2004 年情報論的学習理論ワークショップ
2004 Workshop on Information-Based Induction Sciences (IBIS2004)
Tokyo, Japan, November 8 - 10, 2004.

Desingularization and the Generalization Error of Reduced Rank Regression in Bayesian Estimation

Miki Aoyagi * Sumio WATANABE †

Abstract: Reduced rank regression, or a three-layer neural network with linear hidden units, is an important learning machine because it extracts the essential information from training samples. However, its generalization error had been left unknown because of its singularities in the parameter space. In this paper, we propose a new method of recursive blowing-ups for densingularization of a learning machine. By applying it to the reduced rank approximation, we show the effectiveness of the method and clarify the asymptotic generalization error of the reduced rank regression.

Keywords Stochastic complexity, reduced rank regression, non-regular learning machines, Bayesian estimate, resolution of singularities

1 Introduction

Reduced rank regression is understood that it can be thought as a three-layer neural network with linear hidden units [3]. This method picks up the essential information from examples of input-output pairs. It is a non-regular statistical model which has a degenerate Fisher information matrix. Therefore the theory of regular statistical models, for example, model selection methods AIC[1], TIC[11], HQ[5], NIC[7], BIC[10], MDL[8], cannot be applied to the reduced rank approximation, as it is non-regular.

Recently, the new method to calculate the asymptotic form of the Bayesian stochastic complexity has been introduced, using the method of resolution of singularities [12, 13, 14].

Let $x \in \mathbf{R}^M$ be an input, $y \in \mathbf{R}^N$ an output and $w \in W \subset \mathbf{R}^d$ a parameter. Consider a learning machine $p(x, y|w)$ and a fixed *a priori* probability density function $\psi(w)$. Assume that the true probability distribution $p(x, y|w_0)$ is contained in the learning model.

Let $X^n = (X_1, ..., X_n)$ and $Y^n = (Y_1, ..., Y_n)$ be arbitrary n training samples which are independently taken from the true probability distribution $p(x, y|w_0)$. The *a posteriori* probability density function $p(w|X^n, Y^n)$ is written by

$$p(w|X^n, Y^n) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^{n} p(X_i, Y_i|w),$$

where

$$Z_n = \int_W \psi(w)\prod_{i=1}^{n} p(X_i, Y_i|w)dw.$$

Then the average inference $p(x, y|X^n, Y^n)$ of the Bayesian distribution is given by

$$p(x, y|X^n, Y^n) = \int p(x, y|w)p(w|X^n, Y^n)dw.$$

Let $G(n)$ be the generalization error or the learning efficiency

$$G(n) := E_n\{\int p(x, y|w_0)\log\frac{p(x, y|w_0)}{p(x, y|X_n)}dxdy\},$$

where $E_n\{\}$ is the expectation value over all sets of n training samples.

Then the average stochastic complexity or the free energy

$$F(n) := -E_n\{\log\int \exp(-nK_n(w))\psi(w)dw\},$$

satisfies

$$G(n) = F(n+1) - F(n),$$

where

$$K_n(w) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p(X_i, Y_i|w_0)}{p(X_i, Y_i|w)}.$$

*Department of Mathematics, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554, tel. 03-3238-3460, fax. 03-3238-3933 e-mail. miki-a@sophia.ac.jp

†Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuda, Midori-ku, Yokohama, 226–8503, e-mail. swatanab@pi.titech.ac.jp

Let $J(z)$ be the zeta function of the learning model

$$J(z) := \int K(w)^z \psi(w) dw,$$

where $K(w)$ is the Kullback distance

$$K(w) := \int p(x,y|w_0) \log \frac{p(x,y|w_0)}{p(x,y|w)} dx.$$

Then, we have

$$G(n) \cong \lambda/n - (m-1)/(n \log n), \quad (1)$$

and

$$F(n) = \lambda \log n - (m-1) \log \log n + O(1), \quad (2)$$

where $-\lambda$ is the maximum pole of $J(z)$, m is its order, and $O(1)$ is a bounded function of n.

The values λ and m can be calculated by using the blowing-up process.

In [16], the upper bound of the constant λ for the reduced rank regression models was obtained, but not the exact value for λ.

In this paper, we use the recursive blowing-ups to obtain the exact values λ and m for the reduced rank regression models, and give the asymptotic form of the generalization energy explicitly. We show that desingularization is effective to analyze zeta functions for learning theory.

2 Resolution of singularities

In this section, we introduce Hironaka's Theorem [6] on the resolution of singularities and the construction of blowing up. The blowing up is the main tool in the resolution of singularities of an algebraic variety. We also show its application in the field of learning theory [12, 13, 14].

Theorem[Hironaka [6]]

Let f be a real analytic function in a neighborhood of $w = (w_1, \cdots, w_d) \in \mathbf{R}^d$ with $f(w) = 0$. There exists an open set $V \ni w$, a real analytic manifold U and a proper analytic map μ from U to V such that

(1) $\mu : U - \mathcal{E} \to V - f^{-1}(0)$ is an isomorphism, where $\mathcal{E} = \mu^{-1}(f^{-1}(0))$,

(2) for each $u \in U$, there are local analytic coordinates $(u_1, \cdots, u_n)$ such that

$$f(\mu(u)) = \pm u_1^{s_1} u_2^{s_2} \cdots u_n^{s_n},$$

where $s_1, \cdots, s_n$ are non-negative integers.

The above theorem is one of analytic versions of the Hironaka's theorem used by Atiyah [2].

Consequently, we have

Theorem[Atiyah [2], Bernstein [4], Sato & Shintani [9]]

Let $f(w)$ be an arbitrary analytic function of variables $w \in \mathbf{R}^d$, $g(w)$ be an arbitrary $C^\infty-$ function with compact support W, where $g(0) > 0$.

Then

$$\zeta(z) = \int_W |f(w)|^z g(w) dw,$$

is a holomorphic function in the right-half plane.

Furthermore, $\zeta(z)$ can be analytically continued to a meromorphic function on the entire complex plane. Its poles are negative rational numbers.

Apply Hironaka's theorem to the Kullback distance $K(w)$. For each $w \in K^{-1}(0) \cap W$, we have a proper analytic map μ from a neighborhood V_w of w to an analytic manifold U_w, which satisfy the above (1) and (2). Then the local integration on V_w of $K(w)^z \psi(w)$ is written by

$$J_w(z) = \int_{V_w} K(w)^z \psi(w) dw$$

$$= \int_{U_w} (u_1^{s_1} u_2^{s_2} \cdots u_n^{s_n})^z \psi(\mu(u)) |\mu'(u)| du.$$

Therefore the poles of $J_w(z)$ can be obtained. For $w \in W \setminus K^{-1}(0)$, there exists a neighborhood V_w of w such that $K(w') \neq 0, w' \in V_w$ and so $J_w(z) = \int_{V_w} K(w)^z \psi(w) dw$ has no poles.

Since the set of parameters W is compact, we obtain the poles and their orders of $J(z)$.

Next we explain the construction of blowing up.

There are three kinds of blowing up, i.e., blowing up at the point, blowing up along the manifold and blowing up with respect to the coherent sheaf of ideals. The blowing up along the manifold includes blowing up at the point. The blowing up with respect to the coherent sheaf of ideals includes blowing up along the manifold.

Here let us explain only the blowing up along the manifold used in this paper.

Define a manifold $\mathcal{M}$ by gluing k open sets $U_i \cong \mathbf{R}^d$, $i = 1, 2, \cdots, k$, $d \geq k$ as follows.

Denote the coordinate of U_i by $(\xi_{1i}, \cdots, \xi_{di})$.

Define the equivalence relation

$$(\xi_{1i}, \xi_{2i}, \cdots, \xi_{di}) \sim (\xi_{1j}, \xi_{2j}, \cdots, \xi_{dj})$$

at $\xi_{ji} \neq 0$ and $\xi_{ij} \neq 0$, by $\xi_{ij} = 1/\xi_{ji}, \xi_{jj} = \xi_{ii}\xi_{ji}, \xi_{hj} = \xi_{hi}/\xi_{ji} (1 \leq h \leq k, h \neq i, j), \xi_{\ell j} = \xi_{\ell i} (k+1 \leq \ell \leq d)$. Set $\mathcal{M} = \coprod_{i=1}^{k} U_i / \sim$.

Also define $\pi : \mathcal{M} \to \mathbf{R}^d$ by

$$\begin{aligned} &U_i \ni (\xi_{1i}, \cdots, \xi_{ni}); \\ &\mapsto (\xi_{ii}\xi_{1i}, \cdots, \xi_{ii}\xi_{i-1i}, \xi_{ii}, \xi_{ii}\xi_{i+1i}, \cdots, \xi_{ii}\xi_{ki}, \\ &\qquad\qquad \xi_{k+1i}, \cdots, \xi_{di}). \end{aligned}$$

This map is well-defined and called the blowing up along

$$\begin{aligned} X &= \{(w_1, \cdots, w_k, w_{k+1}, \cdots, w_d) \in \mathbf{R}^d \mid \\ &\qquad\qquad w_1 = \cdots = w_k = 0\}. \end{aligned}$$

The blowing map satisfies

(1) $\pi : \mathcal{M} \to \mathbf{R}^d$ is proper,

(2) $\pi : \mathcal{M} - \pi^{-1}(X) \to \mathbf{R}^d - X$ is isomorphic.

3 Application to reduced rank regression models

In this section, we show how to obtain the maximum pole of the zeta function of learning models in the case of the reduced rank regression models.

Let

$$\begin{aligned} \{w = (A, B) \quad | \quad &A \text{ is an } H \times M \text{ matrix}, \\ &B \text{ is an } N \times H \text{ matrix}\} \end{aligned}$$

be the set of parameters.

We define the norm of any matrix $T = (t_{ij})$ by $||T|| = \sqrt{\sum_{i,j} |t_{ij}|^2}$.

Denote the input value by x with a probability density function $q(x)$. Assume that all eigenvalues of the $M \times M$ matrix $\mathcal{X} = (\int x_i x_j q(x) dx)$ are positive numbers. That is, $\mathcal{X}$ is a positive definite.

Then the output value y of the reduced rank regression model is given by

$$y = BAx + (\text{noise}).$$

Consider the statistical model

$$p(y|x, w) = \frac{1}{(\sqrt{2\pi})^N} \exp(-\frac{1}{2}||y - BAx||^2).$$

Assume that the *a priori* probability density function $\psi(w)$ is a $C^\infty-$ function with compact support W where $\psi(A_0, B_0) > 0$ and that the true parameters w are $w = (A_0, B_0)$.

Lemma 1 *There exist $c_1 > 0$ and $c_2 > 0$ such that*

$$c_1||BA - B_0A_0||^2 \leq K(w) \leq c_2||BA - B_0A_0||^2. \quad (3)$$

Proof.

Put

$$q(x, y) = p(y|x, (A_0, B_0))q(x).$$

Then we have the Kullback information

$$\begin{aligned} K(w) &= \int q(x, y) \log \frac{p(y|x, (A_0, B_0))}{p(y|x, w)} dxdy \\ &= \frac{1}{2} \int ||(BA - B_0A_0)x||^2 q(x) dx. \end{aligned}$$

Put $S = BA - B_0A_0 = (s_{i,j})$ and let Q be an orthogonal matrix such that $Q^t \mathcal{X} Q$ is diagonal.

Then, we have

$$\begin{aligned} K(w) &= \frac{1}{2} \int ||Sx||^2 q(x) dx \\ &= \frac{1}{2} \int \sum_i (\sum_j s_{ij} x_j)^2 q(x) dx \\ &= \frac{1}{2} \sum_{i, j_1, j_2} s_{ij_1} s_{ij_2} \int x_{j_1} x_{j_2} q(x) dx \\ &= \frac{1}{2} \mathrm{Tr}(S\mathcal{X}S^t) = \frac{1}{2} \mathrm{Tr}(SQQ^t\mathcal{X}Q(SQ)^t). \end{aligned}$$

Since we assume all eigenvalues of $\mathcal{X}$ are positive numbers, there exist $c_1 > 0$ and $c_2 > 0$ such that

$$\begin{aligned} &c_1 \mathrm{Tr}(SQ(SQ)^t) = c_1 \mathrm{Tr}(SS^t) \\ \leq \quad &K(w) \leq c_2 \mathrm{Tr}(SQ(SQ)^t) = c_2 \mathrm{Tr}(SS^t). \end{aligned}$$

Since $\mathrm{Tr}(SS^t) = ||S||^2$, we complete the proof.

Q.E.D.

Lemma 2 *[15]*

Let $f(w)$, $f_1(w)$, $f_2(w)$ be an analytic function of variables $w \in \mathbf{R}^d$. Let $g(w)$, $g_1(w)$, $g_2(w)$ be a $C^\infty-$ function with compact support W.

Put

$$\zeta(z) = \int_W |f(w)|^z g(w) dw.$$

Denote the maximum pole of $\zeta(z)$ by $-\Lambda(f, g)$.

Then if $|f_1| \leq |f_2|$ and $g_1 \geq g_2$, we have

$$\Lambda(f_1, g_1) \leq \Lambda(f_2, g_2).$$

Furthermore, for any number $a \in \mathbf{R} - \{0\}$,

$$\Lambda(af, g) = \Lambda(f, ag) = \Lambda(f, g).$$

Lemma 1 and Lemma 2 yield that the zeta function can be written as follows:

$$J(z) = \int_W ||BA - B_0A_0||^{2z}\psi(w)dw.$$

Main Theorem

Let r be the rank of B_0A_0.

The maximum pole $-\lambda$ of $J(z)$ is

$$\max\{-\frac{(N+M)r - r^2 + s(N-r)}{2} - \frac{(M-r-s)(H-r-s)}{2} \mid 0 \leq s \leq \min\{M+r, H+r\}\}.$$

Furthermore, $G(n)$ and $F(n)$ in Equation (1) and (2) are given by using the following maximum pole $-\lambda$ of $J(z)$ and its order m:

(1) Let $N+r \leq M+H$, $M+r \leq N+H$ and $H+r \leq M+N$.

(a) If $M+H+N+r$ is even, then $m=1$ and

$$\lambda = \frac{-(H+r)^2 - M^2 - N^2}{8} + \frac{2(H+r)M + 2(H+r)N + 2MN}{8}.$$

(b) If $M+H+N+r$ is odd, then $m=2$ and

$$\lambda = \frac{-(H+r)^2 - M^2 - N^2}{8} + \frac{2(H+r)M + 2(H+r)N + 2MN + 1}{8}.$$

(2) Let $M+H < N+r$. Then $m=1$ and

$$\lambda = \frac{HM - Hr + Nr}{2}.$$

(3) Let $N+H < M+r$. Then $m=1$ and

$$\lambda = \frac{HN - Hr + Mr}{2}.$$

(4) Let $M+N < H+r$. Then $m=1$ and

$$\lambda = \frac{MN}{2}.$$

In order to prove Main Theorem, we need the following three lemmas.

Let $\mathrm{Mat}(H', M')$ be the set of $H' \times M'$ matrices with real values.

Lemma 3 *Let U be a neighborhood of $\theta_0 \in \mathbf{R}^\ell$. Also let $T_1(\theta)$, $T_2(\theta)$, $T(\theta)$ be any functions from U to $\mathrm{Mat}(N', H')$, $\mathrm{Mat}(N', M')$, $\mathrm{Mat}(H', M')$ respectively.*

Assume that the function $||T(\theta)||$ is bounded.

Then, there exist positive numbers $\alpha > 0$ and $\beta > 0$ such that

$$\begin{aligned}\alpha(||T_1||^2 + ||T_2||^2) &\leq ||T_1||^2 + ||T_2 + T_1T||^2 \\ &\leq \beta(||T_1||^2 + ||T_2||^2).\end{aligned}$$

Proof.

Since $||T(\theta)||$ is bounded, there exists $\beta > 3$ such that

$$\begin{aligned}&||T_1||^2 + ||T_2 + T_1T||^2 \\ &\quad \leq (||T_1||^2 + 2||T_2||^2 + 2||T_1T||^2) \\ &\quad \leq \beta(||T_1||^2 + ||T_2||^2).\end{aligned}$$

Also, there exists $\gamma > 3$ such that

$$\begin{aligned}||T_2||^2 &\leq 2(||T_2 + T_1T||^2 + || - T_1T||^2) \\ &\leq 2(||T_2 + T_1T||^2 + \gamma||T_1||^2),\end{aligned}$$

and hence

$$\begin{aligned}||T_1||^2 + ||T_2||^2 &\leq 2||T_2 + T_1T||^2 + (2\gamma+1)||T_1||^2) \\ &\leq (2\gamma+1)(||T_2 + T_1T||^2 + ||T_1||^2).\end{aligned}$$

By putting $\alpha = 1/(2\gamma+1)$, we complete the proof.

Q.E.D.

Lemma 4 *Let U be a neighborhood of $\theta_0 \in \mathbf{R}^\ell$. Also let $T(\theta)$ be any function from U to $\mathrm{Mat}(H', M')$.*

Let P_0, Q_0 be any regular $M' \times M'$, $H' \times H'$ matrices, respectively.

Then there exist positive numbers $\alpha > 0$, $\beta > 0$ such that

$$\alpha||T||^2 \leq ||P_0TQ_0||^2 \leq \beta||T||^2.$$

Proof.

There exists $\beta > 0$ such that

$$||P_0TQ_0||^2 \leq \beta||T||^2.$$

Also, there exists $\gamma > 0$

$$||T||^2 = ||P_0^{-1}P_0TQ_0Q_0^{-1}||^2 \leq \gamma||P_0TQ_0||^2.$$

By putting $\alpha = 1/\gamma$, we complete the proof.

Q.E.D.

Lemma 5 *Put*

$$\Psi = ||BA - B_0A_0||^2.$$

Then there exist a function Ψ' and an a priori probability density function $\psi'(w')$ such that

(a)

$$\begin{aligned}\Psi' &= ||C_1||^2 + ||C_2||^2 \\ &\quad +||C_3||^2 + ||B_4A_4||^2 \end{aligned} \tag{4}$$

where C_1 is an $r \times r$ matrix, C_2 is an $(N-r)\times r$ matrix, C_3 is an $r \times (M-r)$ matrix, A_4 is an $(H-r)\times(M-r)$ matrix and B_4 is an $(N-r)\times(H-r)$ matrix,

(b) $\psi'(w')$ is a $C^\infty-$ function with compact support W', where $w' = (C_1, C_2, C_3, B_4, A_4)$ and $\psi'(0) > 0$,

(c) the maximum pole of $\int_W \Psi^z\psi dw$ is equal to the one of $\int_{W'} \Psi'^z\psi' dw'$.

Proof.

Since the rank B_0A_0 is r, there exists regular matrices P_0, Q_0 such that $P_0B_0A_0Q_0 = \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}$, where E is the $r \times r$ unit matrix.

Change variables from B, A to B', A' by $B' = P_0^{-1}B$ and $A' = AQ_0^{-1}$.

Then

$$\Psi = \left\| P_0(B'A' - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix})Q_0 \right\|^2.$$

Let $A' = \begin{pmatrix} A_1 & A_3 \\ A_2 & A_4 \end{pmatrix}$ and $B' = \begin{pmatrix} B_1 & B_3 \\ B_2 & B_4 \end{pmatrix}$, where

A_1 is an $r \times r$ matrix,
A_3 is an $r \times (M-r)$ matrix,
A_2 is an $(H-r) \times r$ matrix,
A_4 is an $(H-r) \times (M-r)$ matrix,
B_1 is an $r \times r$ matrix,
B_3 is an $r \times (H-r)$ matrix,
B_2 is an $(N-r) \times r$ matrix,
B_4 is an $(N-r) \times (H-r)$ matrix.

Let $U_{(A',B')}$ be a sufficiently small neighborhood of any point (A', B') with

$$B'A' - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} = 0.$$

Since the rank $\begin{pmatrix} B_1 & B_3 \end{pmatrix}\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ is r, we can assume A_1 is regular. Thus we can change the variables from B_1, B_2 to C_1, C_2 by $C_1 = B_1A_1 + B_3A_2 - E$ and $C_2 = B_2A_1 + B_4A_2$. Also changing the variables from A_4 to A'_4 by $A'_4 = -A_2A_1^{-1}A_3 + A_4$ and from A_3 to A'_3 by $A'_3 = A_1^{-1}A_3$ gives

$$B'A' - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} C_1 & C_1A'_3 + A'_3 + B_3A'_4 \\ C_2 & C_2A'_3 + B_4A'_4 \end{pmatrix}.$$

By changing the variables from A'_3 to A''_3 by $A''_3 = A'_3 + B_3A'_4$, we obtain

$$\Psi = \left\| P_0 \begin{pmatrix} C_1 & C_1(A''_3 - B_3A'_4) + A''_3 \\ C_2 & C_2(A''_3 - B_3A'_4) + B_4A'_4 \end{pmatrix} Q_0 \right\|^2.$$

By Lemma 2 and Lemma 4, the maximum pole of $\int_{U_{(A',B')}} \Psi^z\psi dw$ is equal to the one of

$$\int_{U_{(A',B')}} \left\| \begin{pmatrix} C_1 & C_1(A''_3 - B_3A'_4) + A''_3 \\ C_2 & C_2(A''_3 - B_3A'_4) + B_4A'_4 \end{pmatrix} \right\|^{2z} \psi dw.$$

Then Lemma 2 and Lemma 3 yield that the maximum pole of $\int_{U_{(A',B')}} \Psi^z\psi dw$ is equal to the one of

$$\begin{aligned} &\int_{U_{(A',B')}} \Psi'^z\psi dw \\ &= \int_{U_{(A',B')}} \left\| \begin{pmatrix} C_1 & A''_3 \\ C_2 & B'_4A'_4 \end{pmatrix} \right\|^{2z} \psi dw. \end{aligned} \tag{5}$$

Let $C_3 = A''_3$, $A_4 = A'_4$ and

$$\psi'(C_1, C_2, C_3, A_4, B_4) = \psi(A, B).$$

The proof follows from the fact that the poles of the above function is same when (A', B') with $B'A' - \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix} = 0$ varies.

Q.E.D

Before proving Main Theorem, let us give some notation.

Since we often change the variables during the blowing up process, it is more convenient for us to use the same symbols a_{ij} rather than a'_{ij}, a''_{ij}, $\cdots$, etc, for the sake of simplicity. For instance,

"Let $\begin{cases} a_{11} = u_{11} \\ a_{ij} = u_{11}a_{ij},\ (i,j) \neq (1,1). \end{cases}$"

instead of

"Let $\begin{cases} a_{11} = u_{11} \\ a_{ij} = u_{11}a'_{ij},\ (i,j) \neq (1,1). \end{cases}$"

Proof of Main Theorem.

Let $A_4 = \begin{pmatrix} a_{11} & \cdots & a_{1,M-r} \\ a_{21} & \cdots & a_{2,M-r} \\ & \vdots & \\ a_{H-r,1} & \cdots & a_{H-r,M-r} \end{pmatrix}$,

$B_4 = \begin{pmatrix} b_{11} & \cdots & b_{1,H-r} \\ b_{21} & \cdots & b_{2,H-r} \\ & \vdots & \\ b_{N-r,1} & \cdots & b_{N-r,H-r} \end{pmatrix}$.

Suppose that C_1, C_2 and C_3 are as in Lemma 4. We need to calculate poles of the following function by using the blowing up process together with an inductive method.

Assume

$$\begin{aligned} \Psi''(s) &= ||C_1||^2 + ||C_2||^2 + ||C_3||^2 + \sum_{i=1}^{s} ||\mathbf{b}_i||^2 \\ &\quad + ||\sum_{i=1}^{s} \mathbf{b}_i D_i + B^{(s+1)} A^{(s+1)}||^2, \end{aligned} \tag{6}$$

where for $i = 1, \cdots, H-r$, $\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ \vdots \\ b_{N-r,i} \end{pmatrix}$,

$B^{(s+1)} = \begin{pmatrix} b_{1,s+1} & \cdots & b_{1,H-r} \\ b_{2,s+1} & \cdots & b_{2,H-r} \\ & \vdots & \\ b_{N-r,s+1} & \cdots & b_{N-r,H-r} \end{pmatrix}$ and

$A^{(s+1)} = \begin{pmatrix} a_{s+1,s+1} & \cdots & a_{s+1,M-r} \\ a_{s+2,s+1} & \cdots & a_{s+2,M-r} \\ & \vdots & \\ a_{H-r,s+1} & \cdots & a_{H-r,M-r} \end{pmatrix}$.

$D_i(a_{kl})$ is a function of the entries of the matrix A_4 excluding the entries of $A^{(s+1)}$. The definition of the function $D_i(a_{kl})$ will be given recursively in Equation (7) below.

Now we apply the induction method to Equation (6).

Let $C_1 = (c_{ij}^{(1)})$, $C_2 = (c_{ij}^{(2)})$ and $C_3 = (c_{ij}^{(3)})$.

Construct the blowing up of Ψ'' in (6) along the submanifold

$\{C_1 = C_2 = C_3 = \mathbf{b}_i = A^{(s+1)} = 0, 1 \le i \le s\}$.

Let $\begin{cases} c_{11}^{(1)} = v, c_{ij}^{(1)} = v c_{ij}^{(1)}, \ (i,j) \neq (1,1), \\ \mathbf{b}_j = v\mathbf{b}_j, 1 \le j \le s, C_2 = vC_2, \\ C_3 = vC_3, A^{(s+1)} = vA^{(s+1)}. \end{cases}$

Substituting them into Equation (6) gives

$$\begin{aligned} \Psi'' &= v^2(1 + \sum_{(i,j)\neq(1,1)} (c_{ij}^{(1)})^2 + ||C_2||^2 + ||C_3||^2 \\ &\quad + \sum_{i=1}^{s} ||\mathbf{b}_i||^2 + ||\sum_{i=1}^{s} \mathbf{b}_i D_i + B^{(s+1)} A^{(s+1)}||^2). \end{aligned}$$

Here the Jacobian is

$$v^{(N+M)r - r^2 + s(N-r) + (M-r-s)(H-r-s) - 1}.$$

Therefore we have the pole

$-((N+M)r - r^2 + s(N-r) + (M-r-s)(H-r-s))/2$.

If we set either of $c_{ij}^{(1)} = u, c_{ij}^{(2)} = u, c_{ij}^{(3)} = u, b_{ij} = u$ for any (i,j), we obtain the same pole by symmetry.

Next let

$$\begin{cases} a_{s+1,s+1} = u, a_{j\ell} = u a_{j\ell} \\ \quad s+1 \le j \le H-r, s+1 \le \ell \le M-r, \\ \quad (j,\ell) \neq (s+1, s+1) \\ \quad C_1 = uC_1, C_2 = uC_2, C_3 = uC_3, \\ \quad \mathbf{b}_i = u\mathbf{b}_i, 1 \le i \le s. \end{cases}$$

We also obtain the same pole by setting $a_{j\ell} = u$ for any (j,ℓ).

Substituting our new variables into Equation (6) implies

$$\begin{aligned} \Psi'' &= u^2(||C_1||^2 + ||C_2||^2 + ||C_3||^2 \\ &\quad + \sum_{i=1}^{s} ||\mathbf{b}_i||^2 + ||\sum_{i=1}^{s} \mathbf{b}_i D_i + \\ &\quad \begin{pmatrix} \mathbf{b}_{s+1} & B^{(s+2)} \end{pmatrix} \begin{pmatrix} 1 & \tilde{\mathbf{a}}_{s+1} \\ \mathbf{a}_{s+1} & A^{(s+2)} \end{pmatrix} ||^2) \\ &= u^2(||C_1||^2 + ||C_2||^2 + ||C_3||^2 + \sum_{i=1}^{s} ||\mathbf{b}_i||^2 \\ &\quad + ||\sum_{i=1}^{s} \mathbf{b}_i D_i + \begin{pmatrix} \mathbf{b}_{s+1} + B^{(s+2)}\mathbf{a}_{s+1} & 0 \end{pmatrix} \\ &\quad + \begin{pmatrix} \mathbf{b}_{s+1} & B^{(s+2)} \end{pmatrix} \begin{pmatrix} 0 & \tilde{\mathbf{a}}_{s+1} \\ & A^{(s+2)} \end{pmatrix} ||^2), \end{aligned}$$

where $\tilde{\mathbf{a}}_{s+1} = (a_{s+1,s+2} \ \cdots \ a_{s+1,M-r})$ and $\mathbf{a}_{s+1} = {}^T(a_{s+2,s+1} \ \cdots \ a_{H-r,s+1})$ (T denotes the transpose).

Denote the first column of D_i by $\mathbf{D}_i$. Let $D_i = (\mathbf{D}_i \ \ D_i')$.

Put $\mathbf{b}_{s+1} = \mathbf{b}_{s+1} + B^{(s+2)}\mathbf{a}_{s+1} + \sum_{i=1}^{s} \mathbf{b}_i \mathbf{D}_i$. Then

$$\begin{aligned} \Psi''/u^2 &= ||C_1||^2 + ||C_2||^2 + ||C_3||^2 + \sum_{i=1}^{s+1} ||\mathbf{b}_i||^2 \\ &\quad + ||\sum_{i=1}^{s} \mathbf{b}_i (D_i' - \mathbf{D}_i \tilde{\mathbf{a}}_{s+1}) + \mathbf{b}_{s+1}\tilde{\mathbf{a}}_{s+1} \\ &\quad + B^{(s+2)}(-\mathbf{a}_{s+1}\tilde{\mathbf{a}}_{s+1} + A^{(s+2)})||^2. \end{aligned}$$

Now let $A^{(s+2)} = -\mathbf{a}_{s+1}\tilde{\mathbf{a}}_{s+1} + A^{(s+2)}$. Then,

$$\Psi''/u^2 = ||C_1||^2 + ||C_2||^2 + ||C_3||^2 + \sum_{i=1}^{s+1} ||\mathbf{b}_i||^2$$

$$+||\sum_{i=1}^{s}\mathbf{b}_i(D_i'-\mathbf{D}_i\tilde{\mathbf{a}}_{s+1})$$
$$+\mathbf{b}_{s+1}\tilde{\mathbf{a}}_{s+1}+B^{(s+2)}A^{(s+2)}||^2.$$

Repeat this whole process by setting

$$D_i=D_i'-\mathbf{D}_i\tilde{\mathbf{a}}_{s+1}\quad(1\leq i\leq s)\text{ and }D_{s+1}=\tilde{\mathbf{a}}_{s+1}. \tag{7}$$

Then s will be replaced by $s+1$ in (6) and so on.

Therefore the poles are

$-((N+M)r-r^2+s(N-r)+(M-r-s)(H-r-s))/2$,

for $s=0,\cdots,\min\{H-r,M-r\}$ and so Main Theorem follows.

Q.E.D.

4 Discussion and Conclusion

In this paper, we introduce a recursive blowing up method to obtain the maximum pole of the zeta functions for the reduced rank regression models.

Note that if the rank r of A_0B_0 is zero, then H, M and N can be permuted in the formula λ of Main Theorem.

Figure 1 shows the graphs of the maximum poles λ with λ-values in y-axis and H-values in x-axis, when $M=N=10$ and $r=0$. It is clear that the curve is not linear.

Significance of the obtained result from the viewpoint of learning theory is as follows.

First, our results enable us to construct mathematical foundation for analyzing and developing the precision of the MCMC method. By the MCMC method, the estimated values of marginal likelihoods for hyperparameter estimation and model selection methods of complex learning models, had been calculated, but the theoretical values were not known. Now, we formulated the theoretical value of marginal likelihoods in this paper. Then we can compare the calculated values and the theoretical values.

Second, we can discuss the model selection problem for Bayesian estimation, although it is still open problem for non-regular models. For regular models, $\lambda=d/2$ and $m=1$, where d is the dimension of the parameter space. In other words, it does not depend on the true distribution. However, non-regular models have λ depending on the true distribution and it is smaller than $d/2$. Non-regular models are better learning machines than regular ones provided that the Bayes estimation is applied.

In general, the algebraic method will lead us to solve the difficult problems of learning theory. In particular, this method can be used to compute the asymptotic forms for all possible cases not only the reduced rank regression models. Our aim is to develop a mathematical theory in that context.

Acknowledgments

This research was supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 16700218.

参考文献

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control.* **19** 716-723.

[2] Atiyah, M. F. (1970). Resolution of singularities and division of distributions. *Comm. Pure and Appl. Math.* **13** 145-150.

[3] Baldi, P. and Hornik, K. (1995). Learning in Linear Networks: a Survey. *IEEE Transactions on Neural Networks.* **6** (4) 837-858.

[4] Bernstein, I. N. (1972). The analytic continuation of generalized functions with respect to a parameter. *Functional Analysis Applications.* **6** 26-40.

[5] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of Royal Statistical Society, Series B.* **41** 190-195.

[6] Hironaka, H. (1964). Resolution of Singularities of an algebraic variety over a field of characteristic zero. *Annals of Math.* **79** 109-326.

[7] Murata, N. J., Yoshizawa, S. G. and Amari, S. (1994). Network information criterion - determining the number of hidden units for an artificial neural network model. *IEEE Trans. on Neural Networks.* **5** (6) 865-872.

[8] Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. on Information Theory.* **30** (4) 629-636.

[9] Sato, M. and Shintani, T. (1974). On zeta functions associated with prehomogeneous vector space. *Annals of Math.* **100** 131-170.

[10] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics.* **6** (2) 461-464.

[11] Takeuchi, K. (1976). Distribution of an information statistic and the criterion for the optimal model. *Mathematical Science.* **153** 12-18 (In Japanese).

[12] Watanabe, S. (1999). Algebraic analysis for singular statistical estimation. *Lecture Notes on Computer Science.* **1720** 39-50.

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

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

[15] Watanabe, S. (2001c). Algebraic geometry of learning machines with singularities and their prior distributions. *Journal of Japanese Society of Artificial Intelligence.* **16** (2) 308-315.

[16] Watanabe, K. and Watanabe, S. (2003). Upper Bounds of Bayesian Generalization Errors in Reduced Rank Regression. *IEICE Trans.* **J86-A** (3) 278-287 (In Japanese).

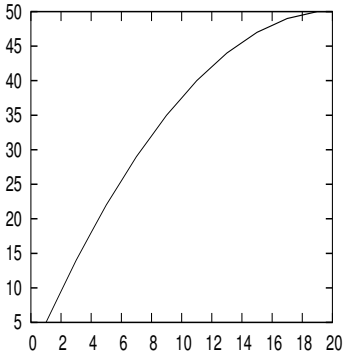

図 1: The curve of λ-values in y-axis and H-values in x-axis, when $M = N = 10$.