純な重み付き平均 $\mu(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^K p_k(\boldsymbol{x};\boldsymbol{\mu}) \exp\left(\tau\mu_k(\boldsymbol{x};\boldsymbol{\beta}_k)\right)$ となる．また，$\tau \to \infty$ のときは $\mu(\boldsymbol{x};\boldsymbol{\theta}) = \max\left(\mu_1(\boldsymbol{x};\boldsymbol{\beta}_1), \ldots, \mu_K(\boldsymbol{x};\boldsymbol{\beta}_K)\right)$ となり，有限値の τ は平均値と最大値（$\tau \to -\infty$ で最小値）の間を調整するパラメータと解釈できる．

本研究では，各クラスタ k における回帰関数を $\mu_k(\boldsymbol{x};\boldsymbol{\beta}_k) = \boldsymbol{\beta}_k^\top(1, \boldsymbol{x}^\top)^\top$ とし，関数 $p_k(\boldsymbol{x};\boldsymbol{\mu}_k)$ について以下のような形を考える．

$$p_k(\boldsymbol{x};\boldsymbol{\mu}) = \frac{\exp\left(-\omega||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right)}{\sum_{\ell=1}^K \exp\left(-\omega||\boldsymbol{x} - \boldsymbol{\mu}_\ell||^2\right)} \tag{4}$$

ここで，$\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top)^\top$，$\omega$ は正のハイパーパラメータである．この関数 (4) は，自由エネルギー $-\dfrac{1}{\omega}\displaystyle\sum_{i=1}^n \log\left(\sum_{k=1}^K \exp(-\omega||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2)\right)$ についての最大化問題の解として得られる（Rose et al., 1990）．自由エネルギーもまた $\phi(z) = \exp(\omega z)$ による擬似線形結合の形式で表されていることに注目すると，(3) における τ の極限と類似の結果が観察される．すなわち，$\omega \to 0$ のとき $p_k(\boldsymbol{x};\boldsymbol{\mu}) = \dfrac{1}{K}$ となり，すべての回帰関数が等しく重みづけられる．また，$\omega \to \infty$ のときは k-means 法と同様の結果が得られる．$0 < \omega < \infty$ のときは，p_k はクラスタ k への所属の度合いを示すような関数と解釈でき，クラスタ間で分散が等しい場合の混合正規分布モデルと捉えることができる．以上より，既存の回帰モデルを拡張したものであることがわかる．

推定上の問題や，数値実験の結果などは，当日報告する．

References

[1] DeSarbo, W.S., Cron, W.L. (1988). *Journal of Classification*, **5**, 249–282.
[2] Jacobs, R.A., et al. (1991). *Neural Computation*, **3**, 79–87.
[3] Laney, D. (2001). http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
[4] O'Driscoll, D.M., et al. (2012). *Sleep*, **35**, 1269–1275.
[5] Omae, K., et al. (2017). *BMC Bioinformatics*, **18**, 308.
[6] Rose, K., et al. (1990). *Physical Review Letters*, **65**, 945–948.
[7] Suk, H., et al. (2014). *PLoS ONE*, **9** (2): e87056.

Extension of Poisson point process based on quasi-linear modeling

Osamu Komori[1], Yusuke Saigusa[2], Shinto Eguchi[3]

[1]University of Fukui
[2]Yokohama City University
[3]The Institute of Statistical Mathematics

Abstract

The investigation to clarify the relationship between habitat distribution of some species and the environmental variables is important for its conservation and and management purpose. To do this, the maximum entropy method (Maxent) or Poisson point process (PPP) is widely employed using the presence-only data. In this paper, we propose an extension of PPP based on the quasi-linear modeling to improve the estimation accuracy. The effect of sampling bias is also considered in our model. Some simulation studies and real data analysis are conducted to show its practical utility.

1 Poisson point process based on quasi-linear modeling

For a study area $\mathcal{A}$ there are m presence locations $\{s_1, \ldots, s_m\}$. The log-likelihood of Poisson point process (PPP) based on a quasi-linear modeling is defined as

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{m} \log(\lambda_\tau(s_i)) - \int_{s\in\mathcal{A}} \lambda_\tau(s) ds, \tag{1}$$

where

$$\lambda_\tau(s) = \exp\left[\frac{1}{\tau}\log\left\{\pi \exp(\tau \boldsymbol{\beta}^\top \boldsymbol{x}(s)) + (1-\pi)\exp(\tau \boldsymbol{\delta}^\top \boldsymbol{z}(s))\right\}\right] \tag{2}$$

and π is a mixing proportion; $\boldsymbol{x}(s) = (1, x_1(s), \ldots, x_p(s))^\top$ and $\boldsymbol{z}(s) = (1, z_1(s), \ldots, z_q(s))^\top$ denote environmental and sampling-bias variables at a location s, respectively. When the shape parameter τ converges to 0, then we have

$$\lim_{\tau\to 0} \lambda_\tau(s) = \exp\left\{\pi \boldsymbol{\beta}^\top \boldsymbol{x}(s) + (1-\pi)\boldsymbol{\delta}^\top \boldsymbol{z}(s)\right\}, \tag{3}$$

which is nothing but the intensity of the original PPP. Clearly, we have the intensity of superposed PPP when $\tau = 1$(Streit, 2010) as

$$\lambda_1(s) = \pi \exp(\boldsymbol{\beta}^\top \boldsymbol{x}(s)) + (1-\pi)\exp(\boldsymbol{\delta}^\top \boldsymbol{z}(s)). \tag{4}$$

Moreover, we have

$$\lambda_{-1}(s) = \frac{1}{\pi \exp(-\boldsymbol{\beta}^\top \boldsymbol{x}(s)) + (1-\pi)\exp(-\boldsymbol{\delta}^\top \boldsymbol{z}(s))} \tag{5}$$

$$\lim_{\tau\to\infty} \lambda_\tau(s) = \exp\big[\max\{\boldsymbol{\beta}^\top \boldsymbol{x}(s), \boldsymbol{\delta}^\top \boldsymbol{z}(s)\}\big] \tag{6}$$

$$\lim_{\tau\to-\infty} \lambda_\tau(s) = \exp\big[\min\{\boldsymbol{\beta}^\top \boldsymbol{x}(s), \boldsymbol{\delta}^\top \boldsymbol{z}(s)\}\big]. \tag{7}$$

In general, the quasi-linear modeling in (2) is formulated by Kolmogorov-Nagumo average (Eguchi & Komori, 2015). See (Omae *et al.*, 2017) for the application to the classification problems.

The study area $\mathcal{A}$ is split into n grid cells, resulting in the approximated log-likelihood (Renner & Warton, 2013)

$$\widetilde{L}_\tau(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} w_i \Big[y_i \log(\lambda_\tau(s_i;)) - \lambda_\tau(s_i) \Big], \tag{8}$$

where $\{s_{m+1}, \ldots, s_n\}$ are the centers of the grid cells containing no presence location, $y_i = I(i \in \{1, \ldots, m\})/w_i$ with a quadrature weight w_i, and $I(\cdot)$ is the indicator function. The weight w_i is a grid cell area divided by the number of locations $\{s_1, \ldots, s_n\}$ contained in the cell. If there is no duplication of locations in each cell, $w_i = |\mathcal{A}|/n$, where $|\mathcal{A}|$ is the area of $\mathcal{A}$.

Here we have

$$\frac{\partial}{\partial \boldsymbol{\beta}} \widetilde{L}_\tau(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} w_i q(s_i)\{y_i - \lambda_\tau(s_i)\}\boldsymbol{x}(s_i) = 0 \tag{9}$$

$$\frac{\partial}{\partial \boldsymbol{\delta}} \widetilde{L}_\tau(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} w_i (1-q(s_i))\{y_i - \lambda_\tau(s_i)\}\boldsymbol{z}(s_i) = 0. \tag{10}$$

where

$$q(s) = \frac{\pi \exp(\tau \boldsymbol{\beta}^\top \boldsymbol{x}(s))}{\pi \exp(\tau \boldsymbol{\beta}^\top \boldsymbol{x}(s)) + (1-\pi)\exp(\tau \boldsymbol{\delta}^\top \boldsymbol{z}(s))}. \tag{11}$$

If $\tau \boldsymbol{\beta}^\top \boldsymbol{x}(s_i)$ has a large positive value, which is often the case where the impact of environmental variables is strong, then $q(s_i)$ becomes nearly 1. On the other hand, if $\tau \boldsymbol{\delta}^\top \boldsymbol{z}(s)$ has a large positive value, which is often the case where the impact of bias sampling is strong, then $q(s_i)$ becomes nearly zero. Hence we expect that $q(s_i)$ can adjust the balance of the impacts of the two variables. This is an extension of the asymmetric logistic regression model (Komori *et al.*, 2016), where the weight function has an important role in adjusting the imbalanceness of the sample sizes of the two populations.

From Small Data to Big Data: Integrating Machine Learning, Physical Model and Uncertainty Quantification for Efficient Polymer Design

○Stephen Wu*, Yukiko Kondo**, Isao Kuwajima[##], Guillaume Lambard*[,##], Kenta Hongo[#,##,%],
Junko Morikawa**, Yibin Xu[##], Ryo Yoshida*[,##]
* The Institute of Statistical Mathematics, Tokyo, 190-8562, Japan
** Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[#] Japan Advanced Institute of Science and Technology, Ishikawa, 923-1292, Japan
[##] National Institute for Materials Science, Tsukuba, 305-0047, Japan
[%] PRESTO, JST, Kawaguchi, Saitama, 332-0012, Japan
○Speaker at the conferenc

1. Introduction

In the past, data could not be efficiently collected, stored and accessed, and calculation required tremendous amount of human power. Especially for material science, data collection comes from experiments that usually take a long period of time, and consume a large amount of financial and human resources. To design materials for a specific purpose, scientists had to plot their research map step-by-step carefully to reduce the probability of significant failures. This traditional approach, which we refer as the small data approach, has led to many successes to understand the physics of material properties at the level of continuum models. However, the problem becomes much harder at the molecular level due to the exponential increase of the complexity. With the advancement of experimental and computational technologies, it is time for material scientists to move from the small data approach to a big data approach. Instead of manually hypothesizing and testing different prediction models for new material discovery, machine learning offers a promising tool for big data analysis to create effective empirical models. Such models can, then, be integrated into a data-driven material design framework.

Similar successes have been seen in various fields, such as image recognition [1,2], natural language processing [3,4], games [5], financial trading [6], and so on. The necessary cost for such an approach is the *large* amount of *informative* data. Here, *large* and *informative* are very subjective concepts depending on the targeted problem. For example, while 250 data points seem to be enough for training a machine learning model to identify liquid crystallinity of five-ring bent-core molecules [7], one would not expect that the same amount of data would be enough to build a model for predicting the thermal conductivity of any amorphous homopolymers. In order to build more general data-driven models for material design, we aim at constructing a bridge that bring us from the existing small data approaches to a big data approach by better exploitation of the existing machine learning techniques. In this talk, we will demonstrate a specific implementation of such an idea using the PolyInfo database [8] for searching high thermal conductivity polymers. Our implementation aims at providing an efficient end-to-end material design process that incorporates Bayesian inversion, machine learning models, and experimental design concepts.

2. Methods

Thermal conductivity of polymers has continuously attracted many attentions for a long period of time [9-12]. Yet, we are far from understanding the underlying mechanism enough for general design purpose. For that, we adopt a rather ambitious vision outlined in [13] that is to perform inverse material design with generation of new molecules. Although we would like to implement the idea directly using the PolyInfo database [10], one of the largest databases of polymers in the world, the data availability of thermal conductivity appears to be significantly less than other material properties: among the over a thousand of data points, only less than a hundred of unique homopolymers data is available with large variance. Given the expensive cost of obtaining new experimental data, we propose an experimental design schedule that integrates existing knowledge and machine learning techniques, in order to increase the probability of new material discovery while we are moving toward the final goal of fully data-driven inverse design. We achieve that by observing a correlation between the glass transition temperature and the thermal conductivity based on the empirical equation in [14]. As a result, we can exploit the machine learning model using the rich data of glass transition and melting temperature in PolyInfo using iqspr, and then, we rank the potential of each candidate by screening through the newly generated molecules using the van Krevelen group contribution method [15], a empirical model that is general enough for our purpose here. We impose uncertainty quantification for the molecules generated from iqspr as well as the screening results in order to have a more intuitive

ranking at the end.

3. Results

In the iqspr package [13], we set the target region to be 320C-800C for glass transition temperature and 430C-800C for melting temperature. Default values are used for the regression model along with a train data set that includes all the homopolymers that have the glass transition temperature values recorded. For the prior model used for generating "homopolymer-like" molecules, we used all of the 14,424 homopolymers available in the PolyInfo database for training. Figure 1 shows four snapshots of the generated molecules. After that, all generated molecules are screened through the van Krevelen group contribution method to pick out candidates with a high probability of achieving high thermal conductivity, as described in Section 2.

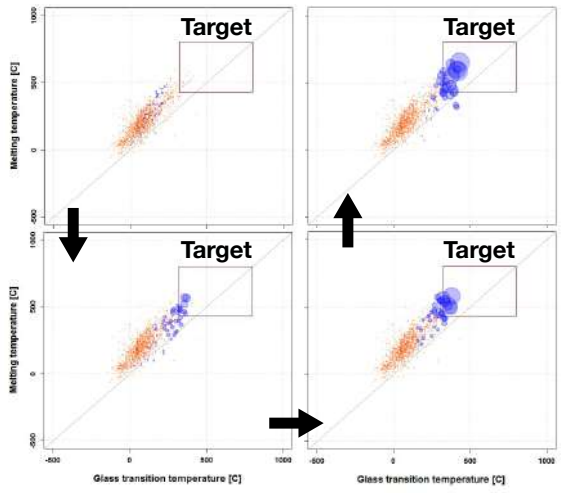

Figure 1 Evolution of homopolymer design of high glass transition and melting temperature using iqspr. Orange dots denote the training data in PolyInfo, and blue circles denote generated homopolymers with radius proportional to the prediction uncertainty.

4. Conclusion

In many scientific applications, we may have already accumulated a relatively large data set, but not yet large enough for direct machine learning application due to the complexity of the problem of interest. We demonstrated a roadmap leading us from the traditional small data approach to the new generation big data approach. We focused on the optimal use of resource during the process, that is to increase the probability of new material discovery by integrating machine learning, physical model and uncertainty quantification techniques. We observed that the resulting molecules from our iqspr simulation exhibit similar patterns, demonstrating the capability of a machine learning to automatically cluster chemical structures relevant to a target material property. However, the resulting candidates after screening the group contribution model were not intuitive to the polymer experts that we have consulted. A key problem with the current approach is the lack of consideration for synthesis ability, which is particular important when we are in the process of creating more data. To tackle this issue, we plan to implement a machine learning classifier to filter out candidates without liquid crystallinity, which is an important property for the ease of synthesis in practice. We are preparing to experimentally test the proposed candidates in the near future. The new results will be used for verification and feedback into our database to help us improving our prediction models.

REFERENCES

[1] M.S.Bartlett, G.Littlewort, C.Lainscsek, I.Fasel, J.Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions", Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, 1(2004), 592-597.

[2] I.Goodfellow, J.Pouget-Abadie, M.Mirza, B.Xu, D.Warde-Farley, S.Ozair, A.Courville, Y.Bengio, "Generative adversarial nets", In NIPS (2014), 2672-2680.

[3] Z.Solan, D.Horn, E.Ruppin, S.Edelman, "Unsupervised learning of natural languages", Proc. of The National Academy of Sciences of the USA, 102(2005), 33:11629-11634.

[4] P.M.Nadkarni, L.Ohno-Machado, W.W.Chapman, "Natural language processing: an introduction", Journal of the American Medical Informatics Association, 18(2011), 5:544-551.

[5] D.Silver et. al, "Mastering the game of Go with deep neural networks and tree search", Nature, 529(2016), 484-489.

[6] D.Enke, S.Thawornwong, "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, 29(2005), 4:927-940.

[7] J.Antanasijević, D.Antanasijević, V.Pocajt, N.Trišovića, K.Fodor-Csorbac, "A QSPR study on the liquid crystallinity of five-ring bent-core molecules using decision trees, MARS and artificial neural networks", RSC Adv., 6(2016), 18452-18464.

[8] PolyInfo, http://polymer.nims.go.jp/index_en.html, accessed on August 16, 2017.

[9] D.R.Anderson, "Thermal Conductivity of Polymers", Chem. Rev., 66(1966), 6:677-690.

[10] C.L.Choy, "Thermal conductivity of polymers", Polymer, 18(1977), 10:984-1004.

[11] E.Algaer, "Thermal Conductivity of Polymer Materials - Reverse Nonequilibrium Molecular Dynamics Simulations", Ph.D. Thesis, Technische Universität, Darmstadt (2010).

[12] G.H.Kim, D.Lee, A.Shanker, L.Shao, M.S.Kwon, D.Gidley, J.Kim, K.P.Pipe, "High thermal conductivity in amorphous polymer blends by engineered interchain interactions", Nature Materials, 14(2015), 295–300.

[13] H.Ikebata, K.Hongo, T.Isomura, R.Maezono, R.Yoshida, "Bayesian molecular design with a chemical language model", Journal of computer-aided molecular design, 31(2017), 4:379-391.

[14] J.Bicerano, "Prediction of Polymer Properties", Chap.14, CRC Press (2002), 503-512.

[15] D.W.van Krevelen, "Properties of Polymers", 3rd Ed., Elsevier Science, Amsterdam (1990).

一般化線形回帰のロバスト化およびスパース化

総合研究大学院大学　川島　孝行
統計数理研究所　藤澤　洋徳

1. はじめに　KL ダイバージェンスに基づく回帰は，次のように定義できる，$D_{KL}(g(y|x), f(y|x;\theta); g(x)) = \int D_{KL}(g(y|x), f(y|x;\theta)g(x)dx$. ただし，$g(y|x)$ および $g(x)$ は，データを生成する分布で，$f(y|x;\theta)$ はパラメトリックモデルである．$g(y|x)$ および $g(x)$ を，それぞれ経験密度関数 $\bar{g}(y|x)$ と $\bar{g}(x)$ で置き換えると，最尤推定に基づく回帰に一致する．また，適切なパラメトリックモデルを選択することで，線形，ロジスティックおよびポアソン回帰といった主要な回帰モデルの多くを一般化線形回帰の枠組みで捉えることができる (Nelder and Wedderburn 1972)．しかし，KL ダイバージェンスは，外れ値に弱い．そこで，我々は，ロバストなダイバージェンスとして知られている，ガンマ・ダイバージェンス (Fujisawa and Eguchi 2008) に基づく回帰を考える．また，これにスパース正則化を組み合わせることで，ロバストかつスパースな一般化線形回帰を達成する．

2. ガンマ・ダイバージェンスに基づくロバストかつスパースな回帰　ガンマ・ダイバージェンスに基づく回帰は，次のように定義できる，

$$\arg\min_{\theta} d_\gamma(g(y|x), f(y|x;\theta); g(x)) = -\frac{1}{\gamma} E_{g(x,y)}[f(y|x;\theta)^\gamma / \left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}].$$

これに，スパース正則化を加えたものは以下で表される．

$$\arg\min_{\theta} d_\gamma(g(y|x), f(y|x;\theta); g(x)) = -\frac{1}{\gamma} E_{g(x,y)}[f(y|x;\theta)^\gamma / \left(\int f(y|x;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}] + \lambda P(\theta). \quad (1)$$

これの経験推定を考える．

$$\arg\min_{\theta} -\log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\theta)^\gamma \Big/ \left(\int f(y|x_i;\theta)^{1+\gamma} dy\right)^{\frac{\gamma}{1+\gamma}}\right\} + \lambda P(\theta). \quad (2)$$

$P(\theta)$ としては，L1 正則化 (Tibshirani 1996)，elastic net (Zou and Hastie 2005)，Group lasso (Yuan and Lin 2006) といった様々な正則化を考えることが可能である．以下に，具体的に，線形回帰・ロジスティック回帰・ポアソン回帰それぞれについての詳細を述べる．

2.1. 線形回帰　(2) の最適化は，通常の方法では，推定アルゴリズムの導出自体が困難となる．そこで，MM アルゴリズム (Hunter and Lange 2004) の考え方を用い，以下で表される最適化が容易な補助関数が得られた．

$$h_{MM}(\theta|\theta^{(m)}) = \frac{1}{2(1+\gamma)}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}\alpha_i^{(m)}\frac{(y_i - \beta_0 - x_i^T\beta)^2}{\sigma^2} + \lambda||\beta||_1,$$

ただし，$\alpha_i^{(m)} = \frac{f(y_i|x_i;\theta^{(m)})^\gamma}{\sum_{l=1}^{n} f(y_l|x_l;\theta^{(m)})^\gamma}$ である．これにより，スパース正則化を加えた場合でも，目的関数値を単調に減少させる推定アルゴリズム の導出を行った．数値実験および実データ（マイクロアレイデータ）の解析においても，比較手法 (Alfons et al. 2013 and Khan et al. 2007) にくらべて非常に大きな改善を見せた (Kawashima and Fujisawa 2017).

2.2. ロジスティック回帰　Kanamori and Fujisawa (2015) では，回帰問題において，外れ値の割合が説明変数 x に依存する場合は，位置・尺度分布に限定 したときのみ，ロバスト性を得ることが可能であった，しかし，(2) に基づく回帰では，そのような仮定は必要なく，ロジスティック回帰の場合でも，ロバスト性を得ることが可能である．また，推定アルゴリズムの導出は，線形回帰の場合と同じように，MM ア

ルゴリズムの一例である，Bohning and Lindsay (1988) の方法を用いて，以下で表される補助関数が得られた．

$$h_{MM}(\theta|\theta^{(m)}) = \frac{\gamma L_{th}}{2n}\sum_{i=1}^{n}\left(\beta_0^{(m)} + x_i^T\beta^{(m)} + \frac{1}{L_{th}}\tilde{\tau}_i^{(m)} - \beta_0 - x_i^T\beta\right)^2 + \lambda||\beta||_1,$$

ただし，$\tilde{\tau}_i^{(m)} = \left(\frac{\exp\left\{y_i(1+\gamma)\tilde{x}_i^T\theta^{(m)}\right\}}{\exp\left\{(1+\gamma)\tilde{x}_i^T\theta^{(m)}\right\}+1}\right)^{\frac{\gamma}{1+\gamma}}\left(y_i - \frac{\exp\left\{(1+\gamma)\tilde{x}_i^T\theta^{(m)}\right\}}{\exp\left\{(1+\gamma)\tilde{x}_i^T\theta^{(m)}\right\}+1}\right)$ および $L_{th}>max\left\{\frac{(1+\gamma)^2}{4+8\gamma}, \frac{(1+3\gamma)^2}{4+8\gamma}-\gamma\right\}$ である．これにより，ロジスティック回帰の場合でも，線形回帰と同じように目的関数値を単調に減少させる推定アルゴリズム の導出を行った．また，アルゴリズムの収束性に関しても，Miral (2013) の結果を用いることで，大域的収束性 が得られた．

2.3. ポアソン回帰　ポアソン回帰の場合でも，Kanamori and Fujisawa (2015) の仮定は必要なく，外れ値の割合が説明変数に依存する場合でも，ロバスト性を得ることが可能 である．しかしながら，ポアソン回帰の場合は，$\left(\int f(y|x_i;\theta)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}$ の項で，超幾何級数の計算が必要となり，(2) に基づく限り，通常の方法では効率的な推定アルゴリズムの導出が困難 となる．そこで，経験推定を経由することなく，直接，(1) を最小化する，確率的最適化の枠組みを用いて，この問題を克服 した．特に，確率的最適化の手法の中でも，非凸かつ滑らかでない目的関数を対象とした，Randomized Stochastic Projected Gradient (Ghadimi et al. 2016) を用いることで，大域的収束性を持つ，オンライン推定アルゴリズム を導出した．

$$\theta^{(t+1)} = \arg\min_{\theta}\left\langle -\frac{1}{m_t}\sum_{i=1}^{m_t}\nabla_\theta \left.\frac{f(y_{t,i}|x_{t,i};\theta)^\gamma}{\left(\int f(y|x_{t,i};\theta)^{1+\gamma}dy\right)^{\frac{\gamma}{1+\gamma}}}\right|_{\theta=\theta^{(t)}}, \theta\right\rangle + \lambda P(\theta) + \frac{1}{2\eta_t}\|\theta-\theta^{(t)}\|_2^2,$$

ただし，m_t は各反復 t における，ミニバッチのサイズを表す．また，ポアソン回帰に限らず，上で述べた回帰にもオンライン推定アルゴリズムを適用することは可能である．これにより，ロバストかつスパースな一般化線形回帰を達成する．

多重代入法を用いたバイアス補正推定量に関する考察

総合研究大学院大学　富田　裕章
統計数理研究所　藤澤　洋徳　　統計数理研究所　逸見　昌之

欠測値を含むデータに対する解析法の 1 つとして Rubin (1987) は 多重代入法（Multiple Imputation : MI）を提唱した．MI では，欠測値を含んだデータに対して，代入法を独立に何度も行い，その度に推定量を計算し，それらを統合して最終的な推定量（統合推定量）を得る．MI には，観測されたデータをすべて利用した推定が可能であること，補完を行ったデータに対する解析は通常のデータに対する解析手法を利用可能であること，繰り返し補完を行うことで推定値の分散も評価可能であること，という利点がある反面，適切に多重代入法を行わないと統合推定量が妥当なものでなくなることが知られている．

古典的には多重代入法が適切（proper）な場合には，統合推定量は一致性を持つことが示されている (Rubin (1987)). 適切なモデルの下で，ベイズ理論に基づく事後分布を用いて代入値を生成した場合には多重代入法が適切であることが知られている．しかしながら，一般の多変量分布では正確な事後分布を生成することは難しく，また適切であるかを確認することも困難である (van Buuren (2012)). また，多重代入法が適切でなくても代入モデルが正しく同定されていれば統合推定量は一致性を持つことが知られている (Little (1992)) が，同定が誤っていた場合には統合推定量が真の値に確率収束することは保証されず，バイアスが発生することがある (Clayton et al. (1998), Robins and Wang (2000))．

本発表では，回帰モデルの推定において，説明変数が欠測している状況下で，多重代入法を利用した時に欠測値を補完するモデルに誤りがあったとしても，補完モデルの誤りによるバイアスを補正して一致性のある推定量を導出する方法（Tomita et al. (2017)）について述べる．具体的に書き下すと以下の通りである．

Y を応答変数とし，X, Z を説明変数とする．ここで，X は欠測が発生しうる変数とし，Z には欠測が発生しないものとした．R を観測指標とし，$R = 1$ のとき X は観測され，$R = 0$ のとき X は観測されないとする．推定量 $\hat{\theta}$ を，

$$\hat{\theta} = \arg\max_{\theta} \left[\sum_{i:R_i=1} \log f(Y_i|X_i, Z_i; \theta) + \sum_{i:R_i=0} \frac{1}{m} \sum_{j=1}^{m} w(X_i^j, Y_i, Z_i) \log f(Y_i|X_i^j, Z_i; \theta) \right] \tag{1}$$

と定義する．ここで，$w(x, y, z)$ は重み付け関数

$$w(x, y, z) = \frac{g(x|y, z)}{h(x|y, z)}$$

である．ただし，$g(x|y, z)$ は $Y = y, Z = z$ のもとでの X の真の条件付き分布をあらわす．一定の条件のもとで，$m, n \longrightarrow \infty$ とすると，推定量 $\hat{\theta}$ は母数との一致性をもつ．

w の計算で用いる X の真の条件付き分布 $g(x|y, z)$ は一般に未知であるため，実際の解析では条件付き密度推定によって推定した分布 $\hat{g}(x|y, z)$ に置き換える．条件付き密度推定法として，カーネル密度推定に基づく手法である最小二乗条件付き密度推定法（Least Square Conditional Density Estimation，LSCDE）(Sugiyama et al. (2010)) を利用する．

解析手順を以下に示す．

Step 1.

欠測した X に対し，何らかの補完モデル $h(x|y,z)$ を仮定し，補完値を独立に m 回発生させる．

Step 2.

LSCDE によって $\hat{g}(x|y,z)$ を推定し，重み関数

$$w(x,y,z) = \frac{\hat{g}(x|y,z)}{h(x|y,z)}$$

を補完した標本 (X_i^j, Y_i, Z_i) すべてについて計算する．

Step 3.

式 (1) に基づいて推定量 $\hat{\theta}$ を求める．

この手法をバイアス補正多重代入法（Bias corrected multiple imputation）と称する．

なお，バイアス補正多重代入法について数値実験や実データ解析によって評価を行っているが，本報告書では省略する．

今後の課題としては，より複雑な一般の欠測問題に対して提案手法の拡張を考えること，信頼区間などを構築する上で必要な推定量の分散を与えることが考えられる．

謝辞

本研究の一部は国立研究開発法人日本医療研究開発機構 (AMED) の臨床研究・治験推進研究事業 (171k0201061h0002) の支援によって行われた．

参考文献

van Buuren, S. (2012) *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*: Chapman and Hall/CRC.

Clayton, D., Spiegelhalter, D., Dunn, G., and Pickles, A. (1998) "Analysis of longitudinal binary data from multiphase sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 60, No. 1, pp. 71–87.

Little, R. J. A. (1992) "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, Vol. 87, No. 420, pp. 1227-1237.

Robins, J. M. and Wang, N. (2000) "Inference for Imputation Estimators," *Biometrika*, Vol. 87, No. 1, pp. 113–124.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*: Wiley, pp.258.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. (2010) "Least-squares conditional density estimation," *IEICE Transactions on Information and Systems*, Vol. 93, No. 3, pp. 583–594.

Tomita, H., Fujisawa, H., and Masayuki, H. (2017) "A bias-adjusted estimator in multiple imputation for missing data," in submission.

最適な半教師付き学習

川喜田雅則（九州大学）・藤澤洋徳（統計数理研究所）

1　はじめに

半教師付き学習において，教師なしデータの有効利用が幾つか提案されている．Kawakita (unpublished) は，密度比を上手く構成することで，教師付き学習を改良する手法を提案した．それを DRESS (Density-Ratio Estimation based on Semi Supervised learning) と名付けた．Kawakita and Takeuchi (2014) は，密度比をより上手く推定する手法を提案した．これらの良さは，平均二乗誤差に基づいたピタゴリアン関係によって，説明することができる．Kawakita and Kanamori (2013) は密度比をノンパラメトリックに推定することを提案した．

本研究では，最適性の観点から，教師なしデータの有効利用を考える．得られた結果から自然に，教師なしデータの新しい有効利用法が見つかる．

2　半教師付き学習

教師付きデータは以下である：

$$\left(\begin{array}{c} x_1 \\ y_1 \end{array} \right), \ldots, \left(\begin{array}{c} x_1 \\ y_1 \end{array} \right) \sim g_{x,y}(x,y).$$

教師なしデータは以下である：

$$x_{n+1}, \ldots, x_N \sim g_x(x).$$

半教師付きデータは上記を合わせたものである．

教師付き学習の例は以下である：

$$\hat{\beta}_0 = \arg\mathrm{solve} \left\{ \beta : \sum_{i=1}^{n} u_\beta(x_i, y_i; \beta) = 0 \right\}.$$

ここで u_β の代表例は線形回帰を意図した $u_\beta(x, y; \beta) = y - \beta^T x$ である．半教師付き学習は半教師付きデータを利用した学習である．本研究では，教師なしデータを上手く利用して，推定量 $\hat{\beta}_0$ を改良する手法を提案する．特に最適性の観点から考察を行う．

3　半教師付き学習の提案例

Kawakita (unpublished) は密度比を利用して次の DRESS を提案した．

$$\hat{\beta}_K = \arg\mathrm{solve}_{\beta} \left\{ \beta : \sum_{i=1}^{n} \frac{g_x(x_i; \hat{\eta}')}{g_x(x_i; \hat{\eta})} u_\beta(x_i, y_i; \beta) = 0 \right\},$$

$$\hat{\eta} = \arg\mathrm{solve} \left\{ \eta : \sum_{i=1}^{n} s_\eta(x_i; \eta) = 0 \right\}, \qquad \hat{\eta}' = \arg\mathrm{solve} \left\{ \eta' : \sum_{i=n+1}^{N} s_\eta(x_i; \eta') = 0 \right\}.$$

ここで，$g_x(x; \eta)$ は $g_x(x)$ を推し量るためのパラメトリックモデルであり，$s_\eta(x; \eta) = (\partial/\partial\eta) \log g_x(x; \eta)$ である．仮に真の分布 $g_x(x)$ が分かっていたとしよう．密度比の分子が $g_x(x_i; \hat{\eta}')$ ではなく $g_x(x)$ であったとしよう．有限標本から得られる密度推定 $g_x(x; \hat{\eta})$ は，$g_x(x)$ とずれている．そのずれを密度比で補正することで全体の推定のずれを補正しようという考えである．この考えは逆確率重み法に通じる考えである．実際に，適当な条件の下では，漸近分散の高次の部分で，$\hat{\beta}_K$ は $\hat{\beta}_0$ を改良していることが，証明できる．

この拡張に関しては，Kawakita and Takeuchi (2014) や Kawakita and Kanamori (2013) で提案されているが，省略する．

4　最適な半教師付き学習

半教師付きデータから作られる推定量 $\hat{\beta}$ は「漸近線形展開」をもち「正則」であるとする．漸近線形展開をもつとは，次が成り立つことである：

$$\sqrt{n}(\hat{\beta}-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_1(x_i,y_i)+\frac{1}{\sqrt{N-n}}\sum_{i=n+1}^{N}\phi_2(x_i)+o_p(1).$$

ただし $\beta_0=\arg\mathrm{solve}\left\{\beta:\mathrm{E}_{g_{x,y}}[u_\beta(x,y;\beta)]=0\right\}$ とする．また，核関数 $\phi_1(x,y)$ と $\phi_2(x)$ は，適当な性質を満たすとする．正則であるとは，セミパラメトリック理論の意味で正則であるという意味である (Tsiatis, 2006)．超有効推定量を排除するための性質であるが，ここでは詳細は省略する．このような仮定の下で次の定理が成り立つ．なお，証明などの詳細は，Kawakita and Fujisawa (2017) を参照されたい．

定理 1　$\hat{\beta}$ は漸近線形展開をもち正則であるとする．そのとき核関数は次の性質を満たす：

$$\phi_1(x,y)=J_{\beta\beta}^{-1}u_\beta(x,y;\beta_0)-\frac{1}{\sqrt{r}}\phi_2(x).$$

ただし，$r=(N-n)/n$，$J_{\beta\beta}=\mathrm{E}_{g_{x,y}}\left[-(\partial u_\beta/\partial\beta)(x,y;\beta_0)\right]$ とする．

定理 2　$\hat{\beta}$ は漸近線形展開をもち正則であるとする．$\hat{\beta}$ の漸近分散を最小にするのは次のときである：

$$\phi_2(x)=\frac{\sqrt{r}}{1+r}J_{\beta\beta}^{-1}\mathrm{E}_{g_{y|x}}[u_\beta(X,Y;\beta_0)|X=x].$$

この定理から，最適な推定量を作るためには，条件付密度関数 $g(y|x)$ に関する期待値を得る必要がある．しかし，これは，半教師付きデータからは推定しにくい．そのため，現実的には，推定しやすいクラスの中で，上記に最も近い核関数を得ることを考えたい．なお，以下の定理では，関数 $\psi(x)$ は適当な基底関数をイメージしている．(たとえば B スプラインやガウシアンカーネルが代表的である．)

定理 3　$\phi_2(x)=B\psi(x)$ と制限されているとする．このクラスの中で漸近分散を最小にする $\hat{\beta}$ では以下が成り立つ：

$$\phi_2(x)=\frac{\sqrt{r}}{1+r}J_{\beta\beta}^{-1}\mathrm{E}_{g_{x,y}}\left[u_\beta(x,y,;\beta_0)\psi(x)^T\right]\mathrm{E}_{g_x}\left[\psi(x)\psi(x)^T\right]^{-1}\psi(x).$$

上記の定理の中にある $\phi_2(x)$ は半教師付きデータから推定可能である．なぜならば，期待値はすべて周辺分布や同時分布の下のため，通常の経験推定が可能だからである．そのような推定量 $\hat{\beta}$ を与える推定方程式は，自然に導出できる．その一つが次に挙げる nDRESS II である：$w(x;\theta)=\exp\{\theta^T\psi(x)\}$,

$$\hat{\beta}_{KF}=\arg\mathrm{solve}\left\{\beta:\sum_{i=1}^{n}w(x_i;\hat{\theta})u_\beta(x_i,y_i;\beta)=0\right\},$$

$$\hat{\theta}=\arg\mathrm{solve}\left\{\theta:\frac{1}{n}\sum_{i=1}^{n}\psi(x_i)w(x_i;\theta)-\frac{1}{N}\sum_{i=1}^{N}\psi(x_i)=0\right\}.$$

なお，定理 3 の性質をもつ推定量 $\hat{\beta}$ を与える推定方程式は，その他にも色々と作ることができる．それについては Kawakita and Fujisawa (2017) を参照されたい．

パラメータ転移学習における汎化誤差の評価

熊谷亘 [*]，金森敬文 [†,*]

理化学研究所 革新知能統合研究センター [*]
名古屋大学 大学院情報学研究科 [†]

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．本発表では $\mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}\right)$ を予測性能の基準値として採用する．このときいくつかの技術的仮定とともに，本研究において新たに導入された局所安定性と転移学習可能性という条件を用いることで，以下の学習バウンドを導出することができる (局所安定性および転移学習可能性の定義に関しては論文 [1] を参照せよ).

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．正則化パラメータ ρ を適切に設定するとき, 以下の不等式が確率 $1-(\delta+2\bar{\delta})$ で漸近的に成り立つ :

$$\mathcal{R}_{\mathcal{T}}\left(\widehat{\boldsymbol{\theta}}_N, \widehat{\mathbf{w}}_{N,n}\right) - \mathcal{R}_{\mathcal{T}}\left(\boldsymbol{\theta}^*_{\mathcal{S}}, \mathbf{w}^*_{\mathcal{T}}\right) \quad \leq \quad c \max\left\{ n^{-1/3}, \left\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*_{\mathcal{S}}\right\|^{2/7} \right\}$$

ここで c は N および n によらない定数である．

本研究に関連する他の理論的結果および数値実験の結果は当日報告する．

参考文献

[1] W. Kumagai, Learning Bound for Parameter Transfer Learning, *NIPS*, (2016).