

階層型モデルの最尤推定の汎化誤差

Generalization error of the maximum likelihood estimation of layered models

福水 健次*

Kenji Fukumizu

Abstract: The statistical asymptotic theory is used in many theoretical results in learning theory. It elucidates the limiting distribution of the maximum likelihood estimator (MLE). However, in layered models such as neural networks, the asymptotic theory does not always hold. The true parameter is unidentifiable, if the true function can be realized by a network of smaller size than the size of the model. In this paper, we calculate the expectation of the generalization error of the MLE in the case of linear neural network model to see the statistical behavior of MLE. We show that the generalization error in unidentifiable cases is larger than what is given by the usual asymptotic theory.

Keywords: ニューラルネット、最尤推定、汎化誤差、識別可能性、縮小ランク回帰

1 まえがき

多層パーセプトロンなどの階層型ニューラルネットワークは、入力から出力への条件付き確率 $p(y|x)$ を、統計的に与えられた学習サンプルから求める、パラメトリックな非線形回帰問題と捉えることができる。この立場に立って、ニューラルネットの学習の解析に統計的漸近理論を使う手法も多く用いられている。たとえばモデル選択の際に検定を行ったり、選択規準として AIC や MDL のような手法を用いる試みが行なわれている。

しかしながら、近年、階層型ニューラルネットにおいては通常の漸近理論が成立しない場合があることが指摘され、そういった場合の推定量の挙動を解析しようとする研究が行なわれている ([1],[2],[3])。このような非正則性は、関数族を定義するパラメータが階層構造を持っているという、モデル固有の性質に根差している。3層ニューラルネットでは、1層目から2層目への結合と、2層目から3層目への結合がある意味で乗法的にモデルを定義している。このようなモデルでは、正解の関数がモデルよりも小さい中間素子数で実現可能ならば、真のパラメータは高次元集合をなし、識別不可能となる。

本論文は、このような、モデルの階層性に由来してパラメータが識別不能になる場合の最尤推定量の汎化誤

差を考察する。このような状況では、通常の漸近理論に基づく議論はすべて再検討を要するが、現在のところ、最尤推定量の漸近分布や汎化誤差については未解決の部分が多い。本論文では、最も簡単な階層型モデルである3層線形ネットワークに対して汎化誤差の期待値を求める。このモデルは単にニューラルネットの単純化であるだけでなく、統計学では縮小ランク回帰 ([4]) と呼ばれ広く用いられている。また、階層性に由来する識別不能性は、統計学でよく用いられるガウス混合モデルや、状況は異なるが ARMA モデルなどでも生じており、本論文は、このような場合の推定量の漸近的性質を明らかにする試みの第一歩である。

2 階層型モデルと識別可能性

一般に、3層ニューラルネットワークはパラメータ θ を持った関数族 $\{f(\cdot; \theta) : \mathbb{R}^L \rightarrow \mathbb{R}^M\}$ で、中間素子が H 個のとき、 $\theta = (w_{11}, \dots, w_{MH}, u_{11}, \dots, u_{HL})$ として

$$f^i(x; \theta) = \sum_{j=1}^H w_{ij} \varphi \left(\sum_{k=1}^L u_{jk} x_k \right), \quad (1 \leq i \leq M) \quad (1)$$

により定義される。ここで、 $\varphi(t)$ は1変数関数であり、 $\tanh(t)$ などがよく使われる。

本論文では、このようなモデルを用いて、入力ベクトル x から出力ベクトル y への関数関係を、ノイズを伴ったサンプル (学習データ) から推定する回帰問題を

*理化学研究所 脳科学総合研究センター. 〒 351-0198 埼玉県和光市広沢 2-1. tel. 048-467-9664, e-mail fuku@brain.riken.go.jp, Brain Science Institute, RIKEN, Hirosawa 2-1, Wako, Saitama, 351-0198, Japan

考察する。入力ベクトル x は確率 $q(x)dx$ に従うとする。推定の対象となる真の関数を $f(x)$ とするとき、入力 x に対する出力 y の条件付き確率 $p(y|x)$ は、

$$y = f(x) + z \quad (2)$$

により定まるとする。ここで、 z は出力に含まれるノイズで、平均0、分散共分散行列 $\sigma^2 I_M$ (I_M は M 次元単位行列) の正規分布 $N(0, \sigma^2 I_M)$ に従う確率変数とする。学習データ $\{(x^{(\nu)}, y^{(\nu)})\}_{\nu=1}^N$ は同時確率分布 $p(y|x)q(x)dx dy$ から独立なサンプルだと仮定する。

ニューラルネットを表わす統計モデルとしては

$$p(y|x; \theta)dx = N(f(x; \theta), \sigma^2 I_M) \quad (3)$$

を仮定する。ここでは簡単のため、出力に加えられるガウスノイズの分散を既知と仮定する。また、真の関数はモデルにより実現可能だと仮定し、真のパラメータを θ_0 で表わす。すなわち、 $f(x; \theta_0) = f(x)$ が成り立つ。

推定量として最尤推定量 (MLE) を扱うことにし、これを $\hat{\theta}$ で表わす。(3) 式のモデルのもとでは、最尤推定は最小2乗誤差推定に一致し、経験誤差

$$\mathcal{E}_{emp} = \sum_{\nu=1}^N \|y^{(\nu)} - f(x^{(\nu)}; \theta)\|^2. \quad (4)$$

を最小にする θ を用いることになる。

推定の精度は汎化誤差で測ることにする。本論文の目的は、MLE の挙動の一側面として、汎化誤差の期待値

$$\mathcal{E}_{gen} \equiv \mathbb{E}_{\{(x^{(\nu)}, y^{(\nu)})\}} [\int \|f(x; \hat{\theta}) - f(x)\|^2 q(x) dx] \quad (5)$$

を漸近的に計算することである。簡単にわかるように、 \mathcal{E}_{gen} は、期待対数尤度と

$$\begin{aligned} \mathbb{E}_{\{(x^{(\nu)}, y^{(\nu)})\}} [\int \int p(y|x)q(x)(-\log p(y|x; \theta)) dy dx \\ = \frac{1}{2\sigma^2} \mathcal{E}_{gen} + Const. \end{aligned} \quad (6)$$

なる関係で結ばれている。

階層型モデルの構造的な顕著な特徴は、設定したモデルよりも少ない中間素子数で真の関数を実現できる場合に、モデルのパラメータの中で真の関数を実現するものが高次元多様体を成すことである。図1からもわかるように、このような場合には、 $\varphi(0) = 0$ として、真の関数は、 $\{\theta \mid w_{i1} = 0 (\forall i), u_{1k} : \text{フリー}\}$ や $\{\theta \mid u_{1k} = 0 (\forall k), w_{i1} : \text{フリー}\}$ といった集合上で実現が可能となる。真の関数を実現するパラメータが1次元以上の多様体の和集合からなるときに、真のパラメータは識別不可能であるということにする。

通常の漸近理論は、正則条件として真のパラメータの識別可能性を要求しており、上述のような状況にはそのまま適用できない。このような場合には MLE は、真の関数を表わす高次元集合に漸近していくことになる。

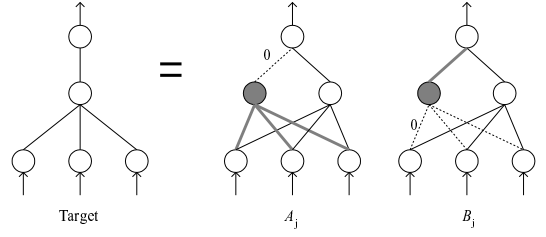


図1: 真のパラメータが識別不可能になる場合

3 線形ニューラルネットワーク

本論文では、階層性に由来するパラメータの識別不可能性を持つ最も簡単なモデルとして、線形ニューラルネットワークを考察の対象とする。中間素子を H 個持つ線形ニューラルネットワーク (LNN) とは、中間素子関数として恒等写像を持つ3層ニューラルネットのことであり、 $H \times L$ 行列 A と $M \times H$ 行列 B を用いて、

$$f(x; A, B) = BAx \quad (7)$$

によって定義される。ここで我々は

$$H \leq M \leq L \quad (8)$$

を仮定する。このとき $f(x; A, B)$ は \mathbb{R}^L から \mathbb{R}^M への線形写像となるが、(8) 式のランクの制限により、モデルは線形写像全体ではなく、ランクが H 以下の線形写像となる。したがって、このモデルで回帰問題を解くことは、単なる線形回帰問題を解くこととは異なっている。このようにランク制限がついた回帰は統計学では縮小ランク回帰 (reduced rank regression) と呼ばれ研究されている ([4])。

(7) 式のパラメータ表現は自明な冗長性を持っている。すなわち、任意の $H \times H$ 正則行列 G に対して、 $(A, B) \mapsto (GA, BG^{-1})$ は写像を変化させない。しかし、この冗長性は、 $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ と書いたとき、 A_1 を単位行列に正規化することによって除去することができる。もし、 BA のランクが H に一致するならば、この正規化によって (A_2, B) の表現は一意に定まる。したがって、このモデルのパラメータ数は $H(L + M - 1)$ に一致する。

簡単な考察により、この正規化を施されたパラメータ空間では、パラメータが識別不可能になることと、 BA のランクが H よりも小さいことが同値であることがわかる。したがって、正解の関数のランクが H に一致する場合には、正規化されたパラメータ空間の中では通常の漸近理論が成立し、この場合の汎化誤差の期待値は、よく知られているように、

$$\mathcal{E}_{gen} = \frac{\sigma^2}{N} \times H(L + M - H) + O(N^{-3/2}) \quad (9)$$

で与えられる。

4 線形ニューラルネットの汎化誤差

4.1 最尤推定量の汎化誤差

線形ニューラルネットに対しては、MLE が陽に書き下せる。以降では学習データを次のように表示する。

$$\begin{aligned} X &= (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})^T, \\ Y &= (\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(N)})^T, \\ Z &= (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)})^T. \end{aligned} \quad (10)$$

Proposition 1 ([5]). $Y^T X (X^T X)^{-1} X^T Y$ の固有値のうち、大きい方から H 個までの固有値に対応する固有ベクトルを並べた $M \times H$ 行列を V_H と書く。このとき、線形ニューラルネットワークの最尤推定量は、

$$\hat{B}\hat{A} = V_H V_H^T Y^T X (X^T X)^{-1} \quad (11)$$

により与えられる。

学習データにはノイズ Z が含まれているので、真のパラメータが識別不可能であっても、MLE は一意に定まることに注意する。この場合の MLE は、真の関数を与える高次元集合のまわりに分布することになる。

汎化誤差の期待値を求める準備として、Wishart 分布について簡単に復習する。 $n \times p$ 行列 S の各行が、 p 次元正規分布 $N(0, \Sigma)$ に従う i.i.d からなるとする。このとき、 $p \times p$ 確率行列 $S = W^T W$ の従う分布を Wishart 分布といい、 $W_p(n; \Sigma)$ と書く。Wishart 分布 $W_p(n; I_p)$ に従う確率行列 S の固有値を $\mu_1 \geq \dots \geq \mu_p \geq 0$ とするとき、その大きい方から q 個までの和の期待値を $\phi(p, n, q)$ で表わすことにする。すなわち、

$$\phi(p, n, q) = E[\mu_1 + \dots + \mu_q]. \quad (12)$$

このとき、線形ニューラルネットの汎化誤差について、以下の定理が成立する。

Theorem 1. 入力分布 $q(x)dx$ の分散共分散行列を正定値とし、真の関数のランクを $r (\leq H)$ とする。このとき、線形ニューラルネットワークの最尤推定量の汎化誤差の期待値は、次式で与えられる。

$$\begin{aligned} \mathcal{E}_{gen} &= \frac{\sigma^2}{N} \{r(L + M - r) + \phi(M - r, L - r, H - r)\} \\ &\quad + O(N^{-3/2}). \end{aligned} \quad (13)$$

(証明は付録 A を参照。)

$\phi(p, n, q)$ の値は、一般には簡単な表示が知られていない。そこで本論文では、期待値の積分が厳密に計算できる簡単なケースと、素子数を無限大にした場合の近似値とを議論し、真のパラメータの識別可能性が汎化誤差にどのような影響を及ぼすかを調べる。

4.2 中間素子が出力素子より 1 個少ない場合

$p = 2$ の場合、 $\phi(2, n, 1)$ は厳密に計算でき、次式で与えられる。($\Gamma(n)$ をガンマ関数。)

$$\phi(2, n, 1) = n + \sqrt{\pi} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \quad (14)$$

(証明略。固有値 $\nu_1 \geq \nu_2 \geq 0$ に対し、 $r = \frac{\nu_1 + \nu_2}{2}$ 、 $\sqrt{\nu_1 \nu_2} = \cos \omega$ と変数変換して $E[\nu_1 - \nu_2]$ を計算する)。

この結果から導かれる興味あるケースは、中間素子数 H が出力素子より 1 個だけ少なく、かつ正解のランクが H よりもさらに 1 だけ小さい場合である。

Theorem 2. $H = M - 1$ かつ $r = H - 1$ のとき、

$$\begin{aligned} \mathcal{E}_{gen} &= \frac{\sigma^2}{N} \{ (M - 1)(L + 1) - 1 + \sqrt{\pi} \frac{\Gamma(\frac{L-r+1}{2})}{\Gamma(\frac{L-r}{2})} \} \\ &\quad + O(N^{-3/2}) \end{aligned} \quad (15)$$

が成立する。

真のパラメータが識別可能、言い換えると $r = H$ であったとすると、(9) 式より、

$$\mathcal{E}_{gen} = \frac{\sigma^2}{N} (M - 1)(L + 1) + O(N^{-3/2})$$

を得る。 $\sqrt{\pi} \Gamma(\frac{L-r+1}{2}) / \Gamma(\frac{L-r}{2}) > 1$ ($L - r \geq 3$) であることから、汎化誤差の期待値は、同じモデルであっても真の関数に依存して異なる値をとり、しかも真のパラメータが識別不可能な場合のほうが大きい値になる。入力次元 L が非常に大きいとすると、Stirling の公式から、 σ^2/N の係数は、識別可能な場合に比べて $O(\sqrt{L})$ という極めて大きな増加を見せる。

4.3 大規模ネットワークの汎化誤差

次に、 L, M, H をすべて同じオーダーで無限大として、汎化誤差の期待値を近似する。 S を Wishart 分布 $W_p(n; I_p)$ に従う確率行列とし、 $\nu_1 \geq \nu_2 \geq \dots \geq \nu_p \geq 0$ を $n^{-1}S$ の固有値とすると、固有値の経験分布を

$$P_n \equiv \frac{1}{p} (\delta(\nu_1) + \delta(\nu_2) + \dots + \delta(\nu_p)), \quad (16)$$

により定義する。ここで $\delta(\nu)$ は一点 ν にのみ確率 1 を持つ Dirac 測度である。 P_n は以下の分布に収束することが知られている。

Proposition 2 ([6]). $0 < \alpha \leq 1$ なる α に対し、 $p/n \rightarrow \alpha$ を満たすように $n \rightarrow \infty, p \rightarrow \infty$ とすると、 P_n の分布関数は殆んどいたるところ

$$\rho_\alpha(u) = \frac{1}{2\pi\alpha} \frac{\sqrt{(u - u_-)(u_+ - u)}}{u} \chi(u) du \quad (17)$$

の分布関数に収束する。ここで $u_{\pm} = (\sqrt{\alpha} \pm 1)^2$ であり、 $\chi(u)$ は $[u_-, u_+]$ の特性関数を表わす。

$\rho_{\alpha}(t)$ は正規化された固有値の頻度分布であるから、大きい方から割合 β ($0 \leq \beta \leq 1$) の固有値の平均値を得るためには、まず β に対応する固有値 u_{β} を

$$\int_{u_{\beta}}^{u_+} \rho_{\alpha}(u) du = \beta$$

によって求め、 u_{β} から u_+ までの固有値の平均値

$$\int_{u_{\beta}}^{u_+} u \rho_{\alpha}(u) du \quad (18)$$

を計算すればよい。ここで $t = \left(u - \frac{u_- + u_+}{2}\right) / (2\sqrt{\alpha})$ と変数変換すると、 t の密度関数は

$$\nu_{\alpha}(t) = \frac{2}{\pi} \frac{\sqrt{1-t^2}}{2\sqrt{\alpha}t + 1 + \alpha}, \quad (19)$$

となる。 t_{β} を $\nu_{\alpha}(t)$ の β -パーセント点、すなわち

$$\int_{t_{\beta}}^1 \nu_{\alpha}(t) dt = \beta \quad (20)$$

と定めると、(18) 式を変数変換することにより

$$\lim_{\substack{n, p \rightarrow \infty \\ p/n \rightarrow \alpha}} \frac{1}{np} \phi(p, n, \beta p) = \frac{2}{\pi} \int_{t_{\beta}}^1 \sqrt{1-t^2} dt \quad (21)$$

を得る。したがって次の定理が得られた。

Theorem 3. 真の関数のランクを r ($r \leq H$) とする。 $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ なる α, β を固定し、 $\frac{M-r}{L-r} \rightarrow \alpha$ と $\frac{H-r}{M-r} \rightarrow \beta$ を満たすように L, M, H, r をすべて無限大に近づけると、

$$\begin{aligned} \mathcal{E}_{gen} &\sim \frac{\sigma^2}{N} \left\{ r(L+M-r) \right. \\ &\left. + (L-r)(M-r) \frac{1}{\pi} \left(\cos^{-1}(t_{\beta}) - t_{\beta} \sqrt{1-t_{\beta}^2} \right) \right\} \quad (22) \end{aligned}$$

と近似される。

t_{β} は陽に解けないが、微分法による初等的な議論により \mathcal{E}_{gen} は r の減少関数であることがわかる。すなわち、同一のモデルを用いた際、真の関数のランクが小さいほど汎化誤差の期待値は大きくなる。

5 計算機シミュレーション

前章の結果を数値的に検証するために計算機シミュレーションを行なった。モデルとして入力 50、出力 30、中間素子 20 個の線形ニューラルネットを固定し、真の関数のランクを 0 から 20 まで変化させて、MLE の汎化

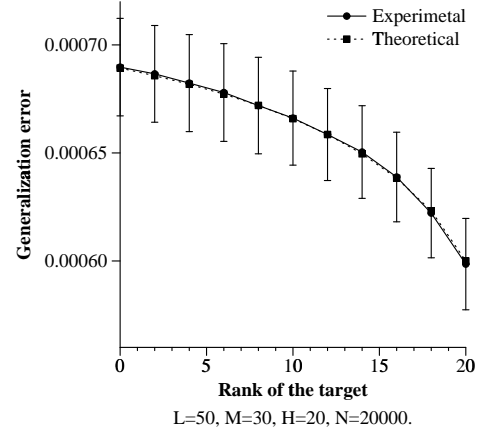


図 2: 正解のランクと汎化誤差

誤差を数値的に求めた。学習データは 20000 個を用い、100 回の試行の汎化誤差の平均とエラーバーを示したものが図 5 である。定理 3 の理論値と実験値は非常によい一致を示している。

本論文では、正解のランクがモデルのランクよりも低い場合を議論したが、現実の問題ではこの条件が完全に満たされることは稀で、むしろ、真の関数を表現するのにほとんど冗長なパラメータが存在している場合が多いと思われる。この場合には、厳密な意味では識別可能な場合の漸近理論が成立するが、漸近理論を適用するために非常に莫大なデータ数が必要となる可能性がある。もしそうであれば、現象を理解するには、真のパラメータが識別不能だと近似したほうがよいかもしれない。このような考察にもとづいて、「ほとんど識別不可能」なケースのシミュレーションを行なった。

モデルとして、2 入力、2 出力の線形ニューラルネットを用意し、真の関数として

$$f(x; \theta_0) = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix} (\varepsilon \ 0)x, \quad (23)$$

を用いた。ここで、 ε は微小な実数であり、 $\varepsilon = 0$ の時に限り真のパラメータが識別不可能となる。1000 個の学習データに対する 100 回の試行による汎化誤差の平均値を図 5 に示す。いまの場合、パラメータ 3 個に対して 1000 個のデータを使っているにも関わらず、小さい ε に対する汎化誤差は、識別可能な場合に漸近理論が与える理論値よりも大きく、むしろ識別不可能な場合の理論値 (図中 \times 印) に近い。このことは、識別不可能な場合の解析が、単に理論的な興味だけでなく、現実には生じる現象を把握する上でも重要であることを示唆している。

6 おわりに

本論文は、モデルの階層性から生じる識別不可能な場合の最尤推定量の挙動について議論するために、最も簡

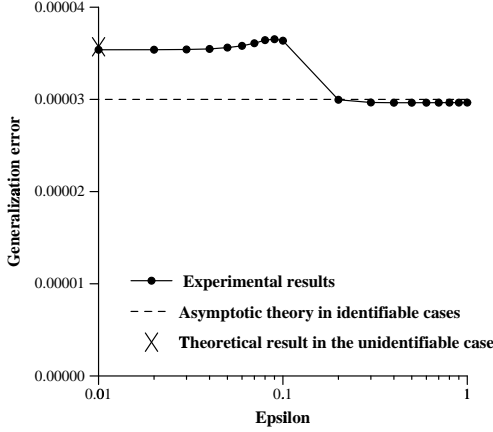


図 3: ほとんど識別不可能な正解に対する汎化誤差

単な階層型モデルである線形ニューラルネットの汎化誤差の期待値を求めた。その結果、真のパラメータが識別不可能な場合の汎化誤差の期待値は、識別可能な場合に通常の漸近理論から求められるものよりも大きくなり、正解のランクが小さいほど汎化誤差が劣化することが明らかとなった。ニューラルネット、混合モデルなど、階層的にパラメータを含むモデルは実際の問題によく応用されており、本論文の事実は、これら階層型モデルの推定量の挙動を再考する必要があることを教えている。

参考文献

- [1] K. Hagiwara, K. Kuno, & S. Usui, “Fisher 情報行列が縮退する場合のニューラルネットワークの学習誤差と汎化誤差について,” シンポジウム「統計的推測理論とその情報論的側面」予稿集, pp. 95-102, 1998.
- [2] K. Fukumizu, “Special statistical properties of neural network learning,” *Proc. NOLTA '97*, pp. 747-750, 1997.
- [3] 渡辺 澄夫, “ベイズ法による階層型統計モデルの推定誤差について,” 信学論 *J81-A*, vol. 10, pp. 1442-1452, 1998.
- [4] G. C. Reinsel & R. P. Velu, *Multivariate Reduced Rand Regression*, Springer: New York, 1998.
- [5] P.F. Baldi and K. Hornik, “Learning in linear neural networks: a survey,” *IEEE Trans. neural networks*, vol.6, no.4, p. 837-858, 1995.
- [6] K. Watcher, “The strong limits of random matrix spectra for sample matrices of independent elements,” *Ann. Prob.*, vol.6, no.1, pp. 1-18, 1978.
- [7] T. Kato, *Perturbation Theory for Linear Operators*, (2nd ed.) Springer: New York, 1976.

A Theorem 1 の証明

真の関数を定めるパラメータを $C_0 = B_0 A_0$ とし、 $\Sigma = E[xx^T]$ とおく。仮定より Σ は正定値である。

$$W = Z^T X (X^T X)^{-1/2} \quad (24)$$

とおくと、 W の各成分は独立に $N(0, \sigma^2)$ に従う。このとき、 $\hat{B}\hat{A} - C_0 = (V_H V_H^T - I_M)C_0 + V_H V_H^T W (X^T X)^{-1/2}$ となるので、

$$\begin{aligned} \mathcal{E}_{gen} = & E_{X,W}[\text{Tr}[V_H V_H^T W (X^T X)^{-1/2} \Sigma (X^T X)^{-1/2} W^T]] \\ & + E_{X,W}[\text{Tr}[C_0 \Sigma C_0^T (I_M - V_H V_H^T)]] \quad (25) \end{aligned}$$

と分解できる。

行列 $X^T X$ に関して、

$$(X^T X)^{1/2} = \sqrt{N} \Sigma^{1/2} + F, \quad X^T X = N \Sigma + \sqrt{N} K$$

と展開する。以下では簡単のため、 $\varepsilon = \frac{1}{\sqrt{N}}$ と書くことにする。このとき

$$T(\varepsilon) \equiv \frac{1}{N} Y^T X (X^T X)^{-1} X^T Y = T^{(0)} + \varepsilon T^{(1)} + \varepsilon^2 T^{(2)} \quad (26)$$

と摂動展開できる。ここに、

$$\begin{aligned} T^{(0)} &= C_0 \Sigma C_0^T \\ T^{(1)} &= C_0 K C_0^T + C_0 \Sigma^{1/2} W^T + W \Sigma^{1/2} C_0^T \\ T^{(2)} &= W W^T + W F C_0^T + C_0 F W^T \quad (27) \end{aligned}$$

である。 $T(\varepsilon)$ の固有空間は、 $C_0 \Sigma C_0^T$ の固有空間が (26) 式の摂動をうけたものである。以下では Kato ([7], Section II) に従い、 $T(\varepsilon)$ の固有値に対応する固有空間への射影子 (以下では固有射影子と呼ぶ) $P_j(\varepsilon)$ を計算する。

(26) 式の主要項 $C_0 \Sigma C_0^T$ のランクは r なので、この行列の正の固有値を $\lambda_1 \geq \dots \geq \lambda_r$ 、それぞれに対応する固有射影子を P_i ($1 \leq i \leq r$)、また、固有値 0 に対応する固有射影子を P_0 とおく。このとき、 $C_0 \Sigma^{1/2}$ の特異値分解からあきらかなように、 \mathbb{R}^L の互いに直交する 1 次元部分空間への射影子 Q_i ($1 \leq i \leq r$) が存在して、

$$\Sigma^{1/2} C_0^T P_i C_0 \Sigma^{1/2} = \lambda_i Q_i \quad (28)$$

とできる。また、次のように射影子 \tilde{Q} を定める。

$$\tilde{Q} = \sum_{i=1}^r Q_i. \quad (29)$$

まず、 λ_i が摂動を受けて生じた固有値を $\lambda_i(\varepsilon)$ ($1 \leq i \leq r$)、対応する固有射影子を $P_i(\varepsilon)$ とおくと、

$$P_i(\varepsilon) = P_i + O(\varepsilon) \quad (30)$$

である。

次に、 $C_0 \Sigma C_0^T$ の固有値 0 が分岐して生じた $T(\varepsilon)$ の固有値を $\lambda_{r+1}(\varepsilon), \dots, \lambda_M(\varepsilon)$ と書く。(26) 式より、確率 1 で $\lambda_{r+1}(\varepsilon) > \dots > \lambda_M(\varepsilon) > 0$ と仮定してよい。それぞれに対応する固有射影子を $P_{r+j}(\varepsilon)$ とし、

$$P_0(\varepsilon) = \sum_{j=1}^{M-r} P_{r+j}(\varepsilon) \quad (31)$$

とおく。 $P_{r+j}(\varepsilon)$ ($1 \leq j \leq M-r$) は、 $T(\varepsilon)P_0(\varepsilon)$ の 0 でない固有値の固有射影子なので、 $P_{r+j}(\varepsilon)$ の摂動展開を得るために、 $T(\varepsilon)P_0(\varepsilon)$ を

$$T(\varepsilon)P_0(\varepsilon) = \sum_{n=1}^{\infty} \varepsilon^n \tilde{T}^{(n)} \quad (32)$$

と展開する。このとき、 $\tilde{T}^{(n)}$ は $P_0, T^{(k)}$, および $I - P_0$ の像空間における $T^{(0)}$ の逆

$$S = \sum_{i=1}^r \lambda_i^{-1} P_i \quad (33)$$

を用いて陽に書くことができる。例えば

$$\begin{aligned} \tilde{T}^{(1)} &= P_0 T^{(1)} P_0 \\ \tilde{T}^{(2)} &= P_0 T^{(2)} P_0 - P_0 T^{(1)} P_0 T^{(1)} S - P_0 T^{(1)} S T^{(1)} P_0 \\ &\quad - S T^{(1)} P_0 T^{(1)} P_0 \end{aligned} \quad (34)$$

となる ($\tilde{T}^{(3)}$ については略。Kato [7], (2.20) を参照)。(28),(29),(33) 式より、

$$\Sigma^{1/2} C_0^T S C_0 \Sigma^{1/2} = \tilde{Q} \quad (35)$$

が成り立つことに注意する。

いま、 $T^{(0)}P_0 = 0$ と Σ の正定値性より $C_0 P_0 = 0$ であるので、さらに (35) 式を用いると、

$$\begin{aligned} \tilde{T}^{(1)} &= 0 \\ \tilde{T}^{(2)} &= P_0 W (I_M - \tilde{Q}) W^T P_0 \end{aligned} \quad (36)$$

を得る。したがって、 $P_{r+j}(\varepsilon)$ は

$$\frac{1}{\varepsilon^2} T(\varepsilon) P_0(\varepsilon) = \tilde{T}^{(2)} + \varepsilon \tilde{T}^{(3)} + \varepsilon^2 \tilde{T}^{(4)} + \dots \quad (37)$$

の固有空間となる。 W は各成分独立に $N(0, \sigma^2)$ に従うが、 $P_0, I_M - \tilde{Q}$ がそれぞれ $M-r, L-r$ 次元の一定の部分空間への射影子であることから、 $\tilde{T}^{(2)}$ は Wishart 分布 $W_{M-r}(L-r; \sigma^2 I_{M-r})$ に従っている。

$P_{r+j}(\varepsilon)$ を

$$P_{r+j}(\varepsilon) = P_{r+j} + \varepsilon P_{r+j}^{(1)} + \varepsilon^2 P_{r+j}^{(2)} + O(\varepsilon^3) \quad (38)$$

と展開すると、 $P_{r+j}^{(n)}$ は $\tilde{T}^{(k)}$ を使って具体的に表現できる (Kato [7], (2.14) 参照)。この具体的な表現を使う

と、 $\tilde{T}^{(2)}$ の正の固有値を $\eta_1 \geq \dots \geq \eta_{M-r}$ とするとき、

$$\begin{aligned} \text{Tr}[C_0 \Sigma C_0^T P_{r+j}^{(1)}] &= 0 \\ \text{Tr}[C_0 \Sigma C_0^T P_{r+j}^{(2)}] &= \frac{1}{\eta_j^2} \text{Tr}[C_0 \Sigma C_0^T (I - P_0) \tilde{T}^{(3)} P_{r+j} \tilde{T}^{(3)} (I - P_0)] \end{aligned} \quad (39)$$

を得る。さらに $\tilde{T}^{(3)}$ の具体的な表示 ([7], (2.20)) から、

$$\begin{aligned} \text{Tr}[C_0 \Sigma C_0^T P_{r+j}^{(2)}] &= \text{Tr}[(T^{(1)} P_0 T^{(2)} - T^{(1)} P_0 T^{(1)} S T^{(1)}) \\ &\quad P_{r+j} (T^{(2)} P_0 T^{(1)} - T^{(1)} S T^{(1)} P_0 T^{(1)}) S] \end{aligned} \quad (40)$$

が得られる。(27) 式より、

$$\begin{aligned} T^{(1)} P_0 T^{(2)} P_{r+j} - T^{(1)} P_0 T^{(1)} S T^{(1)} P_{r+j} \\ = C_0 \Sigma^{1/2} W^T P_0 W W^T P_{r+j} - C_0 \Sigma^{1/2} W^T P_0 W \tilde{Q} W^T P_{r+j} \\ = \eta_j C_0 \Sigma^{1/2} W^T P_{r+j} \end{aligned} \quad (41)$$

であるので、結局 (41),(40),(35) 式より

$$\begin{aligned} \text{Tr}[C_0 \Sigma C_0^T P_{r+j}^{(2)}] &= \text{Tr}[C_0 \Sigma^{1/2} W^T P_{r+j} W \Sigma^{1/2} C_0^T S] \\ &= \text{Tr}[P_{r+j} W \tilde{Q} W^T] \end{aligned} \quad (42)$$

となる。 W の各成分が正規分布に従うことと、 \tilde{Q} と $I_M - \tilde{Q}$ が直交することより、 P_{r+j} と $W \tilde{Q} W^T$ は独立である。したがって (25) 式の第 2 項は

$$\begin{aligned} \sum_{j=H+1-r}^{M-r} \text{E}_{X,W} [\text{Tr}[P_{r+j} W \tilde{Q} W^T]] + O(\varepsilon^3) \\ = \sigma^2 \varepsilon^2 r (M - H) + O(\varepsilon^3) \end{aligned} \quad (43)$$

に一致する。

一方、(25) 式の第 1 項は

$$\begin{aligned} \varepsilon^2 \text{E}_{X,W} [\sum_{i=1}^r \text{Tr}[P_i W W^T] \\ + \sum_{j=1}^{H-r} \text{Tr}[P_{r+j} W W^T]] + O(\varepsilon^3) \end{aligned} \quad (44)$$

に一致する。 P_i は定行列であり W の各成分は独立に $N(0, \sigma^2)$ に従うので、

$$\text{E}_{X,W} [\sum_{i=1}^r \text{Tr}[P_i W W^T]] = \sigma^2 r L \quad (45)$$

となる。また、

$$\begin{aligned} \text{Tr}[P_{r+j} W W^T] \\ = \text{Tr}[P_{r+j} W \tilde{Q} W^T] + \text{Tr}[P_{r+j} (W W^T - W \tilde{Q} W^T)] \\ = \text{Tr}[P_{r+j} W \tilde{Q} W^T] + \eta_j \end{aligned} \quad (46)$$

であるが、 η_j が $W_{M-r}(L-r, \sigma^2 I_{M-r})$ の大きい方から j 番目の固有値であることから、

$$\begin{aligned} \text{E}_{X,W} [\sum_{j=1}^{H-r} \text{Tr}[P_{r+j} W W^T]] = \sigma^2 \{r(H-r) \\ + \phi(M-r, L-r, H-r)\} \end{aligned} \quad (47)$$

を得る。(43),(45),(47) 式により定理は証明された。□