ベイズ決定理論に基づく予測の漸近評価について
～特定不能性を有する階層モデル族への拡張～

知識情報処理研究　　3603R038-4　宅味丈夫
指　　導　松嶋敏泰教授

On the Asymptotics of the Prediction based on Bayes Decision Theory
～An Extension to Non-identifiable Hierarchical Model Class～

by Takeo Takumi

1　はじめに

与えられたデータから次に発生する未知のデータを予測する問題は統計理論における主なテーマであり，工学のみならず，自然科学・社会科学など非常に広い分野において研究がなされている．例えば，株価や需要などの時系列予測や音声・画像認識などのパターン認識問題など予測問題と捉えられる対象は多岐にわたる．

統計的な予測問題においては，まず，データの背後にそれらを発生させる未知の真の確率分布が存在すると考える．ただし，これを全く未知な対象とみるのではなく，何らかの確率構造を仮定する場合が多い．ここで，確率分布の形状（モデル）は既知でパラメータを未知とする問題設定，さらに，真のモデルは未知であるが，モデルの属する集合（モデル族）のみは既知とする問題設定が代表的である．後者は前者の一般化となっており，本研究では後者の場合を対象として考えていくものとする．

このような問題設定のもとで，様々な予測手法が考えられるが，その一つとして，ベイズ決定理論に基づく予測（ベイズ最適な予測）がある．この手法はパラメータ及びモデルの事後分布で平均化した予測分布とよばれる分布を計算することで達成される．

ベイズ最適な予測においては，任意のデータ数に対して事前分布に対する平均的な意味での最適性が保証される．しかし，実際には一つの真の確率分布からデータが発生すると考えれば，予測分布が真の確率分布に近いかどうかで，予測の精度が変わってくると考えられる．従って，予測分布が真の確率分布に近づくのか，また近づくとしたらどの程度の速さで近づいていくのかについての評価が必要である．

このような課題については，従来より漸近評価，すなわちデータ数が大きくなっていくときの評価が様々な形で行われている [1][2][3]．情報処理技術が発達した現在，データマイニング技法が脚光を浴びるなど，データ数が大量に入手できる場面が多くなってきている．そのため，漸近的状況を考えることが現実的にも重要であるといえる．

本研究では，階層構造を持ったモデル族（階層モデル族）におけるベイズ最適な予測の漸近評価について考えていく．階層モデル族とは，モデル間に包摂関係が成立するようなモデル族のことであり，マルコフモデルや重回帰モデルさらには混合正規分布やニューラルネットワークなど実用上有用な多くの統計モデルを含んでいる．

マルコフモデルや重回帰モデルなどの階層モデル族においては，各モデルについてパラメータと分布の一対一対応関係が保証される．このような階層モデル族を特定可能な階層モデル族と定義する．このとき，パラメータの事後確率密度が漸近的に正規分布に従うという性質（漸近正規性）を用いることでベイズ最適な予測について漸近評価することが可能である [2]．

しかし，混合正規分布・ニューラルネットワークなどの階層モデル族においては，あるモデルを，それに対して過大なパラメータ数を持つモデルで表現したときにパラメータと分布との一対一対応関係が成り立たない．このような階層モデル族を特定不能性を有する階層モデル族と定義する．このとき，パラメータの事後確率密度の漸近正規性が成り立たず，従来と同等の方法で評価を行うことは不可能となる．

そこで，本研究では，新たに [3] で紹介された手法を応用することで，特定不能性を有する階層モデル族に拡張してベイズ最適な予測の漸近評価を行う．

2　予測問題

母集団から観測された長さ n のデータ系列 $x^n = \{x_1, \cdots, x_n\}$ が与えられたもとで次のデータ x_{n+1} を予測する問題を考える[1]．

$\mathcal{X}$ を離散あるいは連続の値をとるデータ空間とし，$x_i \in \mathcal{X}(i = 1, \cdots, n+1)$ とする．また，x^n のデータ空間としては $\mathcal{X}^n$ と表現する．

$x \in \mathcal{X}$ は未知の真の確率分布 $p^*(x)$ に従って発生するものとする．ここで，$m \in \mathcal{M}$ をモデルを表す離散ラベル，$\theta_m \in \Theta_m$ を m における d_m 次元のパラメータとして，$p(x|\theta_m, m)$ と表される確率モデルについて考える．ここで，$\mathcal{M}$ は有限個のラベル集合，Θ_m は d_m 次元ユークリッド空間の部分集合とする．このとき，真のモデルを表すラベル m^*，m^* のもとでの真のパラメータ $\theta^*_{m^*}$ が存在して，$p^*(x) = p(x|\theta^*_{m^*}, m^*)$ と表されるものとする．ただし，m^*，$\theta^*_{m^*}$ はともに未知である．

[1] 本研究における結果は，説明変数と被説明変数がある場合にも容易に拡張可能である．

3 ベイズ決定理論に基づく予測

まず，予測に対する良し悪しを測る尺度として損失関数を定義する．損失関数は目的に応じて様々なものが用いられるが，以下が代表的である．ここで，$\hat{x}$ を x の予測値，$\hat{p}(x)$ を $p(x)$ の予測値とする[2]．

2 乗誤差損失

$$Loss_1 = (x_{n+1} - \hat{x}_{n+1})^2. \tag{1}$$

0-1 損失

$$Loss_2 = \begin{cases} 0, & x_{n+1} = \hat{x}_{n+1} \\ 1, & x_{n+1} \neq \hat{x}_{n+1} \end{cases}. \tag{2}$$

対数損失

$$Loss_3 = \log \frac{p(x_{n+1}|x^n, \theta^*_{m^*}, m^*)}{\hat{p}(x_{n+1}|x^n)}. \tag{3}$$

また，損失関数に対しデータの真の分布について期待値をとったリスク関数を定義する．

$$\begin{aligned} &Risk \\ &= \int_{\mathcal{X}^n} \left\{ \int_{\mathcal{X}} Loss \times p(x_{n+1}|x^n, \theta^*_{m^*}, m^*) dx_{n+1} \right\} \\ &\qquad \times p(x^n|\theta^*_{m^*}, m^*) dx^n. \end{aligned} \tag{4}$$

予測は損失関数もしくはリスク関数を最小にするように構成したいが，全てのモデル，パラメータに対し，一様に最小にすることは不可能である．ベイズ規準においては，リスク関数に対し，パラメータ及びモデルの事前分布で期待値をとった以下のベイズリスク関数を最小化することを考える．

$$BR = \sum_{m \in \mathcal{M}} \int_{\Theta_m} Risk \times w(\theta_m|m) d\theta_m \times P(m). \tag{5}$$

ベイズリスク関数を最小にする予測をベイズ最適な予測と定義する．ここで，各損失関数に対するベイズ最適な予測は以下のように与えられる．

2 乗誤差損失

$$\begin{aligned} \hat{x}_{n+1} = \int_{\mathcal{X}} x_{n+1} \Bigg\{ &\sum_{m \in \mathcal{M}} \int_{\Theta_m} p(x_{n+1}|x^n, \theta_m, m) \\ &\times w(\theta_m|m, x^n) d\theta_m \times P(m|x^n) \Bigg\} dx_{n+1}. \end{aligned} \tag{6}$$

0-1 損失

$$\begin{aligned} \hat{x}_{n+1} = \arg \max_{x_{n+1} \in \mathcal{X}} \Bigg\{ &\sum_{m \in \mathcal{M}} \int_{\Theta_m} p(x_{n+1}|x^n, \theta_m, m) \\ &\times w(\theta_m|m, x^n) d\theta_m \times P(m|x^n) \Bigg\}. \end{aligned} \tag{7}$$

対数損失

$$\begin{aligned} \hat{p}(x_{n+1}|x^n) = &\sum_{m \in \mathcal{M}} \int_{\Theta_m} p(x_{n+1}|x^n, \theta_m, m) \\ &\times w(\theta_m|m, x^n) d\theta_m \times P(m|x^n). \end{aligned} \tag{8}$$

以下，(8) の $\hat{p}(x_{n+1}|x^n)$ を予測分布とよぶ．ここで，(6),(7) には，この予測分布が含まれていることがわかる．すなわち，ベイズ最適な予測を導出するうえで予測分布が中心的な役割を持っていることがわかる．

4 評価基準

ベイズ最適な予測においては，任意のデータ数に対して事前分布に対する平均的な意味での最適性が保証される．しかし，実際には一つの真の確率分布からデータが発生すると考えれば，予測分布が真の確率分布に近いかどうかで，予測の精度が変わってくると考えられる．従って，予測分布が真の確率分布に近づくのか，また近づくとしたらどの程度の速さで近づいていくのかについての評価が必要である．

この課題について，本研究では，予測分布の真の分布に対する対数損失 (3) を評価基準として考える[3]．そして，データ数が十分に大きい場合における漸近評価について考えていく．

ここで，

$$\begin{aligned} \hat{p}(x^n) = &\sum_{m \in \mathcal{M}} \int_{\theta_m \in \Theta_m} p(x^n|\theta_m, m) \\ &\times w(\theta_m|m) d\theta_m \times P(m), \end{aligned} \tag{9}$$

とすると，

$$\log \frac{p^*(x_{n+1}|x^n)}{\hat{p}(x_{n+1}|x^n)} = \log \frac{p^*(x^{n+1})}{\hat{p}(x^{n+1})} - \log \frac{p^*(x^n)}{\hat{p}(x^n)}, \tag{10}$$

なる関係が成立するため，まず，$\log p^*(x^n) - \log \hat{p}(x^n)$ を評価すればよいことがわかる．

5 階層モデル族

$m \in \mathcal{M}$ によって表現できる確率分布の集合を，

$$\mathcal{H}_m = \{p(\cdot|\theta_m, m) | \theta_m \in \Theta_m\}, \tag{11}$$

とする．ここで，階層モデル族とは，$d_{m_1} < d_{m_2} < \cdots$ を満たすモデル $m_1, m_2, \cdots \in \mathcal{M}$ に対して，

$$\mathcal{H}_{m_1} \subset \mathcal{H}_{m_2} \subset \cdots, \tag{12}$$

なる包含関係が成立するモデルの集合を指す．この関係は $\mathcal{M}$ 内で全順序関係であっても半順序関係であってもよいものとする．

また，$p^*(\cdot) \in \cup_{m \in \mathcal{M}} \mathcal{H}_m$ を仮定しているので，

$$m_0 = \arg \min_m \{d_m | \exists \theta_m, p(\cdot|\theta_m, m) = p^*(\cdot)\}, \tag{13}$$

[2] 予測問題の中には，x の分布 $p(x)$ を予測する問題も存在する．

[3] 損失でなくリスクを評価基準と考えることもできる．

とすることで，$\theta^*_{m_0}$ が存在して，$p^*(\cdot) = p(\cdot|\theta^*_{m_0}, m_0)$ と表すことができる．また，$\mathcal{H}_{m_0} \subseteq \mathcal{H}_m$ なる m が存在すれば，θ^*_m が存在して，$p^*(\cdot) = p(\cdot|\theta^*_m, m)$ と表すこともできる．

例 1 (マルコフモデル). m を有限マルコフモデルの次数，θ_m をその遷移確率ベクトルすると，全順序の階層構造を持つモデル族となる． □

例 2 (重回帰モデル). m をどの説明変数を取り入れるかを表すラベル，θ_m を m における回帰係数・誤差分散とすると，半順序の階層構造を持つモデル族となる． □

例 3 (混合正規分布). 混合正規分布の確率分布は，

$$p(x|\theta_{m_K}, m_K) = \sum_{k=1}^{K} a_k f(x|b_k), \qquad (14)$$

の形で表現される．ここで，a_k は混合比であり，各要素 $f(x|b_k)$ はパラメータ $b_k = (\mu_k, \sigma_k)$ の正規分布である．これは，全順序の階層構造を持つモデル族となる． □

6 従来研究：特定可能な階層モデル族に対する漸近評価

各モデル m についてパラメータと分布の一対一対応関係が成立すると仮定する．すなわち，特定可能な階層モデル族を対象として考える．マルコフモデルや重回帰モデルは，特定可能な階層モデル族の代表例である．このとき，適当な正則条件[4]のもとで，事後確率密度の漸近正規性が成り立つ．これより，以下の結果が得られる．

補題 1. [2] 事後確率最大推定量を $\tilde{\theta}_m$ とする．このとき，

$$w(\tilde{\theta}_m|x^n, m) = \Big(\frac{n}{2\pi}\Big)^{\frac{d_m}{2}} \sqrt{\det I(\tilde{\theta}_m|m)} + o\big(n^{\frac{d_m}{2}}\big),\ a.s. \qquad (15)$$

となる[5]．ここで，$I(\theta_m|m)$ は m におけるフィッシャー情報行列である． □

補題 1 とベイズの公式より，次の定理を得る．

補題 2. [1][2] $\mathcal{H}_{m_0} \subseteq \mathcal{H}_m$ なる m において，

$$\log \frac{p^*(x^n)}{\hat{p}(x^n|m)} = \frac{d_m}{2}\log n + o(\log n), a.s. \qquad (16)$$

となる．また，それ以外の m において，

$$\log \frac{p^*(x^n)}{\hat{p}(x^n|m)} = C_m n + o(n),\ a.s. \qquad (17)$$

となる（C_m は m に依存して決まる定数）．ここで，

$$\hat{p}(x^n|m) = \int_{\Theta_m} p(x^n|\theta_m, m) w(\theta_m|m) d\theta_m, \qquad (18)$$

である． □

[4] 本論または文献 [2] 参照

[5] $a.s.$ は確率 1 で成立することを表す．

この結果は，あるモデル m を用いて予測を行った場合の累積損失を評価したものと解釈することができる．

まず，(16) は，真の分布に対してパラメータ数が過大なモデルを適用して予測を行った場合の累積損失の評価に対応している．ここで，右辺の結果は，データ数を n，モデル m のパラメータ数を d_m として，漸近的に $d_m/2 \cdot \log n$ が主要項になってくることを示している．

一方で，(17) は，真の分布に対してパラメータ数が過小なモデルを適用して予測を行った場合の累積損失の評価に対応している．ここで，右辺の結果は，漸近的に $C_m n$ が主要項になってくることを示している．

以上の議論から，真の分布を表す最小次数のモデル m_0 を用いて予測を行った場合に，漸近的に累積損失が最小になることがわかる．ただし，m_0 は事前には未知であるため，ベイズ最適な予測においてはモデルの事後分布について重み付けを行っている．ここで，定理 1 の結果を用いると，漸近的にモデルの事後分布が真の分布を表す最小次数のモデルに集中することを示すことができ，次の定理を得る．

定理 1. [2]

$$\log \frac{p^*(x^n)}{\hat{p}(x^n)} = \frac{d_{m_0}}{2}\log n + o(\log n),\ a.s. \qquad (19)$$

となる． □

これは，ベイズ最適な予測が，真の分布を表す最小次数のモデルを用いて予測を行った場合と漸近的に等価となることを示している．また，(10) より，ベイズ最適な予測分布の真の分布に対する損失は漸近的に d_{m_0}/n となることがわかる．すなわち，真の分布を表す最小次数のモデルのパラメータ数に依存して，収束の速さが決まることがわかる．

7 従来研究における問題点

混合正規分布やニューラルネットワークなどの階層モデル族においては，あるモデルを，それに対して過大なパラメータ数を持つモデルで表現したときにパラメータと分布との一対一対応関係が成り立たない．すなわち，特定不能性を有する階層モデルとなっている．逆に，あるモデル m に対して，真の分布がより少ないパラメータ数のモデルで表現可能な場合に，θ^*_m は一意に定まらない．

このとき，θ^*_m の各点においてフィッシャー情報行列のランクが縮退し，事後確率密度における漸近正規性は成立しない．そのため，補題 1 が破綻する．これより，従来の解析法を，パラメータが特定不能となる場合を含む一般の階層モデル族に対してそのまま適用することができないことがわかる．従って，別の解析法が必要となる．

例 4. 真の分布 $p^*(x) = f(x|b^*)$ を，過大なパラメータ数を持つモデル $p(x|a, b_1, b_2) = (1-a)f(x|b_1) + af(x|b_2)$ で表現した場合，真の分布と等価な分布を与えるパラメータ集合は，$\{a = 0, b_1 = b^*\} \cup \{a = 1, b_2 = b^*\} \cup \{b_1 = b_2 = b^*\}$ となる．これらの各点において，フィッシャー情報行列のランクが縮退することは容易に確認できる． □

8 主結果：特定不能性を有する階層モデル族に対する漸近評価

本研究においては，特定不能性を有する階層モデル族に対象を拡張して評価を行う．まず，各モデル m について次の条件が成り立つと仮定する．

条件 1 (事前分布に関する条件).
$\mathcal{H}_{m_0} \subseteq \mathcal{H}_m$ なる m において，θ_m^* の近傍が Θ_m に含まれる． □

条件 2 (対数尤度に関する条件).

1. $\psi(x,\theta_m) = \log p^*(x) - \log p(x|\theta_m, m)$ は θ_m について解析関数（高階微分可能でテイラー展開がある定義域で絶対収束する関数）であり，Θ_m を含む複素空間の開集合 Θ_m^c 上の解析関数に解析接続できる．
2. $$\int_{\mathcal{X}} \sup_{\theta_m \in \Theta_m^c} |\psi(x,\theta_m)|^2 p^*(x) dx < \infty, \qquad (20)$$
が成立する． □

注意 1. 例えば，混合正規分布において混合比が 0 になるような場合には，条件 2 を満たさないことが知られている．ただし，$\psi(x,\theta_m)$ の替わりに $\phi(x,\theta_m) = p(x|\theta_m,m)/p^*(x) - 1$ と置き換えることでこの問題を回避できることが示されている [4]． □

これらの仮定のもとに，[3] では $\mathcal{H}_{m_0} \subseteq \mathcal{H}_m$ なる m において，$\log p^*(x^n) - \log \hat{p}(x^n|m)$ に対する解析を行っている．本研究では，その解析結果を拡張することで，次の補題を得た．

補題 3. 条件 1，2 が成り立つとき，$\mathcal{H}_{m_0} \subseteq \mathcal{H}_m$ なる m において，

$$\log \frac{p^*(x^n)}{\hat{p}(x^n|m)} = \lambda_m \log n + o(\log n), \lambda_m \leq \frac{d_m}{2}, \ a.s. \qquad (21)$$

となる．また，それ以外の m においては，(17) となる．

（**証明**） 本論または [5] 参照． □

また，これを用いて以下の定理を得る．

定理 2. 条件 1，2 が成り立つとき，

$$\log \frac{p^*(x^n)}{\hat{p}(x^n)} = \lambda \log n + o(\log n), \lambda \leq \frac{d_{m_0}}{2}, \ a.s. \quad (22)$$

となる．

（**証明**） 本論または [5] 参照． □

さらに，以下の条件を仮定する．

条件 3.

$$\theta_{m_0} \neq \theta'_{m_0} \ \Leftrightarrow \ p(\cdot|\theta_{m_0}, m_0) \neq p(\cdot|\theta'_{m_0}, m_0), \quad (23)$$

が成り立つ． □

注意 2. 例えば，混合正規分布はパラメータの大小で順序をつけることで条件 3 を満たすことが確認できる． □

定理 3. 条件 1，2，3 が成り立つとき，

$$\log \frac{p^*(x^n)}{\hat{p}(x^n)} = \frac{d_{m_0}}{2} \log n + o(\log n), \ a.s. \qquad (24)$$

となる．

（**証明**） 本論または [5] 参照． □

9 考察

まず，補題 3 の補題 2 に対する相違点について考える．補題 3 では，対象となる累積損失に対して，その主要項が $d_m/2 \cdot \log n$ と同等もしくはそれ以下となることを示している．この事実は，フィッシャー情報行列が縮退することからも直感的に理解できる．ただし，λ_m がどこまで小さくなるかについての保証はない．

次に，補題 3 の結果を用いることで，定理 1 に対応する結果，定理 2 が得られる．これは，ベイズ最適な予測分布の累積損失に対して，その主要項が $d_{m_0}/2 \cdot \log n$ と同等もしくはそれ以下になることを示している．

さらに，条件 3，すなわち，真のモデルを表す最小次数のモデルにおいてパラメータと分布の一対一対応関係が成り立つ場合には，補題 3 における λ_m に対して，下界 $d_{m_0}/2$ が保証される．このとき，定理 3 を得るが，これは定理 1 と同様の結果となっている．

10 まとめ

本研究では，従来では扱われてこなかった特定不能性を有する階層モデル族においてもベイズ最適な予測分布が真の分布に漸近的に収束していくことを示した．また，収束速度も d_{m_0}/n もしくはそれ以上であることを示した（定理 2）．さらに，混合正規分布のように，真の分布を表す最小次数のモデルにおいてパラメータと分布の一対一対応関係が成り立つ場合に限定すれば，収束速度は従来と同様であることも示した（定理 3）．

なお，今後の課題としては，漸近式における $o(\log n)$ 項の詳細な解析について検討していきたい．

参考文献

[1] B. S. Clarke and A. R. Barron, “Information-theoretic asymptotics of bayes methods,” *IEEE Trans. on Inform. Theory*, vol.36, pp.453-471, 1990.

[2] M. Gotoh, T. Matsushima, and S. Hirasawa, “A generalization of B. S. Clarke and A. R. Barron’s asymptotics of bayes codes for FSMX sources,” *IEICE Trans. Fundamentals*, vol.E81-A, pp.2123-2132, 1998.

[3] S. Watanabe, “Algebraic analysis for non-identifiable learning machines,” *Neural Computation*, vol.13, pp.899-933, 2001.

[4] 渡辺澄夫，山崎啓介，青柳美輝,“混合正規分布の特異点の非解析性について,” 信学技法 NC, vol.50, pp.41-46, 2004.

[5] 宅味丈夫，須子統太，松嶋敏泰，“階層モデルにおけるベイズ予測の漸近評価に関する一考察，” 第 27 回情報理論とその応用シンポジウム予稿集, of , pp.639-642, 2004.