科研費シンポジウム

「機械学習・統計学・最適化の数理と AI 技術への展開」

科学研究費・基盤研究（A）「大規模複雑データの理論と方法論の革新的展開」

研究代表者：青嶋 誠 (筑波大学)．課題番号：20H00576

日時： 2020 年 12 月 18 日 (金)〜19 日 (土)

場所： Zoom によるオンライン開催

開催責任者： 金森 敬文，川島 孝行 （東京工業大学 情報理工学院 数理・計算科学系）

プログラム：

12/18 (金)

9:50 - 10:00 オープニング

基盤 A 代表 青嶋 誠 (筑波大)
東工大情報理工学院長・DSAI 研究推進体代表 横田 治夫 (東工大)

10:00 - 11:10 セッション 1

簡単な数理モデルによる新型コロナウイルス感染症流行の予測と社会的制御
土谷 隆 (政策研究大)
仮想的な 0 近傍への外挿とその収束レートについて
奥野 彰文 (統数研/理研 AIP)，下平 英寿 (京大/理研 AIP)

11:20 - 12:30 セッション 2

Active learning for distributionally robust level set estimation
稲津 佑 (名工大)
分散を伝播させる BNN を用いた不確実性計算の高速化とその応用
前 佑樹 (株式会社デンソー)，熊谷 亘 (東京大/理研 AIP)，金森 敬文 (東工大/理研 AIP)

13:45 - 14:55 セッション 3

パレートフロンティアの情報量に基づく多目的ベイズ最適化
烏山 昌幸 (名工大)
確定的な動きを利用したマルコフ連鎖モンテカルロ法の紹介
鎌谷 研吾 (統数研)

15:05 - 16:55 セッション 4: 基盤 A シンポジウム＆東工大 DSAI 研究推進体セミナー合同開催

学習を定める不変量「実対数閾値」の定義と性質
渡辺 澄夫 (東工大)
ニューラルネットワークにおけるテンソルネットワーク表現
林 浩平 (PFN)
深層学習の統計神経力学と情報幾何
甘利 俊一 (理研)

12/19 (土)

10:00 - 11:10 セッション 5

CNN による同変的写像の普遍近似定理

熊谷 亘 (東京大/理研 AIP)，三内 顕義 (理研 AIP)

概念とは群である —学習における前方誤り訂正の必要性—

得丸 久文

11:20 - 12:30 セッション 6

多項ロジットモデルに基づく公的統計データ及び企業データの統計的マッチング・データ融合

高部 勲 (総務省/統数研)，山下智志 (統数研)

二変数間相互作用を考慮した一般化加法モデルとその効率的な学習

上月 正貴 (東京大)，松島慎 (東京大)

13:45 - 14:55 セッション 7

深層展開に基づく反復アルゴリズムの収束加速 —理論的解釈とその応用—

和田山 正 (名工大)

身体動作と音楽の同調性によるストリートダンスの評価手法の提案

上野 寛幸 (早稲田大)，鈴木 量三朗 (有限会社シンビー)，豊泉 洋 (早稲田大)

15:05 - 16:55 セッション 8

高次元データにおける異常値の検出について

中山 優吾 (京都大)，矢田 和善 (筑波大)，青嶋 誠 (筑波大)

Gamma-divergence に基づく変数選択について

伊森 晋平 (広島大/理研 AIP)，橋本 真太郎 (広島大)

事前分布強調型の情報量規準の開発とその WAIC との比較

二宮 嘉行 (統数研)

16:55 - 17:00 クロージング

基盤A代表 青嶋 誠 (筑波大)

開催責任者 金森 敬文 (東工大)

簡単な数理モデルによる新型コロナウイルス感染症流行の予測と社会的制御

政策研究大学院大学　土谷　隆

1. はじめに

新型コロナウイルス感染症の流行が始まってから早くも一年が経過しようとしている．講演者は，日本での流行の比較的初期から，簡単な数理モデルで本感染症の解析を行い [1]，ささやかながら情報を発信してきた [2]．ここでは，モデルと解析の概要について説明し，解決すべき統計的問題，この感染症の被害を社会的に最低限に抑えるための望ましい社会的戦略や今後の課題について問題提起した．

2. モデル

考えている国や自治体の中での t 日における未感染者の比率を $S(t)$, 患者の比率を $I(t)$, 治った人 (+亡くなった人) の比率を $R(t)$ とする．常に $S(t)+I(t)+R(t)=1$ である．考えている集団の大きさを N 人とし，以下の仮定を置く．(仮定 1) 未感染者が十分に多く，周囲が未感染者ばかりである時に，時点 t において 1 感染者が 1 日に他人に感染させる人数を $\beta(t)$ と記し，感染力と呼ぶ．集団の大きさが N である時に, t 日目の感染者数は $I(t)N$ であるが，　この時，t 日目に新たに発生する感染者の人数は $\beta(t)I(t)S(t)N$ で与えられるものとする.（仮定 2）各感染者は，感染した日から D 日経つと治って回復者となり，感染力を失う．

この 2 つの仮定の下で，感染の推移は以下の 3 つの漸化式で書ける．

$$\begin{aligned} S(t+1) &= S(t)-\beta(t)I(t)S(t) \\ I(t+1) &= I(t)+\beta(t)I(t)S(t)-\beta(t-D)I(t-D)S(t-D), \\ R(t+1) &= R(t)+\beta(t-D)I(t-D)S(t-D). \end{aligned}$$

さらに，次の仮定を置く：(仮定 3) 新型コロナウイルスに感染して最終的に抗体を有するようになる者の中で，自治体に感染者として把握される者 (行政的感染者という) の比率は一定である．この比率を $1/C$ と置く．（仮定 4）行政的感染者が感染してから行政から発表されるまでに W 日かかるものとする．

時点 t での新規行政的感染者数を $P(t)$ と記す．ある時点 t での新規感染者は, 未感染者の減少数 $N(S(t-1)-S(t))$ で与えられる．そして，これらの感染者の内 $1/C$ が行政に把握される．把握されるのは，感染してから W 日後であるとしている．したがって，

$$P(t+W)=\frac{1}{C}(N(S(t-1)-S(t)))$$

が成立する．

「実効再生産数」R_t は, 一人の感染者が感染中に他の感染者に移す人数の期待値である．本モデルの場合には，“一度感染した者は D 日感染者でいてその間は他人に感染させる” という単純化した想定を取っているので，$R_t=\beta(t)S(t)D$ である．

3. 推定結果と感染の特徴

感染の状況を決める上で重要なパラメータは，C と D である．D は，データや医学的知見から推定されるべきものであり [1] では推定結果として $D=15$ を得ている．C の推定は難しい．本モデル

では東大の抗体検査の結果に基づいて C は 23 としているが，C は控えめに見積もっても 10 以上はあると考えられる．また，W は 9 月までは 12 日, その後は 9 日 としている．推定結果を図 1 に示す．

4. ハンマーアンドダンスかグレートリセットか

本感染症への社会的対応としてよく言われるのが，ハンマーアンドダンスである．これは，感染が拡大してきたら社会・経済活動を縮小して感染抑止策をとって，感染が収まってきたら，社会・経済的活動を再開する，ということである．日本では，2020 年 4 月の緊急事態宣言は別として，感染抑止の積極策はとらずに各自の自粛に任せる，という形での対応がなされ，その一方で，Go To キャンペーンなどの経済活動を優先する形での施策が行われてきた．これが，日本流のハンマーアンドダンスである．

ハンマーアンドダンスの欠点は，新規陽性者数のコントロールを行うという行政的誘導がないことである．常に，減少しだしたところで，社会・経済活動を再開するため，ピークを迎えた後の定常レベルのコントロールが難しく，波を超えるごと（流行がピークを迎えるごと）に，定常的に発生する陽性者数が増大する傾向が見られる．その分医療システムにかかる負担は増大し，いくつかの波のあとに，医療システムが持ちこたえられなくなり，崩壊する可能性が高い．

一方，如何に激しく流行していても，3 週間程度の厳しい行動規制を行うと，ウイルスをほぼゼロとすることができるのであるから，流行が拡大してきたら 3 週間の厳しい行動規制で感染者をほぼゼロにした上で，引き続いて再び流行するまで，3ヶ月程度「3 蜜 OK の」自由な社会・経済活動を行っていくことを繰り返して，活力を維持しつつ社会を運営していくことができるのではないか，と筆者は考えている．いわばグレートリセットを厭わずに繰り返すものである．これは，論文 [1] で提案している，活動期と自粛期を繰り返す政策を徹底的に追求して深化させる形のものとなる．

直観的には，感染症がいかに流行していても，定数時間で 0 にできるのであるから，グレートリセットを繰り返す政策が（幅広い社会的効用関数の範囲で）ある種の最適性を持っていてもおかしくはないと考えられる．このような施策の可能性と合理性を数理的に検討することは興味深い問題である．

参考文献

[1] 土谷隆：新型コロナウイルス感染症の広がりに関する一考察. 政策研究大学院大学ディスカッション・ペーパー 20-04, 2020 年 5 月 30 日. (改訂版がオペレーションズ・リサーチに掲載予定.)
[2] 土谷隆： http://www3.grips.ac.jp/˜tsuchiya .

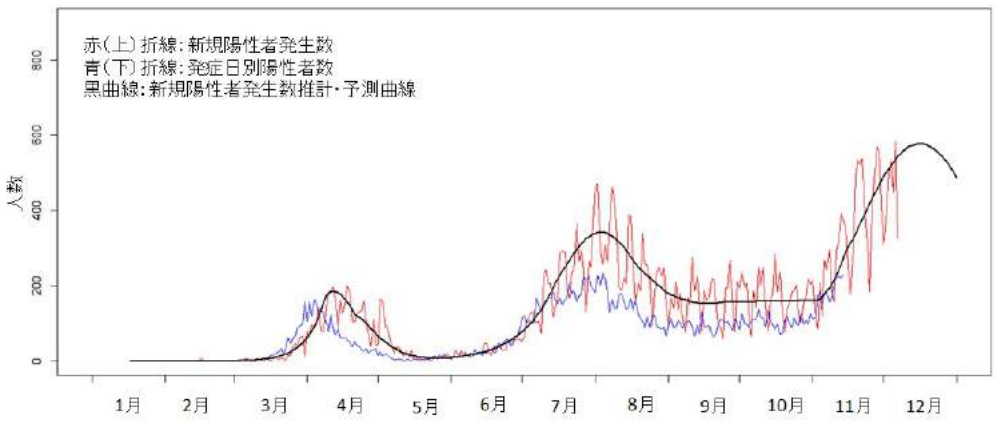


図 1: 東京都の感染者発生数推定結果（2020 年）

仮想的な 0 近傍への外挿とその収束レートについて *

奥野彰文 [1,3], 下平英寿 [2,3]

[1] 統計数理研究所, [2] 京都大学大学院 情報学研究科, [3] 理化学研究所 AIP センター

1 はじめに

分類問題における教師あり学習に幅広く用いられる手法の一つに k 近傍法 (k-Nearest Neighbour, 略称 k-NN; Fix and Hodges (1951)) がある．k 近傍法ではまずクエリ近傍の k 個のデータベクトルを検索し，対応するラベルの平均によりクエリの各ラベル確率を予測する．最大の確率を持つラベルを出力するプラグイン型の分類を行うと，k 近傍法の統計的一致性，すなわち誤分類確率が最適値に収束することが Cover and Hart (1967) などにより示されており，またその収束の速さ (収束レート) が Chaudhuri and Dasgupta (2014) などにより示されている．

一方で，任意のラベル確率予測器を用いてプラグイン型の分類を行う場合の最適な収束レートが Audibert and Tsybakov (2007) により与えられており，ラベルの条件付き期待値が非常に滑らかな場合には k 近傍法は最適レートを達成できない．

k 近傍法にはバイアスとバリアンスのトレードオフがある：k を大きくするとクエリから遠いデータベクトルのラベルが考慮されるようになり，バイアスが増大してしまう．本研究では，いくつかの k について k 近傍法を行い，予測されたラベル確率を $k=0$ に仮想的に外挿することで漸近的なバイアスを減少させ，最適な収束レートを達成する**マルチスケール k 近傍法**を提案する．

2 問題設定

$\mathcal{X} \subset \mathbb{R}^d$ を非空なコンパクト集合とし，$(X,Y) \in \mathcal{X} \times \{0,1\}$ を分布 $\mathbb{Q}$ から生成される確率変数とする．サンプル $\mathcal{D}_n := \{(X_i,Y_i)\}_{i=1}^n$ とクエリ (X_*,Y_*) は独立に分布 $\mathbb{Q}$ に従うとする．$\mathcal{D}_n, X_*$ を用いてクエリとなるデータベクトル $X_* \in \mathcal{X}$ のラベル $Y_* \in \{0,1\}$ を予測する分類器 $\hat{g}_n : \mathcal{X} \to \{0,1\}$ を学習し評価する．

分類器の評価には誤分類確率 $L(g) := \mathbb{P}_{X_*,Y_*}(g(X_*) \neq Y_*)$ を用いた excess risk

$$\mathcal{E}(\hat{g}_n) := \mathbb{E}_{\mathcal{D}_n}(L(\hat{g}_n)) - \inf_{g:\mathcal{X}\to\{0,1\}} L(g)$$

を用いる．Excess risk $\mathcal{E}(\hat{g}_n)$ のサンプル数 n についてのオーダーを**収束レート**と呼ぶ．

3 k 近傍法

クエリからのユークリッド距離により添え字を並べ替える：$\|X_{(1)} - X_*\|_2 \leq \|X_{(2)} - X_*\|_2 \leq \cdots \leq \|X_{(n)} - X_*\|_2$．ユーザの指定するパラメータ $k \in \mathbb{N}$ と和が 1 となる重み $w_1, w_2, \ldots, w_k \geq 0$ を用いた

$$\hat{\eta}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) = \sum_{i=1}^{k} w_i Y_{(i)}$$

を重み付き k 近傍推定量 (k-NN estimator) と呼び，$w_1 = w_2 = \cdots = w_k = k^{-1}$ の場合単純に k 近傍推定量と呼ぶ．プラグイン型の分類器 $\hat{g}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) := 1\left(\hat{\eta}_{k,\boldsymbol{w}}^{(k\mathrm{NN})}(X_*) \geq 1/2\right)$ を特に重み付き k 近傍分類器 (weighted k-NN classifier) と呼ぶ．以降ではクエリ X_* を固定し記号を省略する．

4 マルチスケール k 近傍法

図 1には各 k での k 近傍推定量を示した．k が大きいほど k 近傍推定量の分散は小さいが，青線で示した真値との乖離 (バイアス) が大きくなる．

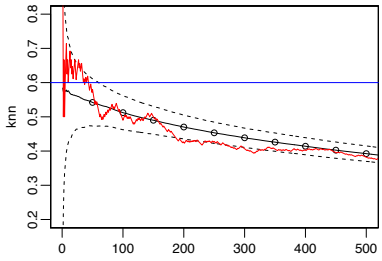


図 1: 赤線：k 近傍推定量，青線：真値 $\eta(X_*) = \mathbb{E}(Y \mid X_*)$，黒線と破線はモンテカルロシミュレーションにおける k 近傍推定量の平均と標本標準偏差を表す．

本研究では，$1 \leq k_1 < k_2 < \cdots < k_V \leq n$ $(V \in \mathbb{N})$ を用いて k 近傍推定量 $\hat{\eta}_{k_1}^{(k\mathrm{NN})}, \hat{\eta}_{k_2}^{(k\mathrm{NN})}, \ldots, \hat{\eta}_{k_V}^{(k\mathrm{NN})}$ を計算し，$(k=)0$ 近傍へ外挿することでバイアスを減少させる**マルチスケール k 近傍法**を提案する．より具体的な手続きとしては

(1) 半径 $r_k := \|X_{(k)} - X_*\|_2$ $(k = k_1, \ldots, k_V)$ を計算．
(2) 半径 r_{k_v} を介した回帰関数 $f(r_{k_v};\theta)$ を用いて k 近傍推定量 $\hat{\eta}_{k_v}^{(k\mathrm{NN})}$ を予測する：

$$\hat{\theta}_{\boldsymbol{k}} := \arg\min_{\theta\in\Theta} \sum_{v=1}^{V} \{\hat{\eta}_{k_v}^{(k\mathrm{NN})} - f(r_{k_v};\theta)\}^2.$$

(3) $r_0 = 0$ として，$k=0$ に外挿する：

$$\hat{\eta}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})} := f(0;\hat{\theta}_{\boldsymbol{k}}).$$

対応するプラグイン型のマルチスケール k 近傍分類器を $\hat{g}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}(X_*) := 1\left(\hat{\eta}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}(X_*) \geq 1/2\right)$ とする．

* 本稿は Okuno and Shimodaira (2020) を基にしている．

次節では次数が偶数の項のみで構成された多項式

$$f(r;\theta) = \theta_0 + \theta_1 r^2 + \theta_2 r^4 + \cdots + \theta_C r^{2C} \quad (1)$$

を回帰関数として用いると，近傍法の収束レートが改善することを紹介する．

5 理論解析

本節では，マルチスケール k 近傍法が通常の k 近傍法の収束レートを改善し，最適レートを達成することを示す．最初に，条件付き期待値 $\eta(x) = \mathbb{E}[Y \mid X = x]$ に関するいくつかの条件を説明する．

定義 1 (α-margin 条件)．ある定数 $L_\alpha \geq 0, \tilde{t} > 0, \alpha > 0$ が存在して $\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq L_\alpha t^\alpha$ $(\forall t \in (0, \tilde{t}], X \in \mathcal{X})$ とできるとき η は α-margin 条件を満たすという．

定義 2 (β-Hölder 条件)．$\mathcal{T}_{q,X_*}[\eta]$ を点 $X_* \in \mathcal{X}$ での関数 η の $q\,(\in \mathbb{N}_0)$ 次テイラー展開とする．ある定数 $L_\beta > 0, \beta > 0$ が存在して $|\eta(X) - \mathcal{T}_{\lfloor\beta\rfloor,X_*}[\eta](X)| \leq L_\beta \|X - X_*\|^\beta$ とできるとき $\eta \in C^{\lfloor\beta\rfloor}(\mathcal{X})$ は β-Hölder 条件を満たすという．

定義 3 (γ-neighbour average smoothness 条件)．集合 $B \subset \mathbb{R}^d$ について定義される関数 $\eta^{(\infty)}(B) := \mathbb{E}(Y \mid X \in B)$ と中心 $X \in \mathcal{X}$ で半径 r の球 $B(X;r)$ について，ある定数 $L_\gamma > 0, \gamma > 0$ が存在して $|\eta^{(\infty)}(B(X;r)) - \eta(X)| \leq L_\gamma r^\gamma$ を満たすとき η は γ-neighbour average smooth であるという．

一般に α, β, γ が大きいほど分類器の収束が早くなる．最後に，X の密度関数について以下の条件を定義すると定理 5が成り立つ．

定義 4 (Strong density assumption)．X の密度関数 μ について，ある定数 $\mu_{\min}, \mu_{\max} \in (0, \infty)$ が存在して $\mu_{\min} \leq \mu(X) \leq \mu_{\max}$ $(\forall X \in \mathcal{X})$ とできるとき μ は strong density assumption (SDA) を満たすという．

定理 5 (k 近傍分類器の収束レート; Chaudhuri and Dasgupta (2014) 定理 4(b))．集合 $\mathcal{X}$ がコンパクトであり，η が α-margin と γ-neighbour average smoothness 条件を満たすとする．SDA が成り立つとし，$k_* := k_n \asymp n^{2\gamma/(2\gamma+1)}$ とすると $\mathcal{E}(\hat{g}_k^{(k\mathrm{NN})}) = O(n^{-(1+\alpha)\gamma/(2\gamma+d)})$.

定理 5は γ-neighbour average smoothness 条件を仮定するが，代わりに β-Hölder 条件を仮定すると，いくつかの条件の下で $\gamma = \min\{\beta, 2\}$ が示せる (Okuno and Shimodaira (2020) 定理 1)．つまり $\beta > 2$ として β-Hölder 条件を仮定すると

$$\mathcal{E}(\hat{g}_k^{(k\mathrm{NN})}) = O(n^{-2(1+\alpha)/(4+d)})$$

となり，β がいくら大きくなっても収束レートがバウンドされてしまう．これは局所線形回帰 (Tsybakov, 2009) と同様の現象である (Hall and Kang, 2005)．

次に，β-Hölder 条件の下でマルチスケール k 近傍法の収束レートを以下の定理 6に示す．

定理 6 (マルチスケール k 近傍分類器の収束レート; Okuno and Shimodaira (2020) 定理 2)．$\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_V) \in \mathbb{R}^V$ を $\ell_1 = 1 < \ell_2 < \cdots < \ell_V < \infty$ なるベクトルとし，$k_{1,n} \asymp n^{2\beta/(2\beta+d)}$ かつ $k_{v,n} \asymp \min\{k \in [n] \mid \|X_{(k)} - X_*\|_2 \geq \ell_v \|X_{(k_{1,n})} - X_*\|_2\}$ とする．回帰関数として (1) を用いる．(i) $\mu, \mu\eta$ がそれぞれ β-Hölder 条件を満たし，(ii) SDA を仮定し，かつ (iii) $C := \lfloor\beta/2\rfloor \leq V - 1$ とすると，推定量が発散しないための条件 (Okuno and Shimodaira (2020) (C-3)) の下で

$$\mathcal{E}(\hat{g}_{\boldsymbol{k}}^{(\mathrm{MS}k\mathrm{NN})}) = O(n^{-(1+\alpha)\beta/(2\beta+d)}). \quad (2)$$

式 (2) で示したマルチスケール k 近傍法の収束レートは，任意のプラグイン型の分類器に関する最適レート (Audibert and Tsybakov, 2007) と一致する．

6 まとめと議論

以上より，提案したマルチスケール k 近傍法は k 近傍法の収束レートを改善し，最適レートを達成することを示した．最後に，本節では提案法と局所多項式回帰，重み付き k 近傍法との比較についての議論をまとめる．

6.1 局所多項式回帰との比較

Audibert and Tsybakov (2007) では，局所多項式回帰 (Tsybakov, 2009) を用いたプラグイン型の分類器が最適レートを達成することを示している．局所多項式回帰はクエリ $X_* \in \mathcal{X} \subset (\mathbb{R}^d)$ の周辺での η のテイラー展開を多項式で推定するため，$1 + d + d^2 + \cdots + d^C$ 個の係数を推定する必要があるが，提案したマルチスケール k 近傍法では (1) に現れる $1 + C$ 個の項の係数を推定するだけでよい．

6.2 重み付き k 近傍法との比較

単純な k 近傍法と回帰を組み合わせ容易に実装することができるマルチスケール k 近傍法は，別の解釈として，回帰を介し間接的に重み付き k 近傍法の**実数値**の重み $w_1, \ldots, w_k \in \mathbb{R}$ を推定しているとみなせる．

重み付き k 近傍法では通常，非負の重みのみを考える．非負の重みのみ考える場合，最適な重みを用いても，収束レートが重みを用いない k 近傍法と同じになってしまうことが知られている (Samworth, 2012)．Samworth (2012) ではさらに excess risk のテイラー展開を最適化する実数の重みを用いれば最適レート (2) が達成できることを示している．

提案したマルチスケール k 近傍法は回帰を介して，Samworth (2012) は excess risk のテイラー展開の最適化を介して重みを決定することから，これら 2 つの手法で得られる実数値の重みは同一ではないが，どちらも最適なレートを達成する．しかし Samworth (2012) の方法ではテイラー展開を重みに関して最適化する必要があり，最適な重みの計算が煩雑になってしまう．提案法には回帰を介して容易に同等な計算を可能にする利点がある．

Active learning for distributionally robust level set estimation

稲津 佑 [1]　岩崎 省吾 [1]　竹内 一郎 [1,2]

[1] 名古屋工業大学 情報工学専攻　[2] 理化学研究所革新知能統合研究センター

1. 概要

評価コストが高い black-box 関数 f が, コントロール可能な変数 $\boldsymbol{x}$ とコントロールできないある分布 P に従う変数 $\boldsymbol{w}$ を用いて $f(\boldsymbol{x},\boldsymbol{w})$ と表されるケースは多い. このとき, $\boldsymbol{w}$ の変動に関してロバストな解を列挙することを考える. ロバスト性の尺度として自然に考えられるもののひとつとして, ある与えられた閾値 h を上回る確率を考えることができる. これは probability threshold robustness (PTR) 尺度と呼ばれ, 産業用ロボットの安全性の保証や金融における chance-constrained optimization 等の応用がなされている. しかしながら, この尺度は分布 P を誤って特定した場合はロバスト尺度として適切でない. そこで, 与えられた候補分布の中での最悪ケースの PTR を考えるという, "distributionally robust PTR measure (DRPTR)" を考える. 本稿では, DRPTR 尺度に関して, ある所望の確率 α を上回る領域, reliable set を効率的に同定する問題を考える. これは, DRPTR に対するレベルセット推定 (LSE) 問題として定式化することができる. 通常の LSE に対してプラクティカルにうまく機能することが知られている MILE と呼ばれる獲得関数を, 本設定に拡張した獲得関数を提案する. また, 数値実験を通し, 提案法の性能を確認する.

2. 設定

関数 $f:\mathcal{X}\times\Omega\to\mathbb{R}$ を, $\mathcal{X}\times\Omega$ で定義された評価コストが高い black-box 関数とする. ただし, $\mathcal{X}$ および Ω は有限集合とする. 各入力 $(\boldsymbol{x},\boldsymbol{w})\in\mathcal{X}\times\Omega$ に対し, 関数 $f(\boldsymbol{x},\boldsymbol{w})$ の値は $f(\boldsymbol{x},\boldsymbol{w})+\varepsilon$ として観測されるとする. ただし, ε は正規分布 $\mathcal{N}(0,\sigma^2)$ に従う, 独立なノイズである. また, $\boldsymbol{w}$ はある未知の分布 $P^\dagger$ に従うものとし, コントロールできない変数とする. ただし, 実験時においては, $\boldsymbol{w}$ の選択は可能とする. また, $\mathcal{A}$ を $P^\dagger$ の候補分布族とする. 本稿では, $\mathcal{A}$ として, $\mathcal{A}=\{\text{p.m.f. } p(\boldsymbol{w}) \mid d(p(\boldsymbol{w}),p^*(\boldsymbol{w}))<\epsilon\}$ を考える. ただし, , $p^*(\boldsymbol{w})$ は実験者が指定した参照分布であり, $d(\cdot,\cdot)$ はある分布間距離である. また, $\epsilon>0$ とする. このとき, 与えられた閾値 h の下, 各 $\boldsymbol{x}\in\mathcal{X}$ に対する DRPTR, $F(\boldsymbol{x})$ を以下で定義する:

$$F(\boldsymbol{x})=\inf_{p(\boldsymbol{w})\in\mathcal{A}}\sum_{\boldsymbol{w}\in\Omega}\mathbb{1}[f(\boldsymbol{x},\boldsymbol{w})>h]p(\boldsymbol{w}).$$

本稿の目的は, ある与えられた確率 $\alpha\in(0,1)$ に対し, $F(\boldsymbol{x})>\alpha$ を満たす $\mathcal{X}$ の部分集合 H (reliable set) を効率的に同定することである:

$$H=\{\boldsymbol{x}\in\mathcal{X}\mid F(\boldsymbol{x})>\alpha\},\ L=\{\boldsymbol{x}\in\mathcal{X}\mid F(\boldsymbol{x})\le\alpha\}. \tag{2.1}$$

2.1. Gaussian process

本稿では, 未知関数 f に対するモデリングとして, Gaussian process (GP) を用いる. 関数 f に対する事前分布に GP, $\mathcal{GP}(0,k((\boldsymbol{x},\boldsymbol{w}),(\boldsymbol{x}',\boldsymbol{w}')))$ を仮定する. ここで, $k((\boldsymbol{x},\boldsymbol{w}),(\boldsymbol{x}',\boldsymbol{w}'))$ は正定値カーネルである. 今, データセット $\{(\boldsymbol{x}_i,\boldsymbol{w}_i,y_i)\}_{i=1}^t$ が与えられた下, f の事後分布は再び GP となり, $f(\boldsymbol{x},\boldsymbol{w})$ の事後平均 $\mu_t(\boldsymbol{x},\boldsymbol{w})$, 事後分散 $\sigma_t^2(\boldsymbol{x},\boldsymbol{w})$ はそれぞれ以下で与えられる:

$$\mu_t(\boldsymbol{x},\boldsymbol{w})=\boldsymbol{k}_t^\top(\boldsymbol{x},\boldsymbol{w})(\boldsymbol{K}_t+\sigma^2\boldsymbol{I}_t)^{-1}\boldsymbol{y}_t,\ \sigma_t^2(\boldsymbol{x},\boldsymbol{w})=k((\boldsymbol{x},\boldsymbol{w}),(\boldsymbol{x},\boldsymbol{w}))-\boldsymbol{k}_t^\top(\boldsymbol{x},\boldsymbol{w})(\boldsymbol{K}_t+\sigma^2\boldsymbol{I}_t)^{-1}\boldsymbol{k}_t(\boldsymbol{x},\boldsymbol{w}).$$

3. 提案法

3.1. Credible interval and LSE

各点 $(\boldsymbol{x},\boldsymbol{w})\in\mathcal{X}\times\Omega$ に対し, 第 t 試行時における $f(\boldsymbol{x},\boldsymbol{w})$ に対する信用区間を $Q_t(\boldsymbol{x},\boldsymbol{w})=[l_t(\boldsymbol{x},\boldsymbol{w}),u_t(\boldsymbol{x},\boldsymbol{w})]$ で定める. ただし, $l_t(\boldsymbol{x},\boldsymbol{w})=\mu_t(\boldsymbol{x},\boldsymbol{w})-\beta_t^{1/2}\sigma_t(\boldsymbol{x},\boldsymbol{w})$, $u_t(\boldsymbol{x},\boldsymbol{w})=\mu_t(\boldsymbol{x},\boldsymbol{w})+\beta_t^{1/2}\sigma_t(\boldsymbol{x},\boldsymbol{w})$ であり, $\beta_t^{1/2}\ge0$ である. 同様に, $\mathbb{1}[f(\boldsymbol{x},\boldsymbol{w})>h]$ に対する信用区間を $Q_t(\boldsymbol{x},\boldsymbol{w})$ に基づき構築する. ここで, 理論的保証のために, ある正の精度パラメータ η

を用いて, $\mathbb{1}[f(\boldsymbol{x},\boldsymbol{w})>h]$ の信用区間 $\tilde{Q}_t(\boldsymbol{x},\boldsymbol{w};\eta)$ を以下のように構築する:

$$\tilde{Q}_t(\boldsymbol{x},\boldsymbol{w};\eta)\equiv[\tilde{l}_t(\boldsymbol{x},\boldsymbol{w};\eta),\tilde{u}_t(\boldsymbol{x},\boldsymbol{w};\eta)]=\begin{cases}[1,1] & \text{if } l_t(\boldsymbol{x},\boldsymbol{w})>h-\eta\\ [0,1] & \text{if } l_t(\boldsymbol{x},\boldsymbol{w})\leq h-\eta \text{ and } u_t(\boldsymbol{x},\boldsymbol{w})>h\\ [0,0] & \text{if } l_t(\boldsymbol{x},\boldsymbol{w})\leq h-\eta \text{ and } u_t(\boldsymbol{x},\boldsymbol{w})\leq h\end{cases}.$$

これを用いて, $F(\boldsymbol{x})$ の信用区間 $Q_t^{(F)}(\boldsymbol{x};\eta)\equiv[l_t^{(F)}(\boldsymbol{x};\eta),u_t^{(F)}(\boldsymbol{x};\eta)]$ を次で定義する:

$$l_t^{(F)}(\boldsymbol{x};\eta)=\inf_{p(\boldsymbol{w})\in\mathcal{A}}\sum_{\boldsymbol{w}\in\Omega}\tilde{l}_t(\boldsymbol{x},\boldsymbol{w};\eta)p(\boldsymbol{w}),\ u_t^{(F)}(\boldsymbol{x};\eta)=\inf_{p(\boldsymbol{w})\in\mathcal{A}}\sum_{\boldsymbol{w}\in\Omega}\tilde{u}_t(\boldsymbol{x},\boldsymbol{w};\eta)p(\boldsymbol{w}).$$

なお, 分布間距離 $d(\cdot,\cdot)$ として $L1$-距離を用いた場合, 上の式は線形計画問題を解くことに相当する. 同様に, $d(\cdot,\cdot)$ として $L2$-距離を用いた場合は 2 次錐計画問題を解くことに相当する. いずれの場合も, 高速に解を求めることができるソルバーが存在するので, これらの分布間距離を用いた場合は $Q_t^{(F)}(\boldsymbol{x};\eta)$ を計算することは容易である. このとき, H および L の推定集合を以下で定める:

$$H_t=\{\boldsymbol{x}\in\mathcal{X}\mid l_t^{(F)}(\boldsymbol{x};\eta)>\alpha\},\ L_t=\{\boldsymbol{x}\in\mathcal{X}\mid u_t^{(F)}(\boldsymbol{x};\eta)\leq\alpha\}. \tag{3.1}$$

また, 未分類集合を $U_t=\mathcal{X}\setminus(H_t\cup L_t)$ とする.

3.2. 獲得関数

ここでは, 次に評価すべき入力点を決定するための獲得関数を与える. 我々は, 先行研究で提案された MILE 獲得関数に基づく獲得関数を提案する. MILE の考え方は, 新たな点が加わったときの分類数と現在の分類数の差が, 期待値的に最も大きくなる点を次の評価点とするという考え方である. 本稿では, 計算コストの都合上, 未分類集合内の点が期待値的にどれだけ新しく H へ分類されるかという基準に基づいた獲得関数を提案する.

点 $\boldsymbol{x}^*$, $\boldsymbol{w}^*$ を新たな入力点とし, $y^*=f(\boldsymbol{x}^*,\boldsymbol{w}^*)+\varepsilon$ が得られたとする. 組 $(\boldsymbol{x}^*,\boldsymbol{w}^*,y^*)$ が追加されたときの, $\eta=0$ とした場合の $F(\boldsymbol{x})$ の信用区間の下端を $l_t^{(F)}(\boldsymbol{x};0|\boldsymbol{x}^*,\boldsymbol{w}^*,y^*)$ と書く. このとき, 以下の獲得関数 $a_t(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を考える:

$$a_t(\boldsymbol{x}^*,\boldsymbol{w}^*)=\sum_{\boldsymbol{x}\in U_t}\mathbb{E}_{y^*}[\mathbb{1}[l_t^{(F)}(\boldsymbol{x};0|\boldsymbol{x}^*,\boldsymbol{w}^*,y^*)>\alpha]]. \tag{3.2}$$

ここで, 先行研究では, オリジナルの MILE 獲得関数に分散項を追加した RMILE と呼ばれる獲得関数を提案している. これにより, オリジナルの MILE の性能を改善できることを示している. 同様の考え方を用いて, 本設定においても, 修正獲得関数を提案する:

Definition 3.1 (提案獲得関数 1)**.** $a_t(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を (3.2) で与えられるものとし, γ を正のパラメータとする. このとき, 提案獲得関数 $a_t^{(1)}(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を以下で定め, 次の評価点を $(\boldsymbol{x}_{t+1},\boldsymbol{w}_{t+1})=\mathrm{argmax}_{(\boldsymbol{x}^*,\boldsymbol{w}^*)\in\mathcal{X}\times\Omega}a_t^{(1)}(\boldsymbol{x}^*,\boldsymbol{w}^*)$ により決定する:

$$a_t^{(1)}((\boldsymbol{x}^*,\boldsymbol{w}^*))=\max\{a_t(\boldsymbol{x}^*,\boldsymbol{w}^*),\gamma\sigma_t(\boldsymbol{x}^*,\boldsymbol{w}^*)\}.$$

ここで, 提案獲得関数は分散項 $\gamma\sigma_t(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を新たに追加しているが, 代わりに, RMILE を追加することもできる. 実際, プラクティカルな性能としてはこちらのほうが優れている:

Definition 3.2 (提案獲得関数 2)**.** $a_t(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を (3.2) で与えられるものとし, γ を正のパラメータとする. このとき, 提案獲得関数 $a_t^{(2)}(\boldsymbol{x}^*,\boldsymbol{w}^*)$ を以下で定め, 次の評価点を $(\boldsymbol{x}_{t+1},\boldsymbol{w}_{t+1})=\mathrm{argmax}_{(\boldsymbol{x}^*,\boldsymbol{w}^*)\in\mathcal{X}\times\Omega}a_t^{(2)}(\boldsymbol{x}^*,\boldsymbol{w}^*)$ により決定する:

$$a_t^{(2)}((\boldsymbol{x}^*,\boldsymbol{w}^*))=\max\{a_t(\boldsymbol{x}^*,\boldsymbol{w}^*),\gamma\mathrm{RMILE}_t(\boldsymbol{x}^*,\boldsymbol{w}^*)\}.$$

本研究における, いくつかの理論結果および数値実験の結果については当日報告する.

謝辞

本研究の一部は，科学研究費 (20H00601, 16H06538), JST CREST (JPMJCR1502), 理化学研究所革新知能統合研究センターの補助を受けて行われた.

分散を伝播させる BNN を用いた 不確実性計算の高速化とその応用

前 佑樹 (株式会社デンソー)　　熊谷 亘 (東京大学/理研 AIP)
金森 敬文 (東京工業大学/理研 AIP)

1　はじめに

環境認識やパスプランニングといった機能コンポーネントの実現に Neural Network (NN) のような予測モデルが使われるようになり，それらコンポーネントを搭載した運転自動化システムの開発が進められている．

運転自動化システムでは，故障に対する安全性に加えて，正常機能そのものの安全性にも対応することが求められる．Safety-critical な車載電子システムにおいては，故障に対する安全性として国際規格である機能安全への対応が従来から求められている．運転自動化システムにおいては，それに加えて，正常機能の性能限界に対する安全性を扱う Safety Of The Intended Functionality (SOTIF) への対応も求められ始めている．

システムレベルでの安全性の達成のため，コンポーネントレベルでは，予測モデルの不確実性を定量化できるベイズ理論や Bayesian Neural Network (BNN) への関心が高まっている．高信頼な運転自動化システムの実現には，テストデータにおいて高い性能を達成するだけでなく，予測分布のような形で不確実性を定量化することも重要である．不確実性を定量化できれば，現実世界で稀にしか起こらないサンプルに対して予測が揺らいでいるかが分かるため，例えば棄却オプションを使って運転手の介入を促すフェイルセーフなシステムを実現できる．また，不確実性の高いサンプルを抽出して性能改善に利用する能動学習も可能となる．

不確実性を容易に定量化できる近似方法として Monte Carlo Dropout (MC-Dropout) [1] が知られているが，リアルタイム性も必要なシステムにとってその方法は計算時間が掛かりすぎる．MC-Dropout は，dropout 層を用いて複数の NN をサンプリングしてそれぞれで順伝播 (フィードフォワード) することで予測分布を得る方法である．MC-Dropout の利用で不確実性を容易に定量化できるが，限られた計算資源下でリアルタイム性も求められる運転自動化システムにとって，複数回のフィードフォワードは計算負荷が大きい．

2　手法の概要

NN 各層の入出力を確率変数と見なしてそれら確率変数の期待値と分散を伝播することで，サンプリングせずに一回のフィードフォワードで予測分布を計算できるようになる [2, 3, 4]．特に，affine 層では独立性を仮定せず，活性化層ではガウス近似で期待値と分散を計算する Variance Propagation Bayesian Neural Network (VPBNN) について，代表的な層における計算を表 1 に示す [4]．NN の各層を期待値と分散を計算する層に置き換えてフィードフォワードすることで，MC-Dropout よりも高速に予測分布が得られるようになる．

表 1 に示した代表的な層の計算を応用したり組み合わせたりすることで，Convolutional Neural Network (CNN) や Recurrent Neural Network (RNN) でも一回のフィードフォワードで予測分布が得られるよう

表 1　層の入出力をそれぞれ x, y としたときの NN および VPBNN における代表的な層での計算．ρ は自信過剰を防ぐために設けた超パラメータであり，バリデーションセットを利用して適切な値に設定できる．

Layer	NN での計算	VPBNN での計算
Affine	$y_i = \sum_j W_{ij} x_j + b_i$	$\mathbb{E}[y_i] = \sum_j W_{ij}\mathbb{E}[x_j] + b_i$ $\mathrm{Var}[y_i] \leq (1-\rho)\sum_j W_{ij}^2 \mathrm{Var}[x_j] + \rho\left(\sum_j \lvert W_{ij}\rvert \sqrt{\mathrm{Var}[x_j]}\right)^2$
Dropout	$y = dx, d \sim \mathrm{Bern}(p)$	$\mathbb{E}[y] = p\mathbb{E}[x]$ $\mathrm{Var}[y] = p\mathrm{Var}[x] + p(1-p)\mathbb{E}[x]^2$
ReLU	$y = \max(0, x)$	$\mathbb{E}[y] \approx \mathbb{E}[x]\Phi(r) + \sqrt{\mathrm{Var}[x]}\phi(r), r = \mathbb{E}[x]/\sqrt{\mathrm{Var}[x]}$ $\mathrm{Var}[y] \approx (\mathbb{E}[x]^2 + \mathrm{Var}[x])\Phi(r) + \mathbb{E}[x]\sqrt{\mathrm{Var}[x]}\phi(r) - \mathbb{E}[y]^2$
Sigmoid	$y = s(x)$	$\mathbb{E}[y] \approx s\left(\frac{\mathbb{E}[x]}{\sqrt{1+0.125\pi\mathrm{Var}[x]}}\right)$ $\mathrm{Var}[y] \approx \mathbb{E}[y](1-\mathbb{E}[y])\left(1 - \frac{1}{\sqrt{1+0.125\pi\mathrm{Var}[x]}}\right)$
Tanh	$y = 2s(2x) - 1$	$\mathbb{E}[y] \approx 2s\left(\frac{2\mathbb{E}[x]}{\sqrt{1+0.5\pi\mathrm{Var}[x]}}\right) - 1$ $\mathrm{Var}[y] \approx (1+\mathbb{E}[y])(1-\mathbb{E}[y])\left(1 - \frac{1}{\sqrt{1+0.5\pi\mathrm{Var}[x]}}\right)$

になる．CNN で使われる畳み込み層や平均プーリング層は affine 層を特化させた層であるため，VPBNN は CNN にも適用できる．同様に，RNN の一種である Long Short-Term Memory (LSTM) は，affine 層，sigmoid 層，tanh 層を組み合わせた層と見なせるため，VPBNN は RNN にも適用できる．

3　数値実験

VPBNN の適用により，LSTM を用いた言語モデリングにおいてキャリブレーション効果が得られること，CNN を用いた Out-of-Distribution 検出において分布外サンプルをより区別できるようになること，不確実性の高いサンプルから選ぶことでラベル付きデータの数を減らせることを確認した．

参考文献

[1] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[2] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-Free Epistemic Uncertainty Estimation Using Approximated Variance Propagation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Alexander Shekhovtsov and Boris Flach. Feed-forward propagation in probabilistic neural networks with categorical and Max layers. In *7th International Conference on Learning Representations*, 2019.

[4] Yuki Mae, Wataru Kumagai, and Takafumi Kanamori. Bayesian Neural Networks with Variance Propagation for Uncertainty Evaluation. submitted.

科研費シンポジウム 機械学習・統計学・最適化の数理と AI 技術への展開
2020 年 12 月 18-19 日 報告書

パレートフロンティアの情報量に基づく多目的ベイズ最適化

烏山 昌幸，名古屋工業大学

概要

複数の目的関数の同時最適化を考える多目的最適化は様々な産業や工業，あるいは科学的な探索問題に頻繁に現れる．ここでは，目的関数自体が未知の状況での多目的最適化を考え，ベイズ最適化に基づいてパレート最適点を探索する問題を考える．特に，単一関数のベイズ最適化で高い探索効率を達成することが知られている情報量に基づくアプローチに着目する．多目的最適化においては，探索の目標がより複雑なため，情報量に基づくアプローチは非常に複雑な近似計算が必要となる方法か，目的関数間のトレードオフを無視するナイーブな方法が知られるのみであった．本発表では，パレート最適点が作る目的関数値の集合であるパレートフロンティアの情報量を考えることで，目的関数間のトレードオフを考慮しつつ，かつ比較的に簡単に評価ができる指標が定義でき，効率の良い探索が可能であることを紹介する．

多目的最適化は多岐に渡る分野の実践的な探索問題に現れる．例えば材料開発では，リチウムイオン電池材料の伝導性と安定性の同時最適化など，複数のパフォーマンスを最適化する新奇材料の発見が望まれている．あるいは，近年発達が目覚ましい深層学習ベースの知能システムのチューニングでは，予測精度とネットワークの計算効率のトレードオフの最適化が発生する．このような問題は一般に，L 個の未知関数 $f^1(\boldsymbol{x}), \ldots, f^L(\boldsymbol{x})$ を同時に最適化する問題として捉えることができる．多目的最適化では通常，各目的関数間にはトレードオフがあり単一の最適化点 $\boldsymbol{x}$ は存在しない．そこで，多くの場合はパレート最適と呼ばれる性質を満たす $\boldsymbol{x}$ の集合の探索を考える．大まかには，パレート最適な $\boldsymbol{x}$ は，目的関数値のベクトル $\boldsymbol{f}_{\boldsymbol{x}} \coloneqq (f^1(\boldsymbol{x}), \ldots, f^L(\boldsymbol{x}))$ を全ての目的関数値について同時に改善する他の $\boldsymbol{x}$ が存在しないような状態として特徴付けられる．また，ここではパレート最適な $\boldsymbol{x}$ の集合が作る $\boldsymbol{f}_{\boldsymbol{x}}$ の集合をパレートフロンティア $\mathcal{F}^*$ とする．

本発表では，パレートフロンティア $\mathcal{F}^*$ に関する情報量 (entropy) を考えることで，効率的に探索点を決定する方法として Pareto-frontier entropy search (PFES) [1] を紹介する．情報量に基づいて，多目的ベイズ最適化を考える先行研究の主なものには [2] と [3] がある．[2] では，パレート最適な $\boldsymbol{x}$ の集合に対する情報量を考えたが，これには非常に複雑な近似計算が必要であり，実装も煩雑なうえ，求めたものが近似前の指標とどの程度乖離するのかも明らかではない．一方で，[3] はそれぞれの目的関数の単一関数としての最適な値の情報量を考える指標を提案したが，これは多目的最適化におけるトレードオフ関係を完全に無視する指標である．PFES においては，情報量ベースの方法が共通で採用する簡単化とサンプリング近似は採用するが，それ以外はほぼ解析的に計算可能な指標に帰着できる．これは，パレートフロンティア $\mathcal{F}^*$ が定義する目的関数値の空間を適切に分割することで，情報量を比較的簡単に評価できるサブブロックに分けることができるために達成される．また，目的関数間のトレードオフも考慮されることになり，上記の先行研究の難点を回避することができる．

また，本発表では，目的関数未知の多目的最適化における「decoupled setting」についても紹介し，PFES がこの設定をうまく扱うことができることも述べる．通常，目的関数の観測を得るときは $\boldsymbol{f}_{\boldsymbol{x}}$ の全ての次元が同時に観測されることを暗に想定するが，この設定では個々の次元 $f^l(\boldsymbol{x})$ が個別に観測できる場合を考える．知る限りこの設定に着目しているのは [2] と PFES[1] だけであるが，実践上非常に重要な可能性を持つ設定と考えられる．例えば，電池材料の安定性と伝導性の最適化の事例を考える．それぞれの性質をシミュレーション計算（量子力学計算など）で得られる場合を考えると，安定性と伝導性は別々のシミュレーション計算で評価される．観測はやはり高コストなため，毎回必ず両方の観測を得るのは必ずしも効率的とは言えない可能性がある．また，それぞれの観測にかかるコストが異なることもある（例えば，伝導度の評価の方が安定性の評価よりも評価コストは高い）．PFES は同時観測の場合の良い性質を decoupled setting でも引き継ぎ，自然な情報量が定義できること示す．

最後に，同時観測と個別観測それぞれの設定について，いくつかのベンチマーク関数での評価を紹介する．また，情報量のアプローチは汎用性が高く，多目的最適化以外の設定にも適用可能であり，その他，複数の fidelity が存在する最適化問題への適用事例 [4] についても時間の許す範囲で簡単に言及し，さらなる発展の可能性を議論する．

[1] S. Suzuki, S. Takeno, T. Tamura, K. Shitara, and M. Karasuyama, “Multi-objective Bayesian optimization using Pareto-frontier entropy,” in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119. PMLR, 2020, pp. 9279–9288.

[2] D. Hernandez-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams, “Predictive entropy search for multi-objective Bayesian optimization,” in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48. PMLR, 2016, pp. 1492–1501.

[3] S. Belakaria, A. Deshwal, and J. R. Doppa, “Max-value entropy search for multi-objective bayesian optimization,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 7823–7833.

[4] S. Takeno, H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama, “Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization,” in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119. PMLR, 2020, pp. 9334–9345.

確定的な動きを利用したマルコフ連鎖モンテカルロ法の紹介

鎌谷 研吾 (ISM，CREST JST)

本発表は２つの部分で構成される．ひとつは，マルコフ連鎖モンテカルロ法に確定的動きを導入する試みの紹介である．Metropolis et al., 1953 により提案された初めてのマルコフ連鎖モンテカルロ法であるランダムウォーク・メトロポリス (Random-walk Metropolis) 法は，ランダムウォークによる提案を補正することで任意の確率分布を近似する手法である．ランダムウォークは近似したい確率分布に対して情報を持たない，無知な提案である．無知な提案に含まれるチューニングパラメータを**オンライン学習**することで，間接的により良い収束を実現する方法を**適合的マルコフ連鎖モンテカルロ (Adaptive Markov chain Monte Carlo) 法**という．適合性を認めることでマルコフ性を失い，解析が極めて煩雑になるが，うまく調整ができれば効果は大きい．

いっぽうで，確率分布の密度関数の勾配を利用して，提案に直接的に知識を持たせる方法が Rossky et al., 1978; Duane et al., 1987 によって提案された．こうした手法は**ハミルトニアン・モンテカルロ**法と呼ばれ，現在でもよく使われている．様々なチューニングパラメータの精密な調整が不可欠である．これらの手法はすべて，メトロポリス・ヘイスティングス (Metropolis–Hastings) 法の一種である．

近年，確率的であるはずのメトロポリス・ヘイスティングス法の提案を，確率的な部分と確定的な部分に明確に分けることで，チューニングパラメータの調整に機械学習の手法をより適用しやすくする試みが Song et al., 2018 などで行われた．それを紹介するにはまず，通常のメトロポリス・ヘイスティングス法を説明する必要がある．

状態空間 $(E, \mathcal{E})$ 上のマルコフカーネル $P(x, A)$ とは $x \in E$ を止めるごとに $P(x, \cdot)$ が $(E, \mathcal{E})$ 上の確率分布になり，$A \in \mathcal{E}$ を止めるごとに $x \mapsto P(x, A)$ が $\mathcal{E}$-可測になるものである．メトロポリス・ヘイスティングス (Metropolis–Hastings) 法は，与えられた確率分布 Π にたいし，

$$\Pi(\mathrm{d}x)P(x, \mathrm{d}y) = \Pi(\mathrm{d}y)P(y, \mathrm{d}x)$$

となるマルコフカーネル P を構成する手法である．上の式を満たすとき，P は Π-対照であるという．もし，マルコフカーネル Q が，ある測度 μ（確率測度ではなくてよい）に対して μ-対照で，$\Pi(\mathrm{d}x) = \pi(x)\mu(\mathrm{d}x)$ と書けるなら，

$$P(x, A) = \int_A Q(x, \mathrm{d}y) \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} + R(x)\delta_x(A)$$

とすると，P は Π-対称になる．ここで $\delta_x(A)$ は A が x を含むなら 1 を，そうでないなら 0 を返す関数である．また，$R(x)$ は $P(x, \cdot)$ をマルコフカーネルにするための調整部分であり，ここでは

$$R(x) = 1 - \int_E Q(x, \mathrm{d}y) \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

と書き下せるが，以下では明示しない．マルコフカーネル Q が μ-対称ではなく，

$$\mu(\mathrm{d}x)Q(x, \mathrm{d}y)J(x, y) = \mu(\mathrm{d}y)Q(y, \mathrm{d}x)$$

なる非負の調整項 $J(x,y)$ が必要なときは,

$$P(x,A)=\int_A Q(x,\mathrm{d}y)\min\left\{1,\frac{\pi(y)}{\pi(x)}J(x,y)\right\}+R(x)\delta_x(A)$$

とすると Π-対称性を持つ.

確定的関数 $\psi:E\to E$ によって $Q(x,A)=\delta_{\psi(x)}(A)$ と書けるときも，もし ψ が μ-保測写像, すなわち

$$\mu(\psi^{-1}A)=\psi(A)$$

となり，さらに $\psi(\psi(x))=x$ を満たすなら,

$$P(x,A)=\delta_{\psi(x)}(A)\min\left\{1,\frac{\pi(y)}{\pi(x)}\right\}+R(x)$$

とすることで Π-対称性をもつマルコフカーネルが構成できる．μ-保測写像でないときも，調整項をつけることで同様の構成が可能だ．発表では，Song et al., 2018 らによる，Generative Adversarial Networks を使った $\psi(x)$ の紹介をする.

本発表のもう一つは，確定的動きとジャンプで構成される，**逐次確定的マルコフ過程 (Piecewise-deterministic Markov Processes)** を使ったモンテカルロ法の紹介である．有限回のジャンプ以外は，確定的な遷移を行う．近年，跳躍粒子サンプリング法 (Peters and With, 2012) や ジグザグサンプリング法 (Bierkens et al., 2019) を始めとして，確定的マルコフ過程を使ったモンテカルロ法が注目されている．データを分割して計算負荷を分散して処理することが可能であることが特徴である．本発表ではそれらの近年の研究成果を発表する.

Bierkens, J., P. Fearnhead, and G. O. Roberts (2019). “The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data”. In: *The Annals of Statistics* 47.3, pp. 1288–1320. ISSN: 0090-5364. DOI: 10.1214/18-aos1715.

Duane, S., A. Kennedy, B. J. Pendleton, and D. Roweth (1987). “Hybrid Monte Carlo”. In: *Physics Letters B* 195.2, pp. 216–222. ISSN: 0370-2693.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 1089-7690. DOI: 10.1063/1.1699114.

Peters, E. A. J. F. and G. de With (2012). “Rejection-free Monte Carlo sampling for general potentials”. In: *Physical Review E* 85.2. DOI: 10.1103/PhysRevE.85.026703.

Rossky, P. J., J. D. Doll, and H. L. Friedman (Nov. 1978). “Brownian dynamics as smart Monte Carlo simulation”. In: *The Journal of Chemical Physics* 69.10, pp. 4628–4633. ISSN: 1089-7690. DOI: 10.1063/1.436415. URL: http://dx.doi.org/10.1063/1.436415.

Song, J., S. Zhao, and S. Ermon (2018). “A-NICE-MC: Adversarial Training for MCMC”. In: *ArXiv*. arXiv: 1706.07561 [stat.ML].

鎌谷 研吾 (Kamatani, Kengo)
190-8562 東京都立川市緑町 10-3 統計数理研究所
E-mail address: kamatani@ism.ac.jp

学習を定める不変量「実対数閾値」の定義と性質

渡辺澄夫（東京工業大学）

1　初めに

実対数閾値は代数幾何学で大切な役割を果たす不変量であって、ベイズ学習の漸近挙動を定めている。本論では、その定義を述べ、数学的な性質を紹介する。

2　定義

d を 1 以上の整数とする。$\mathbb{R}^d$ から $\mathbb{R}$ への非負の解析関数 $f(x)$ が与えらえたとき、$f(x)$ の零点を含む有界な開集合 U についてゼータ関数を

$$\zeta(z) = \int_U f(x)^z dx \quad (z \in \mathbb{C})$$

と定義すると、$\zeta(z)$ は $\mathrm{Re}(z) > 0$ で正則関数であり、複素平面全体に有理型関数として一意に解析接続することができる。また、その極はすべて負の有理数である。最も原点に近い極を $(-\lambda)$ とし、その位数を m とするとき、$\lambda > 0$ を**実対数閾値**と呼び、m を**多重度**と呼ぶ。

注意. 上記の定義で λ と m は有界な開集合 U に一般には依存する。依存するものとして定めたほうが応用上は便利である。依存しないように定めたい場合には U に関して inf を取る。

注意. 任意の解析関数 $x = g(y)$ とそのヤコービ行列式 $|g'(y)|$ を用いて

$$\zeta(z) = \int_{g^{-1}(U)} f(g(y))^z |g'(y)| dy$$

が成り立つことから、実対数閾値は**双有理不変量** であることがわかる。

3　幾何学的な意味

$t \geq 0$ とする。$V(t)$ を有界な開集合 U に含まれる集合 $f(x) \leq t$ の体積とする。すなわち

$$V(t) = \text{集合 } \{x \in U \,;\, f(x) \leq t\} \text{ の体積}$$

このとき $V(0) = 0$ であるが、統計学や多くの分野において $t \to +0$ のときの $V(t) \to +0$ の早さを調べたくなるときがある。それは実対数閾値によりある定数 $c > 0$ が存在して

$$V(t) = c\, t^{\lambda} (-\log t)^{m-1} + \text{小さいオーダー}$$

であることがわかる。すなわち、実対数域値 λ は解析関数 $f(x)$ の零点の近傍の体積次元のような量である。

注意. もしも $f(x)$ の零点が 1 点であり、零点におけるヘッセ行列が可逆ならば、$\lambda = d/2, m = 1$ である。もしも $f(x) = 0$ が特異点を持たない k 次元多様体であれば $\lambda = (d-k)/2, m = 1$ である。

注意. $f(x) = 0$ が特異点を持つとき、λ と m は特異点の幾何学的な性質により定まる。関数 $f(x)$ が座標変換により変数毎の積で表されるとき**正規交差**と呼ばれるが、正規交差であるときには実対数閾値は容易に計算できる。任意の解析関数は**特異点解消写像**によって正規交差に変換できるので、このことを用いて実対数閾値を求めることができる。

4　学習理論への応用

数学的な導出は他書 [1] にあるので、ここでは実対数閾値が結果としてどこに現れるかを紹介する。神経回路網や混合正規分布のように階層的な構造を持つモデルを分析する際に必要になる。

(1) ベイズ統計においてデータを発生した分布と予測分布のカルバック・ライブラ情報量の平均値は**平均汎化誤差**と呼ばれている。データの数を n とするとき平均汎化誤差は λ/n である。

(2) 対数周辺尤度の符号反転を漸近近似することで得られる情報量規準 BIC は、対数尤度関数が 2 次関数で近似できないときには正しくない。BIC における（パラメータ数/2) を λ に置き換えたものが正しいものである。

(3) 情報量規準 AIC の補正項におけるパラメータの個数は一般には実対数閾値ではなく、対数尤度の揺らぎを表す量**特異揺らぎ**である。特異揺らぎは実対数閾値とは異なる双有理不変量であるが、実対数閾値の数理に基づいて定義することができる。

(4) 学習理論に現れる実対数閾値や特異揺らぎは、データを発生した分布と学習モデルの両方に依存するが、どちらもデータだけから推定することが可能であり、情報量規準 WAIC, WBIC が導出されている。

5 具体的な例

5.1 特異点近傍の面積

解析関数 $f(x,y) = (x^2 - y^3)^2$ のときの特異点の近傍を描いたものが図 1 である。

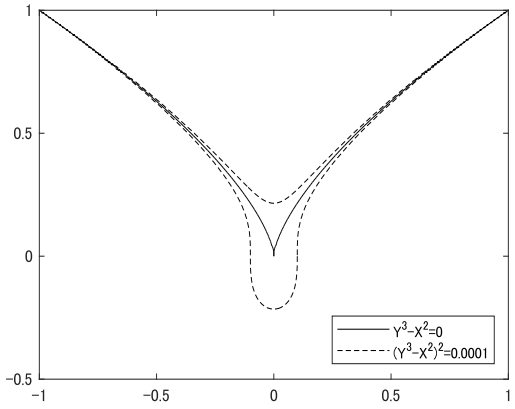


図 1: $f(x,y) = (x^2 - y^3)^2$ として、 $f(x,y) = 0$ と $f(x,y) = 0.0001$ の曲線を描いたもの。代数曲線 $f(x,y) = 0$ は 2 次元的な集合であるが原点は特異点である。特異点の近傍では $f(x,y) \approx 0$ となる集合が広がっていることがわかる。特異点近傍の体積は実対数閾値によって定まる。この例では実対数閾値は 5/12 である。

5.2 統計的行列分解の例

ある行列 A_0B_0 に正規雑音が加わって n 個の独立なデータ行列が得られたとき、 AB を用いて推測を行う問題を考える。ゼータ関数はフロベニウスノルム $\| \quad \|$ を用いて

$$\zeta(z) = \int_U ||AB - A_0B_0||^{2z} dAdB$$

である。このゼータ関数から定まる実対数閾値は [2, 3] で導出されている。

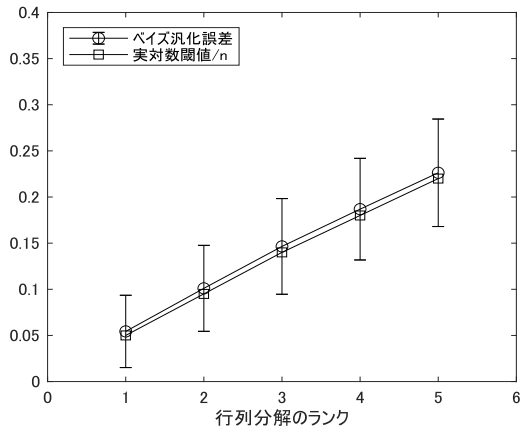


図 2: 統計的行列分解における汎化誤差と実対数閾値から計算される理論値 λ/n の比較。独立な行列の個数が $n = 100$ で A B が $10 \times H$ $H \times 10$ の場合を実験した。漸近領域とは考えられないが、実験値と理論値はよく一致している。

6 結論

双有理不変量である実対数閾値の定義と性質を述べた。

参考文献

[1] S.Watanabe. Mathematical theory of Bayesian statistics. CRC Press, 2018.

[2] M.Aoyagi. Stochastic complexities of reduced rank regression in Bayesian estimation. Neural Networks. Vol.18, pp.923-933, 2005.

[3] N.Hayashi. The exact asymptotic form of Bayesian generalization error in Latent Dirichlet Allocation. arxiv.org/pdf/2008.01304.pdf. 2020.

科学研究費研究集会報告書

深層学習の統計神経力学と情報幾何

甘利俊一　東京大学名誉教授・理化学研究所栄誉研究員

本報告では以下の発表を行った。深層学習は巨大化して素晴らしい能力を誇っている。しかし、これまでは種々の工夫と試行錯誤を重ね、コンピュータの驚くべき性能に依拠して素晴らしい成果を発揮してきたにすぎない。近年、その仕組みと性能を保証する理論が後付けとは言え出始めている。パラメータ数 P が十分大きい時は、非線形コスト関数の局所解はすべて大域解のレベルまで下がるという理論や、汎化誤差に関してパラメータ数をデータ観測数に対してさらに大きく増やしていくと汎化誤差が再び減少に転ずるという、ダブルディセントがそれである。

近年、神経接核理論が現れ、理論が大きく進んだ。これは、P が十分に大きいときには、ランダムに選んだ初期値のごく近傍に最適解があること、したがって深層学習は関数空間で扱えば線形近似できるという、驚くべき内容を含んでいた。しかし理論は難解で、その本質を理解するには新しい視点が必要である。高次元幾何学がこれを与える。これが本報告の趣旨である。

初めに深層学習のモデルを regression の場合に提示する。この場合、入出力関係はパラメータ θ で規定される確率分布で表される。パラメータの空間は Fisher 情報行列をリーマン計量とする多様体をなす。例題を基とする学習は、オンラインでは、確率勾配降下法で記述される。しかし、真の再急降下方向は、Fisher 情報を用いた自然勾配法であるから、Fisher 計量が重要な役を果たす。

パラメータ θ は、それが指定する出力関数 $f(x, \theta)$ に対応するから、関数の空間を考えれば、パラメータの空間は関数の空間に写像される。学習はパラメータの空間で行われるから、これは自然に関数の空間の学習に移せる。しかし、深層学習で実現できる関数、すなわちパラメータ空間の関数空間への写像の像は多様体をなさず、複雑な形状をしている。すなわち、二つのパラメータ θ と θ‘とが同じ関数を表す同値関係が現れ、パラメータ空間をこの同値類で割ったものが、関数空間に表れる。通常のパラメータ学習は、パラメータ空間で起こるが、これを関数の空間で見ると関数の学習になる。

神経接核理論は関数の空間で学習を論じ、結合のパラメータをランダムに選べば、どのランダムに選ばれた初期値に対しても、そのごく近傍に例題学習の最適解があることを示した。考えてみれば、パラメータが平均 0 で独立なガウス分布であればそれは規格化すれば、初期値は半径 1 の球面上に一様に分布している。そのどの点に対しても、その近傍に正解があるというのであれば、正解は球面上のどこにでもあるということになる。これはパラメータ空間の同値関係で球面上の多くの点が同値で結ばれているということである。

まず、単純パーセプトロンモデルを用い、この意味することを考える。この場合パラメータ空間に観測由来の同値類が現れ、これが同値類に対応する零空間を生み出し初期パラメータからなるの球面を切断する。多次元球面上の一様分布を、同値類で割られた実効低次元空間に射影すると、原点周りで分散が極めて小さいガウス分布に縮退する。この正解の一つを、同値類で球面に逆射影すれば、それは球面のほとんどいたるところに写る。神経接核理論の意味するところである。

これを深層回路に拡張する。深層回路では、Fisher 情報行列は最大と最小の大きく異なる固有値を持つから、たとえ線形近似でもこのために収束が遅くなる。ランダム回路の場合に Fisher 情報行列を計算しよう。驚くことに、この場合成分ニューロンごとにブロックに分けると、非対角ブロックの値が 0 に収束する。だから、ブロック対角行列を用いて自然勾配法を適用すれば、行列の逆転に要する手間は極めて大きく減少する。しかし、Fisher 情報行列の逆行列は、この性質を持たない。

観測数が有限なときには、期待値で定義される Fisher 行列でなく、経験分布の期待値で定義する経験 Fisher 行列がある。これはランクが縮退しているが、観測数を無限に増やせば真の Fisher 行列に収束する。そこで、経験 Fisher 情報行列の一般逆行列を用いた経験自然勾配法が出てくる。近年、ニューロンごとの対角ブロックを用いた経験自然勾配学習法が良い性能を示すことが明らかになった。ここではこの場合は新しい損失関数を定義し、Adam の一般化である Super Adam が定義できることを示した。

文献

S. Amari, Any target function exists in a neighborhood of any sufficiently wide rand network: A geometric perspective. Neural Computation, 32, 1413-1447, 2020.

CNNによる同変的写像の普遍近似定理

熊谷 亘[1,2]，三内 顕義[2]

[1]東京大学，[2]理化学研究所

1 Introduction

Deep neural networks have been widely used as models to approximate underlying functions in various machine learning tasks. The expressive power of fully-connected deep neural networks was first mathematically guaranteed by the universal approximation theorem in Cybenko (1989), which states that any continuous function on a compact domain can be approximated with any precision by an appropriate neural network with sufficient width and depth. Beyond the classical result stated above, several types of variants of the universal approximation theorem have also been investigated under different conditions.

Among a wide variety of deep neural networks, convolutional neural networks (CNNs) have achieved impressive performance for real applications. In particular, almost all of state-of-the-art models for image recognition are based on CNNs. These successes are closely related to the property that performing CNNs commute with translation on pixel coordinate. That is, CNNs can conserve symmetry about translation in image data. In general, this kind of property for symmetry is known as the *equivariance*, which is a generalization of the *invariance*. When a data distribution has some symmetry and the task to be solved relates to the symmetry, data processing is desired to be equivariant on the symmetry. In recent years, different types of symmetry have been focused per each task, and it has been proven that CNNs can approximate arbitrary equivariant data processing for specific symmetry. These results are mathematically captured as the universal approximation for equivariant maps and represent the theoretical validity of the use of CNNs.

In order to theoretically correctly handle symmetric structures, we have to carefully consider the structure of data space where data distributions are defined. For example, in image recognition tasks, image data are often supposed to have symmetry for translation. When each image data is acquired, there are finite pixels equipped with an image sensor, and an image data is represented by a finite-dimensional vector in a Euclidean space R^d, where d is the number of pixels. However, we note that the finiteness of pixels stems from the limit of the image sensor and raw scenes behind image data x are thought to be modelled by elements $\tilde{x}$ in $R^{\mathcal{S}}$ with continuous space coordinates $\mathcal{S}$, where $R^{\mathcal{S}}$ is a set of map from $\mathcal{S}$ to R. Then, $\tilde{x}$ is regarded as a functional representation of data x. To appropriately formulate data symmetry, we treat both typical data representation in finite-dimensional settings and functional representation in infinite-dimensional settings in a unified manner.

2 Related Works

Functional Representation and Symmetry. Gordon et al. (2019) point out that, although data symmetry cannot be treated in the set of finite pixels, it can be represented by a group action on continuous space coordinates $\mathcal{S}$. Moreover, the authors also show that functional representations $\tilde{x}$ instead of image data x are suitable for analyzing and considering data processing compatible with data symmetry. As a related study about symmetry-preserving processing, Finzi et al. (2020) propose

group convolution of functional representations and investigate practical computational methods such as discretization and localization.

Universal approximation of group invariant/equivariant networks. Yarotsky (2018) focused on the action of compact groups and provided the network's architectures based on polynomial layers. He also describes a convnet-like model that is a universal approximator of equivariant transformations for the group SE(2). Maron et al. (2019) considered a G-invariant model for an arbitrary subrepresentation G of a symmetry group, using an invariant tensor network. Sannai et al. (2019); Ravanbakhsh (2020) considered a G-equivariant model and proved universal approximation property of them by attributing it to the results of Maron et al. (2019). Thereby, they proved *universal approximation theorem of invariant/equivariant networks* for compact and finite group actions.

3 Our Contributions

Our contributions are summarized as follows. First, we clarify some of the basic properties of equivariant maps. In particular, we show that equivariant maps' generators play an essential role in the analysis of equivariant maps. Second, we formulate neural networks in general settings. Because of this formulation, we can handle finite- and infinite-dimensional settings in a unified manner. Third, we provide a theorem that can convert FNNs to CNNs. Using this theorem, we can induce the proofs of universal approximation theorems for equivariant maps by CNN to those for continuous maps by FNNs. In particular, we obtain the first universal approximation theorems for equivariant maps in infinite-dimensional settings. Fourth, we also provide a novel universal approximation theorem for Lipschitz maps by FNNs in an infinite-dimensional setting. This result enables us to derive universal approximation theorems for equivariant maps for non-compact groups while almost all of the existing works could not handle the action of non-compact groups.

References

[1] G. Cybenko (1989) "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, Vol. 2, pp. 303–314.

[2] M. Finzi, S. Stanton, P. Izmailov, and A. G. Wilson (2020) "Generalizing Convolutional Neural Networks for Equivariance to Lie Groups on Arbitrary Continuous Data," *arXiv preprint arXiv:2002.12880*.

[3] J. Gordon, W. P. Bruinsma, A. Y. Foong, J. Requeima, Y. Dubois, and R. E. Turner (2019) "Convolutional conditional neural processes," *arXiv preprint arXiv:1910.13556*.

[4] H. Maron, E. Fetaya, N. Segol, and Y. Lipman (2019) "On the Universality of Invariant Networks," *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97.

[5] S. Ravanbakhsh (2020) "Universal Equivariant Multilayer Perceptrons," *arXiv preprint arXiv:2002.02912*.

[6] A. Sannai, Y. Takai, and M. Cordonnier (2019) "Universal approximations of permutation invariant/equivariant functions by deep neural networks," *arXiv preprint arXiv:1903.01939*.

[7] D. Yarotsky (2018) "Universal approximations of invariant maps by neural networks," *arXiv preprint arXiv:1804.10306*, URL: https://arxiv.org/abs/1804.10306.

概念は群である―学習における前方誤り訂正の必要性

得丸久文

1．はじめに：デジタル言語の物理層・論理層三段階進化と生物学的基盤

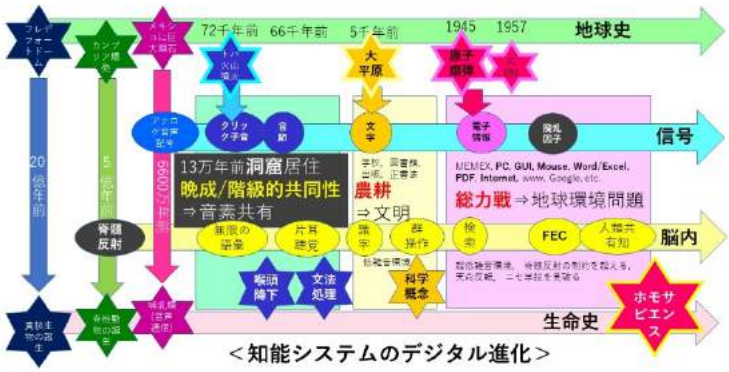
デジタル言語学は，筆者が最古の現生人類洞窟を訪問して始まった学際研究である．6万6千年前に喉頭降下が起きて母音の共鳴が生まれる声道を獲得して，論理成分である音素とモーラを備えた音節を使った言語処理と知能のデジタル第一進化がおきた．第二進化は文字の発明，第三進化は電子的ビットの発明であり，「時空を超える音節」，「検索に応答する音節」を獲得した．

言語と知能を司る生物基盤は，大脳皮質ではなく，5億3000万年前のカンブリア爆発のときに脊椎動物とともに生まれて脊髄反射を司る脳室・脳脊髄液中の免疫細胞ネットワークである．

この生物機構をデジタル信号の特性にうまく適応させて，論理層で第一進化として文法処理が実現し，第二進化は概念，第三進化は前方誤り訂正が期待されている．(表1/図1参照)

表1/図1　知能の三段階デジタル進化　© Kumon Tokumaru 2020

	引き金	誕生	獲得したもの	場所	時期
1	喉頭降下	音節(音素とモーラ)	無限の語彙、 無意識の文法処理	南アフリカ	66千年前
2	農耕と王朝支配	文字(消えない音節) 僧院・大学という低雑音環境	文明(知識の連続的発展) 概念(群として作用する)	大平原	5千年前
3	総力戦と米ソ軍拡競争	電子化(対話する音節)、きわめて大量だが信頼性の保証がない言語情報	PC, インターネット，www, 検索エンジン、**前方誤り訂正(FEC)**	アメリカ	20世紀末



＜知能システムのデジタル進化＞

低雑音環境で生まれた概念の群性と誤り訂正の必要性

大規模複雑データの最大の問題は信頼性の保証がないところにある．ネット上で得られる言語情報は互いに矛盾していて，学際的な統合も行われていないため，そのまま鵜呑みにするととんでもない混沌が生じる．概念は群であるが，そのことが理解されていないうえに，ヒトの真社会性によって歪められて伝えられている．言語のデジタル性を活かして，概念の前方誤り訂正を実施する必要がある．

2.1 概念の誕生と数学的群性

文字が発明され，言語情報が時空を超えて共有され，世代を超えて連続的に発展するようになって文明が生まれ，僧院や修道院など俗世間から隔絶した低雑音環境のなかで概念が生まれた．情報理論では信号対雑音比(S/N)が重要で，雑音(N)が0に漸近するときS/Nは無限大になる．

言葉記号は，脊髄反射の1対1の論理「If A then B」で，言葉と関連する記憶を想起させる．これに対して概念は，論理が数学的群の1対全の論理に進化し，二次元的広がりを意味とする．そのために例外が意味をもち，概念操作の結果も有意な群となる．概念を学習・使用する際に，群性を確認すれば，概念の誤りや歪を正すことができる．ピアジェは群性体という言葉を

用いて，合成性，可逆性，連合性，同一性，同義性の5つの条件を概念に求める．（「知能の心理学」）

2.2 脊髄反射の制約と真社会性由来の重荷

言語処理と知能を脊髄反射回路で処理するために，概念が群であると知らずに，1対1の反射の論理をあてはめる人が多い．脊髄反射回路は，一度刷り込まれると訂正する機構がなく，入力信号の真偽を吟味することもできない．脊髄反射は生命維持のための装置だから，成人して衣食足りて知的好奇心を失う人もいる．デジタル言語を獲得したおかげで知的進化したヒトは，自らを「学ぶサル」と自覚し，死ぬまで真剣に学び続けるのがよい．

また，音素記憶の刷り込みがヒトの真社会性（階級的共同性）に由来することで，共同体の内部秩序が真実・真理に優先し，有無をいわさぬ上意下達の傾向をもたらし，概念を鵜呑みに学習する傾向を生んでいる．さらに，真社会性由来のゼノフォビア（外国人嫌い）が言語の基盤にあるため，敵を攪乱させるためあえて偽概念が広められていると思しき事例がいくつかみられる．

これらが概念を正しく使う上できわめて深刻な障害を生みだしている．

2.3 概念の前方誤り訂正の必要性

前方誤り訂正は，デジタル通信に固有の技術で，受信者が信号相互の論理的整合性をもとに通信回線上で生じた誤りを自力で訂正する．言語情報においては，著者本人が書いたままの真正なデータを確認してから読み始める必要がある．著者生存中に印刷された出版物（著者校正をともなう）は真正とみなせるが，著者没後の出版は確認する必要がある．ほかに第三者による改ざんや隠蔽，ゴーストライター作品や剽窃にも注意する必要がある．

著者の思い違いや限界は，情報源誤りである．論文や科学書は，読者が著者の実験や観察を仮想現実的に追体験できる規則にもとづいて書かれているので，読者は細部にまで気を配り，丁寧に繰り返し熟読し，著者のしたこと，みたこと，考えたことを吟味する．また，諸分野の科学の発展にも配慮して，著者に最新の科学的知識を提供するくらいの気持ちになって，必要ならば結論を改めるくらいでなければならない．

生命体の自己増殖は，一個の有精卵が次々に細胞分裂して，細胞間のネットワークを構築しつつ複雑化する．「個体発生は系統発生をくり返す」というヘッケルの法則，複雑さを増殖するフラクタルな機構がある．科学概念でも，人類的・系統的進化を個体でもくり返すつもりで，できるだけ簡素で単純な事実からはじめて，徐々に概念の複雑次数を高めていくことがのぞましい．

この前方誤り訂正を学習に取り込めば，はじめのうちは不備のある概念が多く手間取るが，一定期間たてば概念は正され，機械学習にも容易に取り込むことができるようになるだろう．

3．動画：学会・研究会に提出した5分・20分の動画

1　言語処理と知能構築を行う脳室内ニューラルネットワークのモバイル仮説（デジタル言語学）　第30回日本神経回路学会(2020年12月)で発表
https://youtu.be/SIUnCj6z_KQ

2　群論理に基づく概念操作 ~ 言葉記号から概念への進化（デジタル言語学）~
(2020年1月25日、信学会HCS研究会）
https://youtu.be/I4aoGi_j2hk

多項ロジットモデルに基づく公的統計データ及び企業データの統計的マッチング・データ融合

総務省統計データ利活用センター・統計数理研究所　高 部　　勲
統計数理研究所　山 下　智 志

1. 統計的マッチングの概要

統計的マッチングは，異なるデータをレコード単位で確率的に結合する技術である（図１）．近年，様々なデータが利用可能になっており，これらを統計的マッチングにより結合できれば，新たに調査やデータ収集等を行うことなく，情報量の多い有用なデータを構築することが可能となる(D'Orazio et al.(2006), Rassler(2002))．

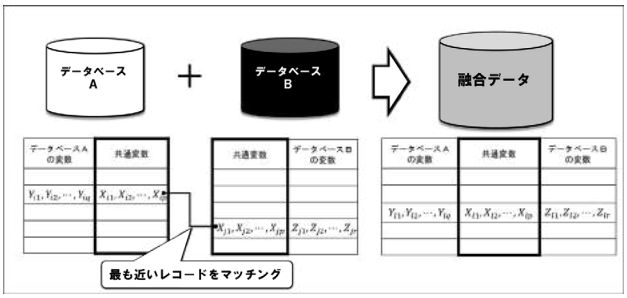


図 1: 統計的マッチングのイメージ

$$d_{ij} = \sum_{p=1}^{P} \beta_p \left| X_{ip} - X_{jp} \right| \quad \text{（距離関数の例：ウエイト付き絶対値距離）}$$

2. 多項ロジットモデルに基づく統計的マッチング

多項ロジットモデルに基づく統計的マッチングの手法は，レコード間の距離を説明変数とする多項ロジットモデルによりマッチングが正しい確率推定し，データの結合を行う方法である（図２）．提案手法により，距離のウエイトを統計的に推定することが可能となる（高部・山下（2018），Takabe and Yamashita (2020)）．距離関数のウエイトは最尤法により推定する。

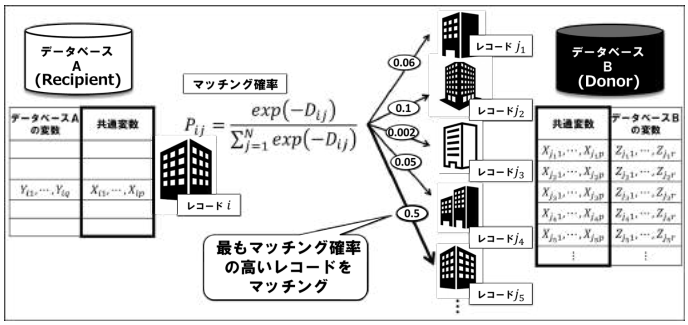


図 2: 多項ロジットモデルに基づく統計的マッチング

$$D_{ij} = \beta_1 \left| X_{i,\text{資本金}} - X_{j,\text{資本金}} \right| + \beta_2 \left| X_{i,\text{従業者数}} - X_{j,\text{従業者数}} \right| + \cdots$$

$$L = \log\left[\prod_{i,j} P_{ij}^{\delta_{ij}} \right] = \sum_{i,j} \delta_{ij} \log\left[P_{ij} \right] \qquad \left(\underset{\beta_1, \beta_2, \cdots, \beta_p}{\operatorname{argmin}} \; L \right)$$

$\delta_{ij} = 1$（レコード i とレコード j が一致），$= 0$（それ以外）

3. データ及び分析方法

提案手法を以下のデータに適用し、企業データの統計的マッチングを行った．

- マッチング元（Recipient）:「帝国データバンク」データ（平成 24 年２月分）
- マッチング先（Donor）:「平成 24 年経済センサス - 活動調査」ミクロデータ
  （※統計法第 33 条による二次的利用の制度に基づき提供を受けたもの。）

3 地域のレコードを完全照合して統合データを作成し、1/3 を学習用データ，2/3 をテスト用データとしてモデルの推定及びマッチングの精度を評価した（マッチング確率が大きい「上位 X 件」のレコードに正解が入っている割合を使用，「X」は、1～100 の範囲で動かす）．最近隣法(Nearest Neighbor Method)を比較対象として、提案手法の性能を評価した。

多項ロジットモデルは適切に推定されており，提案手法は，従来の方法よりもマッチングの精度の観点から優れた性能を示した．

	A県		B県		C県	
従業員数	1.0630	***	1.1215	***	1.0744	***
	(34.021)		(46.898)		(47.554)	
資本金額	0.8024	***	0.8064	***	0.7507	***
	(54.745)		(72.644)		(68.587)	
売上高	0.9288	***	0.7894	***	0.7321	***
	(64.071)		(61.920)		(60.766)	
産業	3.5074	***	3.3997	***	3.4925	***
	(62.438)		(73.641)		(78.185)	
開設年	1.4813	***	1.3427	***	1.4784	***
	(34.944)		(38.697)		(43.628)	
地域（市・郡）	9.4954	***	9.0213	***	8.1163	***
	(18.940)		(22.027)		(37.931)	
経営組織（株式・有限会社）	4.4628	***	3.8687	***	3.7897	***
	(20.164)		(29.407)		(27.880)	
初期対数尤度	-37943		-46662		-52430	
対数尤度	-9350		-15574		-19070	
疑似決定係数	0.7536		0.6662		0.6363	
自由度調整済疑似決定係数	0.7534		0.6661		0.6361	

*** p < 0.001, ** p < 0.01, * p < 0.05, - p < 0.01　( )はｔ値

表 1: 多項ロジットモデルの推定結果

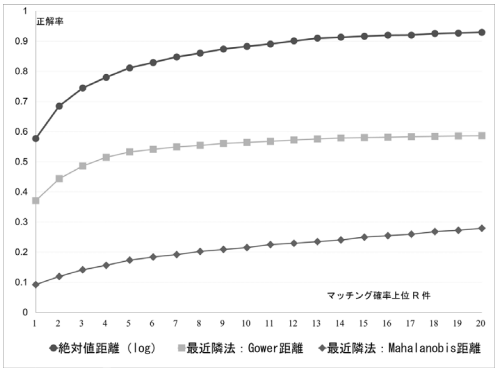


図 3: マッチングの精度比較

4. 提案手法の改善①：ハンガリー法に基づくマッチング精度の改善

マッチング確率に基づいて機械的にマッチングを行った場合，同一のマッチング先レコードに複数のレコードを結びつけてしまう可能性がある．そこで，マッチング確率を重みとした２部グラフのマッチングの問題を考えて，ハンガリー法を適用し，１対１の最適マッチングを行った．精度の改善はそれほど大きくはなかった。

5. 提案手法の改善②：Recipient と Donor を逆転したモデルに基づく手法

通常のモデルと， Recipient 及び Donor の役割を逆転させた形で推定したモデルにより算出したマッチング確率を適当なウエイトで加重平均したものを新たな距離として，マッチングを行った（図 4）．その結果，従来の手法よりも正解率が向上することが確認できた．

更なる精度改善に向けて検討中．

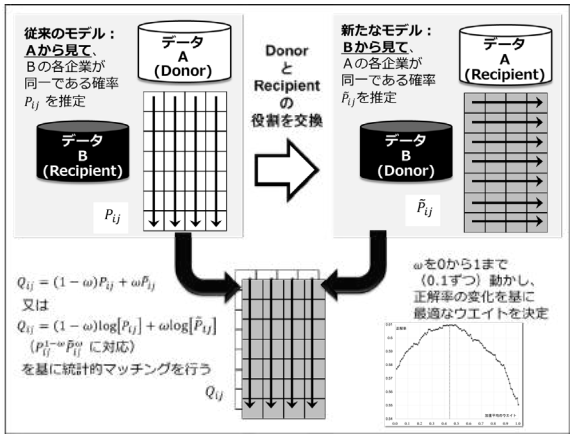


図 4: Donor と Recipient の転置処理

参 考 文 献

[1] 高部勲・山下智志 (2018). 多項ロジットモデルを用いた新たな統計的マッチング手法の提案. 統計学, 115, 1-16.

[2] Takabe, I., and Yamashita, S. (2020). New Statistical Matching Methods Using Multinomial Logistic Regression Model. In I. Tadashi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto & M. Vichi, (Eds.), Advanced Studies in Classification and Data Science (pp. 265-274). Springer.

[3] D'Orazio, M., M. Di Zio & M. Scanu (2006), Statistical Matching: Theory and Practice, Wiley

[4] Rässler, S. (2002), Statistical Matching, Springer

二変数間相互作用を考慮した一般化加法モデルとその効率的な学習

東京大学　上月正貴
東京大学　松島慎

2020 年 12 月 5 日

概要

意思決定への利用や，公平性の観点から，機械学習においては予測の正確さのみならず，予測の解釈性，すなわち，入力の特徴量が予測に与える影響の分かりやすさも同等に重要である．一般化加法モデルは非線形関数の足し合わせで，目標変数と説明変数それぞれの非線形な関係を捉えることが可能なモデルであるが，性能向上に寄与すると考えられる複数変数間の相互作用と目標変数の関係は考慮されていない．本研究では視覚的にとらえることができる二変数間の相互作用をモデルに組み込むことで解釈性を保ちつつその予測が正確な手法とその学習アルゴリズムを提案する．

機械学習アルゴリズムにおいて，一般に，モデルがもつパラメータ数が増えるにつれて予測性能が向上するが，入力 $\boldsymbol{x}$ に対する学習済みモデルの予測 $\tilde{y}$ の解釈性，すなわち予測するうえで重要な入力 $\boldsymbol{x}$ の特徴量が不明瞭になる傾向にある．

本発表で対象とする一般化化加法モデルは，入力を $\boldsymbol{x} \in \mathbb{R}^d$ とすると予測関数が $f_{\mathrm{GAM}}(\boldsymbol{x}) = f_1(x_1)+\cdots+f_p(x_p)+b$，ここで，$x_p$ は $\boldsymbol{x}$ の第 p 要素，f_j は非線形関数で b はバイアス項，である．各特徴量と出力 y の非線形な関係をとらえることで性能向上が期待されると同時に，$f_j(x_j)$ に関しては線形なため各特徴量の予測への影響が解釈しやすい．この特徴から，肺炎診断援助のために一般化加法モデルを利用した研究 [2] や，植生分布に与える気候の影響の応答曲線の形を説明するために一般化加法モデルを利用した研究 [10]，鳥の生息数の時間変化を一般化加法モデルでモデリングした研究 [5] がある．

二変数間の相互作用を考慮した一般化加法モデルには $\mathrm{GA}^2\mathrm{M}$ [7] がある．この手法のモデル関数は $f_{\mathrm{GA^2M}}(\boldsymbol{x}) = f_{\mathrm{GAM}}(\boldsymbol{x}) + \sum_{(j,k)\in\mathcal{I}} f_{j,k}(x_j, x_k)$ である．学習は大きく三段階に分けられ，まず一般化加法モデル $f_{\mathrm{GAM}}(\boldsymbol{x})$ を学習し，次に，残差を小さくすることが期待される相互作用のペアの集合 $\mathcal{I}$ を算出し，最後に，それらを用いて相互作用の一般化加法モデル $\sum_{(j,k)\in\mathcal{I}} f_{j,k}(x_j, x_k)$ を学習する．この手法は，1 変数と 2 変数の扱いが異なり，，出力 y が特徴量間の相互作用のみから生成されている場合に，正しく知識発見できないことが考えられる．

本発表では 1 変数と 2 変数を区別せずに，データの生成過程に関与している特徴量と相互作用のみを用いる一般化加法モデルを構築する．したがって，提案手法（Interaction GAM，IGAM）のモデルは次である：

$$f_{\mathrm{IGAM}}(\boldsymbol{x}) = \sum_{1\le j\le k\le d} f_{j,k}(x_j, x_k) \tag{1}$$

このモデルに対して，その学習を効率化するための正則化項として全変動ノルム [8] を拡張した二変数全変動ノルム $\|g(\cdot,\cdot)\|_{\mathrm{TV}} = \inf\left\{\sum_s |w_s| \;\middle|\; g(x,y) \le \sum_s w_s \mathbb{1}(x'_s \le x \le x''_s)\mathbb{1}(y'_s \le y \le y''_s)\right\}$ を提案する．ここで x'_s, x''_s, y'_s, y''_s は s で決まる値である．さらにこの正則化項を用いた目的関数の最適化は大規模 L_1 正則化問題に帰着することを利用すると，目的関数は $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ を損失関数として

$$P(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 + C\sum_{i=1}^{n} \ell\left(\sum_{s,j,k} w_{s,j,k}\mathbb{1}(x'_{j,s} \le x_{ij} \le x''_{j,s})\mathbb{1}(x'_{k,s} \le x_{ik} \le x''_{k,s}), y_i\right) \tag{2}$$

となる．ここで，x_{ij} は訓練集合の i 番目の入力 $\boldsymbol{x}_i$ の j 番目の要素，$x'_{j,s}, x''_{j,s}, x'_{k,s}, x''_{k,s}$ は s で決まる値である．学習では，まず各特徴量と相互作用の目的関数の評価値の改善期待値 $|\nabla_{w_{s,j,k}}(P(\boldsymbol{w}) - \|\boldsymbol{w}\|_1)| = \frac{\partial}{\partial w_{s,j,k}} \sum_{i=1}^{n} \ell(f_{\mathrm{IGAM}}(\boldsymbol{x}_i), y_i)$ が $\frac{1}{C}$ より大きいものの中で最大の値をとる特徴量を特徴行列 Φ に追加し，入力を Φ，出力を $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ とする L_1 正則化問題（部分問題）を解き，パラメータ $\boldsymbol{w}^{(t)}$ を得るという処理を $|\nabla_{w_{s,j,k}} P(\boldsymbol{w})| < \frac{1}{C}$ となるまで繰り返す．

	次元	データ数	IGAM	GA^{2}M	pyGAM [*1]
人工データ [6]	10	10,000	$\mathbf{5.97 \times 10^{-2}}$	1.24×10^{-1}	6.32×10^{-2}
Boston [*2]	13	506	$\mathbf{2.44 \times 10^{0}}$	3.32×10^{0}	2.88×10^{0}
pyrim [*2]	27	74	$\mathbf{4.75 \times 10^{-2}}$	6.08×10^{-2}	7.67×10^{-2}
bodyfat [*2]	14	252	$\mathbf{1.38 \times 10^{-3}}$	3.71×10^{-3}	7.06×10^{-2}
mpg [*2]	7	392	2.50×10^{0}	$\mathbf{2.30 \times 10^{0}}$	2.32×10^{0}

表 1　二乗平均平方根誤差.

ここで特徴量を選択する問題は，$O(d^2)$ 個の Maximum Rectangle 問題であり，各 Maximum Rectangle 問題の計算量は $O(n^2 \log n)$ [4] であるため，各反復の計算量は O(部分問題の最適化の計算量) + $O(d^2 n^2 \log n)$ である．

複数の回帰問題のデータセットに対する提案手法の二乗平均平方根誤差が表 1 である．

mpg データセット以外の次元が 10 以上のものでは幅広いデータ数で優位であることが確認された．この要因として，特徴量の数 d に対して相互作用の数が $\frac{d(d-1)}{2}$ あることから d が大きくなるにつれて相互作用のデータの生成への寄与が特徴量の寄与より大きくなることが考えられる．そのため，次元が小さい mpg では，まず一般化加法モデルを学習する GA^{2}M，特徴量と特徴量の組み合わせのフィッティングを行う pyGAM がより正確なモデルを学習できたと考えられる．しかし，pyrim のように次元が大きくなると GA^{2}M は 1 次元の特徴量で構成し，固定した一般化加法モデルが真のデータ生成から乖離しているため残差を小さくするよう相互作用を用いて特徴量の組を入力とするモデルを追加したとしてもその性能は提案手法の IGAM ほどに改善されないと考えられる．

本発表では，機械学習アルゴリズムが有する予測性能と予測の解釈性のトレードオフを超えることを目標に，視覚的に把握できる二変数間の相互作用を扱える一般化加法モデルとその学習を効率化するための正則化項である二変数全変動ノルムを提案し，その性能の有効性を回帰問題で示した．

参考文献

[1] Libsvm data: Regression. `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html`. Accessed: 2020-12-01.

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721 – 1730, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] C. Cortés, J.M. Díaz-Báñez, P. Pérez-Lantero, C. Seara, J. Urrutia, and I. Ventura. Bichromatic separability with two boxes: A general approach. *Journal of Algorithms*, 64(2):79 – 88, 2009.

[5] Rachel M Fewster, Stephen T Buckland, Gavin M Siriwardena, Stephen R Baillie, and Jeremy D Wilson. Analysis of population trends for farmland birds using generalized additive models. *Ecology*, 81(7):1970–1984, 2000.

[6] Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580, 2004.

[7] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623 – 631, New York, NY, USA, 2013. Association for Computing Machinery.

[8] Shin Matsushima. Statistical learnability of generalized additive models based on total variation regularization, 2018.

[9] Daniel Servén and Charlie Brummitt. pygam: Generalized additive models in python, March 2018.

[10] Thomas W Yee and Neil D Mitchell. Generalized additive models in plant ecology. *Journal of vegetation science*, 2(5):587–602, 1991.

[*1] pyGAM [9] ライブラリを用いて 2 つの特徴量の相互作用をテンソル積でフィッティングするモデル

[*2] LIBSVM [3] の LIBSVM Data: Regression [1] にあるものを利用した．

深層展開に基づく反復アルゴリズムの収束加速
—理論的解釈とその応用—

名古屋工業大学
和田山 正

1　本稿について

本稿は，科研シンポジウム「機械学習・統計学・最適化の数理と AI 技術への展開」における著者の講演の概要を報告するものである。

2　本講演の背景

近年，無線通信技術に関する研究分野においては，5G システムの先にある “Beyond 5G”, または “6G” に関する研究の方向性を模索する動きが活発化しつつあり，各種国際会議においても 6G までの今後 10 年の流れを見越した講演が増えつつある。このような流れの中で，産業界 (基地局を製造するメーカー，サービスを提供するキャリアなど)・関連国際学会ともに来る 6G システムの基軸となる技術として AI/ML を重要視しているようである。この理由としては，6G での達成目標（5G を上回る高速大容量, 低遅延, 多ユーザ収容) を充足させることは，現状の技術の延長線だけでは厳しい見通しがあり，それを乗り越えるためのひとつの，そして最も有望な方策として AI/ML 技術に強い期待が持たれているためだと考えられる。

無線ネットワークは，無線物理層，ネットワーク層といった階層構造を持ち，そのすべての階層で AI/ML の利活用が検討されている。無線物理層は，情報の符号化・変調・復調などを扱う階層であり，その中で特に重要な技術として MIMO(Multiple-Input Multiple-Output) アンテナ技術とそれに関わる信号処理技術が挙げられる。MIMO 復調問題は複素線形観測に関する逆問題と考えることができ，圧縮センシングやスパース正則化などとも深く関係する推定問題と見ることができる。高い推定精度を持ち，高速に動作をするアルゴリズムが求められており，古典的な問題ながら現在に至るまで活発に研究が進められている重要な研究課題となっている。近年，この問題に対して深層展開 (Deep Unfolding) を適用したデータ駆動型アルゴリズムが注目されている。著者のグループでは，近接勾配法に対して深層展開を適用することで，優れたデータ駆動的アルゴリズム (スパース信号再現 [4], MIMO 復調 [5]) が構成できることを示してきた。

3　深層展開

深層学習技術は，深層ニューラルネットワークの学習に利用できるだけではなく，入出力を伴う “微分可能な反復型アルゴリズム” の内部パラメータ最適化に適用可能である。この考え方は微分可能プログラミング (differential programming) として知られ，最近，制御・物理・常微分方程式/偏微分方程式の数値的解法などの諸分野でその応用が活発に検討されている。深層展開は，この微分可能プログラミングの考え方を一般の反

復型アルゴリズムに適用したものと見ることができる。従前から知られている優れた反復型アルゴリズム (例えば近接勾配法) を基礎として，その内部に学習可能パラメータを埋め込むことで，データに基づく学習可能性を持つ柔軟な派生アルゴリズムをデータ駆動的に構成できる。

本講演の前半では，深層展開の考え方と特徴，そして適用事例を紹介する [2]。深層展開の代表的なメリットとして，1) 多くの反復型アルゴリズムにおいて収束の加速が生じる点，2) 学習されたパラメータについてある程度の解釈が可能な点が挙げられる。本講演の後半では，収束加速の要因となる学習後パラメータに関する解釈の試み (チェビシェフステップに基づくスペクトル半径制御) とそこで得られれたチェビシェフステップの応用について概説する [1][2]。チェビシェフステップの理論では，チェビシェフ多項式の根であるチェビシェフ根が収束速度を支配するスペクトル半径の制御について重要な働きをしている。深層展開における学習プロセスにより見いだされたステップサイズはチェビシェフステップに非常に近いものであることが実験的に確認されており，この結果は，深層展開によるパラメータ学習の妥当性を理論面から強く支持している。

4　文献案内

本講演の内容に関する詳細な議論と実験結果は文献 [1] に含まれている。深層展開に関するチュートリアル的解説とチェビシェフステップに関する平易な解説は [2] にまとめられている。Github レポジトリ [3] では無線通信研究者向けに行ったチュートリアル講演の資料と深層展開関連の実装コード (PyTorch) を公開している。本グループの深層展開に関する代表的な論文として，スパース信号再現アルゴリズム TISTA に関する文献 [4], MIMO 復調問題に関する深層展開の適用事例 [5] を挙げておく。応用的観点からの深層展開に関する最近の包括的レビューについては文献 [6] を参照されたい。

謝辞

本グループにおける深層展開に関する研究は同僚の高邉賢史氏と共同で進めている。日頃より議論いただく同氏に感謝したい。本研究の一部は科研費基盤 (B)19H02138 の補助を受けている。

参考文献

[1] S. Takabe and T. Wadayama, “Convergence acceleration via Chebyshev step: plausible interpretation of deep-unfolded gradient descent,” arXiv:2010.13335, 2020.

[2] 和田山正, 高邉賢史, “(解説論文) 深層展開に基づく信号処理アルゴリズムの設計 –収束加速とその理論的解釈–, ” 電子情報通信学会 Fundamentals Review, 14 巻 1 号 p. 60-72, 2020.

[3] 和田山正, “無線通信のための深層学習の基礎,” MIKA2019 チュートリアル講演資料, `https://github.com/wadayama/MIKA2019`, 2019.

[4] D. Ito, S. Takabe, and T. Wadayama, “Trainable ISTA for sparse signal recovery,” IEEE Trans. Signal Processing, vol. 67, no. 12, pp. 3113-3125, Jun., 2019.

[5] S.Takabe, et al., “Trainable projected gradient detector for massive overloaded MIMO channels: data-driven tuning approach”, IEEE Access, July 2019.

[6] A. Balatsoukas-Stimming and C. Studer “Deep unfolding for communications systems: a survey and some new directions,” `https://arxiv.org/abs/1906.05774`, 2019.

身体動作と音楽の同調性によるストリートダンスの評価手法の提案

上野　寛幸 [1], 鈴木　量三朗 [2], 豊泉　洋 [1]
[1] 早稲田大学大学院基幹理工学研究科, [2] 有限会社シンビー
e-mail：tchaikovsky_1107@akane.waseda.jp

1　はじめに

2024 年パリ五輪からストリートダンスの 1 つであるブレイクダンスが五輪種目に追加される．しかし, ストリートダンスの採点基準はあいまいであり, 他のスポーツと同様に公平なジャッジが可能なのかという点が不安視されている．そこで, 本研究では AI によるストリートダンスの自動採点アルゴリズムを提案し, 個人の好みに左右されない公平なジャッジを目指す．

2014 年に NEC ソリューションイノベータ株式会社が, カラオケの採点と同様にお手本との一致率から採点するダンス採点装置の特許を取得した [1]．しかしダンスは技と音楽の整合性やオリジナリティがより重要であるため, お手本との一致率に加えてこれらの点を考慮したアルゴリズムがダンスの採点には望ましい．

2　ダンス

ダンス $\boldsymbol{D}$ は時空間内の身体の位置と音楽の軌道によって数学的に記述できる [2]．以下では, $\boldsymbol{D} = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}_{t=0}^{T}$ によって定義する．$\boldsymbol{x}_t$ は時刻 t における音楽の特徴量ベクトルで , $\boldsymbol{y}_t$ は時刻 t における姿勢を表すベクトルからなるダンス動作である．この $\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ の列を照らし合わせることで, ダンス動作と音楽の調和を評価することができる．本研究ではダンスジャンルの中でも, 音楽の特徴とダンス動作の関連が顕著な pop ダンスについて扱う．

3　pop ダンス

pop ダンスの動きは主に, 以下のアイソレ, ストップ, ヒットの 3 要素からなる．

アイソレ: 身体のある 1 部位のみを独立して動かす．

ストップ: 急に身体の動きを止める．

ヒット: 筋肉を一瞬緊張させて身体を震わせる．

特にストップとヒットは強い音に合わせて用いられ, 音楽の特徴量と比較しやすい．ここではストップに着目したダンスの評価を考える．

4　ストップの定義

AI にダンスを評価させる上で, まず前述のストップを認識させる必要がある．そこで本研究では, 骨格推定アルゴリズムの一つである OpenPose[3] を用いて身体の動き $\boldsymbol{y}_t$ を数値化する．OpenPose は動画を入力すると, 1 フレーム毎に人体の 25 部位の座標を出力する．座標が出力される各部位を keypoint と呼び, それぞれ図 1 のようにナンバリングされる．

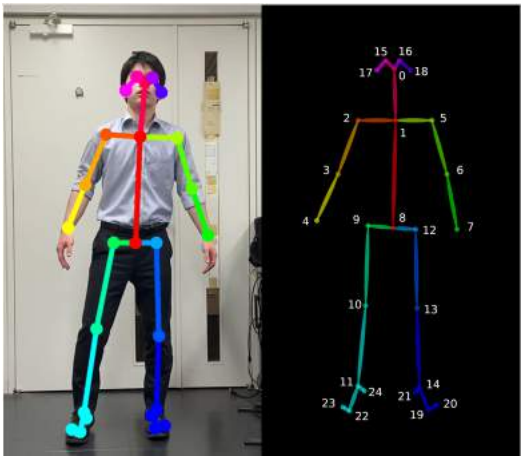


図 1. openpose による身体部位の計測

ここで, i 番の keypoint の t フレーム目での姿勢をベクトル $\boldsymbol{y}_i(t)$ とおき, i 番の keypoint が t 番目のフレームと $t+1$ 番目のフレーム間に移動した距離を $\Delta y_i(t)$ と定義する．これは単位時間当たりの移動距離なので, 2 フレーム間の平均の速さと考えられる．以下, これを用いてストップを定義する．

$$\Delta y_i(t-1) \geq v_m \tag{1}$$

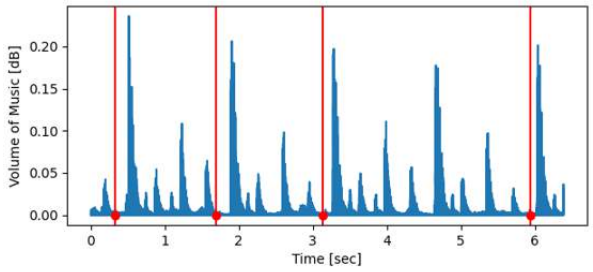
$$0 \leq \Delta y_i(t) \leq v_s \tag{2}$$

(1) 式より $t-1, t$ フレーム間での速さは十分大きく, (2) 式より次のフレーム間では速さは十分小さいつまり静止しているとみなせることから, ストップタイムでストップが生じていると考えられる．ここでの v_m, v_s は入力動画の fps, 解像度, 人間の映り込む大きさによって異なる閾値である．

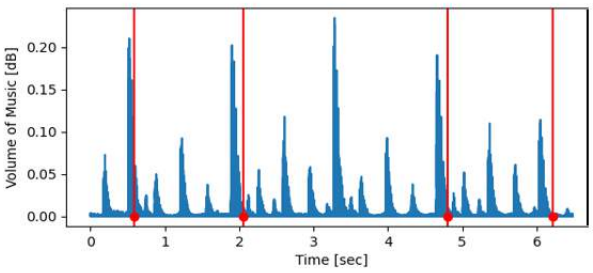
5　ストップタイムの妥当性

pop ダンスにおいてストップやヒットを強い音に合わせて行うことを「オンビート」, 強い音より少し早く行うことを「早取り」, 少し遅く行うことを「遅取り」と呼ぶ [4].

図 2 は同じ条件下で早取りを行ったダンス動画 (a) と遅取りを行ったダンス動画 (b) の音量 x_t をそれぞれ時系列グラフで表し, 身体の動き $\boldsymbol{y}_t$ から (1), (2) 式で検出したストップタイムを記したものである. (a) では音量が極大値をとる少し前にストップタイムが現れ, (b) では音量が極大値をとる少し後にストップタイムが現れる. このように音楽 $\boldsymbol{x}_t$ と身体動作 $\boldsymbol{y}_t$, 特にストップの関係が明確化できる.



(a) 早取り (低評価)



(b) 遅取り (高評価)

図 2. (a) 早取りと (b) 遅取りにおける音量とストップタイム (赤いライン:$v_m = 6.55$, $v_s = 5$) の関係

6　結論

本研究では, 骨格推定アルゴリズム OpenPose を用いて身体の動きを数値化し, 単位時間当たりの移動距離によってストップタイムを定義した. ダンス動画から計算されたストップタイムが, 動画内で行ったストップの時刻と一致していることから, ストップタイムによって pop ダンスの 1 要素であるストップを数学的に定義できることが明らかになった.

図 2 より, 音量が極大値をとる時刻とストップタイムの前後関係から音の早取り, 遅取りも数学的に定義できる. 早取りは音を聞けていないという理由から悪い評価を受ける [4]. 音量が極大値をとる時刻とストップタイムの関係からダンスを評価できると考えられる.

ダンス動画からダンスを採点する手法を開発するために, 本研究で得られた考察から今後の課題をまとめる.

- 本研究ではヒューリスティックに決めていた (1), (2) 式の閾値 v_m, v_s を学習器によって決定する.
- 音量が極大値をとる時刻とストップタイムを入力とするダンスの評価関数を学習器を用いて作成する.
- オリジナリティに関する評価法を考案する.

参考文献

[1] “身体動作採点装置, ダンス採点装置, カラオケ装置及びゲーム装置”, https://patents.google.com/patent/WO2014162787A1/ja.

[2] Ruozi Huang, Huang Hu, WeiWu, Kei Sawada, Mi Zhang, Dance Revolution: Long Sequence Dance Generation with Music via Curriculum Learning, 2020.

[3] “openpose/output.md at master, CMU-Perceptual-Computing-Lab/openpose, GitHub”, https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md.

[4] “ダンスの先生つれづれブログ”, https://dancenoteacher.blog/dance-instructor/i-summarized-the-reasons-for-dance-early-and-late-and-tips-on-how-to-fix-it.

高次元データにおける異常値の検出について

京都大学・情報学研究科　　中山　優吾（Yugo Nakayama）
筑波大学・数理物質系　　矢田　和善（Kazuyoshi Yata）
筑波大学・数理物質系　　青嶋　誠（Makoto Aoshima）

1　はじめに

本講演では，高次元小標本データに対する異常値の検出を考えた．本講演では，Yata and Aoshima (2019) や Nakayama et al. (2020) が提案している主成分スコアによるクラスタリングを応用して，高次元小標本データに対して異常値検出を提案した．

2 つの d 次元分布を Π_1, Π_2 と名付け，それぞれ平均 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ と，共分散行列 $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ をもつと仮定する．ここで，$\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ とする．それらの混合分布から n (≥ 2) 個のデータを無作為に抽出し，データ行列を $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ とする．ただし，$n_i = \#\{j | \boldsymbol{x}_j \in \Pi_i,\ j = 1, ..., n\}$ とおく．ここで，$\#A$ は集合 A の要素の個数とし，$n_i \geq 1$ とする．$\boldsymbol{I}_n$ を n 次の単位行列，$\mathbf{1}_n = (1, ..., 1)^T$ として，$\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$ とする．カーネル関数 $k(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$ を (j, j') 成分にもつグラム行列 $\boldsymbol{K}$ を考え，その中心化グラム行列を $\boldsymbol{K}_0 = \boldsymbol{P}_n\boldsymbol{K}\boldsymbol{P}_n$ とし，$\boldsymbol{K}_0$ の固有値分解を

$$\boldsymbol{K}_0 = \sum_{i=1}^{n-1} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^T \quad (\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, ..., \hat{u}_{in})^T, \quad \|\hat{\boldsymbol{u}}_i\|^2 = 1)$$

とする．$\boldsymbol{x}_j$ の (基準化した) 第 i 主成分スコアを $s_{ij} = \sqrt{n}\hat{u}_{ij}$ とおく．

2　主成分スコアにおける高次元一致性

本節では，ガウシアンカーネル

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\gamma) \ \ (\gamma > 0)$$

を用いたカーネル PCA を考える．$\kappa_\mu = \exp(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/\gamma)$，$\kappa_\Sigma = \exp(-|\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|/\gamma)$，$\Delta_\kappa = 1 + \kappa_\Sigma^2 - 2\kappa_\mu\kappa_\Sigma$ とおく．このとき，次の定理が成り立つ．

定理 1 (Nakayama et al., 2020)**.** 正則条件と次を仮定する．

(A-i) $\limsup_{d\to\infty} \dfrac{(1-\kappa_\Sigma^2)}{\Delta_\kappa n_1} < 1 \quad when\ n_2 \geq 2.$

このとき，$d \to \infty$ で次が成り立つ．

$$s_{1j} = \begin{cases} \sqrt{n_2/n_1} + o_P(1) & when\ j = 1, ..., n_1, \\ -\sqrt{n_1/n_2} + o_P(1) & when\ j = n_1 + 1, ..., n. \end{cases} \tag{1}$$

定理 1 から Nakayama et al. (2020) は第 1 主成分スコアの符号から高次元データのクラスタリングを提案した．(A-i) が成り立たないとき，次の定理が成り立つ．

定理 2 (Nakayama et al., 2020)．正則条件と次を仮定する．

(A-ii) $\liminf_{d\to\infty} \dfrac{(1-\kappa_\Sigma^2)}{\Delta_\kappa n_1} \geq 1 \quad when\ n_2 \geq 2.$

このとき，$d \to \infty$ で次が成り立つ．

$$s_{n_2 j} = \begin{cases} \sqrt{n_2/n_1} + o_P(1) & when\ j = 1, ..., n_1, \\ -\sqrt{n_1/n_2} + o_P(1) & when\ j = n_1 + 1, ..., n. \end{cases};\ and \tag{2}$$

$$\sum_{j=n_1+1}^{n} \frac{s_{ij}^2}{n} = 1 + o_P(1) \quad and \quad s_{ij} = o_P(1) \quad for\ j = 1, ..., n_1;\ i = 1, ..., n_2 - 1$$

異常値が 1 つと想定できる場合は，$n_1 = 1$ として考えることで (1) と (2) から異常値の検出が可能である．本講演では，この主成分スコアの一致性を異常値検出に応用した．当日は，一般のカーネル関数の場合と異常値が複数ある場合についても言及し，主成分スコアの符号に基づき異常値検出が可能であることを理論的に示し，その性能を数値実験と実データ解析を用いて確認した．

参考文献

[1] Nakayama, Y., Yata, K. and Aoshima, M. (2020). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings, submitted.

[2] Yata, K. and Aoshima, M. (2019). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scand. J. Stat.*, 47: 899–921.

γ-divergence に基づく変数選択について

伊森 晋平 [1,2], 橋本 真太郎 [1]

1: 広島大学 大学院先進理工系科学研究科

2: 理研 AIP 数理統計学チーム

1 はじめに

変数選択は統計学において重要な問題のひとつであり，様々な手法やその理論的性質が古くから研究されている．特に近年は変数の数が多い高次元データに対する変数選択が注目されている．一方で，観測データに対して外れ値の混入が懸念される場合，外れ値に対して頑健な変数選択手法が望ましい．外れ値に対して頑健な手法として，γ-divergence を用いた推定方法 (Eguchi and Fujisawa, 2008) が知られており，近年盛んに研究されている．特に Kawashima and Fujisawa (2017) では L_1 正則化と γ-divergence を組み合わせた手法が提案されており，スパース推定を通した外れ値に頑健な変数選択ができると考えられる．本研究では高次元データにおける有効な変数選択手法の一つである greedy-type アルゴリズムについて，γ-divergence と組み合わせることで外れ値に頑健な変数選択手法の構築を試みる．

2 Greedy-type アルゴリズム

本研究で扱うモデルと変数選択手法について説明する．$y \in \mathbb{R}$ を目的変数，$\boldsymbol{x} \in \mathbb{R}^p$ を説明変数とし，真の回帰モデルとして，$\boldsymbol{x}$ を与えた下での y の条件付き確率密度関数を

$$q(y|\boldsymbol{x}) = (1-\varepsilon)\phi(y; \mu_0(\boldsymbol{x}), \sigma_0^2) + \varepsilon\delta(y|\boldsymbol{x})$$

とする．ただし，$\phi(y; \mu_0(\boldsymbol{x}), \sigma_0^2)$ はターゲットの分布であって，平均が $\mu_0(\boldsymbol{x})$, 分散が σ_0^2 である正規分布の確率密度関数を表す．また，$\delta(y|\boldsymbol{x})$ は外れ値の分布であり，$\varepsilon \in [0, 1)$ は外れ値の比率である．一方で，候補モデル $M \subset \{1, \ldots, p\}$ を

$$p_M(y|\boldsymbol{x}) = \phi(y; \mu_M(\boldsymbol{x}), \sigma_0^2)$$

とする．ただし $\mu_M(\boldsymbol{x}) = \alpha_M + \boldsymbol{\beta}_M^\top \boldsymbol{x}_M$, $\boldsymbol{x}_M$ は M に基づく $\boldsymbol{x}$ の部分ベクトルであり，α_M と $\boldsymbol{\beta}_M$ は回帰係数である．簡単のため，分散パラメータ σ_0^2 は既知としている．このとき，Kawashima and Fujisawa (2017) より，γ-divergence に基づく α_M と $\boldsymbol{\beta}_M$ の推定量は次の関数を最小化することで得られる：

$$-\frac{1}{\gamma}\log\iint p_M(y|\boldsymbol{x})^\gamma q(y|\boldsymbol{x})q(\boldsymbol{x})dyd\boldsymbol{x} + \frac{1}{1+\gamma}\log\iint p_M(y|\boldsymbol{x})^{1+\gamma}q(\boldsymbol{x})dyd\boldsymbol{x}.$$

実際にはパラメータ α_M と $\boldsymbol{\beta}_M$ に依存する部分のみを経験推定した

$$-\log\left\{\frac{1}{n}\sum_{t=1}^{n}\exp\left\{-\frac{\gamma(y_t-\alpha_M-\boldsymbol{\beta}_M^\top\boldsymbol{x}_{t,M})^2}{2\sigma_0^2}\right\}\right\}$$

の最小化によってパラメータの推定量 $\hat{\alpha}_M$ と $\hat{\boldsymbol{\beta}}_M$ を求める．ただし $\{(y_t,\boldsymbol{x}_t)\}_{t=1}^n$ は $(y,\boldsymbol{x})$ と同じ分布に従う独立標本であり，$\boldsymbol{x}_{t,M}$ は $\boldsymbol{x}_t$ の M に基づく部分ベクトルを表す．

p が大きく，候補モデルが多い場合，全ての候補モデルに対して当てはまりの良さを計算することは実用的でない．そこで，orthogonal greedy algorithm (Ing and Lai, 2011) を参考に以下のルールでモデル M に変数を加えることを考える．

$$\hat{j}_M=\arg\max_{j\notin M}\max_{\alpha,\beta_j}\log\left\{\frac{1}{n}\sum_{t=1}^{n}\exp\left\{-\frac{\gamma(u_t(M)-\alpha-\beta_j x_{t,j})^2}{2\sigma_0^2}\right\}\right\}.$$

ただし $u_t(M)=y_t-\hat{\alpha}_M-\boldsymbol{x}_{t,M}^\top\hat{\boldsymbol{\beta}}_M$ であり，初期値は $M=\emptyset$, $u_t(\emptyset)=y_t$ と定める．

参考文献

1. Fujisawa, H. & Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis, 99(9), 2053–2081.

2. Kawashima, T. & Fujisawa, H. (2017). Robust and sparse regression via γ-divergence. Entropy, 19(11), 608.

3. Ing, C. K. & Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. Statistica Sinica, 1473–1513.

事前分布強調型の情報量規準の開発とその WAIC との比較

二宮 嘉行
統計数理研究所 数理・推論研究系

1　序論

汎化性能の高い Bayes 予測分布を与えるための情報量規準として，真の分布との Kullback-Leibler ダイバージェンスを評価することで作られたのが DIC であるが，統計的漸近理論になるべく頼らないようにしようという意図があるためか，特殊な評価法に依存していた．その中，統計的漸近理論に基づき，理論的にも実践的にも優れた情報量規準として Watanabe (2010) により提案されたのが WAIC である．MCMC で事後分布を評価すれば同時に与えられるため実装は容易であり，また特異モデルにおいても保証されているため，今後ますます用いられていくことが予想される現代的な情報量規準である．いま，

$$\tilde{z}, z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} f(\cdot \mid \boldsymbol{\theta}); \qquad \boldsymbol{\theta} \sim \pi(\cdot \mid \boldsymbol{\eta})$$

という設定において $\boldsymbol{z} = (z_1, \ldots, z_n)$，$\boldsymbol{\theta}$ の事後分布に基づく期待値を $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\eta}}(\cdot)$ とし，特に $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\eta}}\{f(z \mid \boldsymbol{\theta})\}$ を $f(z \mid \boldsymbol{z}, \boldsymbol{\eta})$ と記すことにすれば，WAIC は汎化損失 $\mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \boldsymbol{z}, \boldsymbol{\eta})\}$ と漸近等価な統計量として

$$-\frac{1}{n}\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\eta}) + \frac{1}{n}\sum_{i=1}^{n} (\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\eta}}[\{\log f(z_i \mid \boldsymbol{\theta})\}^2] - \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{\eta}}\{\log f(z_i \mid \boldsymbol{\theta})\}^2)$$

で与えられる．そして，この値を最小化することで，例えば事前分布を選択する，つまり $\boldsymbol{\eta}$ を与えることができる．次節では，小西 (2000) で提案されている GIC をベースとした情報量規準を導く．この両者をリッジ回帰などの最もシンプルな設定で比較することが，ここでの目的である．

2　別タイプの情報量規準の開発

本節では，$\boldsymbol{\theta}$ の事前分布は $\pi(\cdot \mid \boldsymbol{\eta})^n$ と書けるものとして話を進める．推定された予測分布と真の分布との Kullback-Leibler ダイバージェンスを小さくすることを念頭に，最尤推定でいうところの平均対数尤度に相当する $\mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \boldsymbol{z}, \hat{\boldsymbol{\eta}})\}$ を情報量規準の源とする．ただし，$\hat{\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\eta}} \sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}, \boldsymbol{\eta})$ である．最大対数尤度に相当するものを最初に考える推定量とし，

$$\frac{1}{n}\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}; \hat{\boldsymbol{\eta}}) - \mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \boldsymbol{z}; \hat{\boldsymbol{\eta}})\}$$

を評価することを考える．$\hat{\boldsymbol{\eta}}$ の収束先を $\boldsymbol{\eta}^*$ とし，通常の TIC と同様の評価法を用いると，これは

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \log f(z_i \mid \boldsymbol{z}; \boldsymbol{\eta}^*) - \mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \boldsymbol{z}; \boldsymbol{\eta}^*)\} \\
&+ \mathrm{tr}\left[\left\{-\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'} \log f(z_i \mid \boldsymbol{z}; \hat{\boldsymbol{\eta}})\right\}^{-1} \left\{\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\eta}} \log f(z_i \mid \boldsymbol{z}; \hat{\boldsymbol{\eta}}) \frac{\partial}{\partial \boldsymbol{\eta}'} \log f(z_i \mid \boldsymbol{z}; \hat{\boldsymbol{\eta}})\right\}\right] \qquad (1)
\end{aligned}$$

と評価される．この第一項目と第二項目を少し細かく見ていく．まとめて考えるため，第一項目の z_i や第二項目の $\tilde{z}$ をとりあえず z と書くことにすれば，和や期待値の中身は

$$\log f(z \mid \boldsymbol{z}; \boldsymbol{\eta}^*) = \log \int f(z \mid \boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\eta}^*)^n \mathrm{d}\boldsymbol{\theta} - \log \int f(\boldsymbol{z} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{\eta}^*)^n \mathrm{d}\boldsymbol{\theta}$$

と書ける．この二つの積分それぞれに対して高次の Laplace 近似を適用すると，$q_{z,\boldsymbol{\eta}}(\boldsymbol{\theta}) = n^{-1}\{\log f(z \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta} \mid \boldsymbol{\eta})\}$，$q_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} q_{z_i,\boldsymbol{\eta}}(\boldsymbol{\theta})$，$\hat{\boldsymbol{\theta}}_{z,\boldsymbol{z},\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\{q_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta}) + n^{-1}\log f(z \mid \boldsymbol{\theta})\}$，$J_{z,\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta}) = -\partial^2\{q_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta}) + n^{-1}\log f(z \mid \boldsymbol{\theta})\}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$，$\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, q_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta})$，$J_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta}) = -\partial^2 q_{\boldsymbol{z},\boldsymbol{\eta}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ として

$$
\begin{aligned}
&nq_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{z,\boldsymbol{z},\boldsymbol{\eta}^*}) - \log(|J_{z,\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{z,\boldsymbol{z},\boldsymbol{\eta}^*})|)^{1/2} + \log f(z \mid \hat{\boldsymbol{\theta}}_{z,\boldsymbol{z},\boldsymbol{\eta}^*}) \\
&- nq_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*}) + \log(|J_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})|)^{1/2} + \mathrm{O_P}(n^{-2})
\end{aligned}
$$

となる．これと $\hat{\boldsymbol{\theta}}_{z,\boldsymbol{z},\boldsymbol{\eta}^*} - \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*} = \mathrm{O_P}(n^{-1})$ であることより

$$
\log f(z \mid \boldsymbol{z}; \boldsymbol{\eta}^*) = \log f(z \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*}) + \mathrm{O_P}(n^{-1})
$$

が成り立ち，これより小西 (2000) は，ベイズ予測と罰則付き最尤推定のバイアス補正項は同じになるものとし，ベイズ予測に対する GIC を提案した．一方，実はそのバイアス補正項は $\mathrm{O_P}(n^{-1})$ であり，つまりその GIC は漸近論で保証されたものとはいえないため，$\mathrm{O_P}(n^{-1})$ の項まで評価すると，それは

$$
\begin{aligned}
&a_{z,\boldsymbol{z},\boldsymbol{\eta}^*} \\
&= \frac{1}{2n}\mathrm{tr}\left[J_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})^{-1}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(z \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*}) - \frac{\partial}{\partial\boldsymbol{\theta}}\log f(z \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\frac{\partial}{\partial\boldsymbol{\theta}'}\log|J_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})|\right\}\right]
\end{aligned}
$$

と書ける．これの z に z_i を入れて期待値をとったものは，$\tilde{z}$ を入れて期待値をとったものと漸近的には同等になり，(1) の第一項目と第二項目を合わせたものは

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log f(z_i \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*}) - \mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\} \\
&= \frac{1}{n}\mathrm{tr}\left[J_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}q_{z_i,\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\frac{\partial}{\partial\boldsymbol{\theta}'}q_{z_i,\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\right\}\right]\{1 + \mathrm{o_P}(1)\}
\end{aligned}
$$

と評価される．つまり，これの主項を用いる GIC は，結果として妥当であることがわかる．ここでは，そこに現れる $\boldsymbol{\eta}^*$ にその一致推定量である $\hat{\boldsymbol{\eta}}$ を代入して用いることにする．この $\mathrm{O_P}(n^{-1})$ の評価より，

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log f(z_i \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*}) - \mathrm{E}_{\tilde{z}}\{\log f(\tilde{z} \mid \boldsymbol{z}; \boldsymbol{\eta}^*)\} \\
&= \frac{1}{n}\left(\mathrm{tr}\left[J_{\boldsymbol{z},\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}q_{z_i,\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\frac{\partial}{\partial\boldsymbol{\theta}'}q_{z_i,\boldsymbol{\eta}^*}(\hat{\boldsymbol{\theta}}_{\boldsymbol{z},\boldsymbol{\eta}^*})\right\}\right] + \sum_{i=1}^{n}a_{z_i,\boldsymbol{z},\boldsymbol{\eta}^*}\right)\{1 + \mathrm{o_P}(1)\}
\end{aligned}
$$

が得られ，これを利用すれば，対数尤度にベイズ推定量を代入したものをベースとした情報量規準も構成できる．

参考文献

[1] 小西貞則. (2000). 統計的モデリングと情報量規準構成の理論. 数学, 52, 128–141.

[2] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11, 3571–3594.