数理解析研究所講究録
第1980巻 2016年 220-233

学習理論と学習係数

日本大学理工学部数学科　青柳美輝, 岡田憲相

Miki Aoyagi and Kensuke Okada
Department of Mathematics, College of Science & Technology, Nihon University,
aoyagi.miki@math.cst.nihon-u.ac.jp; cskn15001@g.nihon-u.ac.jp;

Abstract

統計学分野の学習理論における目的は，真の分布から発生する多量のデータセットから，そのデータを発している情報源の真の分布を再生・推測することである．generalization errorは，推定された分布と真の分布とのエントロピーに関する誤差を表している．従って，与えられたデータから求められるtraining errorから，generalization errorを推定することは重要である．機械学習における文字認識，画像認識，音声認識，遺伝子解析などでは，データの情報源の確率分布は，正規分布に従うような単純なものではない．それらの機械学習においては，特異学習モデルと呼ばれる複雑な確率分布を表現できる階層構造・内部構造をもつ神経回路網，混合正規分布や縮小ランクなどが利用されている．

モデル選択法である，"Widely applicable information criterion" (WAIC), は，Akaike information criterion (AIC) を一般化したもので，特異学習モデルにも適用可能である．ベイズ推測における学習係数は，特異学習モデルの学習効率をあたえるものでWAIC法の重要な役割を果たしている．学習係数は，代数幾何の分野では，カルバック関数のlog canonical thresholdとして知られている．この論文では，近年得られた学習係数に関する結果とそれらを得るために必要な定理を紹介する．

1 はじめに

Y を滑らかな多様体，Z を Y の closed subscheme, Y 上の 0 でない解析函数 f に対して，log canonical threshold は解析的に複素数体上

$$c_Z(Y,f) = \sup\{c : |f|^{-c} \text{ is locally } L^2 \text{ in a neighborhood } Z\},$$

実数体上

$$c_Z(Y,f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ in a neighborhood } Z\},$$

と定義される．$c_Z(Y,f)$ は f に関するゼータ関数の最大の極でもある．代数幾何・代数解析では主に代数閉体上での log canonical threshold の研究が行われている．また，低次元での研究が主である (Kollár[13], Mustata[15]). 例えば，複素体上での log canonical threshold は，代数解析における f の Bernstein-Sato 多項式 $b(s)$ の最大の根であることが知られている．

一方，学習理論における学習係数は，ある情報量の実数体上の log canonical threshold とその位数で与えられる．従ってそのまま複素体上の定理を学習係数に適用することができない．例えば，複素数体上の log canonical threshold は 1 より小さいが，実数体上ではそうとは限らない．学習理論における情報量の log canonical threshold は，ほとんどが 1 より大きい．ある関数族に関しては，実数体上の log canonical threshold の方が多くの情報を持っていることが知られている．このように学習係数を求めることは，数学的観点からも興味のある問題である．

この論文では，学習理論においてよく用いられる混合正規分布，三層ニューラルネットワーク，混合二項分布のベイズ推測に関する学習効率を与える Vandermonde matrix singularities と，線形回帰モデルに ARD 法を適用した場合における log canonical threshold に関する結果を述べる．

log canonical threshold は広中の特異点解消定理により，原理的には有限の手続きにより求められるが，具体的に求めるのは難しいとされている．計算機で代数計算により行う方式も提案されているが，Vandermonde matrix singularities は，パラメータを含んでいるため，確定された多項式の特異点解消よりも高度な面を含んでいる．更なる困難な問題点として，特異点が孤立していない・ニュートン図形が退化している等があげられる ([12], [25]).

2 WAIC 法と学習理論

$x \in \mathbf{R}^N$ を確率変数，$q(x)$ を真の確率密度関数とし，$q(x)$ に従う n 個の独立なサンプルを $x^n := \{x_i\}_{i=1}^n$ とする．学習モデル $p(x|w)$ とその事前分布 $\psi(w)$ が与えられているものとする．ここで，パラメータ空間は $W(\ni w) \subset \mathbf{R}^d$ とする．学習理論の目的は，データセット x^n から $p(x|w)$ を用いて，真の分布 $q(x)$ を再生・推定することである．

ベイズ学習では，事後確率 $p(w|x^n)$ を

$$p(w|x^n) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^n p(x_i|w)$$

で定義する．ここで Z_n は正規化定数

$$Z_n = \int_W \psi(w)\prod_{i=1}^n p(x_i|w)\mathrm{d}w$$

である．

ここで，定数 β をもちいて，

$$E_w[f(w)] = \frac{\int dw f(x)\psi(w)\prod_{i=1}^n p(x_i|w)^\beta}{\int dw \psi(w)\prod_{i=1}^n p(x_i|w)^\beta}$$

と定義する．β は，inverse temperature とよばれ，通常 $\beta = 1$ である．

これよりベイズ推測を $p(x|X^n) = E_w[p(x|w)]$ と定義する．

確率密度関数 $p(x), q(x)$ に対して，カルバック距離 $K(q||p)$ を

$$K(q||p) = \sum_x q(x)\log\frac{q(x)}{p(x)},$$

経験カルバック距離 $K_n(q||p)$

$$K_n(q||p) = \frac{1}{n}\sum_{i=1}^n \log\frac{q(x_i)}{p(x_i)}$$

で定義する．$K(p||q)$ は，常に非負関数で，$q(x) = p(x)$ の時に限り $K(q||p) = 0$ となる．

ここで，4 個の誤差，Bayesian generalization error, B_g, Bayesian training error, B_t, Gibbs generalization error, G_g, and Gibbs training error, G_t を次のように定義する．

$$B_g = K(q(x)||E_w[p(x|w)]))$$

$$B_t = K_n(q(x)|E_w[p(x_i|w)])$$

$$G_g = E_w[K(q(x)||p(x|w))]$$

$$G_t = E_w[K_n(q(x)||p(x|w))]$$

Bayesian generalization error は，真の分布を予測分布がどのくらい近似しているかを表したもので最も重要な値である．

$\lambda \in \mathbf{Q}$ を learning coefficient(学習係数), $\nu \in \mathbf{R}$ を singular fluctuation とする. これらは, birational invariant な数である. 正則モデルでは, パラメータの次元を d とすると, $\lambda = \nu = d/2$ が成立する.

論文 [22, 23, 24] において以下の関係が示されている:

$$E[B_g] = \frac{\lambda + \nu\beta - \nu}{n\beta} + o(\frac{1}{n})$$

$$E[B_t] = \frac{\lambda - \nu\beta - \nu}{n\beta} + o(\frac{1}{n})$$

$$E[G_g] = \frac{\lambda + \nu\beta}{n\beta} + o(\frac{1}{n})$$

$$E[G_t] = \frac{\lambda - \nu\beta}{n\beta} + o(\frac{1}{n})$$

これらは, 特異点解消および超関数を用いて示される.

従って,

$$E[B_g] = E[B_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

および

$$E[G_g] = E[G_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

を得る.

真の分布の期待値を除いたものを

$$BL_g = -\sum_x q(x) \log E_w[p(x|w)]$$

$$BL_t = -\frac{1}{n}\sum_{i=1}^{n} \log E_w[p(x_i|w)]$$

$$GL_g = -E_w[\sum_x q(x) \log p(x|w)]$$

$$GL_t = -E_w[\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|w)]$$

とする. このとき,

$$E[BL_g] = E[BL_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

$$E[GL_g] = E[GL_t] + 2\beta(E[G_t] - E[B_t]) + o(\frac{1}{n})$$

となる.

この2式が WAIC 法である. Bayesian および Gibbs generalization error が, 真の分布の情報を用いずに, Bayesian および Gibbs training error から推測できることを示している. Training error は, 観測データ x_i および学習モデル p を用いて求められる. 実際の応用や実験では, 通常, 真の分布が不明で, 観測データのみが得られている. したがって観測データからの真の分布の推定やモデル選択に WAIC 法は有効である.

Bayesian および Gibbs training error の差は,

$$n\beta(E[G_t] - E[B_t]) \to \nu$$

となる.

学習係数 λ は $K(q(x)||p(x|w))$ の log canonical threshold であり，θ をその位数とすると，それらの値を用いて，値 ν は，理論的に次で与えられる．

$$\nu = \frac{1}{2} E_\xi \frac{\int_0^\infty dt \sum_{u^*} \int du \xi(u) t^{\lambda-1/2} e^{-\beta t + \beta\sqrt{t}\xi(u)}}{\int_0^\infty dt \sum_{u^*} \int du t^{\lambda-1/2} e^{-\beta t + \beta\sqrt{t}\xi(u)}}, \tag{1}$$

ここで，$\xi(u)$ は，特異点解消した空間上で定義された経験過程で，平均0分散2のガウス分布となる確率変数である．$\sum_{u^*}$ は，λ および θ が得られる局所座標の和である．

今まで，特異モデルの学習係数は，数例においてその上限が得られていたにすぎず情報科学において長い間未解明であったが，近年，縮小ランクモデルの場合 [8] や，出力層が1個である三層ニューラルネットワークの場合 [7, 1]，1次元の混合正規分布の場合 [3]，Restricted Boltzmann machine [5] の場合の学習係数について解明された．

論文 [6, 4] では，帰納的に行う特異点解消の中心となる多様体の適当な選択によって，Vandermonde matrix 型特異点の学習係数のバウンドが得られている．

また，論文 [20, 21, 27] においては naive Bayesian networks や directed tree models with hidden variables における学習係数が得られている．

得られた結果は複雑な学習モデルの選択やハイパーパラメータの設計において，周辺尤度の値の理論値を与えるため，MCMC 法の精度を解明することや [16, 17]，精度の改良法，モデル選択解析に応用されている [10, 11]．

3 Log canonical threshold

* を付加した記号 a^*, b^*, w^* などは定数を表すものとする．

定義 1 $\mathbb{C}^d$ または $\mathbb{R}^d$ における w^* の十分小さな近傍を U，f を U 上の0でない正則関数または実解析関数とする．$\psi(w)$ をコンパクトサポートを持つ C^∞ 関数で，$\psi(w^*) \neq 0$ を満たすものとする．このとき，f の w^* および ψ に関する log canonical threshold を，$\mathbb{C}^d$ 上では，

$$c_{w^*}(f, \psi) = \sup\{c : |f|^{-c} \text{ is locally } L^2 \text{ in a neighborhood of } w^*\}$$

$\mathbb{R}^d$ 上では，

$$c_{w^*}(f, \psi) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ in a neighborhood of } w^*\}$$

と定義する．また，$\theta_{w^*}(f, \psi)$ をその位数とする．

$\psi(w^*) \neq 0$ ならば，特に，$c_{w^*}(f) = c_{w^*}(f, \psi)$ および $\theta_{w^*}(f) = \theta_{w^*}(f, \psi)$ とおく．これらの値は ψ に依存しないからである．

次の定理1は，複素数体において，関数を超平面に制限したときの log canonical threshold 値に関する定理である．

定理 1 ([19], [13]) $f(w_1, \cdots, w_d, w_{d+1})$ を原点の近傍における**正則関数**とする．g を $g = f|_{w_{d+1}=0}$ とおく．すなわち，f を $w_{d+1} = 0$ に制限した関数を g とする (または，H を超曲面として，f を H に制限した関数を $g = f_H$ とする)．

このとき，$c_0(g) \leq c_0(f)$.

この定理は，実解析関数では成り立たない．たとえば，反例として，例1があげられる．

例 1

$$f(w_1, w_2, w_3, w_4, w_5, w_6) = (w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6 - 1)^2. \tag{2}$$

とおく．このとき，$c_{(0,0,0,0,0,1)}(f) = 1/2$, であるが，$c_0(f(w_1, w_2, w_3, w_4, w_5, 1)) = 5/4$ となる．

しかし，斉次多項式の場合は，以下の様に成立する．

定理 2 ([4]) $f_1(w_1,\ldots,w_d)$, ..., $f_m(w_1,\ldots,w_d)$ を 次数 n_i の $w_1,\cdots,w_d$ に対する斉次多項式とする. $f_1'(w_2,\ldots,w_d) = f_1(1,w_2,\ldots,w_d)$, ..., $f_m'(w_2,\ldots,w_d) = f_m(1,w_2,\ldots,w_d)$ とおく. $w_1^* \neq 0$ であれば

$$c_{(w_1^*,\cdots,w_d^*)}(f_1^2+\cdots+f_m^2) = c_{(w_2^*/w_1^*,\cdots,w_d^*/w_1^*)}(f_1'^2+\cdots+f_m'^2).$$

また, 斉次多項式の場合は, 以下の定理が成り立つ.

定理 3 ([4]) $f_1(w_1,\ldots,w_d)$, ..., $f_m(w_1,\ldots,w_d)$ を $w_1,\cdots,w_j (j \leq d)$ に対する次数 n_i の斉次多項式とする. さらに, ψ を C^∞ 関数で, $\psi_{(0,\cdots,0,w_{j+1}^*,\cdots,w_d^*)} \geq \psi_{(w_1^*,\cdots,w_d^*)}$ および $(0,\cdots,0,w_{j+1}^*,\cdots,w_d^*)$ の近傍で $w_1,\cdots,w_j$ に関して, 斉次であるとする.

このとき,

$$c_{(0,\cdots,0,w_{j+1}^*,\cdots,w_d^*)}(f_1^2+\cdots+f_m^2,\psi) \leq c_{(w_1^*,\cdots,w_j^*,w_{j+1}^*,\cdots,w_d^*)}(f_1^2+\cdots+f_m^2,\psi)$$

が成り立つ.

一般に $w_0 \in \mathbb{R}^d$ が

$$f_i(w_0) = \frac{\partial f_i}{\partial w_j}(w_0) = 0, 1 \leq i \leq m, 1 \leq j \leq d$$

を満たしたとしても

$$c_{w_0}(f_1^2+\cdots+f_m^2,\psi) \leq c_{w^*}(f_1^2+\cdots+f_m^2,\psi)$$

が成り立つとは限らない.

例 2 $f_1 = x(x-1)^2$, $f_2 = (y^2+(x-1)^2)((y-1)^6+x)$, $f_3 = (z^2+(x-1)^2)((z-1)^6+x)$ とする. このとき, $x=1, y=0, z=0$ の時に限り, $f_1 = f_2 = f_3 = \frac{\partial f_1}{\partial x} = \frac{\partial f_2}{\partial y} = \frac{\partial f_2}{\partial x} = \frac{\partial f_3}{\partial z} = \frac{\partial f_3}{\partial x} = 0$ であるが, $c_{(1,0,0)}(f_1^2+f_2^2+f_3^2) = 1/4+1/4+1/4 > c_{(0,1,1)}(f_1^2+f_2^2+f_3^2) = 1/2+1/12+1/12$.

4 Vandermonde matrix 型特異点の Log canonical threshold

補題 4 ([2, 3, 14]) U を $w^* \in \mathbb{R}^d$ の近傍, $\mathbf{J}$ を U で定義された実解析関数 $f_1,\ldots,f_n$ で生成されるイデアルとする.
(1) $g_1^2+\ldots+g_m^2 \leq f_1^2+\cdots+f_n^2$ ならば $c_{w^*}(g_1^2+\cdots+g_m^2) \leq c_{w^*}(f_1^2+\cdots+f_n^2)$.
(2) $g_1,\ldots,g_m \in \mathbf{J}$ ならば $c_{w^*}(g_1^2+\cdots+g_m^2) \leq c_{w^*}(f_1^2+\cdots+f_n^2)$. 特に, $g_1,\ldots,g_m$ が $\mathbf{J}$ の生成元ならば $c_{w^*}(f_1^2+\cdots+f_n^2) = c_{w^*}(g_1^2+\cdots+g_m^2)$.

定義 2 w^* の近傍 U で定義された実解析関数 $f_1,\cdots,f_m$ から生成されるイデアルを $\mathbf{J}$ とする. このとき, $c_{w^*}(\mathbf{J}) = c_{w^*}(f_1^2+\cdots+f_m^2)$ とする.

この定義は, Lemma 4 より矛盾なく定義できる.

定義 3 $Q \in \mathbf{N}$ を固定する.
$b_1^* = \cdots = b_{i-1}^* = 0$, $b_i^* \neq 0$ のとき, $\gamma_i = \begin{cases} 1 & Q \text{ が奇数} \\ |b_i^*|/b_i^* & Q \text{ が偶数} \end{cases}$ に対して,
$[b_1^*, b_2^*, \cdots, b_N^*]_Q = \gamma_i(0,\cdots,0,b_i^*,\cdots,b_N^*)$ と定義する.

定義 4 $Q \in \mathbf{N}$ を固定する.

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1H} & a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a_{21} & \cdots & a_{2H} & a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & \vdots & & & \vdots & \\ a_{M1} & \cdots & a_{MH} & a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix}, I=(\ell_1,\ldots,\ell_N)\in {\mathbf{N}_{+0}}^N$$

$$B_I=(\prod_{j=1}^N {b^*_{1j}}^{\ell_j}, \prod_{j=1}^N {b^*_{2j}}^{\ell_j}, \cdots, \prod_{j=1}^N {b^*_{Hj}}^{\ell_j}, \prod_{j=1}^N {b^*_{H+1,j}}^{\ell_j}, \cdots, \prod_{j=1}^N {b^*_{H+r,j}}^{\ell_j})^t$$

$$B=(B_I)_{\ell_1+\cdots+\ell_N=Qn+1, 0\leq n\leq H+r-1}$$

$$=(B_{(1,0,\cdots,0)}, B_{(0,1,\cdots,0)}, \cdots, B_{(0,0,\cdots,1)}, B_{(1+Q,0,\cdots,0)}, \cdots)$$

とする (t は行列の転置を表す).

a_{ki}, b_{ij} $(1\leq k\leq M, 1\leq i\leq H, 1\leq j\leq N)$ は，定数 a^*_{ki}, b^*_{ij} の近傍で定義された変数とする.

$\mathbf{J}$ を AB のすべての成分から生成されるイデアルとする.

$\mathbf{J}$ で定められる特異点を Vandermonde matrix 型特異点とよぶ．簡単のため，$1\leq j\leq r$ に対して，

$$(a^*_{1,H+j}, a^*_{2,H+j}, \cdots, a^*_{M,H+j})^t\neq 0, (b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N})\neq 0$$

および $j\neq j'$ に対して，

$$[b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N}]_Q\neq [b^*_{H+j',1}, b^*_{H+j',2}, \cdots, b^*_{H+j',N}]_Q$$

を仮定する.

この論文では，次のように定義する.

$$A_{M,H}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} \\ a_{21} & a_{22} & \cdots & a_{2H} \\ & & \vdots & \\ a_{M1} & a_{M2} & \cdots & a_{MH} \end{pmatrix}, B_{H,N,I}=\begin{pmatrix} \prod_{j=1}^N {b_{1j}}^{\ell_j} \\ \prod_{j=1}^N {b_{2j}}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N {b_{Hj}}^{\ell_j} \end{pmatrix}$$

$B^{(Q)}_{H,N}=(B_{H,N,I})_{\ell_1+\ldots+\ell_N=Qn+1, 0\leq n\leq H-1}$.

さらに $\mathbf{a}^*=\begin{pmatrix} a^*_{1,H+1} \\ \vdots \\ a^*_{M,H+1} \end{pmatrix}$ および

$$(A_{M,H}, \mathbf{a}^*)=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a^*_{1,H+1} \\ a_{21} & a_{22} & \cdots & a_{2H} & a^*_{2,H+1} \\ & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a^*_{M,H+1} \end{pmatrix}$$

としておく.

次の定理は $c_0(||A_{M,H}B^{(Q)}_{H,N}||^2)$ および $c_0(||(A_{M,H-1}, \mathbf{a}^*)B^{(Q)}_{H,N}||^2)$ の値がわかれば，すべての Vandermonde matrix 型特異点の log canonical threshold がわかることを示している.

定理 5 ([3]) U を

$$w^*=\{a^*_{ki}, b^*_{ij}\}_{1\leq k\leq M, 1\leq i\leq H, 1\leq j\leq N}$$

の近傍とし，変数 $w=\{a_{ki}, b_{ij}\}_{1\leq k\leq M, 1\leq i\leq H, 1\leq j\leq N}$ は U 内に値を取るとする.

$(b^{**}_{01}, b^{**}_{02}, \cdots, b^{**}_{0N}) = (0, \ldots, 0)$ とおく.
ここで

$$[b^{*}_{i1}, b^{*}_{i2}, \cdots, b^{*}_{iN}]_Q \neq 0, \ for\ i = 1, \ldots, H + r$$

の中で，異なるベクトルを $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N})$, ..., $(b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ とする．すなわち，

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \cdots, b^{**}_{r'N})\ ;\ [b^{*}_{i1}, \cdots, b^{*}_{iN}]_Q \neq 0, i = 1, \ldots, H + r\}.$$

このとき r' は，一意的に定まり，仮定より $r' \geq r$ である．
$1 \leq i \leq r$ に対して，$(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^{*}_{H+i,1}, \cdots, b^{*}_{H+i,N}]_Q$, とする．

$$[b^{*}_{i1}, \cdots, b^{*}_{iN}]_Q = \begin{cases} 0, & 1 \leq i \leq H_0 \\ (b^{**}_{11}, \cdots, b^{**}_{1N}), & H_0 + 1 \leq i \leq H_0 + H_1, \\ (b^{**}_{21}, \cdots, b^{**}_{2N}), & H_0 + H_1 + 1 \leq i \leq H_0 + H_1 + H_2, \\ \vdots & \\ (b^{**}_{r'1}, \cdots, b^{**}_{r'N}), & H_0 + \cdots + H_{r'-1} + 1 \leq i \leq H_0 + \cdots + H_{r'}, \end{cases}$$

および $H_0 + \cdots + H_{r'} = H$ としておく．
このとき

$$c_{w^*}(||AB||^2) = \frac{Mr'}{2} + c_{w_1^{(0)*}}(||A_{M,H_0} B^{(Q)}_{H_0,N}||^2)$$

$$+ \sum_{\alpha=1}^{r} c_{w_1^{(\alpha)*}}(||(A_{M,H_\alpha - 1}, \mathbf{a}^{(\alpha)*}) B^{(1)}_{H_\alpha,N}||^2) + \sum_{\alpha=r+1}^{r'} c_{w_1^{(\alpha)*}}(||A_{M,H_\alpha - 1} B^{(1)}_{H_\alpha - 1,N}||^2).$$

ここで，$w_1^{(0)*} = \{a^{*}_{k,i}, 0\}_{1 \leq i \leq H_0}$,

$w_1^{(\alpha)*} = \{a^{*}_{k,H_0 + \cdots + H_{\alpha-1} + i}, 0\}_{2 \leq i \leq H_\alpha}$ および $\mathbf{a}^{(\alpha)*} = \begin{pmatrix} a^{*}_{1,H+\alpha} \\ \vdots \\ a^{*}_{M,H+\alpha} \end{pmatrix}$ *for* $\alpha \geq 1$.

以下の定理は，学習係数の上界を与える．

定理 6 ([4])
bound_1

$$= \min\{\frac{(H - i + 1)N + d_i(s) + d'_i(s) + d''_i(s)}{2(count(i, s, k(s)) - 1)Q + 2} : 1 \leq i \leq s, 1 \leq k(1), \cdots, k(s) \leq N, 1 \leq s \leq H\}$$

ここで
$count(i, s, j) = \#\{i_1 : i \leq i_1 \leq s, k(i_1) = j\}$, $C(i, s) = \#\{count(i, s, j) = 0, 1 \leq j \leq N\}$,

$$\begin{aligned} d_i(s) &= (N - 1)Q \sum_{s_1 = i}^{s} (count(i, s_1, k(s_1)) - 1) \\ d'_i(s) &= M(i - 1)\{(count(i, s, k(s)) - 1)Q + 1\} \\ &\quad + QM \sum_{\substack{s_1 = i, \\ count(i,s,k(s)) > count(i,s_1,k(s_1))}}^{s-1} (count(i, s, k(s)) - count(i, s_1, k(s_1))), \end{aligned}$$

$$d''_i(s) = \begin{cases} 0, & if\ count(i, s, k(s)) = 1, \\ (H - s)\{C(i, s)Q + (N - 1)Q(count(i, s, k(s)) - 2)\}, & \\ \quad if\ count(i, s, k(s)) \geq 2, N - 1 \leq M, & \\ (H - s)\{C(i, s)Q + MQ(count(i, s, k(s)) - 2)\}, & \\ \quad if\ count(i, s, k(s)) \geq 2, C(i, s) \leq M < N - 1, & \\ (H - s)\{MQ(count(i, s, k(s)) - 1)\}, & \\ \quad if\ count(i, s, k(s)) \geq 2, M \leq C(i, s). & \end{cases}$$

とする。

さらに，$\mathrm{bound}_2 = \dfrac{NH + \sum_{i=0}^{k'-1} MQ(k'-i)\langle {N+Qi \atop N-1} \rangle}{2+2Qk'}$ とする．

ここで $k' = \max\{i \in \mathbb{Z}; NH \geq M\sum_{i'=0}^{i-1}(1+Qi')\langle {N+Qi' \atop N-1} \rangle\}$ である．

また

$$\mathrm{bound}_3 = \frac{MH}{2}$$

とする．

このとき

$$c_0(||A_{M,H}B_{H,N}^{(Q)}||^2) \leq \min\{\mathrm{bound}_1, \mathrm{bound}_2, \mathrm{bound}_3\}$$
$$c_0(||(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q)}||^2) \leq \min\{\mathrm{bound}_1, \mathrm{bound}_2\}$$

次に，$H = 1, 2, 3$ の場合の真の値を与える．$\lambda = c_0(\|A_{M,H}B_{H,N}^{(Q)}\|^2)$, とし，$\theta$ をその位数とする．また，$\lambda' = c_0(\|(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q)}\|^2)$, および θ' をその位数とする．

定理 7 **Case 1** $H = 1$.

1. $\lambda = \min\{\frac{M}{2}, \frac{N}{2}\}, \theta = \begin{cases} 1, & \text{if } M \neq N, \\ 2, & \text{if } M = N, \end{cases}$
2. $\lambda' = \frac{N}{2}, \theta' = 1.$

Case 2 $H = 2$.

1. $M > N+1$ ならば $\lambda = \lambda' = N,\ \theta = \theta' = 1.$
2. $M = N+1$ ならば $\lambda = \lambda' = N,\ \theta = \theta' = 2.$
3. $M = N$ ならば $\lambda = \lambda' = \frac{2N+Q(2N-1)}{2(Q+1)},\ \theta = \theta' = 1.$
4. $M \leq N-1$ ならば $\lambda = M,\ \theta = 1.$
5. $N-Q+1 \leqq M \leq N-1$ ならば $\lambda' = \frac{2N+Q(2N-1)}{2(Q+1)},\ \theta' = 1.$
6. $M = N-Q$ ならば $\lambda' = \frac{N+M}{2},\ \theta' = 2.$
7. $M \leq N-Q-1$ ならば $\lambda' = \frac{N+M}{2},\ \theta' = 1.$

Case 3 $H = 3$.

1. $M > N+2$ ならば $\lambda = \lambda' = \frac{3N}{2},\ \theta = \theta' = 1.$
2. $M = N+2$ ならば $\lambda = \lambda' = \frac{3N}{2},\ \theta = \theta' = 2.$
3. $M = N+1$ ならば $\lambda = \lambda' = \frac{3N+(3N-1)Q}{2(Q+1)},\ \theta = \theta' = 1.$
4. $M = N$ ならば $\lambda = \lambda' = \frac{3N+(3N-2)Q}{2(Q+1)},\ \theta = \theta' = 2.$
5. $M = N-1$ ならば $\begin{cases} \lambda = \frac{3-Q+3M(Q+1)}{2(Q+1)},\ \theta = 1 \textit{ for } Q > 3, \\ \lambda = \frac{3M}{2},\ \theta = 2 \textit{ for } Q = 3, \\ \lambda = \frac{3M}{2},\ \theta = 1 \textit{ for } Q < 3. \end{cases}$

6. $M < N-1$ ならば $\lambda = \frac{3M}{2}$, $\theta = 1$.

7. $M = N - S$ $(S = 1, 2, \cdots)$ ならば

$$\begin{cases} \lambda' = \frac{S(3+Q)-2Q+3M(Q+1)}{2(Q+1)}, & \theta' = 1 \ for \ Q > S, \\ \lambda' = \frac{2M+N}{2}, & \theta' = 2 \ for \ Q = S, \\ \lambda' = \frac{2M+N}{2}, & \theta' = 1 \ for \ Q < S. \end{cases}$$

$N = 1$ の場合は, 次のような結果が得られている.

定理 8 ([2]) $N = 1$ のとき $\lambda = \lambda' = \frac{MQk(k+1)+2H}{4(1+kQ)}$. ここで、$k = \max\{i \in \mathbb{Z}; 2H \geq M(i(i-1)Q + 2i)\}$ である.

$$\theta = \begin{cases} 1, & if \ 2H > M(k(k-1)Q + 2k) \\ 2, & if \ 2H = M(k(k-1)Q + 2k) \end{cases}$$

$$\theta' = \begin{cases} 1, & if \ M = H = 1 \\ 1, & if \ 2H > M(k(k-1)Q + 2k) \\ 2, & if \ 2H = M(k(k-1)Q + 2k), H > 1 \end{cases}$$

5 ARD 法

次に, ARD 法を適用した場合の学習係数について考察する. 特に, 脳活動の計測における MEG (Magnetoencephalography) 線形モデルに ARD 法を適用した場合の学習係数を求める.

5.1 MEG

脳活動の計測においては fMRI, EEG, MEG といった様々な計測方法が行われている. MEG とは脳磁図・脳磁計と呼ばれ, 脳内神経活動により発生する磁場を観測し神経活動を観測する装置である [26].

MEG の特徴として,

1. 非侵襲的である.

2. 高い時間分解能を有する.

3. 電流源推定のための逆問題を解く必要がある.

ということが挙げられる. 逆問題を解くにあたって, 計測に用いるセンサ数に対し, 電流源が膨大な数であるため不良設定問題になってしまう. そのため不要な電流源を削減することが必要になるが, この対策として考えられているのが, ARD 法である.

次の様な MEG 線形回帰モデルを考える.

$$y = Va' + \epsilon$$

$$y \in \mathbb{R}^L, V \in \mathbb{R}^{L \times M}, a' \in \mathbb{R}^M, \epsilon \in \mathbb{R}^L$$

ここで y, a', ϵ はそれぞれ観測磁場, 電流源, 観測誤差である. V はリードフィールド行列と呼ばれるもので, センサ, 電流双極子の位置や方向によって求められる. ここに次で述べる ARD(Automatic Relevance Determination) 事前分布を導入することで不必要な電流源の削除を試みる.

ARD とは関連度自動決定と呼ばれるものであり, 一部のパラメータを 0 にすることにより疎 (sparce) な解を得る方法である [9].

次の様な時系列データセット

$$D = \{a'^{(u)}, y^{(u)} \ \ u = 1, 2, \cdots, U\}$$

が与えられたとき, 入力と出力との間に不必要な変数 (重み) が存在するとき, それを削除することを考える.

先の MEG モデルに次の様な事前分布を導入する.

$$p(a'|a) = \prod_{i=1}^{M} \mathcal{N}(a'_i|0, \alpha_i^{-1})$$

ここで α はハイパーパラメータ, α_i は各重みに対応するハイパーパラメータであり, $\mathcal{N}(a'_i|0, \alpha_i^{-1})$ で平均 0, 分散 α_i^{-1} の正規分布を表す.

これらのハイパーパラメータについての周辺尤度を最大化することにより多くの α_i が 0 に収束し, 結果として不要な重みが削除されることとなる.

MEG モデルの確率密度関数は,

$$p(y|a') = \mathcal{N}_L(y; Va', \sigma_y^2 I_L)$$

で与えられる. ここで論文 [18] の結果より,

$$p(\{y^{(n)}\}|\{a^{(u)}\}, b) = \prod_{u=1}^{U} \mathcal{N}_L(y^{(u)}; V \sum_{m=1}^{M} b_m a_m^{(u)} \mathbf{e}_m, \sigma_y^2 I_L)$$

となる. ここで $\mathbf{e}_m$ は m 次成分が 1, 他の成分は 0 の単位ベクトルである.

l を V のランクとする. $\mathbf{v}_1, \ldots, \mathbf{v}_M$ を V の構成列ベクトルとする :

$$V = (\mathbf{v}_1, \ldots, \mathbf{v}_M).$$

$R(i_1, \ldots, i_k)$ を $\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_k}$ で生成されるベクトル空間:

$$R(i_1, \ldots, i_k) = \{c_i \mathbf{v}_{i_1} + \cdots + c_k \mathbf{v}_{i_k} \ : \ c_1, \ldots, c_k \in \mathbf{R}\}$$

として, その $R(i_1, \ldots, i_k)$ に含まれる V の列ベクトルの個数を $L(i_1, \ldots, i_k)$ とおく :

$$L(i_1, \ldots, i_k) = \#\{\mathbf{v}_i \ : \ \mathbf{v}_i \in R(i_1, \ldots, i_k)\}.$$

ここで

$$M_k = \max_{i_1, \ldots, i_k} L(i_1, \ldots, i_k).$$

とする.

定理 9 $\sum_{u=1}^{U} ||V \sum_{m=1}^{M} b_m a_m^{(u)} \mathbf{e}_m||^2$ の log canonical threshold は

$$\min\{\frac{(\ell - 1)U + M - M_{\ell-1}}{2} \ : \ \ell = 1, \ldots, l+1\}.$$

ここで, $M_0 = \#\{\mathbf{v}_k \ : \ \mathbf{v}_k = 0\}$ および $M_l = M$.

(証明)

k に関する帰納法によって, ブローアップを行い, 以下の関数が得られることを示す.

$$(*) \qquad \Psi^*(k) = w_1^2 \cdots w_k^2 \sum_{u=1}^{U} (a'^2_{1u} + \cdots + a'^2_{ku} + ||V_k \begin{pmatrix} b_{k+1} a_{k+1,u} \\ \vdots \\ b_{M'(k)} a_{M'(k)u} \end{pmatrix} ||^2)$$

ヤコビアンは

$$\prod_{\ell=1}^{k} w_\ell^{(\ell-1)U + M'(\ell-1) - (\ell-1) - 1} dbdada'dw.$$

$M'(\ell)$ は，以下の証明内で現れる Eq. (3) および Eq.(4) で順次定義される．

(Step 1)

$\mathbf{v}_1 \neq 0, \ldots, \mathbf{v}_{M-M_0} \neq 0, \mathbf{v}_{M-M_0+1} = 0, \ldots, \mathbf{v}_M = 0$ と仮定する．

$$M'(0) = M' = M - M_0 \tag{3}$$

とおく．

Ψ を部分多様体 $\{b_i = 0, 1 \leq i \leq M'\}$ に沿ってブローアップする．

$b_1 = w_1$, $b_i = w_1 b_i$, $i = 2, \ldots, M'$ とおく．

$\mathbf{v}_1 \neq 0$ より，$M \times M$ 正則行列 P_1 が存在して，$P_1 \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

$\begin{pmatrix} v_{1i}(1) \\ \mathbf{v}_i(1) \end{pmatrix} = P_1 \mathbf{v}_i$, $i = 2, \ldots, M'$, とおく．$\mathbf{v}_i(1)$ は $M-1$ 次元ベクトルである．$\mathbf{v}_2(1) \neq 0, \ldots, \mathbf{v}_{M'(1)}(1) \neq 0$, $\mathbf{v}_{M'(1)+1}(1) = \cdots = \mathbf{v}_{M'}(1) = 0$ を仮定する．$M'(1) \leq M'$ は明らか．

$V_1 = (\mathbf{v}_2(1), \ldots, \mathbf{v}_{M'(1)}(1))$ とすると，

$$P_1 V \begin{pmatrix} a_{1u} \\ b_2 a_{2u} \\ \vdots \\ b_{M'} a_{M'u} \end{pmatrix} = \begin{pmatrix} a_{1u} + (v_{12}(1), \ldots, v_{1M'}(1)) \begin{pmatrix} b_2 a_{2u} \\ \vdots \\ b_{M'} a_{M'u} \end{pmatrix} \\ V_1 \begin{pmatrix} b_2 a_{2u} \\ \vdots \\ b_{M'(1)} a_{M'(1)u} \end{pmatrix} \end{pmatrix}.$$

a_{1u} から a'_{1u} へ $a'_{1u} = a_{1u} + (v_{12}(1), \ldots, v_{1M'}(1)) \begin{pmatrix} b_2 a_{2u} \\ \vdots \\ b_{M'} a_{M'u} \end{pmatrix}$ により変数変換する．

補題 4 より，Ψ は 次の式 $\Psi^*(1)$ を考察すればよい：

$$(*) \qquad \Psi^*(1) = w_1^2 \sum_{u=1}^{U} (a'^2_{1u} + ||V_1 \begin{pmatrix} b_2 a_{2u} \\ \vdots \\ b_{M'(1)} a_{M'(1)u} \end{pmatrix}||^2).$$

ここでヤコビアンは

$$\prod_{\ell=1}^{k} w_1^{M'-1} dbdada'dw.$$

(Step 2)

$\Psi^*(k)$ を部分多様体 $\{a_{ui} = 0, i = 1, 2, \ldots, k, u = 1, \ldots, U, b_{k+1} = \ldots = b_{M'(k)} = 0\}$ に沿ってブローアップする．

$a_{ui} = w_{k+1} a_{ui}$, $b_{k+1} = w_{k+1} b_{k+1}, \ldots, b_{M'(k)} = w_{k+1} b_{M'(k)}$ とおく．

$a_{11} = 1$ ならば，$\frac{(\ell-1)U + M'(\ell-1) - (\ell-1)}{2}$, $\ell = 1, \cdots, k+1$ は， log canonical threshold の候補となる．

$b_{k+1} = 1$ とおく．$\mathbf{v}_{k+1}(k) \neq 0$ なので，$(M-k) \times (M-k)$ 正則行列 P_{k+1} が存在して，

$P_{k+1} \mathbf{v}_{k+1}(k) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

$\begin{pmatrix} v_{k+1,i}(k+1) \\ \mathbf{v}_i(k+1) \end{pmatrix} = P_{k+1}\mathbf{v}_i(k)$, $i = k+2, \ldots, M'(k)$, とおく. $\mathbf{v}_i(k+1)$ は $M-k-1$ 次元ベクトル.

$$\mathbf{v}_{k+2}(k+1) \neq 0, \ldots, \mathbf{v}_{M'(k+1)}(k+1) \neq 0, \mathbf{v}_{M'(k+1)+1}(k+1) = \cdots = \mathbf{v}_{M'(k)}(k+1) = 0 \quad (4)$$

とおく.

$V_{k+1} = (\mathbf{v}_{k+2}(k+1), \ldots, \mathbf{v}_{M'(k+1)}(k+1))$ とすれば,

$$P_{k+1}V_k \begin{pmatrix} a_{k+1,u} \\ b_{k+2}a_{k+2,u} \\ \vdots \\ b_{M'(k)}a_{M'(k)u} \end{pmatrix}$$

$$= \begin{pmatrix} a_{k+1,u} + (v_{k+1,k+2}(k+1), \ldots, v_{k+1,M'}(k+1)) \begin{pmatrix} b_{k+2}a_{k+2,u} \\ \vdots \\ b_{M'(k)}a_{M'(k)u} \end{pmatrix} \\ V_{k+1} \begin{pmatrix} b_{k+2}a_{k+2,u} \\ \vdots \\ b_{M'(k+1)}a_{M'(k+1)u} \end{pmatrix} \end{pmatrix}.$$

となる. $a_{k+1,u}$ から $a'_{k+1,u}$ へ

$$a'_{k+1u} = a_{k+1u} + (v_{k+1,k+2}(k+1), \ldots, v_{k+1,M'(k)}(k+1)) \begin{pmatrix} b_{k+2}a_{k+2,u} \\ \vdots \\ b_{M'(k)}a_{M'(k)u} \end{pmatrix}$$

によって変数変換すれば, 同様に補題 4 より, $\Psi^*(k+1)$ を考察すればよい.

(Step 3)

最後に $M'(k)-k$ の値を次のように変換する.

$\mathbf{0}$ を $M-k$ 次元 0 ベクトルとする.

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & P_k \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & P_3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & P_2 \end{pmatrix} P_1 V$$

$$= \begin{pmatrix} 1 & v_{12}(1) & v_{13}(1) & & & \cdots & & & v_{1M}(1) \\ 0 & 1 & v_{23}(2) & & & \cdots & & & \cdots \\ 0 & 0 & 1 & & & \cdots & & & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & & & & \\ 0 & 0 & 0 & \cdots 1 & v_{kM}(k) & \cdots & \cdots v_{kM'(k-1)}(k) & 0 & \cdots 0 \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{v}_{k+1}(k) & \cdots \mathbf{v}_{M'(k)}(k) & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}$$ な

ので $R(1, \ldots, k) = \{c_i\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k : c_1, \ldots, c_k \in \mathbf{R}\}$ とおけば, $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_{M'(k)} \notin R(1, \ldots, k)$ および $\mathbf{v}_{M'(k)+1}, \ldots, \mathbf{v}_M \in R(1, \ldots, k)$.

従って $M - (M'(k) - k) = \#\{\mathbf{v}_i : \mathbf{v}_i \in R(1, \ldots, k)\}$ である.

Q.E.D.

謝辞

この研究は科学研究費補助金 22540224 を受けた.

References

[1] M. Aoyagi. The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities*, 1501:153–167, 2006.

[2] M. Aoyagi. Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network. *International Journal of Pure and Applied Mathematics*, 52(2):177–204, 2009.

[3] M. Aoyagi. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Communications in Statistics - Theory and Methods*, 39(15):2667–2687, 2010.

[4] M. Aoyagi. Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9):3714–3733, 2013.

[5] M. Aoyagi. Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *Journal of Algebraic Statistics*, 4(1):30–57, 2013.

[6] M. Aoyagi and K. Nagata. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity. *Neural Computation*, 24(6):1569–1610, 2012.

[7] M. Aoyagi and S. Watanabe. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II*, 10:2112–2124, 2005a.

[8] M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18:924–933, 2005b.

[9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, LLC, 2006.

[10] M. Drton. Conference lecture: Reduced rank regression. *Workshop on Singular Learning Theory, AIM 2011, http://math.berkeley.edu/~critch/slt2011/*, 2011.

[11] M. Drton. Conference lecture: Bayesian information criterion for singular models. *Algebraic Statistics 2012 in the Alleghenies at The Pennsylvania State University, http://jasonmorton.com/aspsu2012/*, 2012.

[12] H. Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Math*, 79:109–326, 1964.

[13] J. Kollár. Singularities of pairs. *Algebraic geometry-Santa Cruz 1995, Proc. Symp. Pure Math., American Mathematical Society, Providence, RI,*, 62:221–287, 1997.

[14] S. Lin. Asymptotic approximation of marginal likelihood integrals. *(preprint)*, 2010.

[15] M. Mustata. Singularities of pairs via jet schemes. *J. Amer. Math. Soc.*, 15:599–615, 2002.

[16] K. Nagata and S. Watanabe. Exchange Monte Carlo sampling from Bayesian posterior for singular learning machines. *IEEE Transactions on Neural Networks*, 19(7):1253–1266, 2008a.

[17] K. Nagata and S. Watanabe. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *International Journal of Neural Networks*, 21(7):980–988, 2008b.

[18] S. Nakajima and S. Watanabe. Analytic solution of hierarchical variational bayes in linear inverse problem. In *Proceedings of the ICANN2006, LNCS*, volume 4132, pages 240–249, 2006.

[19] T. Ohsawa and K. Takegoshi. On the extension of L^2 holomorphic functions. *Math. Zeitschrift*, 195:197–204, 1987.

[20] D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian networks. *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 438–445, 2002.

[21] D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, 6:1–35, 2005.

[22] S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001a.

[23] S. Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8):1049–1060, 2001b.

[24] S. Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, 2010.

[25] S. Watanabe, K. Hagiwara, S. Akaho, Y. Motomura, K. Fukumizu, M. Okada, and M. Aoyagi. *Theory and Application of Learning System.* Morikita Press, 2005.

[26] T. Yoshioka, M. Sato, S. Kajiwara, and K. Toyama. Estimation of current distribution in brain from meg data based on variational bayes. *Technical report of IEICE, C2003*, 149:95–100, 2003.

[27] P. Zwiernik. An asymptotic behavior of the marginal likelihood for general Markov models. *Journal of Machine Learning Research*, 12:3283–3310, 2011.