

# 識別不能性を持つモデルにおける最尤推定量の挙動

福水 健次

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,

E-mail: fuku@brain.riken.go.jp, <http://www.islab.brain.riken.go.jp/~fuku>

## 概要

パラメトリック推定における最尤推定量は、漸近的に正規分布に従うことが知られているが、ニューラルネットワーク、混合分布など階層的なパラメトライゼーションを持つモデルでは漸近理論の正則条件のひとつであるパラメータの識別可能性が必ずしも満足されず、最尤推定量の漸近的挙動が明らかではない。本研究では、3層線形ニューラルネット(縮小ランク回帰)において、パラメータが識別不能な場合に、最尤推定量の期待対数尤度の期待値を求めた。その結果、正則条件が成立する場合とは異なり、期待対数尤度が真のランクに依存することがわかった。

## 1 はじめに

パラメトリック推定における最尤推定量の挙動は、漸近的に正規分布に従うことが知られており、それに基づいた統計的手法が、モデル選択をはじめとして数多く用いられている。しかしながら、応用上よく用いられるモデルの中には、漸近理論の前提となる正則条件が満たされないものが存在し、そういった場合の最尤推定量の挙動については未解決な部分が多い。

例えば、近年広く用いられるようになった、多層パーセプトロンなどの階層型ニューラルネットは、パラメトリックな非線形回帰と捉えることができるが、1層目から2層目への結合と、2層目から3層目への結合が乗法的にモデルを定義しているため、正解の関数がモデルよりも少ない中間素子数で実現可能ならば、真のパラメータは識別不能となり、漸近理論は修正を迫られる。そういった場合の最尤推定量の挙動を解析しようとする試みは行なわれている([1])が、最尤推定量の漸近的挙動は、完全には解明されていない。

パラメータの階層的構造に由来する識別不能性は、ニューラルネットに限らず、混合モデル、縮小ランク回帰、若干状況は異なるがARMAモデルなど、広範囲に存在している。本論文は、このような、パラメータが識別不能になる場合の最尤推定量の挙動を考察する第一歩として、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、縮小ランク回帰と同一のものである。

## 2 階層型モデルと識別可能性

### 2.1 回帰問題におけるニューラルネットワーク

3層ニューラルネットについて簡単に説明する。中間素子を  $H$  個持つ3層ニューラルネットとは、パラメータ  $\theta = (v_1, \dots, v_H, w_1, \dots, w_H)$  を持った関数族  $\{f(\cdot; \theta) : \mathbb{R}^L \rightarrow \mathbb{R}^M\}$  で、

$$f(x; \theta) = \sum_{j=1}^H v_j \varphi(x; w) \quad (1)$$

により定義される。ここで、 $\varphi(x; w)$  は  $L$  次元パラメータ  $w$  を持つ  $L$  変数関数であり、 $\tanh(w^T x)$  などがよく使われる。

本稿では、このようなモデルを、入力変数  $x$  から出力変数  $y$  への条件付期待値を推定する回帰問題に用いる場合を考察する。入力変数  $x$  は確率  $q(x)dx$  に従い、 $x$  に対する出力変数  $y$  は、

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

により定まるとする。ここで、 $f(x)$  は推定対象となる真の関数であり、 $z$  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  $p(y|x)q(x)dx dy$  からの独立なサンプルと仮定する。したがって、ニューラルネットが表現する条件付確率のモデルは

$$p(y|x; \theta) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2} \|y - f(x; \theta)\|^2\right) \quad (3)$$

となる。本稿では簡単のため、ノイズの分散  $\sigma$  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  $\theta_0$  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、

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

を最小にする。(4) 式を経験誤差と呼ぶ。推定の精度は、汎化誤差の期待値である

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

で測ることとする。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

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

なる関係で結ばれているので、期待対数尤度の期待値を漸近的に考察していることになる。

(1) 式のような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、パラメータは識別不能となり、真の関数を実現するパラメータが高次元多様体を成すことである。図 1 からわかるように、モデルよりもひとつ少ない中間素子数によって真の関数を実現できる場合には、

$$\{\theta \mid v_1 = 0, v_2 = \zeta, w_1 : \text{フリー}\} \quad \text{または} \quad \{\theta \mid w_1 = w_2 = u, v_1 + v_2 = \zeta\}$$

といった 1 次元以上の連続集合上で真の関数を実現可能となる。

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。このような識別不能性は、ニューラルネットに限らず、様々なモデルで見受けられる。例えば、ガウス混合分布で、結合の係数と、各ガウス分布のパラメータの両方が変化し得ると、ニューラルネットとほぼ同じ識別不能性が生じる。また、ARMA で零点と極の位置が一致する場合にも同様の識別不能性が存在する。

## 2.2 線形ニューラルネットワーク

本論文では、識別不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 (LNN) あるいは縮小ランク回帰を考察の対象とする。中間素子を  $H$  個持つ LNN とは、 $\varphi(x; w) = w^T x$

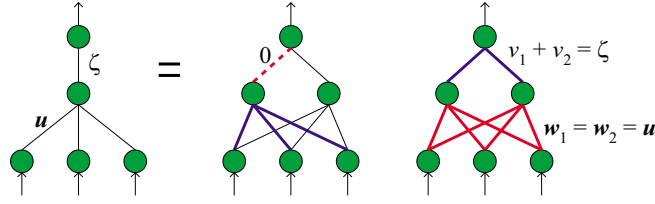


図 1: 真のパラメータが識別不能になる場合 ( 左 : 真の関数 , 右 2 つ : モデルによる実現 )

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、条件 (8) により、モデルはランクが  $H$  以下の線形写像全体となる ( 縮小ランク回帰 )。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。

(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。

### 3 線形ニューラルネットの汎化誤差

#### 3.1 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 が陽に解ける。以降では学習データを次のように表す。

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

命題 1.  $Y^T X (X^T X)^{-1} X^T Y$  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能であっても、MLE は一意に定まる。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。

Wishart 分布  $W_p(n; I_p)$  に従う確率行列  $S$  の固有値を  $\mu_1 \geq \dots \geq \mu_p \geq 0$  とし、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。すなわち、 $\phi(p, n, q) = \mathbb{E}[\mu_1 + \dots + \mu_q]$ 。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。

定理 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。

$$\mathcal{E}_{gen} = \frac{\sigma^2}{N} \{r(L + M - r) + \phi(M - r, L - r, H - r)\} + O(N^{-3/2}). \quad (12)$$

(略証を付録に与える。)

$\phi(p, n, q)$  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、ある条件の下でこれを計算し、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。

### 3.2 中間素子が出力素子より 1 個少ない場合

$p = 2$  の場合、 $\phi(2, n, 1)$  は初等的に計算でき、 $\Gamma(n)$  をガンマ関数として、 $\phi(2, n, 1) = n + \sqrt{\pi} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}$  と与えられる。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。

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

$$\mathcal{E}_{gen} = \frac{\sigma^2}{N} \left\{ (M - 1)(L + 1) - 1 + \sqrt{\pi} \frac{\Gamma(\frac{L-r+1}{2})}{\Gamma(\frac{L-r}{2})} \right\} + O(N^{-3/2}) \quad (13)$$

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、 $\sigma^2/N$  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。

### 3.3 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。Wishart 分布  $W_p(n; I_p)$  に従う確率行列  $S$  に対し、 $n^{-1}S$  の固有値  $\nu_1 \geq \nu_2 \geq \dots \geq \nu_p \geq 0$  の経験分布を

$$P_n \equiv \frac{1}{p} (\delta(\nu_1) + \delta(\nu_2) + \dots + \delta(\nu_p)) \quad (14)$$

により定義する。 $\delta(\nu)$  は 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。

命題 2 ([2]).  $0 < \alpha \leq 1$  なる  $\alpha$  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

$$\rho_\alpha(u) = \frac{1}{2\pi\alpha} \frac{\sqrt{(u - u_-)(u_+ - u)}}{u} \chi(u) du \quad (15)$$

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  $\beta$  ( $0 \leq \beta \leq 1$ ) の固有値の平均値を得るためには、まず  $\beta$  に対応する固有値  $u_\beta$  を  $\int_{u_\beta}^{u_+} \rho_\alpha(u) du = \beta$  によって求め、 $u_\beta$  から  $u_+$  までの固有値の平均値  $\int_{u_\beta}^{u_+} u \rho_\alpha(u) du$  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

$$\nu_\alpha(t) = \frac{2}{\pi} \frac{\sqrt{1-t^2}}{2\sqrt{\alpha}t + 1 + \alpha}, \quad (16)$$

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

$$\int_{t_\beta}^1 \nu_\alpha(t) dt = \beta \quad (17)$$

と定めると、変数変換により次の定理を得る。

**定理 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  $\alpha, \beta$  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、

$$\mathcal{E}_{gen} \sim \frac{\sigma^2}{N} \left\{ r(L+M-r) + (L-r)(M-r) \frac{1}{\pi} \left( \cos^{-1}(t_\beta) - t_\beta \sqrt{1-t_\beta^2} \right) \right\} \quad (18)$$

と近似される。

$t_\beta$  は陽に解けないが、微分法により初等的に  $\mathcal{E}_{gen}$  が  $r$  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。

## 4 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。入力 50、出力 30、中間素子 20 個の線形ニューラルネットを用意し、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図 4 左に示した。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、微小な特異値をもつ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となるが、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  $f(x; \theta_0) = \begin{pmatrix} \varepsilon \\ 0 \end{pmatrix} (\varepsilon 0)x$  を用いた。ここで、 $\varepsilon$  は微小な正数であり、 $\varepsilon = 0$  の時に限り真のパラメータが識別不能となる。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 4 右に示す。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  $\varepsilon$  に対する汎化誤差は、むしろ識別不能な場合の理論値 ( 図中  $\times$  印) に近い。このことは、識別不能な場合の解析が、単に理論的な興味だけでなく、現実を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。

## 5 おわりに

本論文は、識別不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

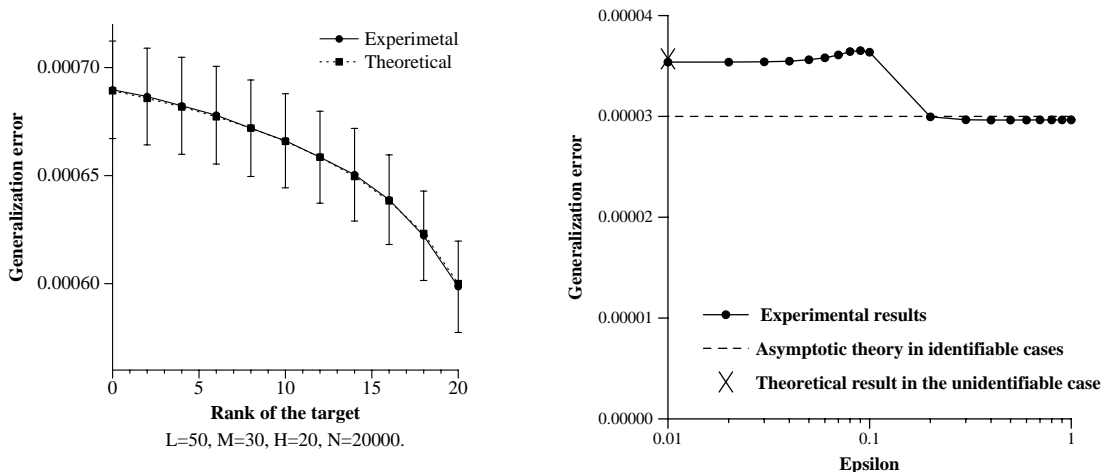


図 2: 計算機シミュレーション：正解のランクと汎化誤差の関係（左図）および、ほとんど識別不可能な正解に対する汎化誤差（右図）

能な場合の汎化誤差の期待値は、識別可能な場合に通常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. W. Wachter, “The strong limits of random matrix spectra for sample matrices of independent elements,” *Ann. Prob.*, vol.6, no.1, pp. 1-18, 1978.
- [3] T. Kato, *Perturbation Theory for Linear Operators*, (2nd ed.) Springer: New York, 1976.

## A 定理 1 の略証

真の関数を定める行列を  $C_0 = B_0 A_0$  とし、 $\Sigma = E[xx^T]$  とおく。仮定より  $\Sigma$  は正定値である。

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

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、

$$\mathcal{E}_{gen} = E_{X,W} [\text{Tr}[V_H V_H^T W (X^T X)^{-1/2} \Sigma (X^T X)^{-1/2} W^T]] + E_{X,W} [\text{Tr}[C_0 \Sigma C_0^T (I_M - V_H V_H^T)]] \quad (19)$$

と分解できる。

行列  $X^T X$  に関して、

$$(X^T X)^{1/2} = \sqrt{N}\Sigma^{1/2} + F, \quad X^T X = N\Sigma + \sqrt{N}K \quad (20)$$

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

$$T(\varepsilon) \equiv \frac{1}{N}Y^T X(X^T X)^{-1}X^T Y = T^{(0)} + \varepsilon T^{(1)} + \varepsilon^2 T^{(2)} \quad (21)$$

と摂動展開できる。ここに、

$$\begin{aligned} T^{(0)} &= C_0 \Sigma C_0^T, & T^{(1)} &= C_0 K C_0^T + C_0 \Sigma^{1/2} W^T + W \Sigma^{1/2} C_0^T \\ T^{(2)} &= W W^T + W F C_0^T + C_0 F W^T \end{aligned} \quad (22)$$

である。 $T(\varepsilon)$  の固有空間は、 $C_0 \Sigma C_0^T$  の固有空間が (21) 式の摂動を受けたものである。以下では Kato ([3], 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。

(21) 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  $\lambda_1 \geq \dots \geq \lambda_r$ 、対応する固有射影子を  $P_i$  ( $1 \leq i \leq r$ )、固有値 0 に対する固有射影子を  $P_0$  とおく。このとき、 $C_0 \Sigma^{1/2}$  の特異値分解から、 $\mathbb{R}^L$  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、

$$\Sigma^{1/2} C_0^T P_i C_0 \Sigma^{1/2} = \lambda_i Q_i \quad (23)$$

とできることがわかる。また、次のように射影子  $\tilde{Q}$  を定める。

$$\tilde{Q} = \sum_{i=1}^r Q_i. \quad (24)$$

まず、 $\lambda_i$  の摂動による固有値を  $\lambda_i(\varepsilon)$  ( $1 \leq i \leq r$ )、対応する固有射影子を  $P_i(\varepsilon)$  とおくと、

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

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。(21) 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、

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

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

$$T(\varepsilon)P_0(\varepsilon) = \sum_{n=1}^{\infty} \varepsilon^n \tilde{T}^{(n)} \quad (25)$$

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

$$S = \sum_{i=1}^r \lambda_i^{-1} P_i \quad (26)$$

を用いて陽に書くことができる。例えば

$$\tilde{T}^{(1)} = P_0 T^{(1)} P_0, \quad \tilde{T}^{(2)} = P_0 T^{(2)} P_0 - P_0 T^{(1)} P_0 T^{(1)} S - P_0 T^{(1)} S T^{(1)} P_0 - S T^{(1)} P_0 T^{(1)} P_0$$

となる ( $\tilde{T}^{(3)}$  については略。Kato [3], (2.20) を参照)。 (23), (24), (26) 式より、

$$\Sigma^{1/2} C_0^T S C_0 \Sigma^{1/2} = \tilde{Q} \quad (27)$$

が成り立つことに注意する。

いま、 $T^{(0)}P_0 = 0$  と  $\Sigma$  の正定値性より  $C_0P_0 = 0$  であるので、さらに (27) 式を用いると、

$$\tilde{T}^{(1)} = 0, \quad \tilde{T}^{(2)} = P_0W(I_M - \tilde{Q})W^TP_0$$

を得る。したがって、 $P_{r+j}(\varepsilon)$  は

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

の固有空間となる。 $W$  は各成分独立に  $N(0, \sigma^2)$  に従うが、 $P_0, I_M - \tilde{Q}$  はそれぞれ  $M - r, L - r$  次元の定部分空間への射影子なので、 $\tilde{T}^{(2)}$  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

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

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 [3], (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、

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

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

$$\text{Tr}[C_0\Sigma C_0^T P_{r+j}^{(2)}] = \text{Tr}[(T^{(1)}P_0T^{(2)} - T^{(1)}P_0T^{(1)}ST^{(1)})P_{r+j}(T^{(2)}P_0T^{(1)} - T^{(1)}ST^{(1)}P_0T^{(1)})S] \quad (28)$$

が得られる。(22) 式より、

$$T^{(1)}P_0T^{(2)}P_{r+j} - T^{(1)}P_0T^{(1)}ST^{(1)}P_{r+j} = \eta_j C_0 \Sigma^{\frac{1}{2}} W^T P_{r+j} \quad (29)$$

が得られるので、結局 (29),(28),(27) 式より

$$\text{Tr}[C_0\Sigma C_0^T P_{r+j}^{(2)}] = \text{Tr}[C_0\Sigma^{1/2}W^T P_{r+j}W\Sigma^{1/2}C_0^T S] = \text{Tr}[P_{r+j}W\tilde{Q}W^T] \quad (30)$$

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 (19) 式の第 2 項は

$$\sum_{j=H+1-r}^{M-r} \text{E}_{X,W}[\text{Tr}[P_{r+j}W\tilde{Q}W^T]] + O(\varepsilon^3) = \sigma^2 \varepsilon^2 r(M - H) + O(\varepsilon^3) \quad (31)$$

に一致する。

一方、(19) 式の第 1 項は

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

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、

$$\text{E}_{X,W}[\sum_{i=1}^r \text{Tr}[P_i W W^T]] = \sigma^2 r L \quad (32)$$

となる。また、

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

であるが、 $\eta_j$  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、

$$\text{E}_{X,W}[\sum_{j=1}^{H-r} \text{Tr}[P_{r+j}W W^T]] = \sigma^2 \{r(H - r) + \phi(M - r, L - r, H - r)\} \quad (33)$$

を得る。(31),(32),(33) 式により定理は証明された。□