博士論文

ベイズ統計理論に基づく確率モデルの推定と予測の漸近的評価に関する研究

A Study on Asymptotic Analysis of Estimation and Prediction based on Bayesian Statistics

後藤正幸
Masayuki GOTOH

早稲田大学 大学院理工学研究科
機械工学専攻 経営システム工学専門分野

目次

第 1 章

序論

1.1 研究背景

刻々と得られるデータを逐次的に学習-予測したり，過去のデータから未来のデータを予測したり，あるいはデータの背後に存在する確率構造を推定する問題は統計理論における主なテーマといえる．統計理論は工学のみならず，農学，医学，心理学，社会学など非常に広い適用分野をもつ強力な解析手法を与えている．統計理論はその豊富な適用範囲もあり，古くから非常に広く研究されてきたが，最近では情報理論や学習理論との接点において新たな研究が展開され始めている．

統計理論では通常，未知の母集団(情報源)に対してある確率分布族(確率モデル族)を仮定する．そして，その中に真の確率分布，あるいは最適な確率分布[1]が存在するものとして議論を進める場合が多い[2]．このとき問題は，“得られたデータ系列から何からの基準で真の確率分布に近いモデルを得る” ということに帰着する．このような問題は，古くから統計的推定・予測問題として広く深く研究されているが，最近では情報理論や学習理論との接点からも様々な方向へと研究が進んでいる．中でも統計的推定と情報理論を関連づけた研究が 1980 年代から活発になり [1, 25, 26, 56, 64, 80, 82, 84, 93, 94, 95, 96, 100, 101]，いくつかの重要な成果を生んでいる．また，学習・予測理論においても，統計理論や情報理論に基づく体系的研究が行われるに至っており [155]，今も新たな研究の側面を切り開いている．

統計理論と情報理論の接点の一つは情報量という概念を用いた体系化である．情報量という概念は，Fisher 情報量や KL 情報量の議論などのように，古くから統計理論においても本質的であることが認識されている．統計理論におい

[1] 何について最適かについては，後の各章において定義する．

[2] 勿論，真の分布が無限の複雑さをもつ確率モデルでないと特定できないといったノンパラメトリックな推定問題などの定式化もある．

て対数損失を用いた理論展開がよく行われるが，これはこのような理論基盤に基づくものといえる．一方，情報量は情報源符号化問題の基礎を与えている．情報源符号化はデータ圧縮のことであり，データの確率的構造を利用して行われるものである．現在実際に使われている圧縮・解凍アルゴリズムのほとんどは，データの確率的構造を推定しながら，逐次的に圧縮を進めるというタイプの符号化である．この観点から見ると情報源符号化の問題も，観測されたデータからその背後に存在する構造を推定したり，未来のデータを予測するという問題に帰着する．品質管理や社会現象の解析などで用いられる統計解析はまさにこの問題を扱っており，とくに統計的モデル化は情報圧縮の理論と密接に関係することが認識されるようになっている．そのため，このような統計的推定や予測の問題をより一般的にとらえ，より広い視野から体系化する基礎研究が重要である．

本研究では，これらの統計理論，情報理論，学習理論の接点において，最近その重要性が再認識されるようになったベイズ統計理論[16, 35]に基づく推定方式を取り上げる．ベイズ統計が今もなおこれらの分野で盛んに研究されているのは，“有限長のデータ系列に対して何らかの意味で最適な推論を構成したい”，という共通の目的に対して，ベイズ決定理論が理論的にも実務的にも良く適合するためである．すなわち，これらは有限長のデータ系列に対してベイズ最適性が保証されている推定方式であり，真の確率モデルが事前分布に従って生じる場合の平均的な意味での最適解を与える．ベイズ統計に基づく統計的推定は古くから議論されているが，特に統計的モデル選択問題において，その理論的な整合性が認識されており，様々な形のベイズ統計理論に基づくモデル選択規準が提案されている．一方，1980年代後半から情報源符号化の分野において，ベイズ符号の研究が活発になってきており，新たな展開が期待されている．ベイズ統計理論に対しては事前確率の設定法や計算量などの課題もあるが，この点についても多くの研究がなされている．

1.2 研究目的

本研究はパラメータ推定問題や未知のデータの予測問題を対象としている．推定や予測という言葉は様々な意味で用いられるが，本研究ではベイズ決定理論で本質的となる損失関数の差異により，

(1) 推定問題：確率モデルを特定するパラメータと推定量間に損失関数を考える問題

(2) 予測問題：

(2-1) データ系列の逐次予測問題：次々に与えられるデータに対する逐次予測の累積予測誤差損失を考える問題

(2-2) 一時点の予測問題：一時点のデータの予測誤差損失を考える問題

のように分類する．このように分類することにより評価すべき損失関数が明確となり，同時に事前分布が仮定できればベイズ決定理論に基づく定式化が可能となる．

ここで，ベイズ最適な推測を方法的観点から以下の2つの場合に分類して考察してみる．

1 確率モデル族の中で最適解を選ぶ問題(統計的モデル選択)

2 確率モデル族の要素全ての混合モデル(予測分布)を用いることを許容する予測問題

1 には確率モデルのパラメータ推定問題も含まれるが，重回帰モデルの説明変数の選択などのように統計的モデル選択として最近でもさかんに議論されている問題を含んでいる．これは確率モデル族の中から適切なモデルを1つ選択する問題である．一方，2 は予測問題に対して，確率モデル上に定義された事前分布あるいは事後分布で平均化したモデル(混合モデル)を用いる方法である．1980年代後半から情報源符号化の分野において，この混合モデルを用いた符号化(ベイズ符号)が議論されるようになり，いくつかの興味深い性質が明らかにされている．

しかし，ベイズ解は事前分布あるいは事後分布に対する平均的な意味での最適性を保証するものである．したがって，“ベイズ最適である解が他の観点からどのような性質を持つか”を明らかにすることが，理論的にも実用的にも必要である．とくに，実際には未知である1つの真の確率構造からデータが発生すると考えれば，この真の確率構造にどの程度の速さで近づくか，という点についての議論が必要といえる．また，ベイズ規準でなくとも，事前分布を仮定するという意味でベイズ統計に基づいている方法も数多く提案されており，これらに本質的な差があるのかどうかを明らかにすることが必要である．さらに，目的に応じて損失関数を変えた場合にどのようなことが起こるか，についても詳しい検討が望まれる．これらの課題に対しては，多くの研究が現在もなされている．しかし，特定の基準と特定のモデル族に限定した結果である場合が多く，より

一般的な性質を体系的に解明しておくことは，ベイズ統計理論の理論的側面と応用的側面の双方の意味で重要となっている．より具体的な研究課題としては，

(1) 1つの確率モデルを選択する方法と混合モデルを用いる方法の差違はどの程度か?

(2) ベイズリスクはどのような値になるか?

(3) 真の確率分布に対してどの程度の推定精度を与えるか?

(4) ベイズ最適なモデル選択は真の確率分布，あるいは最適な確率分布を特定できるのか?

(5) 真の確率分布，あるいは最適な確率分布を選択できる確率はどの程度か?

(6) 目的に合った損失関数をどのように設定し，計算するか?

(7) 損失関数を変えた場合にどのようなことが起こるか?

などがあげられ，現在も多くの研究がなされている．しかし，これらの研究は特定の基準と特定のモデル族に限定した結果である場合が多く，より一般的な性質を解明しておくことは，ベイズ統計理論の理論的側面と応用的側面の双方の意味で重要となっている．

そこで，本研究ではベイズ統計理論に基づく推定方式に対し，漸近的側面から解析，評価を行うことにより，上記の(1)〜(7)の課題について新たな成果を得ることを目的とする．本研究で議論する漸近的評価とは，データ数が大きくなっていくときの性質を考えるという意味である．統計理論はそもそもデータから母集団の構造を推測しようとする目的があるので，データ数が増えたときに正しく母集団構造を特定できるのかという問題意識から，古くから漸近理論が議論されている．さらに情報処理技術が発達した現在，データマイニング技術が脚光を浴びるなど，データ数は大量に採取できる場面が多くなってきており，そのような漸近的状況を考えることが現実問題に対しても有用となってきている．またデータ圧縮の問題などにおいても，大容量のデータファイルを圧縮したいという問題意識があるので，漸近的評価は非常に重要といえる．

ベイズ統計に基づく推測の漸近的な評価が目的であるので，まず第一に推測で考えている損失関数による評価を考えるべきである．第3章，第4章，第5章，第8章では主にこの立場から，ベイズ統計理論に基づく推測の漸近的な評価を行っている．また第二として，ある損失関数に対してベイズ最適に構成された

推測方法が他の評価基準に対してどのような性能をもつか，という点を考えることも，統計的推測の漸近的性質を解明するという意味で価値がある．第6章，第7章では主にこの立場から，推定や予測を目的としてベイズ最適に構成されたモデル選択の性質として，真のモデルを選択できない選択誤りを取り上げて解析している．

漸近的評価の際，従来と異なる解析の切り口として確率モデルのパラメータの事後確率密度が漸近的に正規分布に慨収束することに着目する．事後確率密度の漸近正規性についてはすでにいくつかの研究があるが，それぞれが個別に議論されている．またその漸近正規性自体が興味の対象となっており，そこから導かれる多くの性質に関してはあまり議論されていない．本研究では，ある一般的な条件のもとで事後密度の漸近正規性を導き，これがベイズ推定の漸近的性質において本質的な役割を演ずることを示す．事後確率密度がサンプル数に対してどのような速さで正規分布に近づくかは，確率モデルの複雑さやFisher情報行列に依存する．しかし，真の確率分布と確率モデルを固定した議論では，サンプル数を大きくすることによりいずれ漸近正規性が成り立ち，これが漸近的特性において本質的な役割を演ずる．すなわち，事後密度の漸近正規性を直接用いることにより，ベイズ統計に基づく推定と予測の漸近性質について多くの結果が(直感的にも見通しよく)導かれるということが，本論文の主な主張の1つである．

1.3 本論文の構成

本論文は以下のように9章から構成されている．

まず，第2章において，本論文で使う統計的な推定と予測の数理的な問題設定について述べ，従来研究をまとめる．さらに，本研究の位置づけ，および結果の新規性と有効性について，その概要を述べる．

第3章，第4章においては，データ系列の逐次学習・予測問題に対して，対数損失関数を考えた場合を取り扱う．これは工学的には情報源符号化の問題と等価となる．J.Rissanen の記述長最小原理(minimum description length: MDL)に基づく規準(以下，MDL規準)[56, 95, 155]が暗に事前分布を仮定していることから，ベイズ統計の枠組みで議論できることを指摘し，ベイズ最適解(ベイズ規準)[25, 80, 100]との差を累積対数損失の面から解析する．これにより，逐次予測問題における一般的な母集団に対し，MDL原理に基づきモデルを1つ選択する方法と混合モデルを用いる方法の差を定量的に示す．前者をMDL規準，後者を

ベイズ規準と呼ぶ．その際，第3章では両者で事前分布が同じ場合，第4章では両者で事前分布が異なる場合を取り扱い，累積対数損失の面からベイズ規準のMDL規準に対する優位性を明らかにする．

第5章では，1990年の定常独立な(independently identically distributed: i.i.d.)パラメトリック情報源に対する B.S.Clarke と A.R.Barron の成果[25]を拡張し，情報源符号化問題への応用上極めて重要な FSMX 情報源[70, 81]に対して，ベイズ規準による逐次予測の累積対数損失，及びKL情報量をリスク関数としたときの漸近式を導出する．これにより，Kullback−Leibler(KL)情報量という一般に広く使われている尺度を用いて，ベイズ規準に基づき混合分布を用いる方法の漸近的評価式が得られる．その漸近式は Clarke と Barron の結果を階層モデル族に拡張したものであり，損失の最小値に対するベイズ解の収束速度が従来よりも精密な形で与えられたことになる．

第6章では統計的モデル選択問題[74]をベイズ決定理論の枠組みから取り扱う．まず，統計的モデル選択問題をベイズ決定理論に基づいて定式化し，これがある損失関数のクラスに対しては一致性，すなわち選択されたモデルは漸近的に最適モデルに収束する性質を持つことを数学的に証明する．この結果，目的が真のモデルの発見にある場合でも，未来のデータの予測にある場合でも，ベイズ決定理論に基づくモデル選択は漸近的に最適なモデルを選択することができることを示す．一致性は統計的モデル選択問題で従来から議論されてきた一つの評価規準であり，ベイズ決定理論に基づくモデル選択の一つの漸近的性能の評価が得られたことになる．

第7章では，真のモデルの推定を目的としたベイズ決定理論に基づくモデル選択を扱う．まず，一般的な条件のもとで，真のモデルを選択できない選択誤り確率の上界を導出する．これにより，真のモデルを特定するという目的に対して，その選択誤り率という基準での評価が与えられたことになる．さらに，ベイズ決定理論に基づくモデル選択の応用例として母集団のもつパラメータの変化位置検出問題を取り上げ，ベイズ最適解を導出し，具体的問題におけるベイズ的モデル選択の有効性を示す．

第8章では，J.Rissanen によって提案された確率的コンプレキシティ(stochastic complexity:SC) [99, 100]の概念の拡張形である，拡張確率的コンプレキシティ(extended stochastic complexity:ESC) [153, 154]の性質を解析する．SCはベイズ最適な符号(ベイズ符号)の符号長を統計的複雑さを測る尺度と考えたものであり，ESCはその損失関数を一般形に拡張した規準である．したがって，第3章 ∼ 第5章が累積対数損失を考えていたのに対し，本章では一般の累積損失を適用した

逐次予測問題を扱っている．ESCの漸近式の導出については，この分野の未解決問題となっており[154]，本論文でその一部が示されたことになる．

最後に，第9章において以上の成果をまとめ，結論と今後の課題について述べる．

第 2 章

問題設定 – 従来研究と課題

2.1 統計理論と情報理論と学習理論

本研究では，ベイズ統計理論に基づく推測の問題を扱っている．その理論自体は統計理論の一部であると考えられるが，実は統計理論は情報理論や学習理論と密接に関係している．

情報理論における情報源符号化の問題を取ってみても，それを統計的推定とみなすこともできれば，未知の情報源に対する学習を行っているとみなすこともできる．実際，統計理論の漸近論において，KL情報量やFisher情報量が本質的となるように，漸近論においては，統計理論と情報理論の扱っている問題はほとんど同じといえる．しかし例えば，実際には統計理論では少数サンプルからの推定など，有限時点を議論しようとする思想も強い．一方で情報理論では対象となるデータは膨大であることが多いので，少数サンプルを扱おうとする考え方はほとんどない．

このような点で同じような問題を扱っていても，考え方や立場の違いが生み出す差異があるのは事実である．しかし，これらをより大きな枠組みからとらえ，より体系的な考察を進めていくことは，本質を議論するためにも必要なことであると考えられる．例えば，J.Rissanenが統計的推測のための規準として持ち込んだ記述長最小原理(MDL原理)は，統計理論と情報理論の狭間に位置するものであり，統一的な観点から議論されるべきである．そこで本研究では，情報源符号化の問題を統計的予測問題の一例とみなすなど，統計的決定理論の枠組みから情報源符号化の問題，統計的モデル選択の問題，逐次的な学習問題をとらえ，統一的な観点から体系的に考察している．

言葉の定義では，統計理論と情報理論，学習理論で違いがあるが，最初はなるべく統計理論の言葉を用いて議論を行う．しかし，記述長最小原理の説明などの際には，それが符号長を統計的推測の規準とするものであるために，情報

源符号化の立場からの説明が必要である．また，第8章の拡張確率的コンプレキシティは統計的学習理論における理論的成果であるので，ここでは学習理論の用語が必要になる場合もある．したがって，以後では必要な際には情報理論や学習理論の用語も併用して説明を加えるが，混同が起きない様になるべく両方の用語を併記する形をとる．

2.2 表記法の定義

$\mathcal{X}$ を離散あるいは連続のデータ空間，すなわち確率変数 X の値域とする．これは情報理論の言葉では情報源アルファベットと呼ばれる．母集団(情報源)から観測された長さ n のデータ系列(観測系列)を $x^n = x_1 x_2 \cdots x_n$ と表記する．ただし，$x_i \in \mathcal{X}, (i = 1, 2, \cdots)$ である．また，情報源からの無限系列を x^∞ とする．$x_i \in \mathcal{X}, i = 1, 2, \cdots, n$ から構成される全ての x^n の集合を $\mathcal{X}^n$ $(n = 1, 2, \cdots, \infty)$ とし，$x^1 = x$ と定義する．x^n も確率変数の実現値であり，この $\mathcal{X}^n$ 上の確率変数を X^n とする．以下の章では，例えば回帰モデルのように，説明変数を X，目的変数を Y のように記述することがあるが，この場合は確率変数 $Z = (X, Y)$ に関して同様の表記，すなわち 長さ n の直積で定義される確率変数 Z^n, Z^n の実現値(長さ n のデータ系列) z^n, Z^n の値域 $\mathcal{Z}^n$ などを用いることにする．

本論文を通して，$P(\cdot)$ は確率を表し，$f(\cdot)$ は確率密度を表すものとする．また，確率と確率密度を区別せずに両方の意味で用いるときは，$p(\cdot)$ を用いる．データ系列は確率分布 $p^*(x^n) = p^*(X^n = x^n), x^n \in \mathcal{X}^n$ に従って発生するものとし，この真の確率分布 $p^*(\cdot)$ は事前には未知であるとする．本論文を通じ，右上添字の $*$ は，データ系列を発生させる真のモデルや分布，あるいはモデル集合内で何らかの意味で最適なモデルやパラメータに対して用いる．後で出てくるように，例えばモデル m の事前確率を $P(m)$ として，$P(x^n) = \sum_m P(x^n|m)P(m)$ いうような記述を用いるが，本来正確を期するためにはこれらを $P_{X^n}(x^n)$ や $P_M(m)$ と記述するべきである．しかし，解析において非常に煩雑な記述となってしまうため，本論文を通じてこれらの添字を省略し，$P(x^n), P(m)$ などと記述する．当然ながらこれらは，$P(\cdot)$ に x^n や m を代入した値ではなく，それぞれ引数をとる変数の確率を表すものとする．

とくに断りがない限り，対数の底は e とし，符号化の問題で扱う(理想)符号長の単位も nat とする．

$k \times 1$ 行列 (k 次元列ベクトル) x と $k \times k$ 行列 J に対して，ノルムを $\|x\| = (x^T x)^{1/2}$，および $\|x\|_J^2 = x^T J x$ と定義する．ここで，x^T は x の転置ベクト

ルである．また k 次元ベクトル θ^k に対して 2 階微分可能な関数 $g(\theta^k)$ について，$\frac{\partial g(\theta^k)}{\partial \theta^k}$ と $\frac{\partial^2 g(\theta^k)}{\partial \theta^k (\partial \theta^k)^T}$ をそれぞれ，$\frac{\partial g(\theta^k)}{\partial \theta_i}$ を第 i 要素としてもつ $k \times 1$ ベクトル，$\frac{\partial^2 g(\theta^k)}{\partial \theta_i \partial \theta_j}$ を i-j 要素としてもつ $k \times k$ 行列とする．これらを簡略的に $g'(\theta^k)$ や $g''(\theta^k)$ とも表す．また，x の関数 $g(x)$, x,y の関数 $h(x,y)$ に対し，$g(x)|_{x=x'}$ は $g(x')$ を，$g(x) = h(x,y)|_{y=y'}$ は $y = y'$ のときに等式 $g(x) = h(x,y')$ がなりたつことをそれぞれ意味するものとする．k 次元ユークリッド空間 $\mathcal{R}^k$ のある部分集合 $\mathcal{A} \subset \mathcal{R}^k$ に対して，その体積を $|\mathcal{A}|$ と記述する．また，可算集合 $\mathcal{B}$ に対して，$|\mathcal{B}|$ はその大きさ(要素数)を表すものとする．

2.3 対象とする確率モデル族

本研究では，広いクラスの確率モデル族として，離散モデル族，パラメトリックモデル族，階層(入れ子型)モデル族を取り上げる．

定義 2.1 (離散モデル族) [26] 確率モデルが離散ラベル λ で条件付けられる場合を考える．すなわち，確率分布モデルが $p(\cdot|\lambda)$ で与えられ，λ の集合 Λ は有限可算集合とする．離散集合で表される $\{p(\cdot|\lambda)|\lambda \in \Lambda\}$ を離散モデル族と定義する．本論文では，データ系列 x^n を発生する真の確率分布がこのモデルクラスに含まれていることを仮定しない．ただし，真の確率分布に対して，何らかの意味で最適なモデル λ^* がモデルクラス Λ に存在するものとする(何の意味で最適であるかについては，その都度定義する). □

例 2.1 (多項分布) 有限離散集合 $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の確率変数 X を考え，多項分布を仮定する．p_i をシンボル i, $i \in \mathcal{X}$, の生起確率と考える．$p_\beta = 1 - \sum_{i=0}^{\beta-1} p_i$ である．ベクトル $p^\beta = (p_0, p_1, \cdots, p_{\beta-1})^T$ は，一つの多項分布モデルを規定する．

モデル λ_1 は一つの多項分布モデル

$$p^\beta(\lambda_1) = (p_0(\lambda_1), p_1(\lambda_1), \cdots, p_{\beta-1}(\lambda_1))^T,$$

を意味するものとする．ただし，$0 < p_0(\lambda_1), p_1(\lambda_1), \cdots, p_{\beta-1}(\lambda_1) < 1$ かつ $0 < \sum_{i=1}^{\beta-1} p_i(\lambda_1) < 1$ である．同様に λ_2, λ_3, $\cdots$ を

$$p^\beta(\lambda_2) = (p_0(\lambda_2), p_1(\lambda_2), \cdots, p_{\beta-1}(\lambda_2))^T,$$

$$p^\beta(\lambda_3) = (p_0(\lambda_3), p_1(\lambda_3), \cdots, p_{\beta-1}(\lambda_3))^T,$$

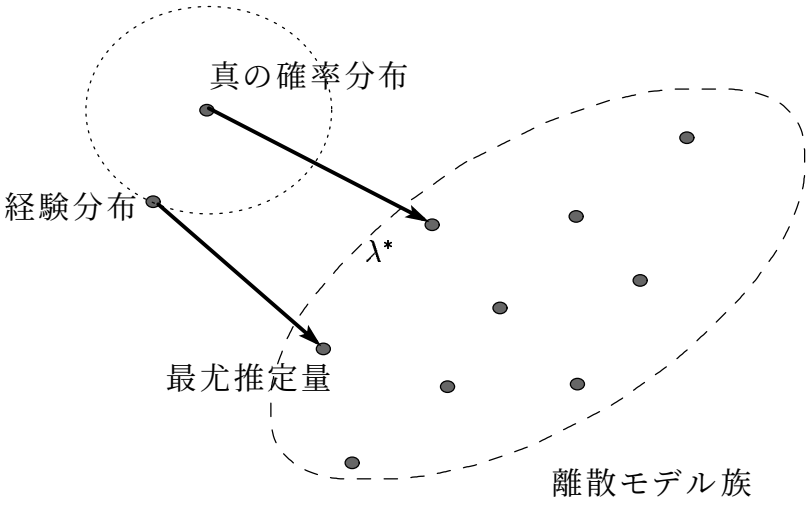

図 2.1: 離散モデル族のイメージ図

.....

として，離散モデルクラス $\Lambda = \{\lambda_1, \lambda_2, \cdots\}$ を構成することができる．例えば，$\beta = 2$ とし，各確率分布モデルを $p^2(\lambda_1) = (1/10, 1/10)^T$, $p^2(\lambda_2) = (1/10, 2/10)^T$, $p^2(\lambda_3) = (1/10, 3/10)^T$, $\cdots$, $p^2(\lambda_7) = (1/10, 8/10)^T$, $p^2(\lambda_8) = (2/10, 1/10)^T$, $p^2(\lambda_9) = (2/10, 2/10)^T, \cdots, p^2(\lambda_{36}) = (8/10, 1/10)^T$ と与えて，離散モデル族 $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_{36}\}$ を構成することができる． □

図2.1に離散モデル族の幾何的なイメージ図を示す．真の確率分布から何らかの規準で最も近いモデルが，その規準に対する最適モデル λ^* である．ちなみに得られたデータ系列の経験分布からKL-情報量で最も近いモデルが最尤推定量となる．

定義 2.2 (パラメトリックモデル族) 確率モデルが k 次元連続パラメータ $\theta^k = (\theta_1, \theta_2, \cdots, \theta_k)^T$ で特徴づけられる場合を考える．すなわち，確率分布モデルが $p(\cdot|\theta^k)$ で与えられ，θ^k は k 次元ユークリッド空間 $\mathcal{R}^k$ の部分集合 Θ^k の要素とする．連続パラメータ θ^k で特定される確率分布モデルの集合 $\{p(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ をパラメトリックモデル族と定義する． データ系列 x^n を発生する真の分布がこのモデルクラスに含まれていることは仮定しない．ただし，真の確率分布に対して，何らかの意味で最適なパラメータ θ^{k*} が Θ^k 内に存在するものとする．以後，これを最適パラメータとよぶ(何の意味で最適であるかについては，その都度定義する). □

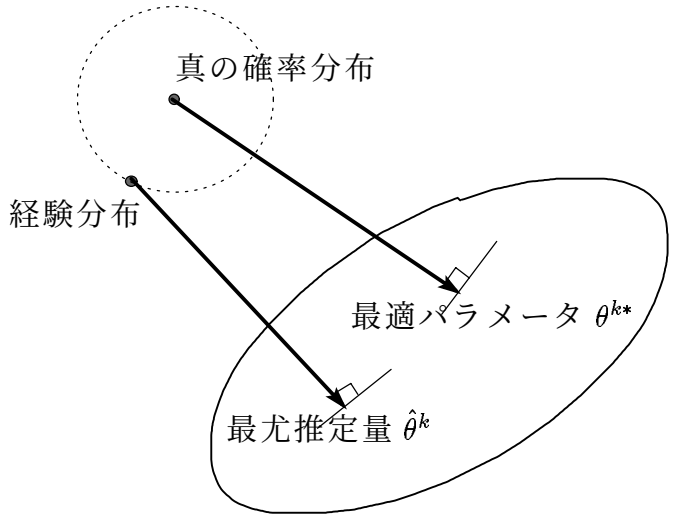

図 2.2: パラメトリックモデル族のイメージ図

例 2.2 (多項分布) 有限離散集合 $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の確率変数 X を考え，多項分布を仮定する．p_i をシンボル $i, i \in \mathcal{X},$ の生起確率と考える．$p_\beta = 1 - \sum_{i=0}^{\beta-1} p_i$ である．$k = \beta$ とし，ベクトル $\theta^k = (\theta_1, \theta_2, \cdots, \theta_\beta)^T = (p_0, p_1, \cdots, p_{\beta-1})^T$ は，一つの多項分布モデルを規定する．この θ^k を $\Theta^k = \{\theta^k \in [0,1]^k : 0 \leq \sum_{i=1}^{k} \theta_i \leq 1\}$ 内の連続パラメータと考えれば，多項分布族は生起確率を連続パラメータとしてもつパラメトリックモデル族である． □

例 2.3 (正規分布) 実数軸上の連続集合 $\mathcal{X} = \mathcal{R}^1$ 上の確率変数 X を考え，正規分布を仮定する．正規分布はその平均 μ，分散 σ^2 を考え，$\theta^k = (\mu, \sigma^2)^T, k = 2,$ とすれば，θ^k を連続パラメータとしてもつパラメトリックモデル族と考えられる． □

図2.2にパラメトリックモデル族の幾何的なイメージ図を示す．真の確率分布から何らかの規準で最も近い分布のパラメータが，その規準に対する最適パラメータ θ^{k*} である．ちなみに得られたデータ系列の経験分布からKL-情報量で最も近い分布のパラメータが最尤推定量となる．

定義 2.3 (階層モデル族) 確率モデルの構造を表す離散ラベル m の集合を $\mathcal{M}$ とする．この m を単にモデルと呼ぶ．各モデル m が k_m 次元連続パラメータ θ^{k_m} を持つ場合を考える．すなわち，確率分布モデルが $p(\cdot|m, \theta^{k_m})$ で

与えられ，θ^{k_m} は k_m 次元ユークリッド空間 $\mathcal{R}^{k_m}$ の部分集合 Θ^{k_m} の要素とする．モデル m と連続パラメータ θ^{k_m} で特定される確率分布モデルの集合 $\mathcal{H} = \{p(\cdot|m,\theta^{k_m})|m \in \mathcal{M}, \theta^{k_m} \in \Theta^{k_m}\}$ を階層モデル族と定義する．$\mathcal{H}^{k_m}$ をモデル m で表現できる確率分布のクラスとする．すなわち，$\mathcal{H}^{k_m} = \{p(\cdot|m,\theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}\}$ である．このとき階層モデル族 $\mathcal{H}$ は

$$\mathcal{H} = \cup_m \mathcal{H}^{k_m}, \tag{2.1}$$

と与えられる．

ここで，$k_{m1} < k_{m2} < k_{m3} \cdots$ をみたすモデル $m1, m2, m3, \cdots \in \mathcal{M}$ に対し，階層構造(入れ子構造)

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots\cdots, \tag{2.2}$$

があることを許容する．この階層関係は，$\mathcal{M}$ 内で全順序関係であっても，半順序関係であってもよい． □

データ系列 x^n を発生する真の分布がこのモデルクラスに含まれていることは仮定しない．ただし，真の確率分布に対して，モデル m に対し何らかの意味で最適なパラメータ $\theta^{k_m^*}$ が Θ^{k_m} 内に存在するものとする．これを以後，モデル m の最適パラメータとよぶ(何の意味で最適かについては，その都度定義する)．図2.3に全順序の階層モデル族の幾何的イメージを示す．各モデルは一つのパラメトリックモデル族を構成し，それらが入れ子構造をなしている．よって，真の確率分布に対して決まる最適なパラメータも，各モデルに対してそれぞれ存在することが分かる．一方，モデル m について考えてみると，$\mathcal{H}^{k_m}$ の入れ子構造により，$\mathcal{H}$ 内で最適な確率分布モデルが複数の $\mathcal{H}^{k_m}$ に含まれてしまう可能性が生じる．そこで，従来から統計的モデル選択問題で扱われてきた定義を採用し，それらの最適な確率分布モデルを含む m のうち，パラメータの次数が最小のものを最適なモデル m^* とする．これは，1) 同じデータ数でパラメータを推定すると次数が小さい方が推定精度が高いこと，2) 真の確率分布に対し，不必要なパラメータを含むモデルは構造解析などの立場から好ましくない，などの理由から受け入れられてきた定義である．

本研究においては，半順序の階層構造をもつモデル族も含めて階層モデル族と定義したが，とくに階層関係が全順序のとき入れ子型モデル族 (nasted model class) と呼ぶ．また解析の難しさの面で考えると，階層構造のない分離型モデル族は比較的容易であり，本研究の結果はこの分離型の場合を含むことが容易に分かる．このため，階層モデル族だけでなく分離構造を含むような広いクラスのモデル族に対して，本研究の結果を適用することができる．

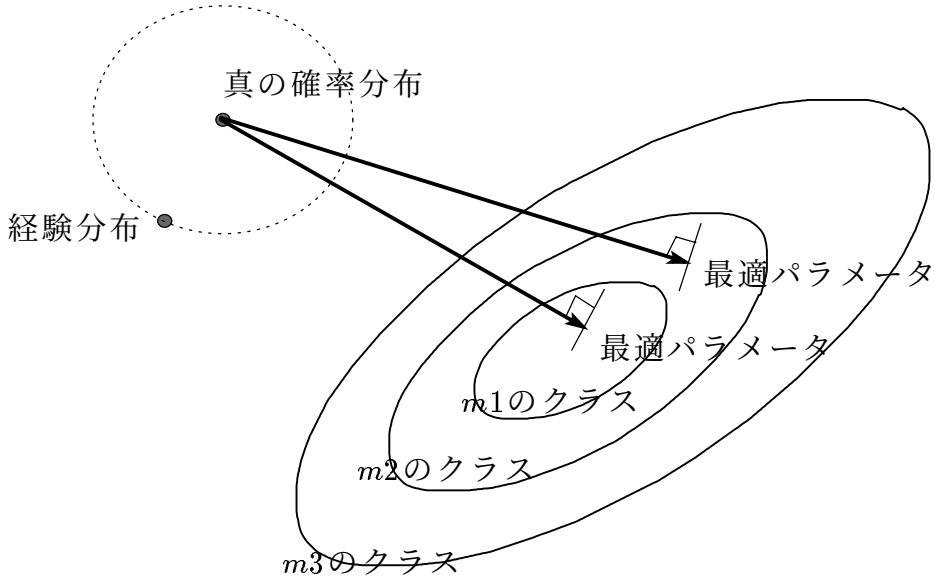

図 2.3: 階層モデル族のイメージ図

例えば，m を有限マルコフ情報源の次数，θ^{k_m} をその遷移確率ベクトルとすると全順序の階層構造をもつモデル族(入れ子型モデル族)となり，重回帰モデルでどの説明変数を取り入れるかを表すラベルを m とし，そのもとでの回帰係数，誤差分散を要素とするベクトルを θ^{k_m} とすれば，これは半順序の階層構造をもつモデル族となる．

例 2.4 (重回帰モデル) 説明変数の集合 $\{X_1, X_2, \cdots, X_r\}$ と目的変数 Y の線形回帰モデルを考える．実際の解析において，どの説明変数が効いているかは未知であり，説明変数の選択が行われる．説明変数の部分集合と目的変数の線形回帰モデルを一つのモデル m と考え，各モデルを

$$m_0 : \quad Y = \alpha_0 + \epsilon, \tag{2.3}$$

$$m_1 : \quad Y = \alpha_0 + \alpha_1 X_1 + \epsilon, \tag{2.4}$$

$$m_2 : \quad Y = \alpha_0 + \alpha_1 X_2 + \epsilon, \tag{2.5}$$

$$\cdots\cdots \qquad \cdots\cdots\cdots\cdots$$

$$m_{2^r-1} : \quad Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_r X_r + \epsilon, \tag{2.6}$$

とすると，モデルクラス $\mathcal{M} = \{m_0, m_1, \cdots, m_{2^r-1}\}$ が構成される．ただし，ϵ は平均 0,分散 σ^2 の正規分布とする．それぞれのモデル m は，その回帰係数 $\alpha_0, \alpha_1, \cdots$

と 誤差分散 σ^2 を連続パラメータとしてもつ．ここで，例えば m_0 は m_1 において $\alpha_1 = 0$ として得られるので，m_0 で表現される確率モデルのクラスは m_1 のそれに包含され，階層構造(入れ子構造)が存在する．この階層構造は，$\mathcal{M}$ 内で全順序関係にあるわけではなく，半順序であることは容易にわかる． □

2.4 統計的推定と予測

本研究では，モデル m, あるいはパラメータ θ^k, θ^{k_m} に対する推測(決定)[1]を目的とし，モデル間，パラメータ間に損失関数を設定している問題を推定問題と定義する．一方，対象とする確率変数 X, あるいは X^n に対する推測(決定)を目的とし，データ間あるいは分布間に損失関数を設定している問題を予測問題と定義する．

前者はモデル集合やパラメータ空間の中から一つを推定結果として算出する．一方，後者はサンプル空間 $\mathcal{X}$ や $\mathcal{X}^n$ の中の一点，あるいはその上に定義される確率分布を予測結果として算出する．したがって，予測と推定では考えている決定の定義域(決定空間)が異なり，予測の良さを測る尺度と推定の良さを測る尺度の定義域も異なっている．

以下では，本論文で扱う3つの問題，すなわち

(1) 推定問題: 損失関数 $L(m : Am)$

(2) 予測問題:

(2-1) データ系列 x^n の逐次予測:
損失関数 $\sum_{t=1}^n L(x_t : Ax)$, $\sum_{t=1}^n L(p(x_t|x^{t-1}, m) : Ap(x_t|x^{t-1}))$

(2-2) 一時点の x の予測:
損失関数 $L(x_{n+1} : Ax)$, $L(p(x_{n+1}, m|x^n) : Ap(x_{n+1}|x^n))$

について述べる．ただし，$L(m : Am)$ はモデル $m \in \mathcal{M}$ と決定 $Am \in \mathcal{M}$ 間の損失関数，$L(x_t : Ax)$ は $x_t \in \mathcal{X}$ と決定 $Ax \in \mathcal{X}$ 間の損失関数，$L(p(x_t|x^{t-1}, m) : Ap(x_t|x^{t-1}))$ は確率分布 $p(x_t|x^{t-1}, m) \in \mathcal{P}$ と決定 $Ap(x_t|x^{t-1}) \in \mathcal{P}$ 間の損失関数である．また，$\mathcal{P}$ は全ての確率分布の集合を表す．すなわち，確率モデルの離散パラメータの空間 $\mathcal{M}$ 上に損失関数が設定されている問題を推定問題，サンプル空間 $\mathcal{X}$ 上，あるいはその上の確率分布の集合 $\mathcal{P}$ 上に損失関数が設定されている問題を予

[1]未知の対象に対して，何らかの情報を与えることを推測という．統計的決定理論では決定と呼ぶ．また，決定の定義域を決定空間と呼ぶ．

測問題としている[2]．確率分布間の損失関数 $L(p(x_t|x^{t-1},m):Ap(x_t|x^{t-1}))$ は，代わりに対数損失 $L(x_t:Ap)=-\log Ap(x_t|x^{t-1})$ のように，$x_t\in\mathcal{X}$ と $Ap(\cdot|x^{t-1})\in\mathcal{P}$ の間に損失関数を考える場合も多い．この場合は，$p(\cdot|x^{t-1},m)$ で期待値を取れば問題は等価となる．問題をこのように損失関数で分類して考えることにより，ベイズ統計に基づく推測の損失関数，すなわち何に対してベイズ最適なのかが明確になる．さらに，その推測の第一に考えられるべき評価関数も明確になる．本研究はベイズ統計に基づく推測の漸近的評価を目的としているので，推測で考えている損失関数を明確に分類して考察することは，結果を体系的に考察するための一つの有効な視点を与えるといえる．

しかし，現実問題への応用場面で使われている方法では，上記3つの問題は必ずしも独立に考察されてはいない．損失関数を統計的推測[3]の目的と考えた場合には，それぞれは別々に議論されるべきものと考えられるが，実際には(2-1)と(2-2)の目的のために(1)の統計的推定が行われることも多い．すなわち，モデルやパラメータの推定は，実際には予測や制御のために行われるものであり，そのために最適な(真のモデルがモデルクラスに存在する場合は真の)モデルとパラメータを推定しておくのが望ましい，という考え方である．しかし本研究では，(1)は “最適なモデルとパラメータを推定することが目的”である場合として，(2-1)と(2-2)の問題と明確に切り分ける．このように統計的推測の目的を明確にすることで，客観的な規準が設定できるためである．ただし，(2-1)や(2-2)の予測を目的とした場合でも，モデル，パラメータの推定量を求め，それらを用いて予測を行う方法を考えることができる．この場合は，精度の高い予測を行うための推定法の構成が問題となり，評価関数は予測誤差損失で測ることができる．以上のように本研究では “漸近的評価” を目的としていることから，推測の評価基準としての損失関数によって問題を分類することにより，結果を体系的に考察している．

表2.1に本研究で扱う問題を分類した結果を示す．予測と推定は決定空間と損失の定義域の差異による分類である．さらに予測は，データ系列に対する累積損失(Cumulative loss)を考えるか，あるいは1時点の損失(Instantaneous loss)を考えるかによって分類される．推定問題で累積損失を考える場合は，(数学的には定式化できるが)現実的な応用例が明確でないため，本研究では扱わない．

[2]ここでは，確率分布が離散パラメータ m で条件付けられている場合を想定したが，これは対象とする確率モデル族で異なるので各章で正しく再定義する．

[3]統計的な推定と予測を合わせて統計的推測と呼ぶ．

表 2.1: 本研究で扱う推定問題と予測問題

		決定空間, 損失の定義域の差異	
		推定	予測
損失関数の差異	x^n に対する累積損失 Cumulative loss		(2-1)
	1時点の損失 Instantaneous loss	(1)	(2-2)

次に，決定を行うタイミングについて述べる．

(1) 推定問題:

$$\begin{array}{ccc} \text{事前知識(モデル族,事前知識)} & & \text{決定} \\ & \text{と} & \Longrightarrow \qquad \hat{m} \\ \text{過去のデータ} x^n & & \end{array}$$

(2-1) データ系列 x^n の逐次予測:

(事前知識) $\Longrightarrow \hat{x}_1$

(事前知識) $+x_1 \Longrightarrow \hat{x}_2$

(事前知識) $+x_1 + x_2 \Longrightarrow \hat{x}_3$

(事前知識) $+x_1 + x_2 + x_3 \Longrightarrow \hat{x}_4$

(事前知識) $+x_1 + x_2 + x_3 + x_4 \Longrightarrow \hat{x}_5$

$\cdots\cdots\cdots\cdots\cdots$

(事前知識) $+x_1 + x_2 + x_3 + \cdots\cdots\cdots + x_{n-1} \Longrightarrow \hat{x}_n$

と逐次予測した場合の $t = 1$ から $t = n$ までの累積予測誤差を議論する

(2-2) データ x の1時点の予測:

$$\begin{array}{ccc} \text{事前知識(モデル族,事前分布)} & & \text{決定} \\ & \text{と} & \Longrightarrow \qquad \hat{x} = \hat{x}_{n+1} \\ \text{過去のデータ} x^n & & \end{array}$$

(2-1)の逐次予測問題では $x^n = x_1 x_2 \cdots x_n$ を逐次的に観測しながら予測を続けるのに対し，(2-2)の一時点の予測問題では x^n はすでに得られたもとでの次のデータの予測を考えている．これらは損失関数による分類であるので，決定の前に観測されたデータ系列が与えられるかどうか(過去のデータの有無)は本質的ではない．しかし，それぞれの問題において “何を大きくしていったときの漸近的状況に興味があるか” を考えると両者とも $n \to \infty$ の場合であるといえる [83]．勿論，x^n の逐次予測問題において，さらに過去の系列 $x^l = x_{-l+1} \cdots x_{-1} x_0$ が与えられている状況を考えても差し支えない．しかし，n を固定して $l \to \infty$ とする問題は，X^n を一つの多次元確率変数とみなしてしまえば一時点の予測問題の結果がそのまま適用できるので，理論的興味は薄い．

表 2.2: 推定問題と予測問題:過去のデータの有無に関して

		推定	予測
x^n に対する累積損失 Cumulative loss	過去のデータ x^l がある	×	×
	過去のデータはない	×	(2-1)
1時点の損失 Instantaneous loss	過去のデータ x^n がある	(1)	(2-2)
	過去のデータはない	×	×

すなわち，(2-1)の逐次予測問題では 逐次予測を長い期間続けていく際に予測期間の累積予測誤差損失がどのようなオーダーとなるかに興味があり，一方(2-2)の一時点の予測問題では観測できる過去のデータ数が大きくなっていくときにどのような性質があるか，に興味があるといえる．したがって，本研究では漸近的性質の評価を目的としているが，いずれの問題においても $n \to \infty$ の状況を考えている．表2.2に，本研究で扱っている問題に対し，過去のデータの有無をいう切り分けも加えた分類を示す.

2.4.1 最適モデルの統計的推定問題

情報源から得られた長さ n のデータ系列 x^n に基づいて，離散モデル族では λ，パラメトリックモデル族では θ^k，階層モデル族ではモデル m とパラメータ θ^{k_m} を推定する問題を考えよう.

こうして推定されたモデルやパラメータはそのまま予測に使うことができるが，本研究では統計的推測の評価尺度という観点から，パラメータ間やモデル間の損失関数を考えている場合を推定問題と定義している．データ系列 x^n から推定された推定量を $\hat{\lambda}, \hat{\theta}^k, \hat{\theta}^{k_m}, \hat{m}$ などと記述し，例えば以下に述べる最尤推定量や事後確率最大推定量，損失最小推定量などの意味で用い，記号については各章で定義する．2種類以上の推定量を扱う場合は，$\tilde{\lambda}, \tilde{\theta}^k, \tilde{\theta}^{k_m}, \tilde{m}$ という記述も推定量の意味で用いる.

推定問題においてはその目的が重要である．真の分布に対する最適値があると考え，その値に何らかの意味で近い推定を行うべきである．しかし，最適値を事前に知る方法はないので，x^n を用いて何らかの規準で推定量を構成することになる．パラメトリックモデル族におけるパラメータの推定量については，多くの研究がなされており，また実際に多くの事例に適用されている．データ系列 x^n の経験分布から KL-情報量の意味で最も近い推定量が最尤推定量[117]，x^n が与えられたもとでの事後確率の最大値を与える推定量が事後確率最大推定量である[16]．また，KL-情報量に限定せず一般の損失関数を用いて，経験分布からの損失を最小にする推定量を損失最小推定量[63]とよぶ．パラメトリックモデル族におけるパラメータの推定方法については，古くからの多くの議論によってほぼ体系化がなされており，本論文ではこのパラメータ推定の問題については議論しない．

本研究で主に取り扱う推定問題は，階層モデル族におけるモデル m の推定問題である．これは統計的モデル選択問題と呼ばれるもので，多くの現実問題において生じる反面，その階層構造のために最大尤度原理が意味をなさず，検定による方法も本質的な問題が生じてしまうという難しさがある．統計的モデル選択問題に関しては，現在でも理論的，実務的の両観点から広く研究されており，モデル m の選択規準として様々なタイプの規準が提案されている．このような規準として代表的なものは，赤池の AIC (Akaike information criterion)[1]，C.Schwarz の BIC (Bayesian information criterion)[102]，J.Rissanen の MDL (minimum description length)[155]などである．例えば AIC は予測を目的とした選択規準であり，先に述べた未来データの予測問題をモデル選択の範疇で議論したものである[117]．また，MDL は x^n を最も圧縮可能なモデルを選択する規準である．すなわち，推定と予測を明確に切り分けるのは難しく，“精度の高い予測を行うための推定”といった議論も行われることが多い．最適モデル m^* の推定を目的とした場合，m の選択に対する 0-1 損失を考えた場合の漸近ベイズ最適を与える BIC がその目的に適した規準といえる．

本研究では，真のモデルを発見を目的としたモデル選択とデータ系列 x^n や未来のデータ x_{n+1} の予測を目的としたモデル選択を，明確に分類して議論を行う．これは，m を一つに決めるという手順は同じであっても，その目的を表す損失関数が異なるためである．したがって，当然ながら両者で性能評価の基準は異なって構わないし，その方が自然であると考えられる．

2.4.2 データ系列の逐次予測問題

未知の母集団から得られる長さ n のデータ系列 x^n を逐次的に予測する問題を考える[83]. 一般に統計とはデータ系列 $x^n = x_1 x_2 \cdots x_n$ を観測して何らかの情報を得る場合を指すことが多いが，例えば情報源符号化の問題や逐次的な学習問題で扱う予測などでは，母集団から逐次的に得られるデータ系列 x^n を逐次圧縮したり，逐次的に予測する問題を扱う必要がある.

ここで，予測には次の2つの形が考えられる[83].

(1) x そのものの予測値を決定する

(2) x の確率分布 $p(x)$ を予測結果として決定する

通常，予測といえば x そのものの予測値を考える場合も多い. すなわち，“明日の天気は晴れである” といった結果の予測である. これに対し，“明日の天気は，晴れが 60%, 曇り 30%, 雨 10% である”というように，その確率分布を予測結果として示す場合はより多くの情報を与えることになる.

ここで，一般的な損失関数を定義しよう. x の確率あるいは確率密度を $Ap(x)$ と予測した場合の損失関数を $L(x, Ap(\cdot))$ とする. もし，x の予測値 Ax を考える場合には，損失関数を $L(x, Ax)$ とする. 以下では記法の簡略化のため，これらの予測を区別しないで，x に対する予測を Ad とし，$L(x : Ad)$ と記述する.

逐次符号化問題や統計的学習問題などの，$x^n = x_1 x_2 \cdots x_n$ を次々に予測していく逐次予測問題において，その累積の予測誤差を評価とする場合を考える. まずある時点 t において，過去の観測データ x^{t-1} から次の観測 x_t を予測する場合を考えよう. 観測データ x^{t-1} に基づく x_t に対する何らかの予測を $Ad(x_t|x^{t-1})$ とし，実際に x_t が観測されたときのその予測誤差損失を $L(x_t : Ad(x_t|x^{t-1}))$ とする. その $t = 1$ から $t = n$ までの累積の予測誤差損失

$$\sum_{t=1}^{n} L(x_t : Ad(x_t|x^{t-1})), \tag{2.7}$$

を定義し，この累積損失を小さくするような予測法 $Ad(\cdot|\cdot)$ を考えると，これは過去の観測が得られない場合のデータ系列 x^n の逐次予測問題ととらえることができる.

実際には，この損失の値は x^n が与えられるまで計算できないので，ベイズ規準[80]や min-max規準[28, 29, 32, 26]などのもとに予測法を構成することになる.

対数損失と逐次予測

確率分布を予測する際の良さを測る尺度である損失関数の代表は，対数損失 (logarithmic loss) $-\log Ap(x^n)$ である．これは，多くの良い性質が知られ，広く受け入れられている損失関数である[83][4].

ここで，実現値 x に対し，確率分布を $AP(x)$ と予測したときの良さを測る尺度として，対数損失関数

$$L(x, AP(\cdot)) = -\log AP(x) \tag{2.8}$$

を考える．この損失は確率理論，統計理論において本質的であり，多くの従来研究はこの損失関数，あるいはこれの期待値をとった平均対数損失をもとに展開されている．平均対数損失は後で述べる KL情報量とエントロピーの和であり，分布間の距離を KL情報量で測ったときの推測を議論することに等しい．対数損失やKL情報量を評価関数とした予測は従来から広く適用，研究されており，この損失関数を議論することは本質的である．

もし，対数損失関数を考え，過去のデータ x^{t-1} のもとで x_t の確率を $AP(x_t|x^{t-1})$ と予測した場合には，

$$-\log AP(x^n) = \sum_{t=1}^{n} \{-\log AP(x_t|x^{t-1})\}, \tag{2.9}$$

となる．ただし，$AP(x^n) = \prod_{t=1}^{n} AP(x_t|x^{t-1})$, $AP(x_1|x^0) = AP(x_1)$ である．

情報源符号化問題，一括予測問題との関連

本研究では，第3章と第4章において，MDL原理に基づく予測を取り扱っている．MDL原理は後で述べるように，符号長という評価尺度を統計的推論に持ち込んだものである．また，MDL規準の導出の際に，一括予測問題という枠組みも必要となってくる．そこで，その準備として情報源符号化の観点から見た予測問題と一括予測問題について述べておく．

逐次予測ではなくデータ系列 x^n をまとめて予測する問題を考えた場合，x^n と予測の間に損失関数が定義されるべきである．このような問題をデータ系列の一括予測問題と呼ぶ．まず，x^n の確率を $AP(x^n)$ と予測した場合の損失関数を $L(x^n, AP(\cdot))$ と定義する．もし，x^n の予測値 $d^n = \hat{x}^n$ を考える場合には，損失関数を $L(x^n, d^n)$ とする．以下では記法の簡略化のため，これらの予測を区別

[4]例えば，分布のパラメータ推定で良く用いられる最尤推定量は対数損失を考えていることになる．

しないで，x^n に対する予測を $Ad(x^n)$ とし，$L(x^n : Ad(\cdot))$ と記述する．当然ながら，逐次予測問題と一括予測問題では決定が異なるので等価な問題ではない．

ここで，実現値 x^n に対し，確率分布を $AP(x^n)$ と予測したときの良さを測る尺度として，対数損失関数

$$L(x^n, AP(\cdot)) = -\log AP(x^n) \tag{2.10}$$

を考えてみる．(2.9)式にあるように，これは逐次予測問題の累積対数損失と一致する．すなわち，対数損失を考えた場合には，データ系列を逐次予測した場合の累積対数損失とデータ系列を一括で予測した場合の損失を考えることは等価となる．

一方，対数損失を評価関数とした予測は情報源符号化問題と密接に関係する．この情報源符号化問題そのものを扱っていると言っても過言ではない．以下にその概略を述べる．

x^n を圧縮する問題は，以下で述べるように逐次予測と一括予測の一例と考えることができる．これは累積対数損失を評価関数としたx^n の予測問題に帰着する．まずは，一括的に与えられた x^n を圧縮する場合を考える．離散のデータ系列 x^n を長さ $C(x^n)$ [nat]で符号化するものとする．このとき，歪みなく符号化するためにはクラフトの不等式[8]

$$\sum_{x^n \in \mathcal{X}^n} \exp\left\{-C(x^n)\right\} \leq 1, \tag{2.11}$$

をみたす必要がある．ここで，$\sum_{x^n} \exp\{-C(x^n)\} < 1$ とすると符号長にロスが生じるので，$\sum_{x^n} \exp\{-C(x^n)\} = 1$ と仮定する．このとき，$C(x^n) = -\log AP(x^n)$ とおくと，$\sum_{x^n} AP(x^n) = 1$ が得られ，$AP(X^n)$ は $\mathcal{X}^n$ 上の確率分布を表すことがわかる．すなわち，符号化の問題は “ X^n の確率分布 $AP(X^n)$ をどのように設定するか ” に帰着する．このように x^n の符号長を $-\log AP(x^n)$ と解釈することができるが，任意の $AP(x^n)$ に対してこれが自然数となる訳ではない．そのため，$-\log AP(x^n)$ は理想符号長と呼ばれる．実際に $AP(x^n)$ が与えられたときに，その確率分布に対して最適な符号化はハフマン符号[56]で達成される．また，計算量が n のオーダで，またシンボルを1文字ずつ逐次符号化していくことを可能とした算術符号[56]でも，ほぼ $-\log AP(x^n)$ 程度の符号長で符号化が可能であり，1文字当たりの平均符号長は $n \to \infty$ のとき $AP(\cdot)$ に対する最適符号長に近づくことが知られている．

この符号化の真の確率分布 $P^*(x^n)$ に対する平均的な冗長度は

$$E^*\left[\log \frac{P^*(X^n)}{AP(X^n)}\right] = \sum_{x^n \in \mathcal{X}^n} P^*(x^n) \log \frac{P^*(x^n)}{AP(x^n)}, \tag{2.12}$$

となり，これを小さくする $AP(x^n)$ が良いことがわかる．ただし，$E^*[\cdot]$ は $P^*(\cdot)$ による期待値を表す．$P^*(x^n)$ が与えられると $-\sum_{x^n \in \mathcal{X}^n} P^*(x^n) \log P^*(x^n)$ は定数であるので，$-\sum_{x^n \in \mathcal{X}^n} P^*(x^n) \log AP(x^n)$ を考えても良いことになる．(2.12)式の冗長度は n 次の KL-情報量，あるいはダイバージェンス(divergence)と呼ばれ，$\forall x^n \in \mathcal{X}^n \quad P^*(x^n) = AP(x^n)$ のとき最小値 0 をとる．したがって，真の確率分布 $P^*(x^n)$ が既知の場合は $-\log P^*(x^n)$ という符号長で符号化すれば最適であるが，現実的に未知な場合が多い．したがって，$P^*(x^n)$ 未知の状況で，$AP(x^n)$ をどのように決定するかが問題となる．あるクラスの情報源全てに対して，1記号当たりの平均符号長，すなわち $-\frac{1}{n}\log AP(x^n)$，あるいはその期待値がその圧縮限界 $\lim_{n\to\infty} \frac{1}{n} E^*[-\log P^*(X^n)]$ に収束するとき，この符号は(そのクラスに対する)ユニバーサル符号と呼ばれる．以上のように，この問題は x^n が出現したときの予測誤差損失を $-\log \frac{P^*(x^n)}{AP(x^n)}$ あるいは $-\log AP(x^n)$ とした場合の x^n の一括予測問題ととらえることができる．

先にも述べた通り，$AP(x_t|x^{t-1}), (t = 1, 2, \cdots, n)$ は $AP(x^n)$ を与えるので，対数損失を仮定した場合には，データ系列 x^n を一括で予測する問題は前述の逐次的に予測する問題と等価になる．すなわち，対数損失は一括の予測損失と逐次予測の累積損失が同じになり，両方を統一的に扱えるという便利さを持ち合わせている[83]．これは x^n を一括で符号化した場合と逐次的に符号化した場合とで，符号長は同じになることを述べている．

2.4.3 一時点のデータの予測問題

情報源から得られた長さ n のデータ系列 x^n に基づいて，次のデータ $x = x_{n+1}$ を予測する問題である．これは，前節で説明したオンラインの予測問題において累積損失を考えず，x^n のもとで次の x_{n+1} の損失のみを考えることに相当する．

予測 $AP(\cdot)$ に対して x が出現したときの損失関数を対数損失 $-\log AP(x)$ とすると，これは x^n が与えられたもとで，次の x_{n+1} を予測符号化する問題と等価である．すなわち，x_{n+1} の予測問題は，過去の系列 x^n が得られたもとで x_{n+1} の確率分布をどのように推定するかに帰着する．前節のデータ系列に対する逐次予測問題との違いは “累積損失を考えるか，それとも次の1シンボルの損失を考えるか” にあるといえる．

2.5 従来研究

前節では統計的推定・予測問題をその目的からまとめた．その中で本研究では 第3章 ~ 第5章で対数損失に対する x^n の予測問題，第6章と第7章で m^* の推定問題と x_{n+1} の予測問題，第8章で一般の損失に対する x^n の予測問題を扱っている．ここでは，これらに対する従来研究の成果をまとめ，本研究への展開について述べる．

2.5.1 統計的モデル選択問題

2.4節で分類したそれぞれの問題に入る前に，一般に統計的モデル選択問題として広く認識されている問題について述べる．本研究では統計的推測の各問題を評価関数である損失関数の面から分類しているが，一方では方法による分類も可能である．階層構造のモデル族において，モデル m を選択する問題はその損失関数に関わらず，統計的モデル選択問題と呼ばれ，広く研究されているテーマの一つである．

本研究では，ベイズ統計に基づく推測の評価を目的としているので，損失関数による分類を行っている．統計的モデル選択問題も，その目的に応じて損失関数を適切に設定し，場面にあった規準を採用するべきである．モデル選択規準による統計的推測の評価を考えた場合，まず第一にその損失関数を議論することが必要であろう．さらにそのモデル選択規準の性質として “他の評価基準に関してはどの程度のパフォーマンスを示すか”という点に興味が生じる．

統計的モデル選択問題が興味の対象となる理由の一つは，非常に広い応用範囲をもつためである．例えば，重回帰分析などの多変量解析においても，必要な変数だけを取り込み，不要な変数を除く操作が行われるが，これも統計的モデル選択問題とみなすことができる．実際にどの重回帰モデルを選択するかについては，固有技術と経験も駆使しながら，F値や寄与率，自由度調整済み寄与率，二重調整済み寄与率，AIC などの規準を用いて適切なモデルを選択することになる．しかし，各規準で最適とみなされるモデルはしばしば異なり，その規準の意味を正確に把握しなければ，最終的にどのモデルを選ぶかについては客観的な判断ができない場合も多い．以上のように，統計的モデル選択問題[74]は，理論的に興味深いテーマであるだけでなく，非常に広範囲の応用を持つ極めて重要な問題の一つである．とくに候補であるモデル族が階層的な構造を持っている場合，次数の低いモデルはより次数の高い(表現能力の高い)モデルに含まれるという性質が問題を難しくしている．このようなモデル選択問題を根

本から考えると，“選択したモデルをどのような目的で用いるか” が重要と考えられる．すなわち，何らかの損失関数，リスク関数が定義され，それをある意味で最適化するモデルを選択する規準が構成できれば，これはモデル選択において有効な情報となる．

次に統計的モデル選択問題として定式化できる問題の例をあげる．

例 2.5 (回帰モデルの変数選択) 重回帰分析[21]では説明変数 $x_1, x_2, \cdots, x_r$ と誤差項 ϵ の線形モデルで目的変数を表現する．通常，説明変数 $x_1, x_2, \cdots, x_r$ には技術的，経験的に考慮して，目的変数に効いてくるであろう変数を取り上げる．しかし，目的変数への影響の強さ(標準回帰係数)は説明変数によって異なり，変数によっては実際には目的変数に影響を与えないことも考えられる．サンプル数が少ないときに，真の標準回帰係数が小さい説明変数を取り込むと予測精度が劣化することがある．また，目的変数に本当は影響しない説明変数を取り込んだ場合も予測精度が劣化する．したがって，これらの説明変数の候補から，目的に応じてモデルに取り込む変数を選択する必要があり[61]，この場合は予測を目的としたモデル選択を議論するべきである．一方，構造解析を目的とした重回帰分析においても，本当に効いている説明変数のみをモデルに取り込むことが必要である．この場合は真のモデルの発見を目的としたモデル選択を議論するべきであろう．以上のことは，線形モデルに限定しない非線形回帰モデルでも同様である．また，時系列を扱う自己回帰モデル(ARモデル)や自己回帰移動平均モデル(ARMAモデル)[53]においても同様のモデル選択が必要である[90]．□

例 2.6 (隠れマルコフモデル(Hidden Markov Model)の選択問題) 隠れマルコフモデル(Hidden Markov Model: HMM)は，音声認識システムにおいて有用なモデルであることが広く認識されている[88]．通常，学習用サンプルからHMMのパラメータを推定し，そのモデルを用いて音声の認識を行うことになる．しかし，音源は連続信号であり，厳密には離散的な状態を考えるHMMに従って発生していないことは明らかである．通常は量子化操作の後で，HMMモデルを適用することになる．したがって，サンプル数が十分多いときは複雑なHMMの方が良い認識率を示すが，逆に少ないときに複雑なモデルを仮定すると認識率が劣化することは明らかである[129]．この問題においては，予測を目的としたモデル選択が必要となる．□

例 2.7 (母集団のパラメータ変化の検出) ある特性値の離散的なカテゴリがあり，それぞれのカテゴリに対してデータが得られる場合を考える．例えば薬効

実験においては，薬品の投入量を数段階(カテゴリ)に設定し，それぞれのカテゴリで薬品の効果に関するデータが与えられる．その効果の現れ方を確率分布で表現できるとき，問題はその効果が投入量に対してどのように変化するかである．投入量を増やしても効果は変化しないかもしれないし，あるカテゴリ間で効果が大きく変化するかもしれない．実験で与えられたデータから，どのカテゴリ間で確率分布のパラメータが変化しているかを推定する問題は，統計的モデル選択問題として定式化できる．

一方，母集団から時系列で与えられるデータをモニタリングし，母集団分布のパラメータの変化を検出する問題がある．統計的品質管理において工程の異常を発見するなどの目的で利用される管理図はこの問題を扱っている．この問題も，変化時点の特定とパラメータ推定を同時に行う統計的モデル選択問題として定式化できる[106, 107].

パラメータの変化を正確に検出したいという場合には，真のモデルの発見を目的としたモデル選択問題として定式化される． □

一般的なモデル選択問題におけるモデル選択規準に関しては，予測を考えた AIC [1], 推定を考えた BIC[102] や MDL[95] の他，多くの情報量規準[92, 111, 131] が提案され，それらの性質が明らかにされている．

2.5.2 一時点のモデル推定問題

ここでは真のモデルの発見を目的としたモデル選択について述べる．

Schwarz の BIC

C.Schwarz は，モデル選択の規準として，

$$\log \int_{\theta^{k_m} \in \Theta^{k_m}} p(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{2.13}$$

という規準を与え，その漸近式を BIC として導出した[102]. その漸近式 $BIC(m)$ は $O(\log n)$ までの項を考え，

$$BIC(m) = \log p(x^n|m, \hat{\theta}^{k_m}) - \frac{k_m}{2} \log n + o(\log n), \tag{2.14}$$

と与えられる．この式に (-1) をかけると，ベイズ符号の符号長となるので，これはパラメータに対して混合をとるタイプの MDL と本質的に等価である．さらに，これをベイズ規準の面から考えると，一様分布 $P(m) = 1/|\mathcal{M}|$ を事前分

布とした場合の m の事後確率 $P(m|x^n)$ を最大化しているのと等価であるので，事後確率に対するモデル選択の誤り率を最小化していることになる．これは，モデル選択の損失として，0-1損失を与えた場合のベイズ最適な決定(ベイズ決定)となっている[78].

この方向の研究としては，Poskitt が パラメータの推定誤差も考慮にいれたベイズ決定に基づくモデル選択法を提案している[92]. しかし，これは特定の目的に特化した規準であり，また式が簡潔になるように損失関数を設定している感がある．

本研究への展開

BIC や MDL はベイズ統計に基礎をおく規準であるが，これは事前分布を仮定しているためである．ここで，もし事前分布が仮定できるならば，ベイズ決定理論[16, 27, 35]に基づき，ベイズ最適なモデル選択を構成することが最適性の面からも合理的であると考えられる[40]. ベイズ規準では様々な損失関数に対し，有限長のデータに対してベイズ最適なモデル選択とパラメータ推定が可能となる．ベイズ決定理論に基づくモデル選択は，ベイズ統計の枠組みで統一した議論が行われるという点で一貫性があり，BIC や MDL でも事前分布が仮定されているように，モデル選択問題に対して非常に相性のよい解法の一つといえる．そこで本研究では第6章と第7章において，様々な現実的な損失関数に対してベイズ最適なモデル選択を定式化する．さらにその性質として，ベイズ最適なモデル選択が一致性をもつ，すなわち漸近的に最適なモデルを選択できるという性質を示す.

2.5.3 データ系列の逐次予測問題 (1): 累積対数損失に対する展開

本論文では，真の確率分布は未知であるが，確率モデル族が与えられている場合を考える．長さ n のデータ系列 x^n を逐次予測する問題に対数損失関数を適用すると，それは前節で述べたように一括予測問題，および情報源符号化問題と本質的に等価となる．この事実を逆手に取り，情報源符号化の観点から導出された規準を，統計的な推定や予測に適用しようとする一連の考え方がある．以下ではその概略を述べる．

ある確率モデルのクラスが与えられたときに，そこから発生するデータをこの確率モデルを用いて符号化するという考え方は，情報源符号化の分野では古くから研究されてきた．その一連の研究において根本に存在する重要な概念

は，J.R.Rissanen の記述長最小原理(minimum description length principle: MDL原理)[56, 155]である．この研究の流れとして，ベイズ符号というベイズ最適な符号が提案され[25, 80, 99]，最近ではベイズ符号の性質，およびその構成法などについて多くの研究がなされている[26, 136, 149]．これらは，本質的にn個のデータに対する累積の対数損失を評価関数とした，x^n の予測を扱っている．そのため，そのまま予測に適用することが可能である．さらに，データ系列 x^n の情報量を測る尺度として MDL原理 を発展させた形の確率的コンプレキシティ(stochastic complexity: SC)[99]についても，損失関数を一般化した拡張確率的コンプレキシティ(extended stochastic complexity: ESC)[153, 154]が提案されるなど，現在多くの研究が進んでいる段階である．これらはすでに情報源符号化という目的のものではなく，統計的学習や予測を目的として導かれている点が重要である．

本研究では，この MDL規準 からベイズ規準，および SC，ESC までの一連の研究の成果を踏まえ，より広い視点からこのような規準による予測の基本的な性質を解明している．以下で，対数損失を扱っているMDL規準(MDL符号)，SC，ベイズ規準の定義を与え，これまでに明らかにされている成果についてまとめる．確率モデルを陽に仮定しないタイプのユニバーサル符号化，例えば定常エルゴード 情報源に対してユニバーサルであることが証明されている Ziv -Lempel 符号[159, 160]などは対象としない．

離散モデル族の場合

まずは MDLが符号化の観点から導かれたことを受けて，情報源符号化の立場からMDL原理の概略をまとめる．また，本節では $\mathcal{X}$ は離散アルファベットとするが，連続でも同様の議論が可能である．$P(x^n|\lambda) = P(X^n = x^n|\lambda)$ をモデル λ が表す確率分布，$C(\lambda)$ を λ を符号化するのに必要な符号長とする．このとき，x^n を λ で符号化すると，その符号長は $-\log P(x^n|\lambda)$ である．これは λ が既知であればクラフトの不等式を等号でみたすが，x^n に対してその都度良い圧縮性能をもつ λ を選ぶとすれば，クラフトの不等式をみたさず，一意復号が不可能となる．そこで，まずどの λ を用いて符号化するかを復号側に知らせるために，λ を一意復号可能な符号長 $C(\lambda)$ で符号化し，次にこの λ を用いて x^n を一括符号化すれば，これは x^n は一意復号可能となる．すなわち，$-\log P(x^n|\lambda) + C(\lambda)$ という符号長で x^n を符号化することができる．MDL符号は，この符号長を最も短くする λ を用いて x^n を符号化する方法である．すなわち，符号長は

$$\min_{\lambda\in\Lambda}\Big\{-\log P(x^n|\lambda) + C(\lambda)\Big\}. \tag{2.15}$$

と与えられる.

ここで，x^n の場合と同様に，$\sum_{\lambda\in\Lambda} e^{-C(\lambda)} = 1$ とすれば，$P(\lambda) = e^{-C(\lambda)}$ をモデル λ の事前確率と解釈することができる．したがって，与えられた x^n に対するMDL規準の対数損失は以下のように定義される.

定義 2.4 (離散モデル族に対するMDL規準) 離散モデル族 Λ に対する MDL規準の累積対数損失，すなわちMDL符号の符号長 $L^{\lambda}_{MDL}(x^n)$ は [5]

$$L^{\lambda}_{MDL}(x^n) = \min_{\lambda\in\Lambda}\{-\log P(x^n|\lambda) - \log P(\lambda)\}, \tag{2.16}$$

で与えられる. □

次式で与えられる $\hat{\lambda}_{MDL}$ は MDL推定量と呼ぶ.

$$\begin{aligned}\hat{\lambda}_{MDL} &= \arg\min_{\lambda\in\Lambda}\{-\log P(x^n|\lambda) - \log P(\lambda)\} \\ &= \arg\max_{\lambda\in\Lambda} P(\lambda|x^n).\end{aligned} \tag{2.17}$$

上式の展開により，離散モデル族に対しては MDL推定量は事後確率最大推定量に一致することがわかる [101]. 事後確率最大推定量は，λ の選択に対して 0-1 損失関数を考えた場合の事後ベイズリスクを最小にするベイズ最適解である.

一方，Λ 上に事前確率 $P(\lambda)$ が仮定されるならば，符号長を損失関数として考えた場合，すなわち対数損失に対するベイズ最適な予測を構成することが可能である．これをベイズ規準による予測と呼ぶ．これを符号化の問題ととらえたとき，このベイズ最適な符号をベイズ符号と呼ぶ．このベイズ最適解は，クラス内の全ての確率モデルの混合モデルの対数損失で与えられる [80]．このことを以下で示そう．λ がデータを発生する真の分布を表すとしたとき，決定 $AP(x^n)$ の冗長度(KL情報量)は

$$R(\lambda, AP) = \sum_{x^n\in\mathcal{X}^n} P(x^n|\lambda)\log\frac{P(x^n|\lambda)}{AP(x^n)}, \tag{2.18}$$

で与えられる．これをリスク関数という．これは，x^n に対する損失を符号長の差 $\log P(x^n|\lambda) - \log AP(x^n)$ とした場合の平均損失である．しかし，事前に λ は未知であるので，事前分布 $P(\lambda)$ に対して平均的に良い予測の構成を考える．すなわち,

$$BR(AP) = \sum_{\lambda\in\Lambda}\sum_{x^n\in\mathcal{X}^n}\left[P(x^n|\lambda)\log\frac{P(x^n|\lambda)}{AP(x^n)}\right]P(\lambda), \tag{2.19}$$

[5] $L^{\lambda}_{MDL}(x^n)$ の上付き記号 λ は離散モデル族に対する符号長であることを意味する．以下で，パラメトリックモデル族，階層モデル族についても，同様な表記法を用いる.

を最小化する $AP(\cdot)$ は事前分布に対して平均的に良い予測を与える．$BR(AP)$ はベイズリスクとよばれる．この最小化はベイズ平均対数損失

$$BL(AP) = \sum_{\lambda \in \Lambda} \sum_{x^n \in \mathcal{X}^n} P(x^n|\lambda)\{-\log AP(x^n)\}P(\lambda), \tag{2.20}$$

の最小化と等価である．これを書き換えると，

$$BL(AP) = \sum_{x^n \in \mathcal{X}^n} P(x^n)\{-\log AP(x^n)\}, \tag{2.21}$$

となる．ただし，$P(x^n) = \sum_{\lambda \in \Lambda} P(x^n|\lambda)P(\lambda)$であり，この $BL(AP)$ を最小化する $AP(x^n)$ は，Shannon の補助定理より，$AP(x^n) = P(x^n)$ で与えられる．すなわち，ベイズ最適な予測は $-\log P(x^n)$ で与えられることがわかる．当然ながら，情報源符号化の言葉で述べると $-\log P(x^n)$ はベイズ符号の符号長を与えているといえる．次節以降で述べるパラメトリックモデル族，階層モデル族に関しても，同様の手順でベイズ符号が構成できる．

定義 2.5 (離散モデル族に対するベイズ規準) 離散モデル族に対するベイズ規準の累積対数損失，すなわちベイズ符号の符号長 $L^{\lambda}_{Bayes}(x^n)$ は

$$L^{\lambda}_{Bayes}(x^n) = -\log \sum_{\lambda \in \Lambda} P(x^n|\lambda)P(\lambda), \tag{2.22}$$

で与えられる． □

離散モデル族に対しては，Rissanen により，同じ事前分布のもとでベイズ規準の累積対数損失がMDL規準のそれを下界することが示されているが [100]，これはパラメトリックモデル族や階層モデル族でも成り立つ一般的性質である．また，Clarke と Barron はi.i.d.情報源に対する離散モデル族を考え，対数損失に対するベイズ規準の漸近的なリスク，ベイズリスク，min-max リスクを解析した [26]．例えばベイズ規準による予測のリスクに対し，

$$E^*\left[L^{\lambda}_{Bayes}(X^n)\right] = E^*\left[-\log P(X^n|\lambda^*)\right] - \log P(\lambda^*) + o(1), \tag{2.23}$$

という漸近式を導いている．ただし，λ^* は第3章で定義する最適なモデルであり，$E^*[\cdot]$ は $P^*(\cdot) = P(\cdot|\lambda^*)$ による期待値を表す．しかし，離散モデル族に限定した議論は，他にはあまりなされていない．これは，応用的側面からパラメトリックモデル族や階層モデル族の方に興味が集中していることや，また離散モデル族に対しては多くの性質が自明に導くことができるためと考えられる．し

かし，離散モデル族に対して，それらの性質を明らかにしておくことは，今後の研究のためにも必要である．

本研究では，Clarke と Barron の成果を受けて，MDL規準とベイズ規準の累積対数損失の差がモデル族に依存して変わってくることを示すために離散モデル族を取り上げ，両規準の累積対数損失の差を解析する．

パラメトリックモデル族の場合

もともとのMDL規準は，パラメトリックモデル族に対して提案され，その累積対数損失の漸近式が Schwarz の BIC と同じ形になることが示された．これが符号化の概念を統計的推測に導入した最初の成果である．一方，ベイズ符号はパラメトリックモデル族に対するベイズ最適な符号として提案され，その後階層モデル族に拡張されている．

まず，符号化の観点から導かれたMDL規準の概略から述べる．パラメトリックモデル族の場合，まず確率モデルのパラメータ θ^k を符号化し，その確率モデルを用いて x^n を $-\log P(x^n|\theta^k)$ という符号長で符号化することを考えればよい．しかし，パラメータ θ^k は連続であるため，それをそのまま符号化することができない．MDL符号では，パラメータ集合 Θ^k を量子化して可算のセル(直方体)に分け，その代表点 $\bar{\theta}^k$ を符号化することを考える．

$\bar{\Theta}_n^k$ を量子化されたパラメータの代表点 $\bar{\theta}^k$ の集合とする．この集合はデータ長 n に依存しても構わない．すなわち，x^n を符号化する際には $\bar{\theta}^k \in \bar{\Theta}_n^k$ とする．量子化セルに番号を付けることを考え，$\Theta_{n,l}^k$ を l 番目のセル内のパラメータの集合とする$(l = 1, 2, \cdots)$．このとき $\cup_l \Theta_{n,l}^k = \Theta^k$，かつ $\Theta_{n,i}^k \cap \Theta_{n,j}^k = \phi, i \neq j$, となるように量子化されるものとする．$\alpha_{n,j}(\bar{\theta}^k)$ を $\bar{\theta}^k$ で代表されるセルの j 番目の辺の量子化幅とすると，セルの個数は高々 $O(1/\max \prod_{j=0}^{k-1} \alpha_{n,j}(\bar{\theta}^k))$ である．このとき，各代表点 $\bar{\theta}^k$ を符号長 $-\log f(\bar{\theta}^k) - \sum_j \log \alpha_{n,j}(\bar{\theta}^k)$で符号化することができる [93]．ここで，$f(\theta^k)$ は Θ^k 上の事前確率密度であり，$\int_{\theta^k} f(\theta^k) d\theta^k = 1$ をみたす．ここで，符号化のロスが生じないように $\bar{\Theta}_n^k$ は $\sum_{\bar{\theta}^k \in \bar{\Theta}_n^k} f(\bar{\theta}^k) \prod_j \alpha_{n,j}(\bar{\theta}^k) = 1$ となるように選ばれているとする．したがって，この符号化は $f(\bar{\theta}^k) \prod_j \alpha_{n,j}(\bar{\theta}^k)$ を $\bar{\Theta}_n^k$ の事前確率と考えて符号化していることに相当する．

ここで，パラメータ集合の量子化の方法を議論しなければならない．量子化は符号長が最小化されるように設定されるべきであるが，これは漸近的な方法

によって可能であり，適当な正則条件のもとで漸近最適な量子化幅は

$$\prod_{j=0}^{k-1} \alpha^*_{n,j}(\bar{\theta}^k) = \frac{1}{\sqrt{n}^k \sqrt{\det J(\bar{\theta}^k)}}, \tag{2.24}$$

で与えられる [56, 95, 136, 146, 155]．すなわち $\alpha^*_{n,j}(\bar{\theta}^k) = O(1/\sqrt{n})$ である．ここで，$J(\theta^k)$ はフィッシャー情報行列 (Fisher information matrix) であり，

$$J(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E_\theta \left[-\frac{\partial^2 \log P(X^n|\theta^k)}{\partial\theta^k (\partial\theta^k)^T} \right], \tag{2.25}$$

で与えられる．ただし，E_θ は $P(\cdot|\theta^k)$ による期待値を表す.

定義 2.6 (パラメトリックモデル族に対するMDL規準) パラメトリックモデル族に対するMDL規準の累積対数損失，すなわちMDL符号の符号長 $L^{\bar{\theta}^k}_{MDL}(x^n)$ は

$$L^{\bar{\theta}^k}_{MDL}(x^n) = \min_{\bar{\theta}^k \in \bar{\Theta}^k_n} \left\{ -\log P(x^n|\bar{\theta}^k) - \log \frac{f(\bar{\theta}^k)}{\sqrt{n}^k \sqrt{\det J(\bar{\theta}^k)}} \right\}, \tag{2.26}$$

で与えられる [56, 95, 136, 146]. □

一方，与えられた事前密度 $f(\theta^k)$ に対してベイズ最適な予測を構成することにより，ベイズ符号が与えられる．

定義 2.7 (パラメトリックモデル族に対するベイズ規準) パラメトリックモデル族に対するベイズ規準の累積対数損失，すなわちベイズ符号の符号長 $L^{\theta^k}_{Bayes}(x^n)$ は

$$L^{\theta^k}_{Bayes}(x^n) = -\log \int_{\theta^k} P(x^n|\theta^k) f(\theta^k) d\theta^k. \tag{2.27}$$

で与えられる [25, 26, 80]．ここで，積分計算は Θ^k 上で行われるものとする．□

パラメータが量子化されたあとを考えると，

$$-\log \sum_{\bar{\theta}^k} P(x^n|\bar{\theta}^k) f(\bar{\theta}^k) \prod_{j=0}^{k-1} \alpha_j(\bar{\theta}^k).$$

という損失の予測法も構成でき，これは量子化後で定義するベイズリスクの意味で最適解を与える．しかし，事前密度が与えられていれば(2.27)式がベイズ最適であり，これは上式で $\alpha_j(\bar{\theta}^k) \to 0$ とすることによっても得られる [93]．すなわち，(2.27)式は累積対数損失を適用した場合のベイズ最適解である．このベイズ最適性は Matsushimaら [80]によって示されたが，それより以前に Rissanen が SC

をこの形で定義している [100]．また，Matsushimaらの論文[80]では，最悪の場合の累積対数損失を最小にするというmin-max規準を達成する予測が，ベイズ解のクラスに存在し，これが $P(x^n|\theta^k)$ を x^n と θ^k 間の通信路と考えた場合の通信路容量を達成する事前分布 $f(\theta^k)$ で与えられることを示している．このような性質から，ベイズ規準はその重要性が情報源符号化の分野で再認識され，多くの研究者によって議論されるようになった．現在ではベイズ統計理論が符号化の問題に対して非常に相性がよいと広く認識されている．

累積対数損失に対するベイズ規準の漸近的な評価としては，最初に Clarke と Barron がベイズ規準の漸近的なリスクを解析し [25],

$$E^*\left[L^{\theta^k}_{Bayes}(x^n)\right] = E^*\left[-\log P(x^n|\theta^{k*})\right] + \frac{k}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det I(\theta^{k*})}}{f(\theta^{k*})} + o(1), \quad (2.28)$$

という漸近式を導いた．ただし，θ^{k*} は真のパラメータであり，$E^*[\cdot]$ は真の分布 $P(\cdot|\theta^{k*})$ による期待値である．さらに，Clarke と Barron はこの議論を押し進め [26]において，漸近的なベイズリスクを解析することに成功し，漸近的にmin-max規準を達成する予測が，ベイズ規準において ジェフリーズの事前分布(Jeffrey's prior)を事前分布として用いた場合に達成されることを明らかとした．しかし，この解析は限られた定常無記憶情報源のクラスに限定されたものであり，この結果をより広い情報源のクラスに拡張する研究が現在も行われている．また，J.Suzuki もベイズ規準の累積対数損失，min-max規準の累積対数損失，max-min規準の累積対数損失について，より詳細な項までの精密な展開式を与えている[126].

パラメトリックモデル族に対する MDL規準とベイズ規準の差に関しては竹内の研究 [136, 137]がある．この論文では，ベイズ規準による混合分布がもとのクラスに必ずしも含まれないことを考慮し，これをもとのクラスに射影した射影ベイズ推定量とMDL原理によって得られる推定量(MDL推定量)の差を解析し，漸近的に両者が一致することが示されている．しかし，MDL規準とベイズ規準の累積対数損失の差については，これまで解析がなされていなかった．

本研究では，パラメータの事後確率密度の漸近正規性を直接用いた解析により，両者の累積対数損失の差を解析している．

階層モデル族の場合

階層モデル族 $\{P(\cdot|m,\theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}, m \in \mathcal{M}\}$ を考える．

$-\log P(m)$ をモデル m の記述長とし，各 m のパラメータの事前密度 $f(\theta^{k_m}|m)$

が与えられたものとする．このとき，階層モデル族に対しては，2つのタイプの MDL 規準を考えることができる．

まず，パラメトリックモデル族の場合と同様に，パラメータ集合を量子化して m と量子化パラメータを符号化し，これらを用いた確率モデルで x^n を符号化する方法を示す．これは，統計的モデル選択問題に対する MDL規準としても提案されているものである．

まず，$\bar{\Theta}_n^{k_m}$ をモデル m のパラメータの代表点 $\bar{\theta}^{k_m}$ の集合，$J(\theta^{k_m}|m)$ をモデル m のFisher情報行列とする．

$$J(\theta^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E_\theta \left[-\frac{\partial^2 \log P(X^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} \right], \tag{2.29}$$

ただし，E_θ は $P(X^n|m,\theta^{k_m})$ による期待値を表す．

x^n を符号化する際には，m と $\bar{\theta}^{k_m} \in \bar{\Theta}_n^{k_m}$ を符号化する．ただし，符号長のロスが生じないように $\sum_{\bar{\theta}^k \in \bar{\Theta}_n^k} f(\bar{\theta}^{k_m}|m) \frac{1}{\sqrt{n}^{k_m}\sqrt{\det J(\bar{\theta}^{k_m}|m)}} = 1$ をみたすように代表点を選んでいるものとする．

定義 2.8 (階層モデル族に対する MDL規準 Type1) パラメータを量子化するタイプのMDL規準の累積対数損失，すなわちMDL符号の符号長 $L_{MDL}^{m,\bar{\theta}^{k_m}}(x^n)$ は

$$L_{MDL}^{m,\bar{\theta}^{k_m}}(x^n) =$$

$$\min_{m\in\mathcal{M},\bar{\theta}^{k_m}\in\bar{\Theta}_n^{k_m}} \left\{ -\log P(x^n|m,\bar{\theta}^{k_m}) - \log f(\bar{\theta}^{k_m}|m) \frac{1}{\sqrt{n}^{k_m}\sqrt{\det J(\bar{\theta}^{k_m}|m)}} - \log P(m) \right\}, \tag{2.30}$$

で与えられる． □

もう一つのタイプの MDL規準は，パラメータについては量子化をせずに混合をとり，

$$-\log P(x^n|m) = -\log \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{2.31}$$

の符号長で x^n を符号化する方法である．このとき，m のみを一緒に符号化すれば一意復号可能な符号が構成できる．

定義 2.9 (階層モデル族に対する MDL 規準 Type2) パラメータについては混合を用い，モデル m のみを選択するタイプの MDL規準の累積対数損失，すなわち MDL符号の符号長 $L_{MDL}^{m,\theta^{k_m}}(x^n)$ は

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = \min_{m\in\mathcal{M}} \left\{ -\log P(x^n|m) - \log P(m) \right\}, \tag{2.32}$$

で与えられる． □

後者の Type 2 の MDL 規準は，m を固定したときのパラメータに対するベイズ規準による予測を構成し，量子化操作を排除している．一方，m の事前確率 $P(m)$，θ^{k_m} の事前密度 $f(\theta^{k_m}|m)$ が与えられれば，ベイズ最適な予測が構成でき，これは全ての m，θ^{k_m} の混合モデルを用いた予測で与えられる．

定義 2.10 (階層モデル族に対するベイズ規準) 階層モデル族に対するベイズ規準の累積対数損失，すなわちベイズ符号の符号長 $L_{Bayes}^{m,\theta^{k_m}}(x^n)$ は，

$$L_{Bayes}^{m,\theta^{k_m}}(x^n) = -\log \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m}, \tag{2.33}$$

で与えられる．ただし，各積分操作は Θ^{k_m} 上でとられる． □

以上で定義してきたように，MDL規準が最も全記述長を短くするような確率モデルを選ぶのに対し，ベイズ規準は仮定したクラス内全ての確率モデルの混合モデルを用いる点が特徴である．

階層構造のモデル族に対しては，統計的モデル選択規準としての MDL原理に関する研究が90年代中頃まで活発になされた．MDL規準が強一致性をもつ選択規準であることを受けて，活発に応用研究もなされ，多くの事例でその有効性が示されている．例えば，確率密度のヒストグラム推定量に適用した場合の収束速度の解析[52]，決定木の選択問題[94]やパラメータの変化位置検出問題[106]など様々な事例に応用することによる有効性の検証などが行われ，多くの成果が得られている．MDL原理は Rissanen 自身によって，いくつかの形に変形されながら確率的コンプレキシティ(stochastic complexity)の概念に発展され[100]，ベイズ統計理論に基づく考察も広く行われている．現在ではベイズ符号も MDL 規準の一つと位置づけられて研究が進んでいる[12]．MDL原理では，“パラメトリックモデル族に対して最小記述長をどのように構成するか” が，現在のところ最大の焦点となっているが，もし事前分布が与えられるのならばベイズ符号が最適であることから，これをパラメトリックモデル族の評価とする流れは自然であると思われる．

また実際の符号化問題では，半階層モデル族である FSMXモデル族に対して，ベイズ符号の効率的な符号化アルゴリズムが示され[70, 81]，ベイズ符号の性質を明らかにすることの重要性が高まっている．このアルゴリズムは符号化のみならず，パターン認識などの予測にそのまま適用可能である点で興味深いものとなっている[114].

一方，符号化における符号長最小という規準は，必ずしも平均符号長に限ったものではない．最悪のケースを最小としようとする min-max 規準，その逆の

max-min 規準を達成する符号化も広く研究されており[28, 29, 32, 26]，ベイズ符号のクラスに min-max符号や max-min符号が存在することが明らかにされている[26, 80]．パラメトリックモデル族に対しては，min-max符号長と max-min符号長が一致し，ジェフリーズの事前分布(Jaffreys' prior)を仮定したベイズ符号が，漸近的に min-max符号，max-min符号を達成することが Clarke と Barron によって示され[26]，その後の多くの研究の土台となっている[138]．また，最近では最尤符号[101]という符号が示され，冗長度に代わる新たな評価として個々の系列 x^n に対するリグレット(regret) [12]の議論が注目されている．パラメトリックモデル族に対するベイズ符号は，この最尤符号とも密接に関連して研究されている[138]．しかし，階層モデル族に対するベイズ符号については現在研究が始められたといってよい段階であり[70]，十分体系化されているとはいえない．

本研究への展開

以上に述べたように，MDL規準とベイズ規準の性質に関しては，独立にそれぞれの性質が解析されている研究が多く，非常に基本的な問題である両者の累積対数損失の差，すなわち符号長の差の解析がなされていなかった．そこで第3章，第4章において，これまであげた3つのモデルクラスに対して，両者の累積対数損失を解析することにより，両規準の性能の差について累積対数損失の面から考察する．

一方，ベイズ規準の累積対数損失の漸近的性質については，パラメトリックモデル族に対して多くの結果が示されているが，階層モデル族に関してはあまり解析がなされていない．また，ベイズ規準による混合確率の実際のアルゴリズムとして，有限マルコフ情報源の一種である FSMXモデル族に対して，効率的なアルゴリズムが提案されている．そこで第5章では，半階層モデル族の一つである FSMXモデル族に対してベイズ規準の漸近的累積対数損失を解析し，個々の系列に対する累積対数損失，平均的な累積対数損失の漸近式を導いている．これにより，FSMX モデル族に対するベイズ規準の漸近的評価が与えられ，最適な平均損失への収束速度が与えられたことになる．これは，情報源符号化問題では，符号長の評価ができたことに相当する．

2.5.4 データ系列の逐次予測問題 (2): 一般の累積損失に対する展開

本節では，本研究で扱う拡張確率的コンプレキシティの概念について述べる．

確率的コンプレキシティの損失一般化

J.Rissanen は MDL 原理 [95]を発展させた概念として，確率的コンプレキシティ(stochastic complexity: SC)[101] を提唱している．これは，真のパラメータが未知であるパラメトリックなクラスが与えられたとき，このクラスを用いてデータ系列を符号化する方法の中での最小記述長として定義できる．すなわち，ある確率モデル族を考えたときのデータ系列の複雑さを，最小記述長で測るという考え方である．しかし，最小記述長といっても，特定のデータ系列に対して性能の良い符号化はいくらでも作ることができるので(当然ながら，他のデータ系列に対しては性能が悪くなる)，何らかの客観的な構成法が必要となる．しかし，パラメータ空間に事前確率が仮定できる場合には，ベイズ最適な符号を考えることは自然であり，いわゆるパラメータに対して混合をとったベイズ符号の符号長が SC となる．最近，K.Yamanishi は SC で考えている対数損失を一般化し，様々な形の損失関数を適用できる形に発展させた[153, 154]．これを extended stochastic complexity (ESC) と呼ぶ．

拡張確率的コンプレキシティ

ここでは，n 組の独立にランダムサンプリングされた確率変数の系列 $x^n = x_1 x_2 \cdots x_n$ を考える．階層モデル族を考え，$m \in \mathcal{M}$ をモデルの構造を決める離散ラベル(モデルと呼ぶ)とし，確率分布のクラスが，$\{p(\cdot|m, \theta^{k_m}) | m \in \mathcal{M}, \theta^{k_m} \in \Theta^{k_m}\}$ で与えられている場合を考える．$\theta^{k_m} \in \Theta^{k_m}$ は k_m 次元パラメータ，Θ^{k_m} はユークリッド空間 $\mathcal{R}^{k_m}$ の部分集合とする．

MDLとその拡張概念といえるSCにはいくつかのタイプがあるが，本研究では次の式 $SC(x^n : m)$ を考える[100].

$$SC(x^n : m) = -\log q(x^n|m), \tag{2.34}$$

$$q(x^n|m) = \int_{\theta^{k_m} \in \Theta^{k_m}} p(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{2.35}$$

ここで，$f(\theta^{k_m}|m)$ は事前密度である．$-\log q(x^n|m)$ は一意復号可能な符号(パラメータについて混合をとるベイズ符号)の符号長であり，これを最小化する m が記述長最小のモデル選択となる．勿論，これは対数損失を考えていることになる．また $-\log q(x^n|m) = -\sum_{t=1}^{n} \log q(x_t|x^{t-1}, m)$ と分解できるので1パスの逐次符号化と同じ符号長となり，m が既知のもとでのベイズ最適性を有している．これに対し，ESCは $L(X : d)$ を決定 d の X に対する損失関数とし，

$$ESC(x^n : m) = -\frac{1}{\lambda} \log \int_{\theta^{k_m} \in \Theta^{k_m}} \exp\left(-\lambda \sum_{t=1}^{n} L(x_t : d_{\theta^{k_m}})\right) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{2.36}$$

で定義される[153]. 例えば，決定 d として X の確率 $P(X)$ を与えるときの損失関数としては，α 損失: $L(X:P) = (1-P(X))^{\alpha}$, Hellinger 損失: $L(X:P) = 1-\sqrt{P(X)}$, 対数損失: $L(X:P) = -\log P(X)$ などをとることができる．ここで，

$$\overline{L}(x_t|x^{t-1},m) = -\frac{1}{\lambda}\log\int_{\theta^{k_m}\in\Theta^{k_m}} \exp\left(-\lambda L(x_t:d_{\theta^{k_m}})\right)\pi(\theta^{k_m}|x^{t-1},m)d\theta^{k_m}, \tag{2.37}$$

$$\pi(\theta^{k_m}|x^{t-1},m) = \frac{\exp\left(-\lambda\sum_{j=1}^{t-1}L(x_j:d_{\theta^{k_m}})\right)f(\theta^{k_m}|m)}{\int\exp\left(-\lambda\sum_{j=1}^{t-1}L(x_j:d_{\theta^{k_m}})\right)f(\theta^{k_m}|m)d\theta^{k_m}}, \tag{2.38}$$

と定義すると，

$$ESC(x^n:m) = \sum_{t=1}^{n}\overline{L}(x_t|x^{t-1},m), \tag{2.39}$$

とかける[153]. これよりオンラインの逐次学習問題に適用できる．ただし，$\pi(\theta^{k_m}|x^{t-1},m)$ はベイズの事後確率ではない．

本研究への展開

K.Yamanishi は ESC の期待値と個々の系列に対し一様に成り立つ上界を導いている．これは，n をデータ数として誤差が $o(\log n)$ の漸近形となっている．これにより，ESC を用いた学習アルゴリズムの漸近的評価が可能となる．すなわち，これまでのベイズ符号に対する研究と同様，ESC に対してもその漸近式を明らかにしておくことは有意義である．しかし，ベイズ符号の符号長に関しては，$O(1)$ の項まで詳細な漸近式が得られているのに対し，ESC に関して得られているのは上界式であり，さらに厳密な漸近式を与えることが未解決の問題となっていた．本研究では第8章において，ESC が導くパラメータ上の分布列 $\pi(\theta^{k_m}|x^n,m)$ がベイズ規則と類似の性質をもつことに着目する．そして，この分布列が漸近正規性をみたすことを示すことにより，概収束と平均収束の意味で $O(1)$ のオーダまで詳細な ESC の漸近式を導く．

2.5.5 過去の観測に基づく一時点の予測問題

与えられた観測データ(学習データ) x^n をもとに，次のデータを予測する問題は，統計学の最も基本的な問題の一つであり，その研究範囲も非常に広い．実務的な応用統計では，観測データから分布のパラメータなどを推定し，その分布をもとにデータを予測することが一般的である．この場合，どのように推定

量を構成すれば精度の良い予測が可能か，という点が問題となる．このように考えると，次節で述べる統計的モデル選択を予測の観点から定式化することができる．一方，ベイズ決定理論に基づく予測を考えると，モデルやパラメータを1つに限定(推定)せず，事後確率で全てを平均化した混合分布(予測分布)がベイズ最適な予測を与える．以下では，AIC とベイズ最適な予測分布についてまとめる．

赤池情報量規準(Akaike information criterion)

予測を考えたモデル選択規準として最も広く認識され，有効性が確認されている規準は，赤池情報量規準(Akaike information criterion: AIC) [1]である．階層モデル族を考え，$\hat{\theta}^{k_m}$ を最尤推定量として，モデル m の AIC は

$$AIC(m) = \log p(x^n|m, \hat{\theta}^{k_m}) - k_m, \tag{2.40}$$

で与えられる．以下では，厳密性は多少犠牲にして，その導出の概略を述べる．最尤推定量 $\hat{\theta}^{k_m}$ を用いた確率分布 $p(\cdot|x^n, m, \hat{\theta}^{k_m})$ のよさを測るために，KL-情報量

$$D(p^*(X_{n+1}|x^n)\|p(X_{n+1}|x^n, m, \hat{\theta}^{k_m}))$$

$$= \int_{x_{n+1}\in\mathcal{X}} \log p^*(x_{n+1}|x^n) dp^*(x_{n+1}) - \int_{x_{n+1}\in\mathcal{X}} \log p(x_{n+1}|x^n, m, \hat{\theta}^{k_m}) dp^*(x^{n+1}), \tag{2.41}$$

を考える．KL-情報量は非負であり，二つの確率分布が一致するときに限り 0 となる．右辺第1項は確率モデル $p(\cdot|m, \hat{\theta}^{k_m})$ に依存しない項なので，第2項のみを考えればよい．第2項の $\hat{\theta}^{k_m} = \hat{\theta}^{k_m}(X^n)$ は X^n により決まる確率変数であるから，その平均的な良さ(リスク関数)は

$$-E^*\left[\log p(X_{n+1}|X^n, m, \hat{\theta}^{k_m})\right] - \int_{x_{n+1}\in\mathcal{X}} \log p(x_{n+1}|x^n, m, \hat{\theta}^{k_m}) dp^*(x^{n+1}|x^n) dp^*(x^n), \tag{2.42}$$

で与えられる．すなわち，これを最小化するモデル m が，KL-情報量の意味で最適なモデルといえる．しかし，このリスクを事前に知ることはできず，知ることができるのは x^n のみである．そこで x^n から，このリスク関数の推定量を求めることを考えよう．その推定量の一つが AIC の $1/n$ で与えられることになる．そのために最尤推定量の漸近正規性が本質的となる．良く知られるように，適当な正則条件のもとで，$\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*})$ の分布は，$n \to \infty$ のとき正規分布 $N(0, \{J^*(\theta^{k_m}|m)\}^{-1} I^*(\theta^{k_m}|m) \{J^*(\theta^{k_m}|m)\}^{-1})$ に法則収束する．ただし，

$$I^*(\theta^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E^* \left[\frac{\partial \log p(X^n|m, \theta^{k_m})}{\partial \theta^{k_m}} \frac{\partial \log p(X^n|m, \theta^{k_m})}{(\partial \theta^{k_m})^T} \right], \tag{2.43}$$

$$J^*(\theta^{k_m}|m) = -\lim_{n\to\infty}\frac{1}{n}E^*\left[\frac{\partial^2 \log p(X^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right], \tag{2.44}$$

である．ただし，E^* は真の分布 $p^*(\cdot)$ による期待値を表す．もし，$p^*(\cdot) = p(\cdot|m,\theta^{k_m^*})$ であれば，$I^*(\theta^{k_m^*}|m) = J^*(\theta^{k_m^*}|m)$ となり，Fisher情報行列となる．

まず，$-\log p(x^n|m,\theta^{k_m^*})$ を $\hat{\theta}^{k_m}$ のまわりでテーラー展開すると，

$$-\log p(x^n|m,\theta^{k_m^*}) \sim -\log p(x^n|m,\hat{\theta}^{k_m}) + \frac{n}{2}(\hat{\theta}^{k_m}-\theta^{k_m^*})^T J^*(\hat{\theta}^{k_m}|m)(\hat{\theta}^{k_m}-\theta^{k_m^*}), \tag{2.45}$$

となる．ここで，$\sim$ は近似式であることを意味する．$\sqrt{n}(\hat{\theta}^{k_m}-\theta^{k_m^*})$ の漸近的な共分散は $\{J^*(\theta^{k_m^*}|m)\}^{-1}I^*(\theta^{k_m^*}|m)\{J^*(\theta^{k_m^*}|m)\}^{-1}$ であるから，$J^*(\hat{\theta}^{k_m}|m) \sim J^*(\theta^{k_m^*}|m)$ を用い，期待値をとれば，

$$-E^*\left[\log p(X^n|m,\theta^{k_m^*})\right] \sim -E^*\left[\log p(X^n|m,\hat{\theta}^{k_m})\right] + \frac{1}{2}\mathrm{Tr}\left(I^*(\hat{\theta}^{k_m}|m)\{J^*(\hat{\theta}^{k_m}|m)\}^{-1}\right), \tag{2.46}$$

が得られる．ただし，Tr はトレースを表す．一方，$-\log p(x_{n+1}|x^n,m,\hat{\theta}^{k_m})$ を $\theta^{k_m^*}$ のまわりでテーラー展開して同様の議論を行えば，(2.42)式の n 倍は

$$-E^*\left[\log p(X^n|m,\theta^{k_m^*})\right] + \frac{1}{2}\mathrm{Tr}\left(I^*(\theta^{k_m^*}|m)\{J^*(\theta^{k_m^*}|m)\}^{-1}\right), \tag{2.47}$$

と近似できるので，再び $J^*(\hat{\theta}^{k_m}|m) \sim J^*(\theta^{k_m^*}|m)$ を用い，(2.46)式とから結局(2.42)式の n 倍の漸近的な不偏推定量として，

$$-\log p(x^n|m,\hat{\theta}^{k_m}) + \mathrm{Tr}\left(I^*(\hat{\theta}^{k_m}|m)\{J^*(\hat{\theta}^{k_m}|m)\}^{-1}\right), \tag{2.48}$$

が導かれる．これば竹内の情報量規準(TIC)と呼ばれる．さらに，真の分布がモデルクラスに十分近ければ，$I^*(\theta^{k_m^*}|m) \sim J^*(\theta^{k_m^*}|m) \sim J(\theta^{k_m^*}|m)$ と近似でき，$\mathrm{Tr}\left(I^*(\theta^{k_m^*}|m)\{J^*(\theta^{k_m^*}|m)\}^{-1}\right) = k_m$ となって AIC が導かれる．

ベイズ最適な予測分布

ここでは，階層モデル族に限定して，対数損失を仮定した場合のベイズ最適な予測法を与える．離散モデル族，パラメトリックモデル族に関しても同様である[80]．

まず，x^n が与えられたという前提のもとで，次の対数損失関数を与える．

$$L(x : Ap|x^n) = -\log Ap(x|x^n). \tag{2.49}$$

これは，決定 $Ap(\cdot|x^n)$ に対して特定の $x_{n+1} = x$ が与えられたときの損失であり，x は事前に分からないので，次の平均損失，すわなちリスク関数を考える．

$$R(Ap,m,\theta^{k_m}|x^n) = -\int_{x\in\mathcal{X}} \log Ap(x|x^n)dp(x|x^n,m,\theta^{k_m}). \tag{2.50}$$

ただし，積分はルベーグ積分である．(2.50)式は先の AIC の導出の際に設定した KL-情報量の右辺第2項であり，ここまでは AIC の考え方と同様である．ここで，m, θ^{k_m} は未知であるが，事前分布 $P(m), f(\theta^{k_m}|m)$ が仮定できるものとしよう．このとき，x^n を知ったもとでの事後分布 $P(m|x^n), f(\theta^{k_m}|x^n, m)$ が得られる．

$$P(m|x^n) = \frac{\int_{\theta^{k_m}\in\Theta^{k_m}} p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m} P(m)}{\sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m}}, \tag{2.51}$$

$$f(\theta^{k_m}|x^n, m) = \frac{p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m)}{\int_{\theta^{k_m}} p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}}. \tag{2.52}$$

この事後確率分布に対する平均リスク，すなわち条件付きベイズリスク

$$\begin{aligned} BR(Ap|x^n) &= \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} R(Ap, m, \theta^{k_m}|x^n) f(\theta^{k_m}|x^n, m) P(m|x^n) d\theta^{k_m}, \\ &= \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} \left[-\int_{x\in\mathcal{X}} \log Ap(x|x^n) dp(x|x^n, m, \theta^{k_m}) \right] f(\theta^{k_m}|x^n, m) P(m|x^n) d\theta^{k_m}, \end{aligned} \tag{2.53}$$

を最小化すれば，これは事前分布が与えられたときの，平均的な意味での最適な予測を与える．さらに，フビニの定理(Fubini's theorem)[23, 105]を利用して積分の順序を入れ替えれば

$$BR(Ap|x^n) = -\int_{x\in\mathcal{X}} \log Ap(x|x^n) dp(x|x^n). \tag{2.54}$$

ただし，

$$p(x|x^n) = \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|x^n, m) P(m|x^n) d\theta^{k_m}, \tag{2.55}$$

となり，シャノンの補助定理[58, 67]より，

$$Ap(x|x^n) = \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|x^n, m) P(m|x^n) d\theta^{k_m}, \tag{2.56}$$

が最適な予測分布として与えられることがわかる．この分布を単に予測分布と呼ぶことがある．これは当然ながら，ベイズ符号における次の1シンボルの符号化関数を表している[80]．この予測法に関しては最近でも，不確実性を含む論理式の推論に適用した研究[78]や，予測分布の計算に含まれる積分計算を近似式で与え，近似予測分布で予測する方法[49]などが提案されている．

本研究への展開

AIC は統計的モデル選択の枠組みで予測を扱っているのに対し，ベイズ決定理論に基づけばモデルを一つ選択するのではなく，全てのモデルの混合分布(予

測分布)が最適となることがわかる．確かにベイズ統計学では推測を事後確率の形で与えるので，その事後確率の平均が良いことは直感的にも理解できる．したがって，ベイズ統計理論の枠組みで考えるならば，予測が目的のときは予測分布を採用し，最適モデルを選択したいという目的の場合には，モデル間に損失関数を設定してモデル選択を行うのが理に適っていると考えることもできる．しかし，現実問題として AIC の考え方にあるように，“予測を考慮して，モデルを一つに特定したい” という考え方も広く受け入れられているのも現状である．

そこで本研究では第6章において，ベイズ決定理論に基づき，一時点の予測に対する損失関数を考えた上で，ベイズ最適なモデル選択を定式化する．さらに，この規準で選択されたモデルは，漸近的に最適モデルと一致するというモデル選択の一致性を証明する．

2.6 本研究の位置づけ

表2.3には，前節で分類した統計的推測の目的(評価関数)という観点からみた分類を示す．すなわち，表2.3は各章で扱っている推測が，どのような損失関数を考えているかを示している．また，表2.4には，確率モデルを1つ選択するか，全ての混合モデルを用いるかという手法的観点からみた分類を示す．

表2.5は，各章でどのような漸近的性質に着目しているかをまとめた表である．各章は，先に述べたように，考えている損失関数で分類することができるが，ある損失関数を考えた推測が他の評価基準に対してどのようなパフォーマンスを示すかを調べることは，その性質を議論する上で有用と考えられる．本研究では，そのような観点から，第6章，第7章において推測方式を構成する際に考えている損失関数と異なる評価基準(一致性，選択誤り率)による評価を行っている．その他の章では，設定した損失関数を用いて漸近的性質を評価している．

最後に事後密度が正規分布で近似できるために必要なサンプル数について述べておく．これはモデルの複雑さに依存するので一般的な議論ができない．すなわち必要サンプル数とモデルの複雑さにはトレード オフの関係があるが，漸近正規性を実務問題に適用して簡便な予測法を構成するような場合には，対象問題と関連して必要サンプル数の議論が必要である．しかし，本研究では漸近的性質の理論的解析のために漸近正規性を用いているので，この目的のためにサンプル数の議論は特に必要ではない．

表 2.3: 統計的推測の目的(損失関数)からみた各章の分類

$(2-1)$ **逐次予測問題**	$(2-2)$ **一時点の予測問題**	(1) **モデル推定問題**
MDL規準とベイズ規準の 累積対数損失の比較 (第3章，第4章) [40, 45] ベイズ規準による予測の 累積対数損失の漸近式 (第5章) [39] 一般の累積損失関数に 対するESCの漸近式 (第8章) [44]		ベイズ決定理論に基づく モデル選択の誤り率 (第7章) [46]
	ベイズ決定理論に基づくモデル選択 の一致性(第6章) [40, 45]	

表 2.4: 手法からみた各章の分類

モデル選択(モデルを1つ選択) ユニバーサルモデリング	**混合モデル(全てのモデルで平均化)**
ベイズ決定理論に基づくモデル選択と一致性の解析(第6章) [40, 45] ベイズ決定理論に基づくモデル選択の選択誤り率の上界(第7章) [46] ESC の漸近式の解析(第8章) [44]	ベイズ規準の累積対数損失の解析 ：FSMX モデルに対する解析(第5章) [39]
MDL規準とベイズ規準を累積対数損失で比較 (第3章，第4章) [42, 43]	

表 2.5: 解析している漸近的性質の差異

各章	解析する漸近的性質，評価基準
MDL規準とベイズ規準の比較 (第3章，第4章) [40, 45] ベイズ規準の漸近的評価 (第5章) [39]	累積対数損失
推定問題と予測問題に対するモデル選択 (第6章) [40, 45]	一致性
推定問題に対するモデル選択 (第7章) [46]	モデル選択の誤り率
ESCの漸近式 (第8章) [44]	一般の累積損失

第 3 章

累積対数損失を評価関数とする予測問題 (1): MDL 規準とベイズ規準の差の解析 (事前分布が同じ場合)

3.1 本章の目的

J.Rissanen によって提案されたMDL原理 [95, 100] は情報源符号化のみならず, データ解析や学習理論など広く適用・研究がなされている [52, 56, 64, 143]. MDL 原理は得られたデータを最も短く符号化することのできる確率モデルを選択する方法であるが, 符号長による評価は対数損失を考えていることと等価であるので, これは対数損失を評価関数とした予測問題を扱っていることになる. 対数損失は予測の損失関数として古くから議論されてきた評価関数であり, そのため多くの予測問題に適用することが可能となっている. 実際, MDL規準を様々な現実問題に適用する研究が行われており, 多くの事例でその有効性が示されている[106].

本研究では MDL 原理に基づく2段階符号(MDL符号)の符号長を規準とした予測を考える. MDL規準はデータの記述長とそれを符号化するための確率モデルの記述長の和を最小にするようなモデルを選択する方法であり, モデル選択としての機能を備えている. また, MDL規準はベイズ統計と深く関係することが明らかにされており, モデルの記述長を定義することはモデル族に暗に事前分布を仮定していることになる. なお, MDL規準で選ばれるモデルは本質的に事後確率最大推定と等価である.

一方, 最近情報源符号化において活発に研究がなされているベイズ規準 [25, 80] はモデル族内全てのモデルの混合モデルを予測に用いる方法である. ベイズ最適な符号はベイズ符号と呼ばれるが, その解は一般的な予測問題に対するベイズ解を符号化に用いるものであり, 本質的には符号化も予測問題の一つととら

えることができる．この観点から，累積対数損失(符号長)はデータの統計的な複雑さを測る尺度である確率的コンプレキシティ(stochastic complexity) [93, 100] としても定義され，統計的モデル化の観点からも多くの研究がなされている [101]. また，ベイズ規準はすなわちベイズ最適性を有するので [80, 93]，同じ事前分布の場合にはベイズ規準による累積対数損失(符号長)がMDL規準のそれを下界することが知られている [100]．さらに，最近情報源符号化の分野では，実際の情報源符号化で重要な有限マルコフ情報源の一種である FSMX情報源に対し，効率的に混合確率を計算するアルゴリズムが提案された [70, 81]．これにより，ベイズ最適な符号(ベイズ符号)の構成が現実的となったが，同時にこのアルゴリズムはパターン認識問題など多くの予測問題に適用可能であり，ベイズ規準による予測の実現性が高まってきている．

MDL規準とベイズ規準の性質や応用については様々な観点から研究がなされている [25, 26, 52, 70, 80, 95, 100, 135, 143]．両規準による予測が漸近的に平均損失の限界を達成するユニバーサルな予測を与えること [25]，またその収束速度も $O(\frac{\log n}{n})$ で最適あること [56]などが明らかにされている．また，MDL原理に基づく推定量とベイズ規準の予測分布をもとのクラスに射影した射影ベイズ推定量の差を解析した研究もあり [136]，それぞれ興味深い結果を与えている．

もともとMDLは符号化という概念を統計的推測に持ち込んだものであり，符号長を統計的推測の評価基準としている．一方，ベイズ規準による予測も，情報源符号化の分野ではベイズ符号として広く認識されている．すなわち，統計的予測と情報源符号化が同じ問題を扱っており，符号長による評価は予測において本質的であるという認識が高まっている．よって，符号長という観点からこれらの規準を考察することは，符号化問題のみならず，この概念を導入した推定や予測の性質を明らかにするためにも必要といえる．しかし，両規準による予測の累積対数損失の差，すなわち符号長の差がどの程度なのか，という問題については，これまで解析がなされていなかった．

そこで本章では，同じ事前分布が仮定されている場合の MDL 規準とベイズ規準による予測の累積対数損失の差，すなわち両符号化の符号長について，定量的な評価を与えることを目的とする．解析の結果，両符号の符号長の差は，離散モデル族に対してはモデルクラスによって $o(1)$ または $O(1)$，パラメトリックモデル族に対しては $O(1)$，階層モデル族に対しては MDL符号の構成法によって $o(1)$ または $O(1)$ のオーダで与えられることを明らかとする．本章のパラメトリックモデル族，階層モデル族の解析においては，パラメータの事後確率密度の漸近正規性が本質的な役割を演じている．

3.2 離散モデル族に対する解析

本節では，離散モデル族に対して (2.16) 式で与えられるMDL規準の累積対数損失と (2.22) 式で与えられるベイズ規準の累積対数損失の差を解析する．これを情報源符号化の言葉で言い直せば，“MDL原理に基づく符号(MDL符号)とベイズ最適な符号(ベイズ符号)の符号長の差を解析している” ということになる．以下では，MDL原理が符号化という概念から出発していることを受けて，情報源符号化の立場からの記述を行う．

符号長を規準としたMDL原理はそのまま推定や予測に適用されるなど，応用統計の分野においても符号長という評価の有効性に関する認識が高まっている．したがって，符号長を議論することによって適用が情報源符号化に限定されるものではない．MDL符号とベイズ符号の符号長の差を定量的に与えることは，対数損失を考えた予測，あるいは符号長による評価を適用した統計的推定や予測の観点からも必要であると考えられる．

また，本章を通じて離散アルファベット $\mathcal{X}$ を考えているが，これを連続アルファベットへ拡張することは容易である．

3.2.1 モデル族の条件

集合 D_0 を $D^*(P^*\|P_\lambda)$ を最小化する λ の集合とする．ただし，$D^*(P^*\|P_\lambda)$ はKL-情報量

$$\begin{aligned} D^*(P^*\|P_\lambda) &= \lim_{n\to\infty}\frac{1}{n}E^*\left[\log\frac{P^*(X^n)}{P(X^n|\lambda)}\right] \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{x^n}P^*(x^n)\log\frac{P^*(x^n)}{P(x^n|\lambda)}, \end{aligned} \tag{3.1}$$

であり，E^* は X^n に対する真の確率分布 $P^*(X^n)$ による期待値とする．$D^*(P^*\|P_\lambda)$ の最小化は

$$H^*(\lambda) = \lim_{n\to\infty}\frac{1}{n}E^*\left[-\log P(X^n|\lambda)\right], \tag{3.2}$$

の最小化と等価である．

$D^*(P^*\|P_\lambda)$ を最小化する λ を λ^* と記述する．ただし，λ^* は唯一とは限らない．すなわち，$|D_0|\geq 1$ であり，λ^* を最適モデルとよぶ．λ^* が唯一とはならない例を次に述べる．例えば，λ_1 と λ_2 が，$P(\cdot|\lambda_1)\neq P(\cdot|\lambda_2)$ であるが $H^*(\lambda_1)=H^*(\lambda_2)$ は満たす場合が存在する．また，クラス内に同じ確率分布を表すモデルが複数存在する場合，すなわち $P(\cdot|\lambda_1)=P(\cdot|\lambda_2)$ となる場合も $H^*(\lambda_1)=H^*(\lambda_2)$ となる．

これらの場合，λ_1 が $H^*(\lambda)$ を最小化すれば，λ_2 も最小化することになるので，$|D_0| \geq 2$ となる．

集合 $D_1 \subset \Lambda$ を $D^*(P^*\|P_\lambda)$ を最小化しない λ の集合とする．すなわち，$\mathcal{M} = D_0 \cup D_1$ である．

条件 3.1

(1) $|D_0| \geq 1$.

(2) $\forall \lambda \in \Lambda$ に対して $P(\lambda) > 0$.

(3) (大数の強法則) $\forall \lambda \in \Lambda$ に対して

$$-\frac{1}{n}\log P(x^n|\lambda) \to H^*(\lambda), \quad a.s. \tag{3.3}$$

ただし，$a.s.$ は概収束(almost sure convergence)を表す．すなわち，$\forall \lambda \in \Lambda$ と $\forall \epsilon > 0$ に対して確率1で，

$$\left| -\frac{1}{n}\log P(x^n|\lambda) - H^*(\lambda) \right| < \epsilon, \tag{3.4}$$

が十分大きな全ての n について成り立つ． □

(3.3)式や(3.4)式の概収束は

$$P^*\left(\left| -\frac{1}{n}\log P(x^n|\lambda) - H^*(\lambda) \right| \geq \epsilon, \text{ infinitely often } \right) = 0, \tag{3.5}$$

であることを意味している[105, 121]．ここで，$P^*(\cdot)$ は真の確率を表す．(3.5)式は $P^*\left(-\frac{1}{n}\log P(x^n|\lambda) \to H^*(\lambda)\right) = 1$ と等価である．もし，

$$\sum_{n=1}^{\infty} P^*\left(\left| -\frac{1}{n}\log P(x^n|\lambda) - H^*(\lambda) \right| \geq \epsilon \right) < \infty, \tag{3.6}$$

が成り立てば，ボレル-カンテリの補題(Borel-Cantelli lemma)[34, 105, 121]より，(3.4)式も成り立つ．これは，x^∞ の部分系列 x^n, $n = 1, 2, \cdots$ のうちの無限個に対して(3.4)式が成り立たないような無限系列は生起しないことを述べている．本論文では，この概収束を例えば，“$\frac{1}{n}\{-\log P(x^n|\lambda)\} \to H^*(\lambda)$, $a.s.$”，あるいは“$n \to \infty$ のとき $|-\frac{1}{n}\log P(x^n|\lambda) - H^*(\lambda)| < \epsilon$, $a.s.$” などと表記する．

(3.3)式が成り立てば，$\forall \lambda_1, \lambda_2 \in \Lambda$ に対して $n \to \infty$ のとき

$$\frac{1}{n}\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)} \to D^*(P_{\lambda_1}\|P_{\lambda_2}), \quad a.s. \tag{3.7}$$

も成り立つ．ここで，

$$D^*(P_{\lambda_1}\|P_{\lambda_2}) = \lim_{n\to\infty} E^*\left[\log\frac{P(X^n|\lambda_1)}{P(X^n|\lambda_2)}\right], \tag{3.8}$$

である．したがって，$\forall c_1 > 0, \forall\lambda^* \in D_0$ と $\forall\lambda \in D_1$ に対し，$n\to\infty$ のとき

$$\exp\{-nD^*(P_{\lambda^*}\|P_\lambda) - nc_1\} < \frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} < \exp\{-nD^*(P_{\lambda^*}\|P_\lambda) + nc_1\},\ \ a.s. \tag{3.9}$$

も成り立つ．定義より $D^*(P_{\lambda^*}\|P_\lambda) > 0$ であるから，$n\to\infty$ のとき

$$\frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \to 0,\ \ a.s. \tag{3.10}$$

が成り立つことがわかる．

次に，$|D_0| \geq 2$ の場合の解析に必要な条件を示しておく．

条件 3.2 $\lambda_1^*, \lambda_2^* \in D_0$ に対し，$P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ であれば，

$$\log\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to +\infty,\ \ a.s. \tag{3.11}$$

または

$$\log\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to -\infty,\ \ a.s. \tag{3.12}$$

が成り立つ． □

条件3.1,(3)より，

$$\frac{1}{n}\log\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to 1,\ \ a.s. \tag{3.13}$$

であるので，条件3.2は $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ となる λ_1^*, λ_2^* に対し $\log\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)}$ が $O(n)$ より小さいオーダーで $+\infty$ か $-\infty$ に発散することを述べている．これは一般的に起こる事実である[1]．ただし，$p(\cdot|\lambda_1^*) = p(\cdot|\lambda_2^*)$ であれば，$\forall n \in \{1, 2, \cdots\}$ に対して

$$\log\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} = 0,\ \ a.s. \tag{3.14}$$

であることは明らかである．

[1]例えば，(3.3) 式において $-\frac{1}{n}\log P(x^n|\lambda) \to H^*(\lambda),\ \ a.s.$ とあるが，これは $\log P(x^n|\lambda) \to nH^*(\lambda)$ と等価ではない．一般に，$|\log P(x^n|\lambda) - nH^*(\lambda)|$ はほとんど確実に ∞ に発散する [34].

3.2.2 離散モデル族の例

例 3.1 (多項分布) $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の多項分布に従うi.i.d.系列を考える．p_i をシンボル $i, i \in \mathcal{X}$ の生起確率とすると，$p_\beta = 1 - \sum_{i=0}^{\beta-1} p_i$ の関係がある．よって，$p^\beta = (p_0, p_1, \cdots, p_{\beta-1})^T$ は一つの確率モデルを表す.

モデル λ_1 は次式の多項分布を表すものとする．

$$p^\beta(\lambda_1) = (p_0(\lambda_1), p_1(\lambda_1), \cdots, p_{\beta-1}(\lambda_1))^T,$$

ただし，$0 < p_0(\lambda_1), p_1(\lambda_1), \cdots, p_{\beta-1}(\lambda_1) < 1$ である．同様に，$\lambda_2, \lambda_3, \cdots$ を

$$p^\beta(\lambda_2) = (p_0(\lambda_2), p_1(\lambda_2), \cdots, p_{\beta-1}(\lambda_2))^T,$$

$$p^\beta(\lambda_3) = (p_0(\lambda_3), p_1(\lambda_3), \cdots, p_{\beta-1}(\lambda_3))^T,$$

$$\cdots\cdots.$$

と定義する．このようにして作ったモデルの集合 $\Lambda = \{\lambda_1, \lambda_2, \cdots\}$ は離散モデル族を構成する．真の確率分布は $p^{\beta *} = (p_0^*, p_1^*, \cdots, p_{\beta-1}^*)^T$ で与えられるものとしよう．ここで，p_i^* はシンボル i の真の生起確率である．このとき，モデル族の中で $\lambda^* \in \Lambda$ を得ることができるが，これは唯一とは限らない.

例えば，$\beta = 2$ に対して $p^2(\lambda_1) = (1/10, 1/10)^T$, $p^2(\lambda_2) = (1/10, 2/10)^T$, $p^2(\lambda_3) = (1/10, 3/10)^T$, $\cdots$, $p^2(\lambda_7) = (1/10, 8/10)^T$, $p^2(\lambda_8) = (2/10, 1/10)^T$, $\cdots$, $p^2(\lambda_{36}) = (8/10, 1/10)^T$ とし，離散モデル族を $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_{36}\}$ とした場合を考える．ここで $p_0^* = p_1^* = p_2^* = 1/3$ であるとすると，$p^2(\lambda_{18}) = (3/10, 3/10)$, $p^2(\lambda_{19}) = (3/10, 4/10)$, $p^2(\lambda_{25}) = (4/10, 3/10)$ の 3 つのモデルに対して，$H(\lambda_{18}) = H(\lambda_{19}) = H(\lambda_{25})$ となり，$D_0 = \{\lambda_{18}, \lambda_{19}, \lambda_{25}\}$ である．

多項分布に対しては，大数の強法則 [34]が成り立つ．すなわち，$n \to \infty$ のとき

$$\frac{n_i}{n} \to p_i^*, \quad a.s. \tag{3.15}$$

となる．ただし，n_i は系列 x^n 内のシンボル i の生起回数である．一方，$\log P(x^n|\lambda)$ は

$$\log P(x^n|\lambda) = \sum_{i=0}^{\beta} n_i \log p_i(\lambda), \tag{3.16}$$

であるので，(3.15)式から，$\forall \lambda \in \Lambda$ に対して $n \to \infty$ のとき，

$$-\frac{1}{n} \log P(x^n|\lambda) \to -\sum_{i=0}^{\beta} p_i^* \log p_i(\lambda), \quad a.s. \tag{3.17}$$

が得られる．ここで，$H^*(\lambda) = -\sum_{i=1}^{\beta} p_i^* \log p_i(\lambda)$ である．したがって，このモデル族は条件3.1, (3)を満たすことがわかる．

また，$\forall\lambda_1, \forall\lambda_2 \in \Lambda$ に対し，

$$\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)} = \sum_{i=0}^{\beta} n_i \log \frac{p_i(\lambda_1)}{p_i(\lambda_2)}, \tag{3.18}$$

であるので，$n_i \to \infty$, $a.s.$ より[2]，$p_i(\lambda_1) \neq p_i(\lambda_2)$ であればほとんど確実に $\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)}$ は $+\infty$ か $-\infty$ に発散することは容易にわかる．したがって，このモデル族は条件3.2をみたしている．

さらに，$D^*(P_{\lambda^*}\|P_\lambda)$ は

$$\begin{aligned} D^*(P_{\lambda^*}\|P_\lambda) &= \sum_{i=0}^{\beta} p_i^* \log \frac{p_i(\lambda^*)}{p_i(\lambda)} \\ &= D^*(P^*\|P_\lambda) - D^*(P^*\|P_{\lambda^*}), \end{aligned} \tag{3.19}$$

で与えられ，

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P(x^n|\lambda^*)}{P(x^n|\lambda)} = D^*(P_{\lambda^*}\|P_\lambda), \quad a.s. \tag{3.20}$$

が成り立つ．一方，定義から $\forall\lambda^* \in D_0$ と $\forall\lambda \in D_1$ に対して，$D^*(P^*\|P_\lambda) - D^*(P^*\|P_{\lambda^*}) > 0$ が成り立っている． □

例 3.2 (有限エルゴード マルコフモデル) $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の有限状態エルゴードマルコフモデルを考える．$p_{i,j}$ を j番目の状態 s_j におけるシンボル i の生起確率とする．ただし，$i \in \mathcal{X}$ かつ $j = 0, 1, \cdots, S$ とする．q_{s_j} を状態 s_j の定常確率とすると，これは $p_{i,j}$, $i \in \mathcal{X}$, $j = 0, 1, \cdots, S$ より一意に与えられる．モデル λ の定めるこれらの確率を $p_{i,j}(\lambda)$ と $q_{s_j}(\lambda)$ とする．

モデル λ は，一つのマルコフモデル $\lambda = \{p_{i,j}(\lambda) | i = 0, \cdots, \beta-1, j = 0, \cdots, S\}$ を定めるものとし，最適モデル λ^* を $\lambda^* = \{p_{i,j}(\lambda^*) | i = 0, \cdots, \beta-1, j = 0, \cdots, S\}$ とする．すなわち，$\lambda^* \in D_0$ である．

初期の状態は既知であるとする．n_0, $\cdots$, n_S を x^n 内で各状態 $s_0, \cdots, s_S$ に到達した回数，$n_{0,j}, n_{1,j}, \cdots, n_{\beta,j}$ を状態 s_j のもとで各シンボル $0, 1, \cdots, \beta$ が生起した回数とする．すなわち，$n = \sum_j n_j$ かつ $n_j = \sum_i n_{i,j}$ である．j番目の状態 s_j におけるシンボル i の真の生起確率を $p_{i,j}^*$, $i = 0, \cdots, \beta-1, j = 0, \cdots, S$, とかく．すなわち，$p_{i,j}^*$ は

$$p_{i,j}^* = \lim_{n\to\infty} \frac{1}{n} E^* [N_{i,j}], \tag{3.21}$$

[2] $\frac{n_i}{n} \to p_i^*$, $a.s.$ であるが，$|n_i - n| \to p_i^*$, $a.s.$ は成り立たない．一般に，$|n_i - n| \to +\infty$, $a.s.$ となる．

で与えられる．ただし，$N_{i,j}$ は確率変数 X^n 内で，各状態 s_j, $j = 0, \cdots, S$ のもとでシンボル $0, 1, \cdots, \beta$ の生起している回数を表す確率変数である．すなわち，

$$\lim_{n\to\infty} E^* \left[\frac{N_{i,j}}{n} \log p_{i,j} \right], \tag{3.22}$$

を最小化する $p_{i,j}$, $0, 1, \cdots, \beta$, $j = 0, \cdots, S$ が，$p^*_{i,j}$ $0, 1, \cdots, \beta$, $j = 0, \cdots, S$ である．$p^*_{i,j}$ から計算される真の定常確率を $q^*_{s_0} \cdots q^*_{s_S}$ と記述する．このとき，エルゴード性から $n \to \infty$ のとき

$$\frac{n_j}{n} \to q^*_{s_j}, \quad a.s. \tag{3.23}$$

$$\frac{n_{i,j}}{n_j} \to p^*_{i,j}, \quad a.s. \tag{3.24}$$

が成り立つ．

モデル族 Λ を多項分布の例と同様に構成しよう．すなわち，各モデルを $p_{i,j} \in \{1/2, 1/3, \cdots\}$ の組み合わせで与える．このとき，$\log P(x^n|\lambda)$ は

$$\log P(x^n|\lambda) = \sum_{i,j} n_{i,j} \log p_{i,j}(\lambda), \tag{3.25}$$

で与えられるので，$n \to \infty$ のとき

$$\begin{aligned} \frac{1}{n} \log P(x^n|\lambda) &= \sum_{i,j} \frac{n_j}{n} \frac{n_{i,j}}{n_j} \log p_{i,j}(\lambda) \\ &\to \sum_{i,j} q^*_{s_j} p^*_{i,j} \log p_{i,j}(\lambda), \quad a.s. \end{aligned} \tag{3.26}$$

が成り立つ．すなわち，このモデル族も条件3.1, (3)を満たすことがわかる．

また，$\forall\lambda_1$, $\forall\lambda_2 \in \mathcal{M}$ に対し，

$$\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_1)} = \sum_{i,j} n_{i,j} \log \frac{p_{i,j}(\lambda_1)}{p_{i,j}(\lambda_2)}, \tag{3.27}$$

であるから，$n_{i,j} \to \infty$, $a.s.$ より，$p_{i,j}(\lambda_1) \neq p_{i,j}(\lambda_2)$ であれば，$\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_1)}$ はほとんど確実に $+\infty$ か $-\infty$ に発散することがわかる．したがって，このモデル族は条件3.2をみたしている．

さらに，$D^*(P_{\lambda^*} \| P_\lambda)$ を

$$D^*(P_{\lambda^*} \| P_\lambda) = \sum_{j=0}^{S} q^*_{s_j} \sum_{i=0}^{\beta} p^*_{i,j} \log \frac{p^*_{i,j}}{p_{i,j}(\lambda)}, \tag{3.28}$$

として，

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P(x^n|\lambda^*)}{P(x^n|\lambda)} = D^*(P_{\lambda^*} \| P_\lambda), \quad a.s. \tag{3.29}$$

が成り立つ．定義より，$\forall\lambda^* \in D_0$ と $\forall\lambda \in D_1$ に対して，$D^*(P^* \| P_\lambda) - D^*(P^* \| P_{\lambda^*}) > 0$ が成り立つ． □

3.2.3 離散モデル族に対する解析

まず, 次の補題を与える.

補題 3.1 条件3.1のもとで, ベイズ符号の符号長は

$$L^{\lambda}_{Bayes}(x^n) = -\log \sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*) - o^+(1), \quad a.s. \tag{3.30}$$

で与えられる. ここで, $o^+(1)$ は正で $n \to \infty$ のとき $o^+(1) \to +0$, $a.s.$ となる項である.

(証明) D_0 の定義より

$$L^{\lambda}_{Bayes}(x^n) = -\log \left\{ \sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*) + \sum_{\lambda \in D_1} P(x^n|\lambda)P(\lambda) \right\}$$

$$= -\log \sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*) - \log\Big\{1 + \frac{\sum_{\lambda \in D_1} P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*)}\Big\}. \tag{3.31}$$

Λ は有限集合であり, また $\forall \lambda^* \in D_0$ と $\forall \lambda \in D_1$ に対して $n \to \infty$ のとき $\frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \to 0$, $a.s.$ となるので,

$$\frac{\sum_{\lambda \in D_1} P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*)} = o^+(1), \quad a.s. \tag{3.32}$$

が成り立つ. したがって,

$$\log\Big\{1 + \frac{\sum_{\lambda \in D_1} P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*)}\Big\} = o^+(1), \ a.s. \tag{3.33}$$

が得られ, 補題が示された. □

モデル族には同じ確率分布を表すモデルが複数存在せず, 最適なモデルが唯一であることが仮定できる場合がある. このとき, $|D_0| = 1$ であり, ベイズ符号の符号長は

$$L^{\lambda}_{Bayes}(x^n) = -\log P(x^n|\lambda^*) - \log P(\lambda^*) - o^+(1), \ a.s. \tag{3.34}$$

で与えられることがわかる.

次に, 本章の結果に対して極めて重要で本質的な定理を示す.

定理 3.1 同じ事前分布のもとで, (2.16)式と(2.22)式の関係は

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) - \log P(\hat{\lambda}|x^n), \tag{3.35}$$

で与えられる．ここで，$\hat{\lambda}$ は事後確率 $P(\lambda|x^n)$ を最大にするモデルである．
(証明) ベイズ規則より

$$P(x^n|\lambda)P(\lambda) = P(\lambda|x^n)P(x^n), \tag{3.36}$$

が成り立つので，対数をとれば

$$-\log P(x^n|\lambda) - \log P(\lambda) = -\log P(\lambda|x^n) - \log P(x^n), \tag{3.37}$$

が得られる．(2.16)式に上式を代入し，$P(x^n)$ がモデルでの最小化に無関係であることに注意すれば，

$$L^{\lambda}_{MDL}(x^n) = -\log P(x^n) - \max_{\lambda\in\Lambda} \log P(\lambda|x^n). \tag{3.38}$$

ここで，$P(x^n)$ はベイズ符号の予測分布であるから，(3.35)式が示された． □

この定理は，ベイズ符号がMDL符号に対して，$-\log P(\hat{\lambda}|x^n)$ だけ符号長を小さくできることを述べており，同じ事前分布のもとでのベイズ符号の優位性を示している．すわなち，$-\log P(\hat{\lambda}|x^n)$ がどの程度の大きさをもつのかを解析できれば，両規準の差が明らかとなる．

通常，最適モデルは唯一と仮定できる場合が多い．そこでまず，$|D_0| = 1$ の場合の符号長の差を与える．

定理 3.2 λ^* が唯一，すなわち $|D_0| = 1$ とする．このとき，条件3.1のもとで(2.16)式と(2.22)式の関係は次式で与えられる．

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) + o^+(1), \quad a.s. \tag{3.39}$$

ここで，$o^+(1)$ は正で，$n \to \infty$ のとき $o^+(1) = +0$，$a.s.$ となる項である．
(証明)

$$P(\lambda|x^n) \propto P(x^n|\lambda)P(\lambda), \tag{3.40}$$

と(3.10)式より，

$$\frac{P(\lambda|x^n)}{P(\lambda^*|x^n)} = o^+(1), \quad a.s. \tag{3.41}$$

が得られる．したがって，

$$1 - P(\lambda^*|x^n) = o^+(1), \quad a.s. \tag{3.42}$$

よって，事後確率 $P(\lambda|x^n)$ を最大化するモデル $\hat{\lambda}$ は，最適モデル λ^* に概収束する．すなわち，十分大きな n に対して $\hat{\lambda} = \lambda^*$ であることが確率1でいえる．よって，$n \to \infty$ のとき $\hat{\lambda}$ は λ^* で置き換えることができ，

$$-\log P(\hat{\lambda}|x^n) = o^+(1), \quad a.s. \tag{3.43}$$

となる．よって，定理が示された． □

この定理は，最適モデルが唯一であるとき，MDL符号とベイズ符号の差は漸近的に 0 に収束し，両者の符号長は漸近的に等価であることを示している．次に，$|D_0| \geq 2$ の場合であるが，この場合は $\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ の場合と $\exists\lambda_1^*, \exists\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ の場合で結果が異なる．

定理 3.3 $|D_0| \geq 2$, かつ $\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ であるとする．このとき，条件3.1のもとで(2.16)式と(2.22)式の関係は次式で与えられる．

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) + O^+(1), \quad a.s. \tag{3.44}$$

ただし，$O^+(1)$ は $n \to \infty$ のとき $O^+(1) \to C$, $a.s.$ を満たす項で，正定数 C は

$$C = -\log \frac{\max_{\lambda^* \in D_0} P(\lambda^*)}{\sum_{\lambda' \in D_0} P(\lambda')}. \tag{3.45}$$

で与えられる．

(証明) ベイズ規則から導かれる

$$-\log P(\lambda|x^n) = -\log P(x^n|\lambda) - \log P(\lambda) + \log P(x^n), \tag{3.46}$$

と補題3.1より，

$$-\log P(\lambda|x^n) = -\log \frac{P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*)} + o^+(1), \quad a.s. \tag{3.47}$$

が成り立つ．ここで，条件3.1, (3)より，$\forall\lambda \in D_1$ に対しては $-\log P(\lambda|x^n) \to +\infty$, $a.s.$ となる．一方，$\lambda^* \in D_0$ に対しては，

$$-\log P(\lambda^*|x^n) = -\log \frac{P(\lambda^*)}{\sum_{\lambda' \in D_0} P(\lambda')} + o(1), \quad a.s. \tag{3.48}$$

となる．ここで，$o(1)$ は $\lim_{n\to\infty} o(1) = 0$, $a.s.$ となる項である．

(3.48)式の右辺第1項はデータ系列長 n に依存しないので，MDL符号のとベイズ符号の関係は

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) - \max_{\lambda^* \in D_0} \log P(\lambda^*) + \log \left\{ \sum_{\lambda^* \in D_0} P(\lambda^*) \right\} + o(1), \quad a.s. \tag{3.49}$$

で与えられる．条件3.1,(2)より，

$$-\max_{\lambda^* \in D_0} \log P(\lambda^*) > -\log \sum_{\lambda^* \in D_0} P(\lambda^*), \tag{3.50}$$

であるから，

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) + O^+(1), \quad a.s. \tag{3.51}$$

が得られた．ここで，$O^+(1)$ は $O^+(1) \to C, a.s.$ となる項である． □

以上の結果は，次のように解釈することができる．MDL原理に基づくモデルの選択は，本質的に事後確率最大のモデルを選択することと等価である．もし，最適モデル $\lambda^* \in D_0$ が唯一であるとすると，$P(\lambda|x^n)$ の最大化により，漸近的に確率1で λ^* を選択することができる．しかし，$|D_0| \geq 2$, かつ $\forall\lambda^*_1, \forall\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) = P(\cdot|\lambda^*_2)$ である場合には，$P(\lambda|x^n)$ の最大化は，漸近的に D_0 の中で事前確率の最大のものを選択することと等価となる．すなわち，$P(\lambda^*|x^n)$ は 1 に収束せず，モデル選択に対するあいまい性が残ることになる．このあいまい性が MDL符号の符号長を増加させるといえる．一方，混合確率 $\sum_\lambda P(x^n|\lambda)P(\lambda)$ を用いるベイズ最適な予測(ベイズ予測)を考えてみる．$\forall\lambda^*_1, \forall\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) = P(\cdot|\lambda^*_2)$ であるとする．

ここで，λ^*_c を考え，$P(\lambda^*_c) = \sum_{\lambda^* \in D_0} P(\lambda^*)$ かつ $P(x^n|\lambda^*_c) = P(x^n|\lambda^*_1)$ とする．さらに $\lambda^* \neq \lambda^*_1$ かつ $\lambda^* \in D_0$ を満たす $\forall\lambda^*$ をモデル族から除き，この新たに作られたモデル族と事前分布に対してベイズ予測を構成する．このとき，新たに構成したベイズ予測の符号長はもとのベイズ予測のそれと等価である．さらに，λ^*_c を唯一の最適モデルと解釈することができる．したがって，ベイズ予測では混合をとるために，漸近的には最適モデルの個数に依存しないといえる．

一方，$\exists\lambda^*_1, \exists\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) \neq P(\cdot|\lambda^*_2)$ の場合には次の定理が導かれる．

定理 3.4 $|D_0| \geq 2$, かつ $\exists\lambda^*_1, \exists\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) \neq P(\cdot|\lambda^*_2)$ であるとする．このとき，条件3.1, 3.2のもとで(2.16)式と(2.22)式の関係は

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) + o^+(1), \quad a.s. \tag{3.52}$$

で与えられる． □

(証明) 定理3.3の証明と同様に，

$$-\log P(\lambda|x^n) = -\log \frac{P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P(\lambda^*)} + o^+(1), \quad a.s. \tag{3.53}$$

が成り立ち，条件3.1, (3)より，$\forall \lambda \in D_1$ に対しては $-\log P(\lambda|x^n) \to +\infty, \ a.s.$ となる．一方，$\lambda^* \in D_0$ に対しては，条件3.2より，D_0 内のうちの1つの λ^* について $-\log P(\lambda^*|x^n) \to 0, \ a.s.$ となり，残りはほとんど確実に $-\infty$ に発散する．したがって，

$$-\log P(\hat{\lambda}|x^n) = o^+(1), \quad a.s. \tag{3.54}$$

となり，定理が得られる． □

これは，漸近的に D_0 の中のどれか1つのモデルの事後確率が1に概収束することを受けている．どのモデルの事後確率が1となるかは 系列 x^∞ に依存して決まる．しかし D_0 の中の各モデルは平均対数損失という観点からは等価であるので，そのうちのどのモデルの事後確率が1に収束するかは漸近的な累積損失に影響を与えない．したがって，この場合 MDL符号とベイズ符号の符号長は漸近的に等価になる．

3.3 パラメトリックモデル族に対する解析

本節ではパラメトリックモデル族 $\{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ に対するMDL符号の符号長,(2.26)式とベイズ符号の符号長,(2.27)式の差を解析する．データ系列 x^n は，真の分布 $P^*(x^n)$ に従うものとし，パラメトリックモデル族にこの真の分布が含まれていることは仮定しない．まず，解析に本質的となる情報行列 $J^*(\theta^k)$ を以下のように定義する．

$$J^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^* \left[-\frac{\partial^2 \log P(X^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \right]. \tag{3.55}$$

ここで，E^* は真の分布 $P^*(\cdot)$ による期待値を表す．定義より，$P^*(\cdot) \in \{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ であれば，$J^*(\theta^{k*}) = J(\theta^{k*})$ が成り立つが，これは真の分布がクラスに含まれない一般的状況では成り立たない．しかし，情報源符号化などで扱う離散確率変数のモデル族のほとんどは，$J^*(\theta^{k*}) = J(\theta^{k*})$ を満たす．また，$H^*(\theta^k)$ を

$$H^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^*[-\log P(X^n|\theta^k)], \tag{3.56}$$

と定義する．平均損失の意味で最適パラメータ θ^{k*} は

$$\theta^{k*} = \arg \min_{\theta^k \in \Theta^k} H^*(\theta^k), \tag{3.57}$$

で与えられる．

データ系列 x^n が与えられたときの最尤推定量を $\hat{\theta}^k$, 事後確率最大推定量を $\tilde{\theta}^k$ で表す.

$$\hat{\theta}^k = \arg \max_{\theta^k \in \Theta^k} \log P(x^n|\theta^k). \tag{3.58}$$

$$\tilde{\theta}^k = \arg \max_{\theta^k \in \Theta^k} \log f(\theta^k|x^n). \tag{3.59}$$

3.3.1 モデル族の条件

パラメータ空間に球 $B_\delta(\theta^{k*}) = \{\theta^k \in \Theta^k | \|\theta^k - \theta^{k*}\| < \delta\}$ を定義する. ここでは次の条件を仮定する.

条件 3.3

(1) (θ^{k*} の一意性) 関数 $-H^*(\theta^k)$ は Θ^k 上で単峰形関数とする. すなわち, θ^{k*} は Θ^k の内部に唯一存在するものとする.

(2) (モデル族の smoothness) Fisher 情報行列 $J(\theta^k)$ は $\forall \theta^k \in \Theta^k$ に対して $0 < C_0 < \det J(\theta^k) < \infty$ を満たす. ただし, C_0 はある定数である. また, $J^*(\theta^{k*}) = J(\theta^{k*})$ とする. さらに, $\det J(\theta^k)$ と $\det J^*(\theta^k)$ は, $\forall \theta^k \in \Theta^k$ に対し,

$$\left\| \frac{\partial \det J(\theta^k)}{\partial \theta^k} \right\| < \infty, \tag{3.60}$$

$$\left\| \frac{\partial \det J^*(\theta^k)}{\partial \theta^k} \right\| < \infty, \tag{3.61}$$

を満たす.

(3) (事前密度の smoothness) $\forall \theta^k \in \Theta^k$ に対し, $f(\theta^k) > 0$ とし, $f(\theta^k)$ は θ^k に対し3回微分可能とする. すなわち,

$$\left| \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} \frac{\partial^3 f(\theta^k)}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| < c_f(\theta^k), \tag{3.62}$$

であり, $c_f(\theta^k)$ は θ^k に依存する定数である.

(4) (推定量の一意性) 尤度関数 $P(x^n|\theta^k)$ は, $\forall x^n \in \mathcal{X}^n$ に対して単峰形, あるいは単調であるとする. また, 事後密度関数 $f(\theta^k|x^n)$ は $\forall x^n \in \mathcal{X}^n$ に対し, Θ^k 内で最大値を唯一もつものとする. すなわち, 最尤推定量 $\hat{\theta}^k$ と事後確率最大推定量 $\tilde{\theta}^k$ は Θ^k 内で唯一に存在する.

(5) (事後確率最大推定量の一致性) 事後確率最大推定量 $\tilde{\theta}^k$ は強一致推定量である，すなわち，$n \to \infty$ のとき

$$\left\| \tilde{\theta}^k - \theta^{k*} \right\| \to 0, \quad a.s. \tag{3.63}$$

とする．

(6) (情報行列の一致性) $\forall \theta^k \in B_\delta(\theta^{k*})$ に対して一様に

$$-\frac{1}{n} \log P(x^n|\theta^k) \to H^*(\theta^k), \quad a.s. \tag{3.64}$$

$$-\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \to J^*(\theta^k), \quad a.s. \tag{3.65}$$

となる $\delta > 0$ が存在する． □

上の条件が成り立つとき，$\forall \theta^k \in \Theta^k$ に対して $\frac{1}{n} \log f(\theta^k|x^n) \propto \frac{1}{n} \log P(x^n|\theta^k) + \frac{1}{n} f(\theta^k) \to \frac{1}{n} \log P(x^n|\theta^k)$, かつ $-\frac{1}{n} \log P(x^n|\theta^k) \to H^*(\theta^k)$, $a.s.$ であるので，もし $\tilde{\theta}^k \to \hat{\theta}^k$, $a.s.$ が成り立たないとすれば，(3.63)式も成り立たない．したがって，式 (3.63) は $\hat{\theta}^k \to \theta^{k*}$, $a.s.$ も意味している．

さらに，従来研究の MDL符号による予測の損失に関する漸近解析では，最尤推定量の漸近正規性が仮定されている [56]が，本稿の解析では直接これを必要としない．しかし，多くの現実的モデルにおいて，最尤推定量の漸近正規性は成り立つ性質といえる．

注意 3.1 条件3.3, (3) については，現実的に考えられる事前分布ではみたされると考えられる．例えば，多項分布族に対する共役事前分布であるディレクレ分布は，この条件を満たす．ディレクレ分布は有限マルコフモデルにも適用できる極めて有用な事前分布である [81]. □

注意 3.2 条件3.3, (5), (6) について考える．情報源符号化に用いられるモデル族，例えば有限マルコフモデルなどでは最尤推定量に対し重複対数の法則 [34]が成り立つ．これは一般の予測で扱う多くのモデル族で成り立つ本質的な性質である．さらにこれを用いて，情報行列に対する重複対数の法則を示せることが多い．次節の例を参照． □

3.3.2 パラメトリックモデル族の例

本節では条件3.3を満たすモデル族の例を示す．これらは，情報源符号化において代表的で有用なモデル族である．

例 3.3 (多項分布族) アルファベット $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の多項分布に従うi.i.d. 系列(定常無記憶系列)を考える．θ_i をシンボル i, $i \in \mathcal{X}$ の生起確率とすると，ベクトル $\theta^k = (\theta_0, \theta_1, \cdots, \theta_{k-1})^T$ は一つの多項分布を規定する．ここで，$k = \beta$ である．したがって，θ^k を $\Theta^k = \{(\theta_0, \theta_2, \cdots, \theta_{k-1})^T | 0 \leq \theta_0, \theta_2, \cdots, \theta_{k-1}, \theta_k \leq 1\}$ 上のパラメータとみなし，一つのパラメトリックモデル族を構成できる．ただし，$\theta_k = 1 - \sum_{i=0}^{k-1} \theta_i$ である．

θ^{k*} が Θ^k の内部に存在するものとすると，情報行列 $J^*(\theta^k)$ は

$$J^*(\theta^k) = -\frac{\partial^2 \sum_{i=0}^{k} \theta_i^* \log \theta_i}{\partial \theta^k (\partial \theta^k)^T}, \tag{3.66}$$

で与えられるので，$J^*(\theta^k)$ の i-i 要素は

$$\frac{\theta_i^*}{(\theta_i)^2} + \frac{\theta_k^*}{\left(1 - \sum_{j=1}^{k-1} \theta_j\right)^2}, \tag{3.67}$$

i-j 要素は

$$\frac{\theta_k^*}{\left(1 - \sum_{j=1}^{k-1} \theta_j\right)^2}, \tag{3.68}$$

となる．ただし，$\theta^{k*} = (\theta_0^*, \theta_1^*, \cdots, \theta_{k-1}^*)^T$ である．したがって，$\det J^*(\theta^k)$ は明らかに微分可能である．また，$\theta^k = \theta^{k*}$ のとき，$\det J^*(\theta^{k*})$ は $\det J(\theta^{k*})$ に一致することもわかる．$J(\theta^k)$ の行列式は

$$\det J(\theta^k) = \frac{1}{\theta_0 \theta_1 \cdots \theta_k}, \tag{3.69}$$

で与えられるので，$\det J(\theta^k)$ は $\theta_0 = \theta_1 = \cdots = \theta_k$ のとき最小化され，その最小値は

$$\min \det J(\theta^k) = (k+1)^{k+1} > 0, \tag{3.70}$$

で与えられる．さらに，$\forall j \in \{0, 1, \cdots, k\}$ に対し $0 < \theta_j < 1$ であるので，$\det J(\theta^k) < \infty$ かつ $\left\| \frac{\partial \det J(\theta^k)}{\partial \theta^k} \right\| < \infty$ であり，条件3.3, (1) が成り立つことがわかる．$\forall j \in \{0, 1, \cdots, k\}$ に対して $\lim_{\theta_j \to 0} \sqrt{\det J(\theta^k)} \to \infty$ であるから，任意の定数 C' に対して $\sqrt{\det J(\theta^k)} < C'$ は成り立たないことがわかる．

尤度関数 $P(x^n | \theta^k)$ は

$$P(x^n | \theta^k) = \prod_{i=0}^{k} (\theta_i)^{n_i}, \tag{3.71}$$

で与えられ，この関数は $\forall x^n \in \mathcal{X}^n$ に対して Θ^k 内で唯一の最大値をもつ．ただし，n_i は x^n におけるシンボル i の生起回数である．最尤推定量は $\hat{\theta}_i = n_i / n$, $i = 0, 1, \cdots, k$ で与えられる．この最尤推定量は $\lim_{n \to \infty} \frac{n_i}{n} \to \theta_i^* > 0$, $a.s.$ を満

たす [34]. すなわち, $\hat{\theta}^k \to \theta^{k*}$, $a.s.$ である. もし Θ^k 上の事前分布としてディレクレ分布を仮定すると, 事後確率最大推定量 $\tilde{\theta}_i$ は $\hat{\theta}_i = \frac{n_i+\gamma_i-1}{n+\sum_j \gamma_j - k - 1}$ で与えられる. ただし, γ_i はディレクレ分布のパラメータである. よって, 条件3.3, (6) が成り立つ. したがってこの場合, $\tilde{\theta}^k \to \hat{\theta}^k \to \theta^{k*}$, $a.s.$ が成り立つ. またもし, $\sup_{\theta^k \in \Theta^k} f(\theta^k) < \infty$ であれば, $\frac{1}{n}\log P(x^n|\theta^k) + \frac{1}{n} f(\theta^k) \to \frac{1}{n}\log P(x^n|\theta^k)$ が $\theta^k \in \Theta^k$ で一様に成り立つので, この場合も $\tilde{\theta}^k \to \theta^{k*}$, $a.s.$ が成り立つ. よってこれらの場合, 条件3.3, (5) が成り立つ.

ここで θ^{k*} が Θ^k の内部に存在するとすると, 条件3.3, (4) も明らかに成り立つ.

次に条件3.3, (6) を考える. $\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta)}{\partial\theta^k(\partial\theta^k)^T}$ は

$$\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta)}{\partial\theta^k(\partial\theta^k)^T} = \frac{\partial^2 \sum_{i=0}^k \frac{n_i}{n}\log\theta_i}{\partial\theta^k(\partial\theta^k)^T}, \tag{3.72}$$

で与えられるので, $\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta)}{\partial\theta^k(\partial\theta^k)^T}$ の i-i 要素は

$$-\frac{n_i}{n(\theta_i)^2} - \frac{n_k}{n\left(1-\sum_{j=0}^{k-1}\theta_j\right)^2}, \tag{3.73}$$

i-j 要素 $(i \neq j)$ は

$$-\frac{n_k}{n\left(1-\sum_{j=0}^{k-1}\theta_j\right)^2}, \tag{3.74}$$

となる. $n_i/n \to \theta_i^*$, $a.s.$ より,

$$\frac{1}{n}\log P(x^n|\theta^k) \to \sum_{i=0}^k \theta_i^* \log\theta_i, \quad a.s. \tag{3.75}$$

$$\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} \to \frac{\partial^2 \sum_{i=0}^k \theta_i^*\log\theta_i}{\partial\theta^k(\partial\theta^k)^T}, \quad a.s. \tag{3.76}$$

が得られる. したがって, $-\frac{1}{n}\log P(x^n|\theta) \to H^*(\theta^k)$, $a.s.$ かつ $-\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta)}{\partial\theta^k(\partial\theta^k)^T} \to J^*(\theta^k)$, $a.s.$ となり, 条件3.3, (6) も示された.

したがって, この多項分布族は条件3.3を満たすことがわかる. さらにこのモデル族では, 最尤推定量の漸近正規性も成り立っている [34]. □

例 3.4 (有限エルゴード マルコフモデル) 情報源アルファベット $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ 上の有限エルゴード マルコフモデルを考える. S 個の有限集合である状態の集合は与えられているものとし, $\theta_{i,j} = p_i^{(s_j)}$ を j 番目の状態 s_j におけるシンボル i の生起確率とする. また, $\theta_{i,j}$ から計算される状態 s_j の定常確率を q_{s_j} とする.

ここで，$i \in \mathcal{X}$ かつ $j = 0, 1, \cdots, S$ である．$\theta^k = (\theta_{0,0}, \cdots, \theta_{\beta-1,S})^T$ は，このマルコフモデルの $k = \beta(S+1)$ 次元パラメータとなる．

$N_{i,j}$ を，確率変数 X^n の中の，各状態 s_j, $j = 0, \cdots, S$ のもとでの各シンボル $0, 1, \cdots, \beta$ の生起回数を表す確率変数としよう．このとき $p^*_{i,j} = \lim_{n\to\infty} \frac{1}{n} E^* [N_{i,j}]$ と定義し，$p^*_{i,j}$ から計算される各状態の定常確率を $q^*_{s_0} \cdots q^*_{s_S}$ とする．このとき，情報行列 $J^*(\theta^k)$ は

$$J^*(\theta^k) = -\frac{\partial^2 \sum_{j=0}^{S} \sum_{i=0}^{\beta} q^*_{s_j} \theta^*_{i,j} \log \theta_{i,j}}{\partial\theta^k (\partial\theta^k)^T}, \tag{3.77}$$

で与えられる．ただし，$\theta_{\beta,j} = 1 - \sum_{i=0}^{\beta-1} \theta_{i,j}$ かつ $\theta^*_{\beta,j} = 1 - \sum_{i=0}^{\beta-1} \theta^*_{i,j}$ である．よって，$\det J^*(\theta^k)$ は明らかに微分可能である．$\theta^k = \theta^{k*}$ のときは $J^*(\theta^{k*}) = J(\theta^k)$ となることも明らかである．$J(\theta^k)$ の行列式は

$$\det J(\theta^k) = \prod_{j=0}^{S} (q_{s_j})^{\beta} \frac{1}{\theta_{0,j}\theta_{1,j}\cdots\theta_{\beta,j}}, \tag{3.78}$$

で与えられる．ここで q_{s_j} は θ^k に依存し，$\det J(\theta^k)$ を一般式の形で θ^k だけの関数として記述することはできない．しかし，マルコフモデルの構造を一つ決めてやれば，q_{s_j} を θ^k で書き下すことが可能となり，$C_0 < \det J(\theta^k) < \infty$ であることが容易にわかる．例えば，$\mathcal{X} = \{0, 1\}$ 上の2元単純マルコフモデルを考える．$S = 1$ であり，s_i は最後のシンボルが i であったことを意味する．このとき q_{s_0} と q_{s_1} はそれぞれ，$\frac{\theta_{0,1}}{\theta_{1,0}+\theta_{0,1}}$ と $\frac{\theta_{1,0}}{\theta_{1,0}+\theta_{0,1}}$ で与えられる．よって $\det J(\theta^k)$ は

$$\det J(\theta^k) = \frac{1}{(\theta_{1,0} + \theta_{0,1})^2 \theta_{0,0}\theta_{1,1}}, \tag{3.79}$$

となる．$\theta_{0,0} = \theta_{1,1} = 1/2$ のとき，この行列式は最小化され，その最小値は 4 である．また，$\theta_{0,0} \to 0$ あるいは $\theta_{1,1} \to 0$ や $\theta_{1,0} + \theta_{0,1} \to 0$ としたとき，$\det J(\theta^k) \to 0$ となるので，$\det J(\theta^k)$ は定数の上界を持たないが，正定数の下界値を持つことがわかる．したがって，このモデルは条件3.3, (2) を満たす.

再び初期の状態は既知であるとしよう．n_0, $\cdots$, n_S と $n_{0,j}, n_{1,j}, \cdots, n_{\beta,j}$ をそれぞれ，各状態と状態 j のもとでの各シンボルの生起回数とする．すなわち，$n = \sum_j n_j$ かつ $n_j = \sum_i n_{i,j}$ である．尤度関数 $P(x^n|\theta^k)$ は

$$P(x^n|\theta^k) = \prod_{j=0}^{S} \prod_{i=0}^{\beta} (\theta_{i,j})^{n_{i,j}}, \tag{3.80}$$

で与えられ，Θ^k 内で唯一の最大値をもつ．この有限マルコフ情報源においても大数の強法則 [34]

$$\frac{n_{i,j}}{n_j} \to \theta^*_{i,j}, \quad a.s. \tag{3.81}$$

が成り立つ．これは，条件3.3, (5) を意味する．また，条件3.3, (1) が成り立てば，条件3.3, (4) も成り立つ．

さらに，対数尤度 $\log P(x^n|\theta^k)$ は

$$\log P(x^n|\theta^k) = \sum_{i,j} n_{i,j} \log \theta_{i,j}, \tag{3.82}$$

で与えられ，(3.81)式が成り立つことから，条件3.3, (6) が成り立つことがわかる．

以上より，有限定常エルゴードマルコフモデルは，条件3.3を満たすことがわかる．このモデルでは最尤推定量の漸近正規性も成り立つ [34]. □

条件3.3は定常独立観測(i.i.d.系列)を仮定していない．情報源符号化の実際で重要となる情報源は記憶をもつ情報源であり，i.i.d.系列では応用範囲が狭い．本研究の仮定はマルコフモデルなどのi.i.d.ではない情報源に対しても成り立っている．

3.3.3 事後確率密度の漸近正規性

符号長の解析の前に，本質的な性質である事後確率密度の漸近正規性を示す．J.Rissanen は最尤推定量の漸近正規性から最尤符号化 [101]の漸近符号長を導いており，漸近正規性はパラメトリックモデルでの予測において本質的と考えられる．しかし，この最尤推定量の漸近正規性は法則収束を意味するのに対し，事後確率の漸近正規性はデータ系列 x^n が与えられたもとでの事後密度の形を議論したものである．

事後確率密度の漸近正規性の議論については，[16],pp.285-297 あるいは [22] の議論が有用である．

補題 3.2 [16],[22] 無限系列 x^∞ を固定して考える．$\tilde{\theta}^k$ を $L_n(\theta^k) = \log f(\theta^k|x^n)$ の(1点で極大となる)極大値とする．すなわち，

$$L_n'(\tilde{\theta}^k) = \left.\frac{\partial L_n(\theta^k)}{\partial \theta^k}\right|_{\theta^k=\tilde{\theta}^k} = \mathbf{0}, \tag{3.83}$$

かつ

$$\Sigma_n = -\left(L_n''(\tilde{\theta}^k)\right)^{-1} = -\left(\left.\frac{\partial^2 L_n(\theta^k)}{\partial \theta^k (\partial \theta^k)^T}\right|_{\theta^k=\tilde{\theta}^k}\right)^{-1}, \tag{3.84}$$

が正定値行列とする．パラメータ空間上に球 $B_\delta(\tilde{\theta}^k) = \{\theta^k \in \Theta^k | \|\theta^k - \tilde{\theta}^k\| < \delta\}$ を定義する．

さらに，次の3つの条件を考える．

(c.1) *"Steepness"* $\lim_{n\to\infty}\bar{\sigma}_n^2 \to 0$, ただし $\bar{\sigma}_n^2$ は Σ_n の最大固有値とする.

(c.2) *"Smoothness"* 任意の正定数 $\epsilon > 0$ に対し, N と $\delta > 0$ が存在し, 全ての $n > N$ と $\theta^k \in B_\delta(\tilde{\theta}^k)$ に対して $L_n''(\theta^k)$ が存在して

$$I - A(\epsilon) \leq L_n''(\theta^k)\left\{L_n''(\tilde{\theta}^k)\right\}^{-1} \leq I + A(\epsilon), \tag{3.85}$$

を満たす. ただし, I は $k\times k$ 基本行列, $A(\epsilon)$ はその最大固有値が $\epsilon \to 0$ のとき 0 に収束するような正定値行列とする.

(c.3) *"Concentration"* $\forall\delta > 0$ に対し, $n \to \infty$ のとき

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} f(\theta^k|x^n)d\theta^k \to 1. \tag{3.86}$$

条件(c.1), (c.2) のもとで,.

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\left(\det\Sigma_n\right)^{1/2} \leq (2\pi)^{-k/2}, \tag{3.87}$$

が成り立つ.

さらに, 条件(c.1), (c.2) に条件 (c.3)を加えた3つの条件は, 事後確率密度の漸近正規性の必要十分条件となる. すなわち, 条件 (c.1), (c.2), (c.3) のもとで

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\left(\det\Sigma_n\right)^{1/2} = (2\pi)^{-k/2}, \tag{3.88}$$

が成り立つ.

また, 条件 (c.1), (c.2), (c.3) のもとで $\phi^k = \Sigma_n^{-1/2}(\theta^k - \tilde{\theta}^k)$ は k-次元正規分布 $f(\phi^k) = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}(\phi^k)^T\phi^k\right\}$ に収束する. ここで, $\Sigma_n^{-1/2}$ は $\Sigma_n^{-1/2}\Sigma_n^{-1/2} = \Sigma_n^{-1}$ を満たす $k\times k$-次元行列である. □

補題3.2は 1本の分布列の収束に関する結果であり, 確率分布の法則収束の議論と同じである. 1つ異なる点は, 密度の収束を述べている点である. 通常の中心極限定理では, 二項分布などの離散分布が漸近的に正規分布に従うことが示されるが, 二項分布は密度をもたないので密度関数の収束は成り立たない. これに対し, (c.1) ~ (c.3) を満たす確率密度関数については, その条件から密度関数の値自体の収束を示すことができる.

以上の議論が一つの分布列に対する議論であったが, 予測の問題では x^n は確率変数の実現値として出現するものととらえるので, 情報源から出てくる x^n それぞれについて, 漸近正規性が成り立つかどうかを議論しなくてはならない. そこで, 以下では以上の漸近正規性を概収束の意味で再構成する. 得られる以下の補題は, 本研究の結果において本質的な役割を演ずる.

補題 3.3 (漸近正規性) 条件3.3のもとで，$\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ の事後確率密度は平均ゼロ，共分散 $\{J^*(\tilde{\theta}^k)\}^{-1}$ の正規分布に概収束する．すなわち，R_ξ を任意の直方体とすると，事後密度 $f(\theta^k|x^n)$ は

$$P(\xi^k \in R_\xi|x^n) = \int_{\xi^k \in R_\xi} f_\xi(\xi^k|x^n)d\xi^k \\ \to \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \int_{\xi^k \in R_\xi} e^{-\frac{1}{2}\|\xi^k\|^2_{J^*(\tilde{\theta}^k)}} d\xi^k, \ a.s. \tag{3.89}$$

を満たす．ここで $f_\xi(\xi^k|x^n)$ は ξ^k の事後密度であり，

$$f_\xi(\xi^k|x^n) = \frac{1}{\sqrt{n}^k} f(\theta^k|x^n), \tag{3.90}$$

で与えられる．さらに，Ω^k を $|\Omega^k| < \infty$ を満たす任意の直方体として，$\xi^k \in \Omega^k$ に対して一様に

$$f_\xi(\xi^k|x^n) \to \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\|\xi^k\|^2_{J^*(\tilde{\theta}^k)}}, \ a.s. \tag{3.91}$$

が成り立つ．

(証明) 節3.6.1を参照． □

ξ^k の事後確率密度 $f_\xi(\xi^k|x^n)$ が，任意の直方体 Ω^k に対して $\xi^k \in \Omega^k$ 内で一様に正規分布に概収束するという結果は，従来の事後確率密度の漸近正規性に関する研究[16, 22, 51, 60, 145]よりも強いことを述べている．

3.3.4 パラメトリックモデル族に対する解析

符号長の差についてまず，次の補題を与えよう．

補題 3.4 $\bar{\xi}^k = \sqrt{n}(\bar{\theta}^k - \tilde{\theta}^k)$ と $\bar{\Xi}^k_n = \{\bar{\xi}^k | \bar{\theta}^k \in \bar{\Theta}^k_n\}$ を定義する．このとき，(2.26)式と(2.27)式の関係は

$$L^{\bar{\theta}^k}_{MDL}(x^n) = L^{\theta^k}_{Bayes}(x^n) - \max_{\bar{\xi}^k \in \bar{\Xi}^k_n} \left\{ \log \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} \right\}, \tag{3.92}$$

で与えられる．

(証明) (2.26)式と(2.27)式より，

$$L^{\bar{\theta}^k}_{MDL}(x^n) = L^{\theta^k}_{Bayes}(x^n) - \max_{\bar{\theta}^k \in \bar{\Theta}^k_n} \left\{ \log \frac{P(x^n|\bar{\theta}^k)f(\bar{\theta}^k)}{\int_{\theta^k \in \Theta^k} P(x^n|\theta^k)f(\theta^k)d\theta^k} \frac{1}{\sqrt{n}^k \sqrt{\det J(\bar{\theta}^k)}} \right\}, \tag{3.93}$$

が成り立つ．ここで，

$$f(\bar{\theta}^k|x^n) = \frac{P(x^n|\bar{\theta}^k)f(\bar{\theta}^k)}{\int_{\theta^k\in\Theta^k} P(x^n|\theta^k)f(\theta^k)d\theta^k}, \tag{3.94}$$

は $\theta^k = \bar{\theta}^k$ における事後確率密度の値である．$\bar{\theta}^k = \tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k$ であることから，補題が成り立つことがわかる． □

補題3.3と補題3.4を用いることにより，$L^{\bar{\theta}^k}_{MDL}(x^n)$ と $L^{\theta^k}_{Bayes}(x^n)$ の漸近的な差が次の定理によって与えられる．

定理 3.5 条件3.3のもとで，(2.26)式と(2.27)式の差は漸近的に

$$L^{\bar{\theta}^k}_{MDL}(x^n) = L^{\theta^k}_{Bayes}(x^n) + O^+(1), \quad a.s. \tag{3.95}$$

で与えられる．ただし，$O^+(1)$ は $n\to\infty$ のときほとんど確実に $0 < C_1 < O^+(1) < C_2 < \infty$ を満たす項である．すなわち $C_1, C_2 > 0$ が存在し，$n\to\infty$ のとき

$$C_1 < L^{\bar{\theta}^k}_{MDL}(x^n) - L^{\theta^k}_{Bayes}(x^n) < C_2, \quad a.s. \tag{3.96}$$

が成り立つ．
(証明) 3.6.2節参照 □

定理3.5の結果については，以下のように解釈を与えることが可能である．(3.92)式の右辺は，事後確率密度を量子化されたセルの領域で積分したものであり，量子化セルの事後確率とみなすことができる．離散モデル族に対する結果でもあったように，もしこの量子化セルの事後確率が 1 に収束すれば，両者の差は 0 に収束するはずである．ここで，パラメータの事後確率は漸近的に共分散 $(1/n)\{J^*(\tilde{\theta}^k)\}$ の正規分布に従う．一方，量子化セルの量子化幅も，この正規分布の主軸方向への標準偏差程度である．すなわち，n が増加してパラメータの事後確率密度が一点に集中してきても，それと同じ早さでセルの体積が小さくなっていき，セル内の事後密度は 1 に収束しない．したがって，真のセルというものは特定できず，MDL符号とベイズ符号の差は漸近的に定数として残ってしまう．これは連続パラメータを離散化した結果生じた性質である．

3.4 階層モデル族に対する解析

階層モデル族に対して，MDL符号とベイズ符号の差を解析する．階層モデル族に対し，

$$P(x^n|m) = \int_{\theta^{k_m} \in \Theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{3.97}$$

を定義しておく.

データ系列 x^n は真の分布 $P^*(x^n)$ に従って発生し，$P^*(x^n)$ はモデル族 $\mathcal{H}$ に含まれることを仮定しない．また，本研究では階層モデル族は階層構造を持たない分離的な族でも構わないものとしているが，結果はこの場合を含んでいる.

モデル m の最適パラメータ $\theta^{k_m^*}$ は

$$\theta^{k_m^*} = \arg \min_{\theta^{k_m} \in \Theta^{k_m}} H^*(m, \theta^{k_m}), \tag{3.98}$$

で定義する．ここで，$H^*(m, \theta^{k_m})$ は

$$H^*(m, \theta^{k_m}) = \lim_{n \to \infty} \frac{1}{n} E^* \left[-\log P(X^n|m, \theta^{k_m}) \right], \tag{3.99}$$

で与えられる．x^n が与えられたもとでの最尤推定量を $\hat{\theta}^{k_m}$，事後確率最大推定量を $\tilde{\theta}^{k_m}$ と記述する.

階層モデル族では入れ子構造がある場合，$H^*(m_1, \theta^{k_{m_1}}) = H^*(m_2, \theta^{k_{m_2}})$ を満たす $m_1, m_2, \theta^{k_{m_1}}, \theta^{k_{m_2}}$ $(m_1 \neq m_2)$ が存在するが，最適モデル m^* は以下で定義する.

$$m^* = \arg \min_{m \in \mathcal{M}} \left\{ k_m \middle| H^*(m, \theta^{k_m^*}) = H^* \right\}. \tag{3.100}$$

ここで H^* は

$$H^* = \min_{P(\cdot|m, \theta^{k_m}) \in \mathcal{H}} H^*(m, \theta^{k_m}), \tag{3.101}$$

で与えられ，m^* は唯一であるとする．また，各モデルのパラメータ空間上に球 $B_\delta(\theta^{k_m^*}|m) = \{\theta^{k_m} \in \Theta^{k_m} | \|\theta^{k_m} - \theta^{k_m^*}\| < \delta\}$ を定義する．最後に情報行列 $J^*(\theta^{k_m}|m)$ を

$$J^*(\theta^{k_m}|m) = -\lim_{n \to \infty} \frac{1}{n} E^* \left[\frac{\partial^2 \log p(X^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} \right], \tag{3.102}$$

と定義する.

3.4.1 モデル族の条件

次に解析に必要な条件を示す．これは前節のパラメトリックモデル族に対する仮定よりも強いものとなっている．すなわち，大数の法則のかわりに，重複対数の法則を仮定する．これは，モデルの事後確率の収束を示すために必要であり，統計的モデル選択の一致性の議論[57]で従来仮定されているものである.

条件 3.4

(1) ($\theta^{k_m^*}$ の一意存在性) $\forall m \in \mathcal{M}$ に対し，$-H^*(m, \theta^{k_m})$ は Θ^{k_m} 上で単峰形であり，Θ^{k_m} の内部に最大値をもつ．すなわち m の最適パラメータ $\theta^{k_m^*}$ が Θ^{k_m} の内部に唯一存在する．

(2) (クラスのSmoothness) $\forall m \in \mathcal{M}$ に対し，$Fisher$情報行列 $J(\theta^{k_m}|m)$ は $\forall \theta^{k_m} \in \Theta^{k_m}$ に対して $0 < C_0 < \det J(\theta^{k_m}|m) < \infty$ を満たす．ただし C_0 はある正定数である．また，$J^*(\theta^{k_m^*}|m) = J(\theta^{k_m^*}|m)$ とする．さらに，$\forall \theta^{k_m} \in \Theta^{k_m}$ に対し， $\det J(\theta^{k_m}|m)$ と $\det J^*(\theta^{k_m}|m)$ は

$$\left\| \frac{\partial \det J(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty, \tag{3.103}$$

$$\left\| \frac{\partial \det J^*(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty, \tag{3.104}$$

を満たす.

(3) (事前分布のSmoothness) $\forall m \in \mathcal{M}$ と $\forall \theta^{k_m} \in \Theta^{k_m}$ に対して $f(\theta^{k_m}|m) > 0$，かつ $f(\theta^{k_m}|m)$ は θ^{k_m} に対して3回微分可能とする．すなわち，

$$\left| \sum_{i=0}^{k_m-1} \sum_{j=0}^{k_m-1} \sum_{l=0}^{k_m-1} \frac{\partial^3 f(\theta^{k_m}|m)}{\partial \theta_i^{k_m} \partial \theta_j^{k_m} \partial \theta_l^{k_m}} \right| < c_f(\theta^{k_m}), \tag{3.105}$$

とする．ただし，$c_f(\theta^{k_m})$ は θ^{k_m} に依存する定数である．

(4) (推定量の存在性) 尤度関数 $P(x^n|m, \theta^{k_m})$ は $\forall x^n \in \mathcal{X}^n$ に対して単峰形，あるいは単調な関数とする．事後確率密度 $f(\theta^{k_m}|x^n, m)$ は $\forall x^n \in \mathcal{X}^n$ に対して，Θ^{k_m} 内で唯一の最大値をもつ．すなわち，$\forall x^n \in \mathcal{X}^n$ と $\forall m \in \mathcal{M}$ に対して，最尤推定量 $\hat{\theta}^{k_m}$ と事後確率最大推定量 $\tilde{\theta}^{k_m}$ が唯一存在する．

(5) (重複対数の法則: 1) $\forall m \in \mathcal{M}$ に対し，

$$\tilde{\theta}^{k_m} = \theta^{k_m^*} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right), \quad a.s. \tag{3.106}$$

(6) (重複対数の法則: 2) $\forall m \in \mathcal{M}$ に対し，

$$-\frac{1}{n} \log p(x^n|m, \theta^{k_m}) = H(m, \theta^{k_m}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right), \quad a.s. \tag{3.107}$$

$$-\frac{1}{n} \frac{\partial^2 \log p(x^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} = J^*(\theta^{k_m}|m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right), \quad a.s. \tag{3.108}$$

が $\theta^{k_m} \in B_\delta(\theta^{k_m^*}|m)$ に対して一様に成り立つような定数 $\delta > 0$ が存在する．

□

条件3.4,(2),(6)より，$\forall m \in \mathcal{M}, \forall \theta^{k_m} \in B_\delta(\theta^{k_m^*})$ に対し，$n \to \infty$ のとき

$$\frac{1}{n}\left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial^3 \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right| < c_p'(m,\theta^{k_m}),\quad a.s. \tag{3.109}$$

が成り立っている．ただし，$c_p'(m,\theta^{k_m}) > 0$ は m と θ^{k_m} に依存する有限値である．

3.4.2 階層モデル族の例

次に条件3.4を満たす階層モデル族の例を示す.

例 3.5 (有限定常エルゴード マルコフモデル) 情報源アルファベット $\mathcal{X} = \{0,1,2,\cdots,\beta\}$ 上の有限定常エルゴードマルコフモデルを考える．マルコフモデルの構造を表す状態の集合は事前に未知であるとする．すなわち，記憶の長さは未知である．しかし，その候補は与えられているものとすると，これは階層モデル族を構成する．m_1 を状態集合 $\{s_0(m_1), s_1(m_1), \cdots, s_{S_{m_1}}(m_1)\}$ を表すモデルとする．S_{m_1} はモデル m_1 の状態数である．同様に m_2, m_3, $\cdots$ を構成する．

例として二元アルファベット $\mathcal{X} = \{0,1\}$ を考えよう．m_1 を状態数2の一次マルコフモデルとする．m_2 を状態数4，記憶の長さ2の二次のマルコフモデルとする．同様に m_3, m_4, $\cdots$, をそれぞれ 3次マルコフモデル，4次マルコフモデル, $\cdots$, とする．このとき $\mathcal{M} = \{m_1, m_2, \cdots, m_M\}$ は階層モデル族を構成する．ここで M はモデルの最高次数である．

各モデル $m \in \mathcal{M}$ は状態遷移確率を表すパラメータ $\theta_{i,j}^{k_m}$ をもつ．ここで，$\theta_{i,j}^{k_m} = p_i^{(s_j(m))}$ はモデル m の，第 j 状態 $s_j(m)$ におけるシンボル i の生起確率を表す．$q_{s_j}(m)$ を状態 $s_j(m)$ の定常確率とする．ここで $i \in \mathcal{X}$ かつ $j = 0,1,\cdots,S_m$, S_m はモデル m の状態数である．$\theta^{k_m} = (\theta_{0,0}^{k_m}, \cdots, \theta_{\beta-1,S_m}^{k_m})^T$ はモデル m のパラメータであり，$k_m = \beta(S_m+1)$ である．

再び初期状態は既知とする．$N_{i,j}(m)$ を確率変数 X^n 内の，状態 $s_j(m)$, $j = 0,\cdots,S_m$, のもとでの各シンボル $0,1,\cdots,\beta$ の生起回数を表す確率変数とする．$p_{i,j}^*(m) = \lim_{n\to\infty}\frac{1}{n}E^*N_{i,j}(m)$ とし，$p_{i,j}^*(m)$ から計算できるモデル m の各状態の定常確率を $q_{s_0(m)}^*(m) \cdots q_{s_{S_m}(m)}^*(m)$ と表す．最適パラメータ $\theta^{k_m^*}$ は $\theta^{k_m^*} = (p_{0,0}^*(m), \cdots, p_{\beta-1,S_m}^*(m))^T$ で与えられる．

重複対数の法則 [34] より，

$$\frac{n_{i,j}(m)}{n} = p_{i,j}^*(m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \tag{3.110}$$

が成り立つので，例3.4と同様の議論により，条件3.4が成り立つことがわかる．

□

3.4.3 階層モデル族に対する結果

定理 3.6 同じ事前分布のもとで，(2.32)式と(2.33)式の関係は

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = L_{Bayes}^{m,\theta^{k_m}}(x^n) - \log P(\hat{m}|x^n), \tag{3.111}$$

で与えられる．ここで，$\hat{m}$ は事後確率 $P(m|x^n)$ を最大化するモデルであり，$P(m|x^n)$ は次式で与えられる．

$$P(m|x^n) = \frac{\int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m}}{\sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m}}. \tag{3.112}$$

(証明) 定理3.1と同様に証明できる． □

この定理も，ベイズ符号の符号長が MDL符号のそれより $-\log P(\hat{m}|x^n)$ だけ損失を小さくできることを示しており，ベイズ符号の有効性を示している．次に $-\log P(\hat{m}|x^n)$ を解析する．

そのために，次の補題を与える．これにより，

$$\left\{ P(x^n|m) = \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m} \middle| m \in \mathcal{M} \right\}$$

を離散モデル族と同様に考えることができる．

補題 3.5 $\forall m \in \mathcal{M}$ に対し，条件3.4を仮定する．このとき $\forall m \neq m^*$, $m \in \mathcal{M}$ に対し，

$$\left| \frac{P(x^n|m)}{P(x^n|m^*)} \right| \to 0, \quad a.s. \tag{3.113}$$

が成り立つ．

(証明) 3.6.3参照． □

ベイズ規則

$$P(m|x^n) \propto P(x^n|m)P(m), \tag{3.114}$$

と(3.113)式から，$\forall m \neq m^*$ に対し

$$\left| \frac{P(m|x^n)}{P(m^*|x^n)} \right| \to 1, \quad a.s. \tag{3.115}$$

が成り立つ.

この結果は事後確率最大のモデルは最適モデルに収束することを示しており, モデル選択における事後確率最大規準の一致性を示している. 事後確率最大規準は C.Schwarz により提案され (BIC), その $\log n$ の項までの漸近式が与えられている [102].

以上より, 以下の定理が与えられる.

定理 3.7 条件3.4のもとで, (2.32)式と (2.33)式の関係は

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = L_{Bayes}^{m,\theta^{k_m}}(x^n) + o^+(1),\ a.s. \tag{3.116}$$

で与えられる. ここで, $o^+(1)$ は $\lim_{n\to\infty} o^+(1) = 0$ を満たす正定数である.

(証明) 補題3.5より

$$P(m^*|x^n) \to 1,\quad a.s. \tag{3.117}$$

であるので, $\hat{m}$ は m^* に漸近的に一致する. よって,

$$-\log P(\hat{m}|x^n) = o^+(1), \tag{3.118}$$

が得られる. □

この定理により, モデル m を一つ選択することによる符号長の増加は, 漸近的に 0 に収束することがわかる. 次にパラメータを離散化してパラメータも符号化する MDL符号を考え, (2.30)式と (2.33)式の差を考える. そのためにまず次の補題を考える.

補題 3.6 各 m に対し, $\bar{\xi}^{k_m} = \sqrt{n}(\bar{\theta}^{k_m} - \hat{\theta}^{k_m})$, $\bar{\Xi}_n^{k_m} = \{\bar{\xi}^{k_m} | \bar{\theta}^{k_m} \in \bar{\Theta}_n^{k_m}\}$, および $L_{MDL}^{\bar{\theta}^{k_m}}(x^n|m)$ を

$$L_{MDL}^{\bar{\theta}^{k_m}}(x^n|m) = \min_{\bar{\theta}^{k_m}\in\bar{\Theta}_n^{k_m}} \left\{ -\log P(x^n|m,\bar{\theta}^{k_m}) - \log \frac{f(\bar{\theta}^{k_m}|m)}{\sqrt{n}^{k_m}\sqrt{\det J(\bar{\theta}^{k_m}|m)}} \right\}, \tag{3.119}$$

と定義する. このとき,

$$L_{MDL}^{\bar{\theta}^{k_m}}(x^n|m) = -\log P(x^n|m) - \max_{\bar{\xi}^{k_m}\in\bar{\Xi}_n^{k_m}} \left\{ \log \frac{f_\xi(\bar{\xi}^{k_m}|x^n,m)}{\sqrt{\det J(\hat{\theta}^{k_m} + \frac{\bar{\xi}^{k_m}}{\sqrt{n}}|m)}} \right\}, \tag{3.120}$$

が成り立つ. ここで, $f_\xi(\xi^{k_m}|x^n,m)$ は $\xi^{k_m} = \sqrt{n}(\theta^{k_m} - \hat{\theta}^{k_m})$ の事後密度であり,

$$f_\xi(\xi^{k_m}|x^n,m) = \frac{1}{\sqrt{n}^{k_m}} f(\theta^{k_m}|x^n,m), \tag{3.121}$$

で与えられる．さらに条件3.4のもとで，$n \to \infty$ のとき

$$0 < C_1 < -\max_{\bar{\xi}^{k_m} \in \bar{\Xi}_n^{k_m}} \left\{ \log \frac{f_\xi(\bar{\xi}^{k_m}|x^n, m)}{\sqrt{\det J(\hat{\theta}^{k_m} + \frac{\bar{\xi}^{k_m}}{\sqrt{n}}|m)}} \right\} < C_2 < \infty, \quad a.s. \tag{3.122}$$

が成り立つ．ただし C_1 と C_2 はある正定数である．

(証明) パラメトリックモデル族に対する結果より明らか． □

この補題より，次の定理が与えられる．

定理 3.8 条件3.4のもとで，(2.30)式と(2.33)式の関係は

$$L_{MDL}^{m,\bar{\theta}^{k_m}}(x^n) = L_{Bayes}^{m,\theta^{k_m}}(x^n) + O^+(1), \ a.s. \tag{3.123}$$

で与えられる．ここで，$O^+(1)$ は $n \to \infty$ のとき $0 < C_1 < O^+(1) < C_2 < \infty$ を満たす項である．

(証明) 補題3.6より，$\forall m \in \mathcal{M}$ に対し

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = L_{MDL}^{m,\bar{\theta}^{k_m}}(x^n) + O^+(1), \tag{3.124}$$

が成り立つ．(3.124)式と定理3.7より，定理が示された． □

以上の議論より，パラメータを量子化して符号化することによる符号長の増加，すなわち予測の累積対数損失の増加分は定数として残ってしまうことがわかり，パラメータの量子化が予測に関しては有効ではないことが明らかとなった．

3.5 考察

ベイズ符号の符号長が MDL符号のそれを下界することはすでに示されていたが[100]，本研究ではその差を漸近解析した．これは，統計的予測問題において，両者の累積対数損失の差を解析したことになる．

MDL符号では 2つの操作が特徴的である．一つはパラメータの量子化であり，一つは確率モデルを選択することである．本章の結果より，最適モデルがモデル族内で唯一であれば，モデルを選択することよりもパラメータの量子化操作の方が符号長への影響は大きいことが明らかとなった．しかし，両者とも符号長を増加させることは明らかである．また，最適モデルがモデル族内で唯一ではない場合にも，ベイズ符号の優位性が示された．これは，最適モデルが一つ

ではない状況でモデルを選択するというMDL符号の操作が符号長を増加させてしまうことを示している．以上の意味で，確率モデルを一つ選択するという方法は必ずしも全ての目的に対して有効ではないことがわかる．MDL原理は予測，あるいは符号化というよりは，統計的モデル選択，あるいはユニバーサルモデリング[101]の観点から議論されるべきと考えられる．

MDL原理をモデル選択に適用する場合，パラメータを量子化するタイプの $L_{MDL}^{m,\bar{\theta}^k m}(x^n)$ よりも，$L_{MDL}^{m,\theta^k m}(x^n)$ を適用するべきと考える．なぜならば，$\forall x^n \in \mathcal{X}^n$ に対して，$L_{MDL}^{m,\theta^k m}(x^n) \leq L_{MDL}^{m,\bar{\theta}^k m}(x^n)$ であり，$L_{MDL}^{m,\bar{\theta}^k m}(x^n)$ は真の意味で最小記述長ではないためである．一方，$-\log P(x^n|m)$ は m を固定した場合のベイズ最適解を与えている．$-\log P(x^n|m) - \log P(m)$ の最小化は，モデルの事後確率 $P(m|x^n)$ の最大化と等価となり，階層モデル族に対して一致性をもつ規準となる．

同じ事前分布を仮定した場合，MDL符号よりもベイズ符号の方が符号長を小さくできるが，もし異なる事前分布を仮定した場合にはその限りではない．すなわち，結果的に有利な事前分布を仮定していた方が優れるような場合が生じるはずである．この議論については次章で考察する．

3.6 補題と定理の証明

3.6.1 補題 3.3 の証明

まず，条件(c.1), (c.2), (c.3) が $n \to \infty$ のときほとんど確実に成り立つことを示し，事後密度関数が概収束の意味で正規分布に収束することを示す.

$f(\theta^k)$ は 3回微分可能であることから

$$\frac{1}{n}L_n''(\theta^k) = \frac{1}{n}\frac{\partial^2 \log P(x^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} + \frac{1}{n}\frac{\partial^2 \log f(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}. \tag{3.125}$$

ここで $\frac{\partial^2 \log f(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}$ は n に依存せず，$-\frac{1}{n}\frac{\partial^2 \log P(x^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} \to J^*(\theta^k),\ a.s.$ であるので，$\theta^k \in B_\delta(\theta^{k*})$ に対して一様に

$$-\frac{1}{n}L_n''(\theta^k) \to J^*(\theta^k),\ a.s. \tag{3.126}$$

したがって，$\theta^k \in B_\delta(\theta^{k*})$ に対して一様に

$$\frac{1}{n}(\Sigma_n)^{-1} \to J^*(\tilde{\theta}^k),\ a.s. \tag{3.127}$$

ただし Σ_n は，

$$\Sigma_n = -\left(L_n''(\tilde{\theta}^k)\right)^{-1} = -\left(\left.\frac{\partial^2 L_n(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right|_{\theta^k=\tilde{\theta}^k}\right)^{-1}, \tag{3.128}$$

で与えられる．$\tilde{\theta}^k \to \theta^{k*}$, $a.s.$ であるから，$n \to \infty$ のとき $C_1 < \det J^*(\tilde{\theta}^k) < C_2$, $a.s.$ を満たす正定数 $C_1, C_2 > 0$ が存在する．$n\Sigma_n \to (J^*(\tilde{\theta}^k))^{-1}$, $a.s.$ であることから，Σ_n の最大固有値が 0 に概収束することは明らかであり，(c.1)がほとんど確実に成り立つ．

次に $\tilde{\theta}^k \to \theta^{k*}$ と(3.126)式より，$L_n''(\theta^k)\{L_n''(\tilde{\theta}^k)\}^{-1}$ は $\theta^k \in B_\delta(\tilde{\theta}^k)$ に対して一様に

$$L_n''(\theta^k)\{L_n''(\tilde{\theta}^k)\}^{-1} \to J^*(\theta^k)\{J^*(\tilde{\theta}^k)\}^{-1}, \quad a.s. \tag{3.129}$$

よって条件より，(c.2)がほとんど確実に成り立つ．

よって，条件(c.1)，(c.2)がほとんど確実に成り立つので，補題3.2より，

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\,(\det \Sigma_n)^{1/2} \le (2\pi)^{-k/2}, \tag{3.130}$$

が成り立つ．

あとは 条件(c.3)を示せばよい．$\forall \theta^k \in B_\delta(\theta^{k*})$ に対して一様に

$$\frac{1}{n}\log P(x^n|\theta^k) + \frac{1}{n}\log f(\theta^k) \to -H^*(\theta^k), \quad a.s. \tag{3.131}$$

が成り立つ．よって，

$$P(x^n|\theta^k)f(\theta^k) = \exp\{-nH^*(\theta^k) + o(1)\}, \quad a.s. \tag{3.132}$$

である．これは $f(\theta^k|x^n) = \frac{P(x^n|\theta^k)f(\theta^k)}{P(x^n)}$ の分子であり，分母は θ^k に依存しない．一方，$\tilde{\theta}^k \to \theta^{k*}$, $a.s.$, $\hat{\theta}^k \to \theta^{k*}$, $a.s.$ であり，θ^{k*} は Θ^k の内部に存在するので，$n \to \infty$ のとき尤度関数 $P(x^n|\theta^k)$ はほとんど確実に単峰形であり，$\tilde{\theta}^k$ も Θ^k の内部に唯一存在する．したがって，$n \to \infty$ のとき，$\forall \epsilon > 0$ に対して

$$\inf_{\theta^k \in B_\delta(\tilde{\theta}^k)} \frac{1}{n}\log P(x^n|\theta^k)f(\theta^k) + \epsilon \ge \sup_{\theta^k \notin B_\delta(\tilde{\theta}^k)} \log P(x^n|\theta^k)f(\theta^k), \quad a.s. \tag{3.133}$$

が成り立ち，$|\Theta^k| < \infty$ より

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} L_n(\theta^k)d\theta^k \to 1, \quad a.s. \tag{3.134}$$

が得られる．これは (c.3) が概収束の意味で成立することを示している．

(c.1), (c.2), (c.3) が $n \to \infty$ のとき，ほとんど確実に成り立つので，

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\,(\det \Sigma_n)^{1/2} = (2\pi)^{-k/2}, \; a.s. \tag{3.135}$$

が成り立つ．さらに $\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ の事後確率密度は，平均ゼロ，共分散 $\{J^*(\tilde{\theta}^k)\}^{-1}$ の正規分布に概収束することもわかる．

次に定理の後半，すなわち事後密度関数が任意の直方体 Ω^k 内で一様に収束していることを示す．テーラー展開により

$$\begin{aligned} f(\theta^k|x^n) &= f(\tilde{\theta}^k|x^n)\exp\{L_n(\theta^k) - L_n(\tilde{\theta}^k)\} \\ &= f(\tilde{\theta}^k|x^n)\exp\left\{-\frac{1}{2}(\theta^k - \tilde{\theta}^k)^T(I + R_n)\Sigma_n^{-1}(\theta^k - \tilde{\theta}^k)\right\}. \end{aligned} \tag{3.136}$$

ただし，R_n は

$$R_n = L_n''(\theta^{k+})\left\{\mathrm{L}_n''(\tilde{\theta}^k)\right\}^{-1} - I, \tag{3.137}$$

で与えられ，θ^{k+} は θ^k と $\tilde{\theta}^k$ を結ぶ線分上の点である．

よって，

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\frac{1}{2}(\xi^k)^T(I + R_n)(\Sigma_n)^{-1}(\xi^k)\right\}. \tag{3.138}$$

任意の直方体 Ω^k に対し，$\sup_{\xi^k\in\Omega^k}\|\xi^k\| < V_\Omega$ を満たす定数 $V_\Omega > 0$ が存在する．したがって，$\theta^k = \tilde{\theta}^k + \frac{\xi^k}{\sqrt{n}} \to \tilde{\theta}^k$ が $\forall\xi^k \in \Omega^k$ に対して一様に成り立つ．これは，$\forall\epsilon > 0$ に対して $n \to \infty$ のとき

$$(1-\epsilon)I \le I + R_n \le (1+\epsilon)I, \quad a.s. \tag{3.139}$$

が $\forall\xi^k \in \Omega^k$ に対して一様に成り立つことを意味する．これは $L_n''(\theta^{k+})\{L_n''(\tilde{\theta}^k)\}^{-1} \to J^*(\theta^{k+})\{J^*(\tilde{\theta}^k)\}^{-1}$, $a.s.$ かつ，$\forall\xi^k \in \Omega^k$ に対して一様に $\|\theta^k - \tilde{\theta}^k\| \to 0$, $a.s.$ であることから自明である．

(3.138)式と(3.139)式より，$\forall\xi^k \in \Omega^k$ に対して一様に

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\frac{1}{2}(\xi^k)^T J^*(\tilde{\theta}^k)(\xi^k)\right\}\{1 + o(1)\}, \quad a.s. \tag{3.140}$$

$\exp\left\{-\frac{1}{2}(\xi^k)^T J^*(\tilde{\theta}^k)(\xi^k)\right\}$ は有界であるので，$\forall\xi^k \in \Omega^k$ に対して一様に

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\frac{1}{2}(\xi^k)^T J^*(\tilde{\theta}^k)(\xi^k)\right\} + o(1), \quad a.s. \tag{3.141}$$

(3.135)式より，証明が完結した． □

3.6.2 定理 3.4 の証明

補題3.3より，$\forall\xi^k \in \Omega^k$ に対して一様に

$$f_\xi(\xi^k|x^n) \to \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{\left\|\xi^k\right\|^2_{J^*(\tilde{\theta}^k)}}{2}}, \quad a.s. \tag{3.142}$$

が成り立っている.

一方, $P(x^n|\theta^k)$ は $n \to \infty$ のとき, ほとんど確実に単峰形である. よって(3.142)式より, 事後密度 $f_\xi(\xi^k|x^n)$ は $n \to \infty$ のとき

$$\max_{\bar{\xi}^k \in \bar{\Xi}^k} f_\xi(\bar{\xi}^k|x^n) \leq \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} + \epsilon_1, \quad a.s. \tag{3.143}$$

を満たす. ここで, $\bar{\xi}^k$ は

$$\bar{\xi}^k = \sqrt{n}\left(\bar{\theta}^k - \tilde{\theta}^k\right). \tag{3.144}$$

で与えられる.

$0 < \sqrt{C_0} < \sqrt{\det I(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k)}$ かつ $f_\xi(\xi^k|x^n)$ は $n \to \infty$ のとき単峰形に近づくので, $\forall \epsilon_2 > 0$ に対し $n \to \infty$ のとき,

$$\max_{\bar{\xi}^k \in \bar{\Xi}^k} \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det I(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k)}} = \max_{\bar{\xi}^k \in (\Omega^k \cap \bar{\Xi}^k)} \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det I(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k)}}, \quad a.s.$$

$$\leq \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2} \min_{\bar{\xi}^k \in (\Omega^k \cap \bar{\Xi}^k)} \sqrt{\det I(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k)}} + \epsilon_2, \quad a.s. \tag{3.145}$$

が成り立つ.

$\sqrt{\det I^*(\theta^k)}$ と $\sqrt{\det I(\theta^k)}$ は θ^k に関して連続微分可能であり, $\tilde{\theta}^k \to \hat{\theta}^k \to \theta^{k*}$, $a.s.$ と $I^*(\theta^{k*}) = I(\theta^{k*})$ が成り立っているから, $n \to \infty$のとき, $\xi^k \in \Omega^k$ に対して一様に

$$\sqrt{\det I^*(\tilde{\theta}^k)} \to \sqrt{\det I(\theta^{k*})}, \quad a.s. \tag{3.146}$$

$$\sqrt{\det I(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\bar{\xi}^k)} \to \sqrt{\det I(\theta^{k*})}, \ a.s. \tag{3.147}$$

が成り立つ. (3.145)式, (3.146)式, (3.147)式より, $\forall \epsilon_3 > 0$ に対し, $n \to \infty$ のとき

$$\max_{\bar{\xi}^k \in \bar{\Xi}^k} \log \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det I(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} \leq \frac{k}{2}\log\frac{1}{2\pi} + \epsilon_3, \quad a.s. \tag{3.148}$$

が成り立つ. したがって, ある定数 $C_1 > 0$ が存在して, $n \to \infty$ のとき

$$C_1 < L^{\bar{\theta}^k}_{MDL}(x^n) - L^{\theta^k}_{Bayes}(x^n), \ a.s. \tag{3.149}$$

が成り立つこと示された. ただし $0 < C_1 < \frac{k}{2}\log 2\pi$ である.

次に，逆側の不等式

$$L_{MDL}^{\bar{\theta}^k}(x^n) - L_{Bayes}^{\theta^k}(x^n) < C_2,\ a.s. \tag{3.150}$$

を示す．条件3.3, (2) より，$\exists K_{\Omega^k} > 0$ が存在して，$n \to \infty$ のとき

$$\frac{\max_{\bar{\xi}^k \in (\Omega^k \cap \bar{\Xi}^k)} f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k)} + \frac{K_{\Omega^k}}{\sqrt{n}}} \leq \max_{\bar{\xi}^k \in (\Omega^k \cap \bar{\Xi}^k)} \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}},\ a.s. \tag{3.151}$$

ただし，K_{Ω^k} は Ω^k に依存する．

(3.142)式より，明らかに $C' > 0$ が存在し，$n \to \infty$ のとき

$$C' < \max_{\bar{\xi}^k \in (\Omega^k \cap \bar{\Xi}^k)} f_\xi(\bar{\xi}^k|x^n) \leq \max_{\bar{\xi}^k \in \bar{\Xi}^k} f_\xi(\bar{\xi}^k|x^n),\ a.s. \tag{3.152}$$

が成り立つ．

(3.146) 式，(3.151) 式，(3.152) 式より，正定数 $C_2 > 0$ が存在して，$n \to \infty$ のとき

$$-\max_{\bar{\xi}^k \in \bar{\Xi}^k} \log \frac{f_\xi(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} < C_2,\ a.s. \tag{3.153}$$

が成り立つ．これは(3.150)式を導く．

(3.149)式と(3.150)式をあわせて定理を得る． □

3.6.3 補題 3.5 の証明

条件3.4,(5)より，

$$\left\| \tilde{\theta}^{k_m} - \theta^{k_m^*} \right\| = O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right),\ a.s. \tag{3.154}$$

であり，これは $\tilde{\theta}^{k_m}$ は $\theta^{k_m^*}$ に対する強一致推定量であることを示している．すなわち $\forall m \in \mathcal{M}$ に対して，$\tilde{\theta}^{k_m} \to \theta^{k_m^*},\ a.s.$ 一方，$\exists \delta > 0$ が存在し，$\forall \theta^{k_m} \in B(\theta^{k_m^*}|m)$ に対して一様に

$$\frac{1}{n} \log P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) \to \frac{1}{n} \log P(x^n|m, \theta^{k_m}), \tag{3.155}$$

が成り立つ．

ベイズ規則から導かれる

$$\log P(x^n|m) = \log \frac{P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m)}{f(\theta^{k_m}|x^n, m)}, \tag{3.156}$$

と補題3.3より，$\forall m \in \mathcal{M}$ に対して

$$\log P(x^n|m) = \log P(x^n|m, \tilde{\theta}^{k_m}) - \frac{k_m}{2}\log\frac{n}{2\pi} - \log\frac{\sqrt{\det J^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)} + o(1),\ a.s. \quad (3.157)$$

よって，$\forall m \in \mathcal{M}$ に対し

$$\log\frac{P(x^n|m)}{P(x^n|m^*)} = \log\frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \Big(\frac{k_m}{2} - \frac{k_{m^*}}{2}\Big)\log\frac{n}{2\pi}$$

$$- \log\frac{f(\tilde{\theta}^{k_{m^*}}|m^*)\sqrt{\det J^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)\sqrt{\det J^*(\tilde{\theta}^{k_{m^*}}|m^*)}} + o(1),\ a.s. \quad (3.158)$$

が成り立つ．

ここで，$\tilde{\theta}^{k_m} \to \theta^{k_m^*},\ a.s.$ より，$f(\tilde{\theta}^{k_m}|m) = O(1),\ a.s.$ かつ $\det J^*(\tilde{\theta}^{k_m}|m) = O(1),\ a.s.$ である．ただし $O(1)$ は $\exists C > 0$ が存在して $n \to \infty$ のとき $|O(1)| \le C\ \ a.s.$ となる項である．

(3.158)式より，もし

$$\log\frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n \to -\infty,\quad a.s. \quad (3.159)$$

が $\forall m \neq m^*$ に対して成り立てば，証明は完結する．

まず，$\log\frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})}$ を評価するために，

$$-\log P(x^n|m, \theta^{k_m^*}) = -\log P(x^n|m, \tilde{\theta}^{k_m})$$

$$-\frac{1}{2}\left(\tilde{\theta}^{k_m} - \theta^{k_m^*}\right)^T \frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\bigg|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \left(\tilde{\theta}^{k_m} - \theta^{k_m^*}\right)$$

$$+ O\Big(\left\|\tilde{\theta}^{k_m} - \theta^{k_m^*}\right\|^3 \bigg|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m} \frac{\partial^3 \log P(x^n|m, \theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\bigg|_{\theta^{k_m}=\theta^{k_m}+}\Big), \quad (3.160)$$

と展開する．

条件3.4, (5),より，

$$-(\tilde{\theta}^{k_m} - \theta^{k_m^*})^T \frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\bigg|_{\theta^{k_m}=\tilde{\theta}^{k_m}} (\tilde{\theta}^{k_m} - \theta^{k_m^*})$$

$$= \sqrt{n}(\tilde{\theta}^{k_m} - \theta^{k_m^*})^T J^*(\tilde{\theta}^{k_m}|m)\sqrt{n}(\tilde{\theta}^{k_m} - \theta^{k_m^*}) + O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \quad (3.161)$$

一方，

$$\left\|\tilde{\theta}^{k_m} - \theta^{k_m^*}\right\| = O\Big(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\Big),\ a.s. \quad (3.162)$$

と (3.109) 式より,

$$\left\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\right\|^3\left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial^3\log P(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}=O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \tag{3.163}$$

したがって,

$$-\log P(x^n|m,\theta^{k_m^*})=-\log P(x^n|m,\tilde{\theta}^{k_m})+\frac{1}{2}\sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})^TJ^*(\tilde{\theta}^{k_m}|m)\sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})$$

$$+O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \tag{3.164}$$

が成り立つ.

条件3.4, (2) と $\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\|=o(1),\ a.s.$ より, $J^*(\tilde{\theta}^{k_m}|m)\to J^*(\theta^{k_m^*}|m),\ a.s.$ が成り立つので,

$$\log\frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})}=O(\log\log n),\quad a.s. \tag{3.165}$$

したがって,

$$\log\frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m^*,\tilde{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n$$

$$=\log\frac{P(x^n|m,\theta^{k_m^*})}{P(x^n|m^*,\theta^{k_{m^*}^*})}-\frac{k_m-k_{m^*}}{2}\log n+O(\log\log n),\quad a.s. \tag{3.166}$$

まず, $k_{m^*}>k_m$ の場合を考える. このとき $H^*(m,\theta^{k_m})>H^*(m^*,\theta^{k_{m^*}^*})$ であるので, 条件3.4, (5), より, $\forall\theta^{k_m}\in\Theta^{k_m}$ に対して

$$\log\frac{P(x^n|m,\theta^{k_m})}{P(x^n|m,\theta^{k_{m^*}^*})}=-Cn+o(n),\quad a.s. \tag{3.167}$$

が成り立つ. ただし, C はある正定数である.

したがって, $\forall m\in\mathcal{M}$ かつ $k_{m^*}>k_m$ のとき

$$\log\frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m^*,\tilde{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n=-Cn-\frac{k_m-k_{m^*}}{2}\log n+o(n)$$

$$\to-\infty,\quad a.s. \tag{3.168}$$

が成り立つ.

次に $k_{m^*}<k_m$ を考える. 条件3.4, (5), より

$$\frac{1}{n}P(x^n|m,\theta^{k_m^*})\to\frac{1}{n}P(x^n|m^*,\theta^{k_{m^*}^*}),\quad a.s. \tag{3.169}$$

であるから，(3.166)式より，$\forall m \in \mathcal{M}$ かつ $k_{m^*} < k_m$ に対し，

$$\log \frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2} \log n = \frac{k_{m^*} - k_m}{2} \log n + O(\log n \log n)$$
$$\to -\infty, \quad a.s. \tag{3.170}$$

が成り立つ．

よって，$\forall k_m \neq k_{m^*}$ に対し (3.159)式が成り立ち，定理が示された． □

3.7 本章の結論

本章では，同じ事前分布のもとで，MDL符号とベイズ符号の符号長の差を解析した．この結果，最適モデルがモデル族内で唯一の場合には，パラメータの量子化を行わないMDL符号とベイズ符号の符号長の差は漸近的に 0 に収束するが，パラメータを量子化するタイプの MDL符号とベイズ符号の符号長の差は定数オーダとなることが明らかとなった．すなわち，パラメータを量子化する操作は，符号長の面からは有効ではないことが示された．また，最適モデルがモデル族内で唯一でない場合には，両符号の差は定数オーダとなり，この場合のベイズ符号の優位性が示された．統計的予測の立場から見ると，これはMDL規準に対するベイズ規準の優位性を述べていることになる．ただし，これは符号長，あるいは累積対数損失の評価による比較であり，現実的なアルゴリズムの計算量に対しては別途議論が必要である．現在のところ，ベイズ符号化は他の方法に比べて計算量が多いといえる．

本章では，同じ事前分布のもとで，両符号の符号長を解析した．両者で異なる事前分布を仮定した場合には，結果的に有利な事前分布，すなわち最適な確率モデルに高い事前確率を割り当てていた方が符号長を小さくする場合が存在することが直感的にも理解できる．次章では，このような場合の符号長の大小関係について解析する．

MDL原理の他の観点からの性質を議論することは今後の課題である．とくに，統計的モデル選択，あるいはユニバーサルモデリングの観点からの研究が重要といえる [1, 57, 76, 92, 102, 111].

第 4 章

累積対数損失を評価関数とする予測問題 (2): MDL 規準とベイズ規準の差の解析 (事前分布が異なる場合)

4.1 本章の目的

本章では，事前分布が異なる場合の(2段階符号による)MDL規準とベイズ規準の累積対数損失，すなわちMDL符号とベイズ符号の符号長の差を解析する．

MDL規準とベイズ規準の性質に対しては様々な視点から研究が行われているが，MDL原理を提唱したRissanen自身によって，同じ事前分布のもとで，ベイズ規準の累積対数損失がMDL規準のそれを下界することが示された [100]．現在では，パラメータについて混合をとったベイズ符号の符号長もMDL原理の発展形である確率的コンプレキシティの一つの形として認識され，現在も多くの研究が行われている [56, 137, 70, 151]．本研究では前章において，同じ事前分布のもとで，同じデータ系列を予測したときのMDL符号とベイズ符号の符号長の差を直接解析し，概収束の意味でその差が高々定数オーダであることを示した．しかし，異なる事前分布を仮定した場合に両符号の差はどのようになるか，という点については解析がなされていない．もともとMDL規準は事前分布という概念を陽に示さずに提案されたものであり，暗黙の内に一様な事前分布を仮定していることが多い．これに対し，ベイズ規準では積極的に事前分布を活用しようとする思想背景がある．

本章では，異なる事前分布を持つ場合に対し，MDL規準とベイズ規準の累積対数損失(符号長)の差異を解析し，両規準による累積対数損失(符号長)の漸近的な大小関係を明らかにする．その結果，モデル族が離散の場合には，漸近的には真のモデルに高い事前分布を仮定した方が優れ，パラメトリックの場合には，異なる事前分布を持つ場合でもその差がある値以内であれば，ベイズ規準の累

積対数損失の方が小さくなることを明らかにする.

4.2 MDL 規準とベイズ規準

まず, 事前分布が異なる場合も含め, MDL規準とベイズ規準を再定義する. 前にも述べたが, 対数損失を考えた予測は情報源符号化という解釈が可能であり, そのためMDL規準やベイズ規準で予測したときの累積対数損失は, そのままそれぞれの符号長と言い換えることができるが, これは符号化の問題だけに適用が限定されるということではない.

以下では, 第3章と同様に, MDL原理がそもそも符号長という評価基準のもとに導出されたことを受けて, 情報源符号化の分野からの展開を行う.

1)離散的な場合: $P(x^n|\lambda)$をx^nを符号化するための符号化確率, $C(\lambda)$ を確率モデル λ を記述するのに必要な符号長とすれば, MDL符号の符号長 $L^{\lambda}_{MDL}(x^n)$ は,

$$L^{\lambda}_{MDL}(x^n) = \min_{\lambda\in\Lambda}\left\{-\log P(x^n|\lambda) + C(\lambda)\right\}, \tag{4.1}$$

となる. λは有限集合Λの要素で, $\sum_{\lambda\in\Lambda} e^{-C(\lambda)} = 1$ を仮定する [93]. この場合, 確率モデル λ の符号長 $C(\lambda)$ の定義は λ の事前確率を $P_M(\lambda) = e^{-C(\lambda)}$ と仮定することと等価である [93, 101]. このとき,

$$L^{\lambda}_{MDL}(x^n) = \min_{\lambda\in\Lambda}\left\{-\log P(x^n|\lambda) - \log P_M(\lambda)\right\}, \tag{4.2}$$

である. これはMDL原理に基づいた予測の x^n に対する累積対数損失を与えており, 確率モデルの推定や予測のために $L^{\lambda}_{MDL}(x^n)$ を規準として用いることができる. 一方, ベイズ符号の符号長 $L^{\lambda}_{Bayes}(x^n)$ は,

$$L^{\lambda}_{Bayes}(x^n) = -\log\sum_{\lambda\in\Lambda} P(x^n|\lambda)P_B(\lambda), \tag{4.3}$$

で表される. ここで, $P_B(\lambda)$ はベイズ符号における離散的なモデル λ の事前確率である.

2) パラメトリックなモデルの場合: 次に確率モデルが (連続の) k 次元未知パラメータで特定されるパラメトリックなクラス $\{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ を考える. MDL規準を与える2段階符号の符号長は,

$$L^{\bar{\theta}^k}_{MDL}(x^n) = \min_{\bar{\theta}^k\in\bar{\Theta}^k}\left\{-\log P(x^n|\bar{\theta}^k) - \log f_M(\bar{\theta}^k)\prod_{j=0}^{k-1}\alpha_{n,j}(\bar{\theta}^k)\right\}, \tag{4.4}$$

で与えられる．$f_M(\theta^k)$ はMDL符号におけるパラメータの事前密度である．ここで，量子化幅は大きすぎても小さすぎても符号長が伸びてしまい，最適な量子化幅が議論できることが知られている．その漸近的に最適な量子化セルはFisher情報行列の主軸方向を向き，

$$\prod_{j=0}^{k-1} \alpha^*_{n,j}(\bar{\theta}^k) = \frac{1}{\sqrt{n}^k \sqrt{\det J(\bar{\theta}^k)}}, \tag{4.5}$$

を満たすような量子化幅 $\alpha^*_j(\bar{\theta}^k), j = 0, 1, \cdots, k-1$ で与えられる [56, 137, 146]. Fisher情報行列 $J(\theta^k)$ は

$$J(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E_\theta \left[-\frac{\partial^2 \log P(X^n|\theta^k)}{\partial\theta^k (\partial\theta^k)^T} \right], \tag{4.6}$$

で与えられる．ただし，E_θ は $P(\cdot|\theta^k)$ による期待値を表す.

一方，ベイズ符号の符号長は

$$L^{\theta^k}_{Bayes}(x^n) = -\log \int_{\theta^k \in \Theta^k} P(x^n|\theta^k) f_B(\theta^k) d\theta^k, \tag{4.7}$$

となる．$f_B(\theta^k)$ はベイズ符号におけるパラメータの事前密度である．(2.27) 式は，事前分布 $f_B(\theta^k)$ に対してベイズ最適な符号(ベイズ符号)を与えている．

3) 階層モデル族の場合: 階層モデル族 $\{P(\cdot|m, \theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}, m \in \mathcal{M}\}$ を考える．第3章で述べたように，階層モデル族に対しては2種類の MDL規準が定義できる．

まず，$\bar{\Theta}^{k_m}_n$ をモデル m のパラメータの代表点 $\bar{\theta}^{k_m}$ の集合，$J(\theta^{k_m}|m)$ をモデル m のFisher情報行列とする．

$$J(\theta^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E_\theta \left[-\frac{\partial^2 \log P(X^n|m, \theta^{k_m})}{\partial\theta^{k_m} (\partial\theta^{k_m})^T} \right], \tag{4.8}$$

ただし，E_θ は $P(X^n|m, \theta^{k_m})$ による期待値を表す.

MDL原理においては，x^n を符号化する際に m と $\bar{\theta}^{k_m} \in \bar{\Theta}^{k_m}_n$ を同時に符号化することを考える．ただし，符号長のロスが生じないように

$$\sum_{\bar{\theta}^k \in \bar{\Theta}^k_n} f_M(\bar{\theta}^{k_m}|m) \frac{1}{\sqrt{n}^{k_m} \sqrt{\det J(\bar{\theta}^{k_m}|m)}} = 1 \tag{4.9}$$

を満たすように代表点を選んでいるものとする．ここで，$f_M(\theta^{k_m}|m)$ はMDL符号における m を与えたもとでのパラメータの事前密度である．

パラメータを量子化するタイプの MDL符号(Type 1)の符号長 $L^{m,\bar{\theta}^{k_m}}_{MDL}(x^n)$ は

$$L^{m,\bar{\theta}^{k_m}}_{MDL}(x^n) =$$

$$\min_{m\in\mathcal{M},\bar{\theta}^{k_m}\in\bar{\Theta}_n^{k_m}}\left\{-\log P(x^n|m,\bar{\theta}^{k_m})-\log f_M(\bar{\theta}^{k_m}|m)\frac{1}{\sqrt{n}^{k_m}\sqrt{\det J(\bar{\theta}^{k_m}|m)}}-\log P_M(m)\right\}, \tag{4.10}$$

で与えられる．ただし，$P_M(m)$ はMDL符号におけるモデルの事前確率である．

もう一つのタイプの MDL符号(Type 2)は，パラメータについては量子化をせずに混合をとり，

$$-\log P_M(x^n|m) = -\log\int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m})f_M(\theta^{k_m}|m)d\theta^{k_m}, \tag{4.11}$$

の符号長で x^n を符号化する方法である．このとき，m のみを一緒に符号化すれば一意復号可能な符号が構成できる．

パラメータについては混合を用い，モデル m のみを選択するタイプの MDL符号の符号長 $L_{MDL}^{m,\theta^{k_m}}(x^n)$ は

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = \min_m\left\{-\log P_M(x^n|m)-\log P_M(m)\right\}, \tag{4.12}$$

で与えられる．

最後に，階層モデル族に対するベイズ符号の符号長 $L_{Bayes}^{m,\theta^{k_m}}(x^n)$ は，

$$L_{Bayes}^{m,\theta^{k_m}}(x^n) = -\log\sum_{m\in\mathcal{M}}\int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m,\theta^{k_m})f_B(\theta^{k_m}|m)P_B(m)d\theta^{k_m}, \tag{4.13}$$

で与えられる．ただし，$P_B(m)$ と $f_B(\theta^{k_m}|m)$ はそれぞれ，ベイズ符号におけるモデルの事前確率，パラメータの事前密度である．

第3章でも述べたが，ベイズ符号は符号化関数として候補となる全ての確率モデルの混合をとる点が特徴である．事前分布が同じ場合，ベイズ符号の符号長がMDL符号の符号長を下界することが Rissanen によっても示されている [100]．また，前章においては事前分布が同じ場合の両符号長の差を漸近解析した．以下では，異なる事前分布を持つ場合を含めて符号長の差を考察する．

4.3 離散モデル族に対する解析

離散的に表現されるモデル族に対するベイズ符号とMDL符号の符号長はそれぞれ(4.2)式,(4.3)式で与えられる．

ここで，離散有限モデル族 Λ について，以下の条件を仮定する．

条件 4.1

(1) $|D_0| \geq 1$.

(2) $\forall \lambda \in \Lambda$ に対して $P_M(\lambda) > 0$, $P_B(\lambda) > 0$.

(3) $\forall \lambda \in \Lambda$ に対して

$$-\frac{1}{n}\log P(x^n|\lambda) \to H^*(\lambda), \quad a.s. \tag{4.14}$$

が成り立つ. □

次に, $|D_0| \geq 2$ の場合の解析に必要な条件を示しておく.

条件 4.2 $\lambda_1^*, \lambda_2 \in D_0$ に対し, $p(\cdot|\lambda_1^*) \neq p(\cdot|\lambda_2^*)$ であれば,

$$\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to +\infty, \quad a.s. \tag{4.15}$$

または

$$\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to -\infty, \quad a.s. \tag{4.16}$$

が成り立つ. □

注意 4.1 第3章で仮定した条件3.1と条件4.1で異なるのは, (2)で $P_M(\lambda)$, $P_B(\lambda)$ が正であれば異なっても構わない点である. 第3章でみたように条件4.1,(3)のもとで, $\forall \lambda^* \in D_0$, $\forall \lambda \in D_1$ に対して, $n \to \infty$ のとき,

$$\left| \frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \right| \to 0, \quad a.s. \tag{4.17}$$

が成り立つ. (4.17)式はすなわち, $\forall \epsilon > 0$ に対し

$$\sum_{n=1}^{\infty} P^*\Big\{ \frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \geq \epsilon \Big\} < \infty, \tag{4.18}$$

が成り立っていることを意味する(Borel-Cantelliの補題 [34]). 条件4.1は条件3.1よりも緩い条件であるので, 第3章で示したモデル族の例は条件4.1を満たすことがわかる. □

次に, まず解析の基本となる定理を示す.

定理 4.1 (4.2)式に対して, 以下の関係が成り立つ.

$$L_{MDL}^{\lambda}(x^n) = -\log \sum_{\lambda \in \Lambda} P(x^n|\lambda) P_M(\lambda) - \log P_M(\hat{\lambda}|x^n). \tag{4.19}$$

ただし, $\hat{\lambda}$ は事前分布$P_M(\lambda)$のもとでの事後確率 $P_M(\lambda|x^n)$ を最大化する確率モデルである.

(証明) 定理3.1より明らか. □

一方，ベイズ符号の符号長に関しては次の補題が成り立つ．

補題 4.1 条件4.1のもとで，ベイズ符号の符号長は

$$L^{\lambda}_{Bayes}(x^n) = -\log \sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P_B(\lambda^*) - o^+(1), \quad a.s. \tag{4.20}$$

で与えられる．ここで，$o^+(1)$ は正で $n \to \infty$ のとき $o^+(1) \to +0$, $a.s.$ となる項である．

(証明) 補題3.1より明らか． □

次に，事前分布が異なる場合について，両規準による予測の符号長の差を解析するためにまず，$-\log P_M(\hat{\lambda}|x^n)$ が実際にどの程度のオーダになるかを考えてみる．まず $|D_0| = 1$ の場合を考えよう．

補題 4.2 条件4.1と $|D_0| = 1$ のもとで，任意の事前分布 $P_M(\lambda)$ に対して，

$$-\log P_M(\hat{\lambda}|x^n) = o^+(1), \quad a.s. \tag{4.21}$$

が成り立つ．ただし，$o^+(1)$ は正でかつ $\lim_{n\to\infty} o^+(1) = 0$, $a.s.$ となる項である．

(証明)

$$P_M(\lambda|x^n) \propto P(x^n|\lambda)P_M(\lambda), \tag{4.22}$$

と(4.17)式より，$n \to \infty$ のとき，$\forall \lambda \neq \lambda^*$ に対して

$$\left| \frac{P_M(\lambda|x^n)}{P_M(\lambda^*|x^n)} \right| \to 0, \quad a.s. \tag{4.23}$$

が成り立つ．従って，$n \to \infty$ のとき，

$$P_M(\lambda^*|x^n) \to 1, \quad a.s. \tag{4.24}$$

も成り立ち，これより

$$-\log P_M(\hat{\lambda}|x^n) = o^+(1), \quad a.s. \tag{4.25}$$

を得る． □

一方，ベイズ符号の符号長に関しては以下が成り立つ．

補題 4.3 条件4.1と $|D_0| = 1$ のもとで，任意の事前分布 $P_B(\lambda)$ に対して，次式が成り立つ．

$$-\log \sum_{\lambda \in \Lambda} P(x^n|\lambda)P_B(\lambda) = -\log P(x^n|\lambda^*) - \log P_B(\lambda^*) - o^+(1), \ a.s. \tag{4.26}$$

(証明) 補題4.1より明らか． □

以上により，異なる事前分布を持つ場合のMDL符号とベイズ符号の符号長に関して，次の補題が成り立つ．

補題 4.4 条件4.1と $|D_0| = 1$ のもとで，(4.2)式と(4.3)式に対して以下の式が成り立つ．

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) - \log \frac{P_M(\lambda^*)}{P_B(\lambda^*)} + o(1), \quad a.s. \tag{4.27}$$

$o(1)$ は $\lim_{n\to\infty} o(1) = 0,\ a.s.$ となる項である．

(証明) 定理4.1より，

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) - \log \sum_{\lambda\in\Lambda} P(x^n|\lambda)P_M(\lambda)$$

$$+ \log \sum_{\lambda\in\Lambda} P(x^n|\lambda)P_B(\lambda) - \log P_M(\hat{\lambda}|x^n),\ a.s. \tag{4.28}$$

が得られるから，補題4.2と補題4.3を適用して(4.27)式を得る． □

この補題4.4と定理4.1より，漸近的なMDL符号とベイズ符号の符号長の差について，次の定理が成り立つ．

定理 4.2 条件4.1と $|D_0| = 1$ のもとで，$n \to \infty$ のとき，以下の不等式が成り立つ．

$P_M(\lambda^*) > P_B(\lambda^*)$ のとき，

$$L^{\lambda}_{MDL}(x^n) < L^{\lambda}_{Bayes}(x^n), \quad a.s.$$

$P_M(\lambda^*) < P_B(\lambda^*)$ のとき，

$$L^{\lambda}_{MDL}(x^n) > L^{\lambda}_{Bayes}(x^n), \quad a.s.$$

$P_M(\lambda^*) = P_B(\lambda^*)$ のとき，両者は漸近的に等価である． □

最後の条件式 $P_M(\lambda^*) = P_B(\lambda^*)$ を，$\forall\lambda \in \Lambda$ に対して $P_M(\lambda) = P_B(\lambda)$ と置き換えれば，任意の x^n $(n = 1,2,\cdots)$に対して $L^{\lambda}_{MDL}(x^n) \geq L^{\lambda}_{Bayes}(x^n)$ となる [100]．もし，$\forall\lambda \in \Lambda$ に対して $P(x^n|\lambda) > 0$ かつ $P(\lambda) > 0$ であれば，$L^{\lambda}_{MDL}(x^n) > L^{\lambda}_{Bayes}(x^n)$ となることも容易にわかる．

この定理より，両規準化では，真のモデル λ^* に高い事前分布を仮定した方の規準が漸近的に符号長が小さくなる．これは，有利な事前分布を仮定した方が優れるという直感と一致する結果といえる．ただし，その符号長の差は $O(1)$ であるので，各々の符号長自体が $O(n)$ であることから，事前分布が異なる場合でも，1記号当たりの平均対数損失は $O(1/n)$ で一致することが分かる．

一方，最適モデルが唯一でない，すなわち $|D_0| > 2$ の場合を考える．この場合，第3章の議論と同様に $\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ の場合と $\exists\lambda_1^*$, $\exists\lambda_2^* \in D_0$に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ の場合で結果が異なる．

補題 4.5 $|D_0| \geq 2$, かつ $\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ であるとする．このとき 条件4.1のもとで，任意の事前分布 $P_M(\lambda)$ に対して，$n \to \infty$ のとき

$$-\log P_M(\hat{\lambda}|x^n) = C_M + o(1), \quad a.s. \tag{4.29}$$

が成り立つ．ただし，正定数 C_M は

$$C_M = -\log \frac{\max_{\lambda^* \in D_0} P_M(\lambda^*)}{\sum_{\lambda' \in D_0} P_M(\lambda')}, \tag{4.30}$$

で与えられ，$o(1)$ は $\lim_{n\to\infty} o(1) = 0$, $a.s.$ となる項である．
(証明) ベイズ規則から導かれる

$$-\log P_M(\lambda|x^n) = -\log P(x^n|\lambda) - \log P_M(\lambda) + \log P_M(x^n), \tag{4.31}$$

と補題3.1より，

$$-\log P(\lambda|x^n) = -\log \frac{P(x^n|\lambda)P_M(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*)P_M(\lambda^*)} + o^+(1), \quad a.s. \tag{4.32}$$

が成り立つ．(4.17)式と $|\Lambda|$ の有界性より，$\forall\lambda \in D_1$ に対しては $-\log P_M(\lambda|x^n) \to \infty$, $a.s.$ となる．一方，$\forall\lambda^* \in D_0$ に対しては $n \to \infty$ のとき

$$-\log P(\lambda^*|x^n) = -\log \frac{P(\lambda^*)}{\sum_{\lambda' \in D_0} P(\lambda')} + o(1), \quad a.s. \tag{4.33}$$

となる．これは $\forall\lambda_1^*, \lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ であることから明らかである． □

定理 4.3 $|D_0| \geq 2$, かつ $\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ であるとする．このとき，条件4.1のもとで(4.2)式と(4.3)式の関係は次式で与えられる．

$$L_{MDL}^{\lambda}(x^n) = L_{Bayes}^{\lambda}(x^n) - \log \frac{P_M(\lambda^{**})}{\sum_{\lambda^* \in D_0} P_B(\lambda^*)} + o(1), \quad a.s. \tag{4.34}$$

ただし λ^{**} は

$$\lambda^{**} = \arg_{\lambda^*} \max_{\lambda^* \in D_0} P_M(\lambda^*) \tag{4.35}$$

で与えられる.

(証明) 定理4.1と 補題4.5より,

$$L^{\lambda}_{MDL}(x^n) = -\log \sum_{\lambda\in\Lambda} P(x^n|\lambda)P_M(\lambda) + C_M + o(1), \quad a.s. \tag{4.36}$$

一方, 補題4.1より,

$$-\log \sum_{\lambda\in\Lambda} P(x^n|\lambda)P_B(\lambda) = -\log \sum_{\lambda^*\in D_0} P(x^n|\lambda^*)P_B(\lambda^*) - o^+(1), \quad a.s. \tag{4.37}$$

$$-\log \sum_{\lambda\in\Lambda} P(x^n|\lambda)P_M(\lambda) = -\log \sum_{\lambda^*\in D_0} P(x^n|\lambda^*)P_M(\lambda^*) - o^+(1), \quad a.s. \tag{4.38}$$

が成り立つ. したがって, $\forall\lambda^*_1, \forall\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) = P(\cdot|\lambda^*_2)$ であることから,

$$\begin{aligned} L^{\lambda}_{MDL}(x^n) &= L^{\lambda}_{Bayes}(x^n) - \log \frac{\sum_{\lambda\in\Lambda} P(x^n|\lambda)P_M(\lambda)}{\sum_{\lambda\in\Lambda} P(x^n|\lambda^*)P_B(\lambda^*)} + C_M + o(1), \quad a.s. \\ &= L^{\lambda}_{Bayes}(x^n) - \log \frac{\sum_{\lambda^*\in D_0} P_M(\lambda^*)}{\sum_{\lambda^*\in D_0} P_B(\lambda^*)} + C_M + o(1), \quad a.s. \\ &= L^{\lambda}_{Bayes}(x^n) - \log \frac{P_M(\lambda^{**})}{\sum_{\lambda^*\in D_0} P_B(\lambda^*)} + o(1), \quad a.s. \end{aligned} \tag{4.39}$$

が得られる. □

上の定理より, MDL符号とベイズ符号の大小関係について以下の定理が成り立つ.

定理 4.4 $|D_0| \geq 2$, かつ $\forall\lambda^*_1, \forall\lambda^*_2 \in D_0$ に対して $P(\cdot|\lambda^*_1) = P(\cdot|\lambda^*_2)$ であるとする. このとき条件4.1のもとで, $n \to \infty$ のとき, 以下の不等式が成り立つ.

$P_M(\lambda^{**}) > \sum_{\lambda^*\in D_0} P_B(\lambda^*)$ のとき,

$$L^{\lambda}_{MDL}(x^n) < L^{\lambda}_{Bayes}(x^n), \quad a.s.$$

$P_M(\lambda^{**}) < \sum_{\lambda^*\in D_0} P_B(\lambda^*)$ のとき,

$$L^{\lambda}_{MDL}(x^n) > L^{\lambda}_{Bayes}(x^n), \quad a.s.$$

$P_M(\lambda^{**}) = \sum_{\lambda^*\in D_0} P_B(\lambda^*)$ のとき, 両者は漸近的に等価である. □

したがって, $|D_0| \geq 2$ のときには, $P_M(\lambda^{**}) > P_B(\lambda^{**})$ であっても, それが $P_M(\lambda^{**}) < \sum_{\lambda^*\in D_0} P_B(\lambda^*)$ の範囲であれば, 漸近的にベイズ符号の符号長の方が

MDL符号のそれよりも小さくなることがわかる．$\forall\lambda_1^*, \forall\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$ という状況は特殊なケースであるが，この場合のベイズ符号の有効性が示されたといえる．

一方，$\exists\lambda_1^*, \exists\lambda_2^* \in D_0$に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ の場合は次の補題が導かれる．

補題 4.6 $|D_0| \geq 2$, かつ $\exists\lambda_1^*, \exists\lambda_2^* \in D_0$に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ であるとする．このとき 条件4.1のもとで，任意の事前分布 $P_M(\lambda)$ に対して，$n \to \infty$ のとき

$$-\log P_M(\hat{\lambda}|x^n) = o(1), \quad a.s. \tag{4.40}$$

となる．

(証明) (3.54)式より明らか． □

したがって，この場合のMDL符号とベイズ符号の符号長の差について以下の定理が導かれる．

定理 4.5 $|D_0| \geq 2$, かつ $\exists\lambda_1^*, \exists\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ であるとする．このとき，条件4.1のもとで(4.2)式と(4.3)式の関係は次式で与えられる．

$$L^{\lambda}_{MDL}(x^n) = L^{\lambda}_{Bayes}(x^n) - \log\frac{P_M(\dot{\lambda})}{P_B(\dot{\lambda})} + o(1), \quad a.s. \tag{4.41}$$

ただし，$\dot{\lambda} \in D_0$ は x^∞ を固定したときの $\hat{\lambda}$ の収束先で $\hat{\lambda} \to \dot{\lambda}|_{X^\infty = x^\infty}$ で与えられる．

(証明) 条件3.2より，$\exists\lambda_1^*, \forall\lambda_2^* \in D_0$ に対し，

$$\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to 0, \quad a.s. \tag{4.42}$$

あるいは

$$\frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to +\infty, \quad a.s. \tag{4.43}$$

となる．したがって，$n \to \infty$ のときほとんど確実に，D_0 内のうちの1つのモデルの事後確率が1に収束する．事後確率が1となるモデル $\dot{\lambda}$ はx^∞ に依存するが，これを用いると

$$\begin{aligned} L^{\lambda}_{MDL}(x^n) &= L^{\lambda}_{Bayes}(x^n) - \log\frac{\sum_{\lambda\in\Lambda} P(x^n|\lambda)P_M(\lambda)}{\sum_{\lambda\in\Lambda} P(x^n|\lambda^*)P_B(\lambda^*)} - \log P(\hat{\lambda}|x^n) \\ &= L^{\lambda}_{Bayes}(x^n) - \log\frac{\sum_{\lambda^*\in D_0} P(x^n|\lambda^*)P_M(\lambda^*)}{\sum_{\lambda^*\in D_0} P(x^n|\lambda^*)P_B(\lambda^*)} - o^+(1), \quad a.s. \\ &= L^{\lambda}_{Bayes}(x^n) - \log\frac{P_M(\dot{\lambda})}{P_B(\dot{\lambda})} - o^+(1), \quad a.s. \end{aligned} \tag{4.44}$$

が得られる． □

よって，以下の定理が得られる．

定理 4.6 $|D_0| \geq 2$, かつ $\exists\lambda_1^*, \exists\lambda_2^* \in D_0$ に対して $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ であるとする．このとき条件4.1のもとで，$n \to \infty$ のとき，以下の不等式が成り立つ．

$P_M(\dot{\lambda}) > P_B(\dot{\lambda})$ のとき，

$$L_{MDL}^{\lambda}(x^n) < L_{Bayes}^{\lambda}(x^n), \quad a.s.$$

$P_M(\dot{\lambda}) < P_B(\dot{\lambda})$ のとき，

$$L_{MDL}^{\lambda}(x^n) > L_{Bayes}^{\lambda}(x^n), \quad a.s.$$

$P_M(\dot{\lambda}) = P_B(\dot{\lambda})$ のとき，両者は漸近的に等価である． □

定理4.6の条件に現れる $P_M(\dot{\lambda})$ と $P_B(\dot{\lambda})$ も確率変数であり，情報源から発生する x^∞ に依存して変化することに注意が必要である．すなわち，この場合は D_0 内で x^∞ に依存して事後確率が1になったモデルのみが漸近的に符号長の大小関係を決定するといえる．もし，$\forall\lambda^* \in D_0$ に対して $P_M(\lambda^*) < P_B(\lambda^*)$ であれば，$n \to \infty$ のとき $L_{MDL}^{\lambda}(x^n) < L_{Bayes}^{\lambda}(x^n)$, $a.s.$ であり，逆に $\forall\lambda^* \in D_0$ に対して $P_M(\lambda^*) > P_B(\lambda^*)$ であれば，$n \to \infty$ のとき $L_{MDL}^{\lambda}(x^n) > L_{Bayes}^{\lambda}(x^n)$, $a.s.$ であることは定理4.6から明らかである．

同じ事前分布の場合には，データ数 n が有限の場合でもベイズ符号の方が符号長を小さくできることが示せたが，このように事前分布が異なる場合には，漸近的な符号長の大小関係しか示すことができない．これは，事前分布が異なる場合には，ある有限長の x^n に対してはこの大小関係が逆転する可能性があるからである．

4.4 パラメトリックモデル族に対する解析

パラメトリックモデル族に対するMDL符号とベイズ符号は(4.4)式，(4.7)式で定義される．まず，解析に本質的となる情報行列 $J^*(\theta^k)$ を以下のように定義する．

$$J^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^* \left[-\frac{\partial^2 \log P(X^n|\theta^k)}{\partial\theta^k (\partial\theta^k)^T} \right], \tag{4.45}$$

ここで，E^* は真の分布 $P^*(\cdot)$ による期待値を表す．定義より，$P^*(\cdot) \in \{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ であれば，$J^*(\theta^{k*}) = J(\theta^{k*})$ が成り立つが，これは真の分布がクラスに含ま

れない一般的状況では成り立たない．しかし，符号化などで扱う離散確率変数のモデル族のほとんどは，$J^*(\theta^{k*}) = J(\theta^{k*})$ を満たす．また，$H^*(\theta^k)$ を

$$H^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^*\Big[-\log P(X^n|\theta^k)\Big], \tag{4.46}$$

と定義する．真の分布から測ったKL-情報量の意味で最適パラメータ θ^{k*} は

$$\theta^{k*} = \arg_{\theta^k} \min H^*(\theta^k), \tag{4.47}$$

で与えられる．また，$P_M(x^n), P_B(x^n)$ を以下のように定義しておく．

$$P_M(x^n) = \sum_{m\in\mathcal{M}} P(x^n|m) P_M(m), \tag{4.48}$$

$$P_B(x^n) = \sum_{m\in\mathcal{M}} P(x^n|m) P_B(m). \tag{4.49}$$

4.4.1 モデル族の条件

本研究で扱うパラメトリックモデルクラスに対して以下を仮定する．

準備として，パラメータ空間に球 $B_\delta(\theta^{k*}) = \{\theta^k \in \Theta^k | \|\theta^k - \theta^{k*}\| < \delta\}$ を定義しておく．

条件 4.3

(1) (θ^{k*} の一意性) 関数 $-H^*(\theta^k)$ は Θ^k 上で単峰形関数とする．すなわち，θ^{k*} は Θ^k の内部に唯一存在するものとする．

(2) (モデル族の smoothness) Fisher 情報行列 $J(\theta^k)$ は $\forall\theta^k \in \Theta^k$ に対して $0 < C_0 < \det J(\theta^k) < \infty$ を満たす．ただし，C_0 はある定数である．また，$J^*(\theta^{k*}) = J(\theta^{k*})$ とする．さらに，$\det J(\theta^k)$ と $\det J^*(\theta^k)$ は，$\forall\theta^k \in \Theta^k$ に対し，

$$\left\| \frac{\partial \det J(\theta^k)}{\partial \theta^k} \right\| < \infty, \tag{4.50}$$

$$\left\| \frac{\partial \det J^*(\theta^k)}{\partial \theta^k} \right\| < \infty, \tag{4.51}$$

を満たす．

(3) (事前密度の smoothness) $\forall\theta^k \in \Theta^k$ に対し，$f_M(\theta^k) > 0$, $f_B(\theta^k) > 0$ とし，$f_M(\theta^k), f_B(\theta^k)$ は θ^k に対し3階微分可能とする．さらに，$\forall\theta^k \in \Theta^k$ に対して，

$$0 < C_1 \le \frac{f_M(\theta^k)}{f_B(\theta^k)} \le C_2 < \infty, \tag{4.52}$$

であるとする．ただし，C_1, C_2 は正定数である．

(4) (推定量の一意性) 尤度関数 $P(x^n|\theta^k)$ は，$\forall x^n \in \mathcal{X}^n$ に対して単峰形，あるいは単調であるとする．また，事後密度関数 $f(\theta^k|x^n)$ は $\forall x^n \in \mathcal{X}^n$ に対し，Θ^k 内で最大値を唯一もつものとする．すなわち，最尤推定量 $\hat{\theta}^k$ と事後確率最大推定量 $\tilde{\theta}^k$ は Θ^k 内で唯一に存在する．

(5) (事後確率最大推定量の一致性) 事後確率最大推定量 $\tilde{\theta}^k$ は強一致推定量である，すなわち，$n \to \infty$ のとき

$$\left\| \tilde{\theta}^k - \theta^{k*} \right\| \to 0, \quad a.s. \tag{4.53}$$

とする．

(6) (情報行列の一致性) $\forall \theta^k \in B_\delta(\theta^{k*})$ に対して一様に

$$-\frac{1}{n} \log P(x^n|\theta^k) \to H^*(\theta^k), \quad a.s. \tag{4.54}$$

$$-\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \to J^*(\theta^k), \quad a.s. \tag{4.55}$$

となる $\delta > 0$ が存在する． □

注意 4.2 第3章で示したように，条件4.3を満たす分布族としては，例えば事前分布にディレクレ分布を仮定した多項分布族，有限マルコフ情報源，FSMX情報源 [1] などがある．これはディレクレ分布の共役性から [16],pp.290-294 の議論からも明らかである [39]. □

4.4.2 事後密度の漸近正規性

第3章で示したように，条件4.3のもとでパラメータの事後密度は漸近正規性を満たす.

補題 4.7 条件4.3のもとで，パラメータに対する事後分布は漸近的に正規分布に概収束する．すなわち，$\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ の事後分布は，平均ゼロベクトル，共分散 $\{J^*(\tilde{\theta}^k)\}^{-1}$ の正規分布に収束する．すなわち，R_ξ を任意の直方体とすると，事後密度 $f(\theta^k|x^n)$ は

$$\begin{aligned} P(\xi^k \in R_\xi|x^n) &= \int_{\xi^k \in R_\xi} f_\xi(\xi^k|x^n) d\xi^k \\ &\to \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \int_{\xi^k \in R_\xi} e^{-\frac{1}{2}\left\|\xi^k\right\|^2_{J^*(\tilde{\theta}^k)}} d\xi^k, \ a.s. \end{aligned} \tag{4.56}$$

[1] ただし，シンボルの正規確率を表すパラメータの真値が 0 や 1 の端点となる場合を除く．このような Fisher 情報行列が無限大となる点では特別な解析が必要となる．

を満たす. ここで $f_\xi(\xi^k|x^n)$ は ξ^k の事後密度であり,

$$f_\xi(\xi^k|x^n) = \frac{1}{\sqrt{n}^k} f(\theta^k|x^n), \tag{4.57}$$

で与えられる. さらに, Ω^k を $|\Omega^k| < \infty$ を満たす任意の直方体として, $\xi^k \in \Omega^k$ に対して一様に

$$f_\xi(\xi^k|x^n) \to \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\left\|\xi^k\right\|^2_{J^*(\tilde{\theta}^k)}}, \ a.s. \tag{4.58}$$

が成り立つ. □

4.4.3 パラメトリックモデル族に対する結果

パラメータを量子化した後を考えれば, これらの仮定のもとで以下の補題が成り立つ.

定理 4.7 $\bar{\xi}^k = \sqrt{n}(\bar{\theta}^k - \tilde{\theta}^k)$ と $\bar{\Xi}_n^k = \{\bar{\xi}^k | \bar{\theta}^k \in \bar{\Theta}_n^k\}$ 定義する. このとき条件4.3のもとで, (4.4)式は次式で与えられる.

$$L^{\bar{\theta}^k}_{MDL}(x^n) = L^{\theta^k}_{Bayes}(x^n) - \max_{\bar{\xi}^k \in \bar{\Xi}_n^k} \left\{ \log \frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} + \log \frac{f_M(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})} \right\}, \tag{4.59}$$

ただし, $f_{\xi,B}(\xi^k|x^n)$ は $f_B(\cdot)$ を Θ^k 上の事前分布としたときの, $\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ で形成される空間上の事後確率密度であり, $f_B(\theta^k|x^n)$ を θ^k の事後確率密度

$$f_B(\theta^k|x^n) = \frac{P(x^n|\theta^k)P_B(\theta^k)}{f_B(x^n)}, \tag{4.60}$$

として

$$f_{\xi,B}(\xi^k|x^n) = \frac{1}{\sqrt{n}^k} f_B(\theta^k|x^n), \tag{4.61}$$

で与えられる [2].

(証明) MDL規準とベイズ符号の定義より,

$$L^{\bar{\theta}^k}_{MDL}(x^n) = L^{\theta^k}_{Bayes}(x^n) - \max_{\bar{\theta}^k \in \bar{\Theta}_n^k} \left\{ \log \frac{P(x^n|\bar{\theta}^k) f_B(\bar{\theta}^k)}{P_B(x^n)} + \log \frac{f_M(\bar{\theta}^k)}{f_B(\bar{\theta}^k)} + \frac{1}{\sqrt{n}^k \sqrt{\det J(\bar{\theta}^k)}} \right\}, \tag{4.62}$$

であることから明らか. □

[2] $\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ という関係より明らか.

この定理より，(4.59)式の右辺第2項によって，MDL符号とベイズ符号による符号長の大小関係が決まることが分かる．以下では，この右辺第2項を漸近的に評価することにより，漸近的な両規準による予測の符号長の差を解析する．

補題 4.8 $r>0$ に対して, $\mathcal{A}(r)$ と $\bar{\mathcal{A}}(r)$ を以下のように定義する．

$$\mathcal{A}(r)=\left\{\bar{\xi}^k \left| \left\|\bar{\xi}^k\right\|_{J^*\left(\tilde{\theta}^k\right)}>r, \tilde{\theta}^k+\frac{1}{\sqrt{n}}\bar{\xi}^k\in\Theta^k\right.\right\}, \tag{4.63}$$

$$\bar{\mathcal{A}}(r)=\left\{\bar{\xi}^k \left| \left\|\bar{\xi}^k\right\|_{J^*\left(\tilde{\theta}^k\right)}\leq r, \tilde{\theta}^k+\frac{1}{\sqrt{n}}\bar{\xi}^k\in\Theta^k\right.\right\}, \tag{4.64}$$

このとき，

$$\tilde{\xi}^k=\arg\max_{\bar{\xi}^k\in\bar{\Xi}_n^k}\left\{\log\frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}}+\log\frac{f_M(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}\right\}, \tag{4.65}$$

とすれば，条件4.3のもとで，$n\to\infty$ のとき

$$\tilde{\xi}^k\in\bar{\mathcal{A}}(r)\cap\bar{\Xi}_n^k,\quad a.s. \tag{4.66}$$

を満たすような $r>0$ が存在する．

(証明) 4.7.1節参照． □

この定理は，(4.59)式の右辺第2項の最大化は，$\bar{\mathcal{A}}(r)$ 内の $\bar{\xi}^k$ のみを考えればよいことを示している．この補題を用いると，次の結果が導かれる．

補題 4.9 条件4.3のもとで，$\forall\epsilon>0$ に対して，$n\to\infty$ のとき

$$\begin{aligned}\frac{k}{2}\log 2\pi-\log\frac{f_M(\theta^{k^*})}{f_B(\theta^{k^*})}-\epsilon\leq L_{MDL}^{\bar{\theta}^k}(x^n)-L_{Bayes}^{\theta^k}(x^n)\\ \leq\frac{k}{2}\log 2\pi+\frac{1}{2}-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}+\epsilon,\quad a.s.\end{aligned} \tag{4.67}$$

が成り立つ．

(証明) 4.7.2節参照． □

ここで，$-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}<0$ は，真のパラメータに対して，MDL符号の方が高い事前分布を仮定していることを意味するが，漸近的には $-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}$ が $\frac{k}{2}\log 2\pi$ をキャンセルする程小さくなければ，ベイズ符号の符号長がMDL符号のそれよりも小さくなることがわかる．よって補題5と補題6より，MDL符号とベイズ符号による予測の符号長の差に関して，次の定理を得る．

定理 4.8 条件4.3のもとで，もし $f_B(\theta^{k*}) < f_M(\theta^{k*})$ であっても，

$$\log \frac{f_M(\theta^{k*})}{f_B(\theta^{k*})} < \frac{k}{2}\log 2\pi, \tag{4.68}$$

の範囲内であれば，$n \to \infty$ のとき

$$L^{\theta^k}_{Bayes}(x^n) < L^{\bar{\theta}^k}_{MDL}(x^n), \quad a.s. \tag{4.69}$$

となる．□

この定理から，真のパラメータに高い事前確率密度を仮定することが符号長(累積対数損失)に対して有利であるが，たとえこの意味でベイズ符号に不利な事前分布が仮定された場合でも，ある範囲内であればやはりベイズ符号の符号長が小さくなるというベイズ符号，すなわちベイズ符号化の優れた特性が理解できる．

4.5 階層モデル族に対する解析

最後に階層モデル族に対して，(4.10) 式，(4.12) 式，(4.13) 式の漸近的関係を示す．

4.5.1 モデル族の条件

条件 4.4

(1) ($\theta^{k^*_m}$ の一意存在性) $\forall m \in \mathcal{M}$ に対し，$-H^*(m, \theta^{k_m})$ は Θ^{k_m} 上で単峰形であり，Θ^{k_m} の内部に最大値をもつ．すなわち m の最適パラメータ $\theta^{k^*_m}$ が Θ^{k_m} の内部に唯一存在する．

(2) (クラスのSmoothness) $\forall m \in \mathcal{M}$ に対し，Fisher情報行列 $J(\theta^{k_m}|m)$ は $\forall \theta^{k_m} \in \Theta^{k_m}$ に対して $0 < C_0 < \det J(\theta^{k_m}|m) < \infty$ を満たす．ただし C_0 はある正定数である．また，$J^*(\theta^{k^*_m}|m) = J(\theta^{k^*_m}|m)$ とする．さらに，$\forall \theta^{k_m} \in \Theta^{k_m}$ に対し， $\det J(\theta^{k_m}|m)$ と $\det J^*(\theta^{k_m}|m)$ は

$$\left\| \frac{\partial \det J(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty, \tag{4.70}$$

$$\left\| \frac{\partial \det J^*(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty, \tag{4.71}$$

を満たす．

(3) (事前分布のSmoothness) $\forall m \in \mathcal{M}$ と $\forall \theta^{k_m} \in \Theta^{k_m}$ に対して $f_M(\theta^{k_m}|m) > 0$, $f_B(\theta^{k_m}|m) > 0$, かつ $f_M(\theta^{k_m}|m)$, $f_B(\theta^{k_m}|m)$ は θ^{k_m} に対して3階微分可能とする. さらに, $\forall m \in \mathcal{M}$, $\forall \theta^k \in \Theta^k$ に対して,

$$0 < C_1 \leq \frac{f_M(\theta^{k_m}|m)}{f_B(\theta^{k_m}|m)} \leq C_2 < \infty, \tag{4.72}$$

であるとする. ただし, C_1, C_2 はある正定数である.

(4) (推定量の存在性) 尤度関数 $P(x^n|m, \theta^{k_m})$ は $\forall x^n \in \mathcal{X}^n$ に対して単峰形, あるいは単調な関数とする. 事後確率密度 $f_M(\theta^{k_m}|x^n, m)$, $f_B(\theta^{k_m}|x^n, m)$ は $\forall x^n \in \mathcal{X}^n$ に対して, Θ^{k_m} 内で唯一の最大値をもつ. すなわち, $\forall x^n \in \mathcal{X}^n$ と $\forall m \in \mathcal{M}$ に対して, 最尤推定量 $\tilde{\theta}^{k_m}$, $f_M(\theta^{k_m}|x^n, m)$ に対する事後確率最大推定量 $\tilde{\theta}_M^{k_m}$, $f_B(\theta^{k_m}|x^n, m)$ に対する事後確率最大推定量 $\tilde{\theta}_B^{k_m}$ がそれぞれ唯一存在する.

(5) (重複対数の法則: 1) $\forall m \in \mathcal{M}$ に対し,

$$\tilde{\theta}^{k_m} = \theta^{k_m^*} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{4.73}$$

(6) (重複対数の法則: 2) $\forall m \in \mathcal{M}$ に対し,

$$-\frac{1}{n}\log p(x^n|m, \theta^{k_m}) = H^*(m, \theta^{k_m}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{4.74}$$

$$-\frac{1}{n}\frac{\partial^2 \log p(x^n|m, \theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = J^*(\theta^{k_m}|m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{4.75}$$

が $\theta^{k_m} \in B_\delta(\theta^{k_m^*}|m)$ に対して一様に成り立つような定数 $\delta > 0$ が存在する.

□

4.5.2 階層モデル族に対する結果

まず, (4.12)式と(4.13)式の関係について次の補題が得られる.

補題 4.10 条件4.4のもとで

$$L_{MDL}^{m,\theta^{k_m}}(x^n) = L_{Bayes}^{m,\theta^{k_m}}(x^n) - \log\frac{f_M(\theta^{k_{m^*}^*}|m^*)P_M(m^*)}{f_B(\theta^{k_{m^*}^*}|m^*)P_B(m^*)} + o(1), \quad a.s. \tag{4.76}$$

が成り立つ.

(証明) 4.7.3節参照. □

この補題より，明らかに以下の定理が成り立つ．

定理 4.9 条件4.4のもとで $n \to \infty$ のとき，(4.12)式と(4.13)式の関係について以下の不等式が成り立つ．

$f_M(\theta^{k^*_{m^*}}|m^*)P_M(m^*) > f_B(\theta^{k^*_{m^*}}|m^*)P_B(m^*)$ のとき，

$$L^{m,\theta^{km}}_{MDL}(x^n) < L^{m,\theta^{km}}_{Bayes}(x^n), \quad a.s.$$

$f_M(\theta^{k^*_{m^*}}|m^*)P_M(m^*) < f_B(\theta^{k^*_{m^*}}|m^*)P_B(m^*)$ のとき

$$L^{m,\theta^{km}}_{MDL}(x^n) > L^{m,\theta^{km}}_{Bayes}(x^n), \quad a.s.$$

□

これは，パラメータに対しては混合をとる MDL符号とベイズ符号を比較した場合，最適モデルと最適パラメータに高い事前確率を仮定していた方が優れるということを示している．

定理 4.10 条件 4.4のもとで，もし $f_B(\theta^{k^*_{m^*}}|m^*)P_B(m^*) < f_M(\theta^{k^*_{m^*}}|m^*)P_M(m^*)$ であっても，

$$\frac{f_B(\theta^{k^*_{m^*}}|m^*)P_B(m^*)}{f_M(\theta^{k^*_{m^*}}|m^*)P_M(m^*)} > \frac{k_{m^*}}{2}\log 2\pi, \tag{4.77}$$

であれば，$n \to \infty$ のとき

$$L^{m,\bar{\theta}^{km}}_{MDL}(x^n) > L^{m,\theta^{km}}_{Bayes}(x^n), \quad a.s. \tag{4.78}$$

が成り立つ． □

4.6 考察

$|D_0| = 1$ の場合の離散モデル族に対する符号長は，両符号のうち最適モデルに高い事前確率を仮定した方の符号長が短くなったが，これは同じ事前分布を仮定した場合にはモデルを一つ選択する方法と全てのモデルの混合を用いる方法による符号長の差が漸近的に0に収束することに起因する．すなわち，同じ事前分布のときにMDL符号とベイズ符号は漸近的等価になるので，事前分布が異なる場合には真のモデルに高い事前確率を仮定した方がより符号長を小さくできるわけである．一方，パラメトリックなモデル族の場合には，真のモデルにMDL符号よりも小さい事前密度を仮定したとしても，ある範囲内であれば漸近

的にベイズ符号が優れることがわかる．これは，パラメータの事後分布の主軸方向の標準偏差と量子化幅が漸近的に等しいので，最適パラメータを含むセルの事後確率は常に1以下の定数で抑えられ，1に収束しないことに起因する．この結果は，連続のパラメータを離散化する操作が必ずしも有効ではないことを示している．これは階層モデル族に対しても同様である．

4.7 補題と定理の証明

4.7.1 補題 4.7 の証明

まず，$\bar{\xi}^k \in \mathcal{A}(r) \cap \bar{\Xi}_n^k$ の場合について考える．このとき $\left\|\bar{\xi}^k\right\|_{J(\tilde{\theta}^k)} > r$ である．補題4.7より，ξ^k の事後密度 $f_{\xi,B}(\xi^k|x^n)$ は $R_\xi(0) \in \{|R_\xi(0)| < \forall R\}$ に対して一様に(4.56)式を満たす．また条件4.3,(1),(4)より尤度関数は $n \to \infty$ でほとんど確実に単峰形であり，さらに条件4.3,(3)より事前密度の導関数は有界であることから，R を十分大きくとれば，$\forall \bar{\xi}^k \in \mathcal{A}(r) \cap \bar{\Xi}_n^k$ と $\forall \epsilon_1 > 0$，$\forall r > 0$ に対して，$n \to \infty$ のとき，

$$f_{\xi,B}(\bar{\xi}^k|x^n) \leq \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{r^2}{2}} + \epsilon_1,\ a.s. \tag{4.79}$$

が成り立つ．さらに，条件4.3, (2), (3) より，$\forall \theta^k \in \Theta^k$ に対して，

$$0 < \sqrt{C_0} \leq \sqrt{\det J(\theta^k)} < \infty, \tag{4.80}$$

$$0 < C_1 \leq \frac{f_M(\theta^k)}{f_B(\theta^k)} \leq C_2 < \infty, \tag{4.81}$$

であることから，$\forall \bar{\xi}^k \in \mathcal{A}(r) \cap \bar{\Xi}_n^k$ と $\forall \epsilon_1 > 0$，$\forall r > 0$ に対して，$n \to \infty$ のとき

$$\frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} \frac{f_M(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})} \leq \frac{C_2}{C_0(2\pi)^{k/2}} \left\{ \sqrt{\det J^*(\tilde{\theta}^k)} e^{-\frac{r^2}{2}} + \epsilon_1 (2\pi)^{k/2} \right\},\ a.s. \tag{4.82}$$

が成り立つ．条件4.3, (5)より $\tilde{\theta}^k \to \theta^{k*}$，$a.s.$ であり，$r > 0$ と $\epsilon_1 > 0$ も任意だったので，この右辺は任意に小さくできることがわかる．

次に，$\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k$ の場合を考える．$\sqrt{\det J(\theta^k)}$ の θ^k に関する微係数が存在するので，$\forall \bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k$ に対して一様に，

$$\left| \sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)} - \sqrt{\det J(\tilde{\theta}^k)} \right| \leq \frac{K_r(\tilde{\theta}^k)}{\sqrt{n}}, \tag{4.83}$$

を満たす $K_r(\tilde{\theta}^k)$ が存在する．一般に，$K_r(\tilde{\theta}^k) > 0$ は $\tilde{\theta}^k$ に依存する．

もし $n \to \infty$ のとき $\sqrt{\det J(\tilde{\theta}^k)} \to \infty$ であるとすると，一般に $K_r(\tilde{\theta}^k) \to \infty$ となってしまう．しかし条件4.3, (5)より，$\theta^{k*} \in \Theta^k$ と $\forall \epsilon_2 > 0$ に対して，$n \to \infty$ のとき

$$\left\| \tilde{\theta}^k - \theta^{k*} \right\|_{J(\theta^{k*})} < \epsilon_2, \quad a.s. \tag{4.84}$$

が成り立つ．従って，$\forall \epsilon_3 > 0$ に対して $n \to \infty$ のとき，

$$K_r(\tilde{\theta}^k) \leq K_r(\theta^{k*}) + \epsilon_3, \quad a.s. \tag{4.85}$$

が成り立つ．ここで，$K_r(\theta^{k*})$ は確率変数である $\tilde{\theta}^k$ には依存せず，真のパラメータ θ^{k*} に依存する[3].

従って(4.81)式とから，$\forall \epsilon_3 > 0$ に対して $n \to \infty$ のとき，

$$\frac{C_1 f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J(\tilde{\theta}^k)} + \frac{K_r(\theta^{k*}) + \epsilon_3}{\sqrt{n}}} \leq \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}} \frac{f_M\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}{f_B\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}, \quad a.s. \tag{4.86}$$

が成り立ち，このとき同様に，

$$\frac{C_1 \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J(\tilde{\theta}^k)} + \frac{K_r(\theta^{k*}) + \epsilon_3}{\sqrt{n}}} \leq \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}} \frac{f_M\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}{f_B\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}, \quad a.s. \tag{4.87}$$

が成り立つ．

ここで補題4.7より，$\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k$ に対して一様に，$\forall \epsilon_4 > 0$ に対して $n \to \infty$ のとき，

$$\left| f_{\xi,B}(\bar{\xi}^k | x^n) - \frac{\sqrt{\det J^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{\left\| \bar{\xi}^k \right\|^2_{J^*(\tilde{\theta}^k)}}{2}} \right| < \epsilon_4, \quad a.s. \tag{4.88}$$

が成り立つ．ここで，条件4.3, (2), (5)より，$J^*(\tilde{\theta}^k) \to J(\theta^{k*}), a.s.$ かつ $J(\tilde{\theta}^k) \to J(\theta^{k*}), a.s.$ である．一方，ξ^k 空間上でのMDLの量子化幅 $\alpha^*_{\xi_j}(\bar{\xi}^k), j = 0, 1, \cdots, k-1$ は，Fisher情報行列の主軸方向を向き，かつ

$$\prod_{j=0}^{k-1} \alpha^*_{\xi_j}(\bar{\xi}^k) = \frac{1}{\sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}} \tag{4.89}$$

を満たすこと，および(4.83)式，(4.88)式とから，$\forall \epsilon_5 > 0$ に対して，$n \to \infty$ のとき，

$$\frac{\sqrt{\det J(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{1}{2}} - \epsilon_5 \leq \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} f_{\xi,B}(\bar{\xi}^k | x^n) \leq \frac{\sqrt{\det J(\tilde{\theta}^k)}}{(2\pi)^{k/2}} + \epsilon_5, \quad a.s. \tag{4.90}$$

---

[3]もちろん，もし $0 < C_0 < \sqrt{\det J(\theta^k)} < C_0' < \infty$, が $\forall \theta^k \in \Theta^k$ に対して仮定できれば，$\forall \tilde{\theta}^k \in \Theta^k$ に対して $K_r(\tilde{\theta}^k) < K_r$ となる $K_r > 0$ が存在する．

が成り立つ．従って，(4.87)式と(4.90)式より，$\forall \bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k$ と $\forall \epsilon_3, \epsilon_5 > 0$ に対して，$n \to \infty$ のとき，

$$\frac{C_1}{\sqrt{\det J(\tilde{\theta}^k)} + \frac{K_r(\theta^{k*}) + \epsilon_3}{\sqrt{n}}} \left\{ \frac{\sqrt{\det J(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{1}{2}} - \epsilon_5 \right\}$$

$$\leq \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}} \frac{f_M\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}{f_B\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}, \quad a.s. \tag{4.91}$$

が成り立つ．

$\epsilon_3, \epsilon_5 > 0$ は任意でかつ，条件4.3, (3)より $n \to \infty$ で $\tilde{\theta}^k \to \theta^{k*}, a.s.$ であるから，$\forall \epsilon_6 > 0$ に対して，$n \to \infty$ のとき，

$$\frac{C_1 e^{-\frac{1}{2}}}{(2\pi)^{k/2}} - \epsilon_6 \leq \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}} \frac{f_M\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}{f_B\left(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}}\right)}, \quad a.s. \tag{4.92}$$

が成り立つ．

r は任意であったこと，(4.82)式において正数 ϵ_1 も任意であること，及び(4.92)式の左辺は正定数で下界されることから，十分大きい r を選べば，$n \to \infty$ のとき，

$$\tilde{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k, \quad a.s. \tag{4.93}$$

が成り立つ． □

4.7.2 補題 4.8 の証明

補題4.8より次式が成り立つ．

$$\max_{\bar{\xi}^k \in \bar{\Xi}_n^k} \log \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} \frac{f_M(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}$$

$$= \max_{\bar{\xi}^k \in \bar{\mathcal{A}}(r) \cap \bar{\Xi}_n^k} \log \frac{f_{\xi,B}(\bar{\xi}^k | x^n)}{\sqrt{\det J(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}} \frac{f_M(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}, \quad a.s. \tag{4.94}$$

$f_M(\theta^k)$, $f_B(\theta^k)$ の連続性と r が定数であること，及び $\tilde{\theta}^k \to \theta^{k*}, a.s.$ より，$\forall \epsilon_7 > 0$ と $\bar{\xi}^k \in \bar{\mathcal{A}}(r)$ に対して，$n \to \infty$ のとき，

$$\left| \frac{f_M(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})}{f_B(\tilde{\theta}^k + \frac{\bar{\xi}^k}{\sqrt{n}})} - \frac{f_M(\theta^{k*})}{f_B(\theta^{k*})} \right| \leq \epsilon_7, \quad a.s.$$

であり，このとき同時に

$$-\max_{\bar{\xi}^k\cap\bar{\Xi}_n^k}\log\frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}}-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}-\epsilon_7\le L_{MDL}^{\bar{\theta}^k}(x^n)-L_{Bayes}^{\theta^k}(x^n)$$

$$\le-\max_{\bar{\xi}^k\cap\bar{\Xi}_n^k}\log\frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}}-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}+\epsilon_7,\quad a.s. \tag{4.95}$$

が成り立つ．補題4.8より $n\to\infty$ のとき

$$\tilde{\xi}^k\in\bar{\mathcal{A}}(r)\cap\bar{\Xi}_n^k,\quad a.s. \tag{4.96}$$

であることと，$\bar{\xi}^k\in\bar{\mathcal{A}}(r)\cap\bar{\Xi}_n^k$ に対しては (4.90) 式が成り立つことから，$\forall\epsilon_8>0$ に対して $n\to\infty$ のとき，

$$\begin{aligned}\frac{k}{2}\log 2\pi-\epsilon_8&\le-\max_{\bar{\xi}^k\in\bar{\Xi}_n^k}\log\frac{f_{\xi,B}(\bar{\xi}^k|x^n)}{\sqrt{\det J(\tilde{\theta}^k+\frac{\bar{\xi}^k}{\sqrt{n}})}}\\&\le\frac{k}{2}\log 2\pi+\frac{1}{2}+\epsilon_8,\quad a.s.\end{aligned} \tag{4.97}$$

が成り立つ．

従って，$\forall\epsilon_9>0$ に対して，$n\to\infty$ のとき，

$$\frac{k}{2}\log 2\pi-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}-\epsilon_9\le L_{MDL}^{\bar{\theta}^k}(x^n)-L_{Bayes}^{\theta^k}(x^n)$$

$$\le\frac{k}{2}\log 2\pi+\frac{1}{2}-\log\frac{f_M(\theta^{k*})}{f_B(\theta^{k*})}+\epsilon_9,\quad a.s. \tag{4.98}$$

が成り立ち，補題が得られた． □

4.7.3 補題 4.9 の証明

定理3.6より，

$$\begin{aligned}L_{MDL}^{m,\theta^{km}}(x^n)&=-\log\sum_{m\in\mathcal{M}}P_M(x^n|m)P_M(m)-\max_{m\in\mathcal{M}}P_M(m|x^n)\\&=L_{Bayes}^{m,\theta^{km}}(x^n)-\log\frac{\sum_{m\in\mathcal{M}}P_M(x^n|m)P_M(m)}{\sum_{m\in\mathcal{M}}P_B(x^n|m)P_B(m)}+o(1),\quad a.s.\end{aligned} \tag{4.99}$$

ここで，ベイズ規則より

$$-\log\sum_{m\in\mathcal{M}}P_M(x^n|m)P_M(m)=-\log\frac{P_M(x^n|m^*)P_M(m^*)}{P_M(m^*|x^n)}, \tag{4.100}$$

$$-\log \sum_{m\in\mathcal{M}} P_B(x^n|m)P_B(m) = -\log \frac{P_B(x^n|m^*)P_B(m^*)}{P_B(m^*|x^n)}, \tag{4.101}$$

である.

ここで，補題3.5より，$\forall m \neq m^*, m \in \mathcal{M}$ に対し，

$$\frac{P_M(x^n|m)}{P_M(x^n|m^*)} \to 0, \quad a.s. \tag{4.102}$$

$$\frac{P_B(x^n|m)}{P_B(x^n|m^*)} \to 0, \quad a.s. \tag{4.103}$$

が成り立ち，ベイズ規則と $|\mathcal{M}|$ の有界性とから，

$$P_M(m^*|x^n) \to 1, \quad a.s. \tag{4.104}$$

$$P_B(m^*|x^n) \to 1, \quad a.s. \tag{4.105}$$

が成り立つ．一方，(3.157)式で見たように，$\forall m \in \mathcal{M}$ に対して

$$\log P(x^n|m) = \log P(x^n|m,\tilde{\theta}^{k_m}) - \frac{k_m}{2}\log\frac{n}{2\pi} - \log\frac{\sqrt{\det J^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)} + o(1),\ a.s. \tag{4.106}$$

が成り立っている．$\tilde{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ に注意して，(4.104)式，(4.105)式，(4.106)式を(4.99)式に代入すれば証明が完結する.

4.8 本章のまとめ

本章では，事前分布の異なる場合について，MDL規準とベイズ規準による予測の累積対数損失の差，すなわちMDL符号とベイズ符号の符号長の差を解析した．その結果，最適モデルが唯一の場合の離散モデル族に対しては，漸近的に真のモデルに高い事前分布を仮定した方が優れるという極めて自然な結果が得られた．一方，パラメトリックなモデル族の場合には，真のパラメータにMDL符号より低い事前分布を仮定したとしても，ある範囲内であれば，漸近的にベイズ符号が優れることが明らかとなった．

通常，MDL符号を構成する際に，その事前分布が暗に仮定されるので，同じ事前分布を用いてベイズ符号を適用すれば MDL符号の符号長よりも良い予測が可能であることは自明である．このため，もし事前分布が既知であるならば，損失の面からはベイズ符号を用いるべきである．しかし，MDL符号とベイズ符号の事前分布が異なる場合でもそれがある範囲内であれば，ベイズ符号の符号長が小さくなる場合が存在するので，この点からもベイズ符号による予測の損失の面からの有効性が示されたといえる．

ベイズ符号による予測では一般に全てのモデルで混合をとる操作が必要となる．あるクラスの情報源と共役な事前分布を仮定すれば，ある程度効率的なアルゴリズムを構成することができるが[81]，計算量の低減についてはさらに研究が進められている段階である．本研究の成果が実際に有用となるためには，例えば情報源符号化の分野では効率的なベイズ符号化アルゴリズムを構成することが必要であり，これらの現実に利用可能な方法を構成することは今後の課題である．

第 5 章

累積対数損失を評価関数とする予測問題 (3): ベイズ規準の漸近的評価

5.1 本章の目的

ベイズ規準 [25, 26]による予測は，ベイズ決定理論に基づくベイズ最適な予測である[80]．予測の良さを測る評価基準としては，古くから広く適用されている対数損失関数がある．累積対数損失を考えた場合には全てのモデルで混合をとった予測分布がベイズ最適な予測となり，予測分布の対数が累積対数損失を与える．与えられたデータに対する累積対数損失は，MDL符号の発展形である確率的コンプレキシティ(stochastic complexity) [100]として定義されるなど，現在の予測の適当な評価基準として多くの研究がなされている．

ベイズ規準の性質に関しては様々な研究がなされているが[25, 26, 70, 80, 100, 101, 137]，中でも重要な成果として，B.S.Clarke と A.R.Barron によってなされた，i.i.d.パラメトリックモデル族に対する平均累積対数損失の漸近解析 [25] およびベイズリスクの漸近式の解析と 漸近的に min-max規準を達成する事前分布の解析 [26]があげられる．この研究の流れは，現在も多くの関連した結果を導いている．

一方，最近文脈木重み付け法(context tree weighting: CTW法)[148]とその拡張形[70, 81]により，現実的な半順序構造の階層モデル族である FSMXモデル族に対し，効率的に予測分布(ベイズ符号の符号化確率)を求めるアルゴリズムが与えられ，ベイズ規準による予測，及びベイズ符号の実現化に向けて，大きな一歩を踏み出した．この方法はパターン認識問題のカテゴリ予測などにも適用可能であり，ベイズ規準による予測の現実的な方法を与えている．このベイズ規準による予測の評価を考えた場合，パターン認識問題などに対する予測問題では，i.i.d.モデルを仮定できる場合が多いので，Clarke と Barron の漸近評価が

そのまま適用可能といえる．しかし，情報源符号化問題などを考えた場合，実用的なFSMXモデル族に対するベイズ符号については次の2点により Clarke と Barron の漸近解析の結果を適用できない．すなわち，1) FSMXモデル族はマルコフモデルの一種であり，i.i.d.情報源ではないこと，2) FSMXモデル族は半順序構造の階層モデル族を構成し，ベイズ符号化アルゴリズムではモデルについても混合をとっているので，一つのパラメトリックモデルとしては扱えないこと，である．

本章では，階層モデル族であるFSMXモデルに対する予測の累積対数損失とその期待値(リスク)を考える．このような記憶のある情報源に対して解析を行うことにより，ベイズ規準による予測の性能評価が広い適用分野に対して可能となる．すなわち本章では，Clarke と Barron のリスクの漸近式の結果を，記憶のある階層モデル族であるFSMXモデルに対し，モデルとパラメータの両方で混合をとった場合のベイズ最適な予測に対して拡張する．この結果の適用例として，FSMX情報源という情報源符号化問題で有用となるモデル族に対するベイズ符号の符号長の評価が得られる．すなわち解析の結果，先のベイズ符号化アルゴリズムで与えられる符号長について一つの評価が与えられることになる．

まず，FSMXモデル族において，ほとんど確実に出現する系列に対するベイズ規準の累積対数損失を解析する．これは考えられる全ての系列を対象としているわけではなく，真の情報源から確率1で選られる系列全てに対して成り立つ性質を議論するという意味である[1]．さらに，ベイズ規準の平均累積対数損失の漸近式を与える．本研究の解析の鍵は，パラメータの事後密度が漸近正規性を満たすことである．Clarke と Barron も [25] において，漸近的符号長の式が漸近正規性から直感的に導かれることを示唆しているが，その証明においてはその漸近正規性を直接用いず，i.i.d.の性質を利用して漸近式を導いている．しかし，事後確率密度の漸近正規性は i.i.d.のモデル族だけでなく，マルコフモデルでも成り立つ性質である [16, 22]．本研究では，まず FSMXモデル族においてパラメータの事後密度の漸近正規性が成り立つことを示し，次にこれを直接用いることにより，ベイズ規準による予測の累積対数損失を解析する．この結果，Clarke と Barron[25]で導かれた漸近式の階層モデル族への拡張形が与えられる．

[1]一方，個々の系列全体に対して一様に成り立つ漸近式は，J.Rissanen によって提唱されているユニバーサルモデリング [101] の概念において本質的な役割を演ずる．本論文では，標本空間内の全ての系列に対して一様に成り立つ漸近式に変わって，真の確率分布から出現する系列全てに対して成り立つ漸近式を議論している．

5.2 準備

本章では，情報源符号化問題で有用となる FSMXモデル族を対象としているので，ベイズ規準による予測をベイズ符号化と限定した記述をする．しかし，この結果は符号化のみに限定されるものではなく，対数損失を考えた予測全般に適用できることに注意する．対数損失を考えた場合の予測の損失と符号長は同じものであり，解釈を変えたものに過ぎないことはこれまで述べてきた通りである．すなわち，本章の結果はそのまま対数損失を考えた予測問題に適用可能である．

5.2.1 FSMX モデル族

前章までと同様に，離散アルファベットを $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$，長さ n の情報系列を $x^n = x_1 x_2 \cdots x_n, (\forall i, x_i \in \mathcal{X})$，長さ n の系列の集合を $\mathcal{X}^n$，無限系列を x^∞ と記述する．

一つの FSMX情報源モデル m は状態の集合で規定され，文脈木と呼ばれる完全$(\beta+1)$進木 $T(m)$ で表現される[2]．以後，一つの FSMXモデル m を単にモデル m と表す．文脈木の一つの枝はシンボル $x \in \mathcal{X}$ を表現しており，一つの葉から根までのパスは FSMXモデルの一つの文脈，または状態を表している． 時点 t における FSMXモデルの状態は過去の系列 x^{t-1} により決まり，モデル m の x^{t-1} から状態への写像を $\varphi_m(x^{t-1})$ とする．$s(m)$ はモデル m の状態を，$S(m)$ はモデル m の状態の集合を表す． $S(m)$ はまた，文脈木 $T(m)$ の葉を表現している．

モデル m はパラメータ空間を構成する．モデル m のパラメータの次元を k_m とし，k_m次元パラメータを θ^{k_m}，パラメータ空間を Θ^{k_m} とする．本研究では $\Theta^{k_m} = (0,1)^{k_m}$ とする．ただし，$(0,1)^{k_m}$ は 1辺が $[0,1]$ の k_m次元直方体の内部を表す.

$\theta^{k_m}_{i,j}$ を状態 $s_j(m)$ のもとでのシンボル i の生起確率を表すパラメータ，$q_{s_j(m)}$ を θ^{k_m} から一意に決まる状態 $s_j(m)$ の定常確率とする．ただし，$i \in \mathcal{X}$ かつ $j = 0, 1, \cdots, |S(m)| - 1$ とする．$\theta^{k_m} = (\theta^{k_m}_{0,0}, \cdots, \theta^{k_m}_{\beta-1,|S(m)|-1})^T$ は $\beta|S(m)|$次元パラメータとみなすことができ，すなわち $k_m = \beta|S(m)|$ であることがわかる．

FSMXモデル m の確率構造は x^{t-1} のもとでの x_t の条件付き確率 $P(x_t|x^{t-1}, m, \theta^{k_m})$ で与えられる．モデル m とパラメータ θ^{k_m} で規定される，x^{t-1} が与えられたも

[2]FSMX モデルはマルコフモデル (Markov model) の一種であり，その確率構造は状態の集合と状態遷移確率で決定される．

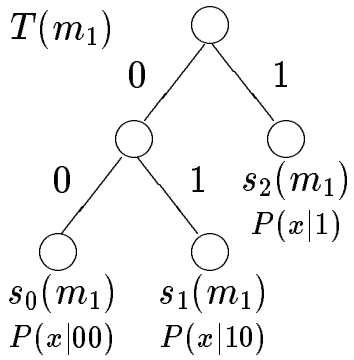

図 5.1: FSMXモデルの例: $T(m_1)$

とでの次のシンボル x_t の生起確率は

$$P(x_t|x^{t-1}, m, \theta^{k_m}) = P(x_t|\varphi_m(x^{t-1}), m, \theta^{k_m}). \tag{5.1}$$

で与えられる.

例として, 図5.1に示した, 二元の場合 ($\beta = 1$) の FSMX モデルの例 (m_1 とする) を考える. もし過去の系列が $x^{t-1} = \cdots 10$ であれば, これは葉ノード s_1 から根へのパスを表すので, 状態は s_1 となる. このモデルにおいて, $S(m_1)$ は $\{s_0(m_1), s_1(m_1), s_2(m_1)\}$ で与えられる. $x^5 = 10010$ が与えられたものとする. 時点 $t = 2$ における状態は $s_2(m_1) = \varphi_{m_1}(1)$ であり, 時点 $t = 3$ における状態は $s_1(m_1) = \varphi_{m_1}(10)$, 時点 $t = 4$ における状態は $s_0(m_1) = \varphi_{m_1}(100)$ と与えられる. FSMXモデル m_1 のパラメータ $\theta^{k_{m_1}}$ は $(\theta_{0,0}^{k_{m_1}}, \theta_{0,1}^{k_{m_1}}, \theta_{0,2}^{k_{m_1}})^T = (P(0|00), P(0|10), P(0|1))^T$ で与えられ, $k^{m_1} = 3$ である.

モデル m の集合を $\mathcal{M}$ で表し, $\mathcal{M}$ は有限集合とする. 第2章で定義した最適モデルを m^*, 最適パラメータ $\theta^{k_{m^*}^*}$ とする. 最適モデル m^* は $\mathcal{M}$ で唯一存在するものとする. $\theta^{k_{m^*}^*}$ から一意に決定される最適モデル m^* の状態集合 $S(m^*)$ の上の定常確率を $q_{j(m^*)}^*$ と記述する.

1つのFSMXモデル m はパラメトリックモデル族を構成するので, これを $\mathcal{H}^{k_m}$ とかく. すなわち, $\mathcal{H}^{k_m} = \{P(\cdot|m, \theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}\}$ である. このとき, 全モデルクラス $\mathcal{H}$ は $\{P(\cdot|m, \theta^{k_m})|m \in \mathcal{M}, \theta^{k_m} \in \Theta^{k_m}\}$ であり,

$$\mathcal{H} = \cup_m \mathcal{H}^{k_m}, \tag{5.2}$$

で定義される. ここで, $m1, m2, \cdots \in \mathcal{M}$ かつ $k_{m1} < k_{m2} < \cdots$ に対し,

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots, \tag{5.3}$$

という入れ子構造が半順序関係で成り立つ.

初期の状態は既知であるものとする. 与えられたデータ系列 x^n に対し, $n_0^{(m)}, \cdots, n_{|S(m)|-1}^{(m)}$ を各状態 $s_0(m), \cdots, s_{|S(m)|-1}(m)$ の生起回数, $n_{0,j}^{(m)}, n_{1,j}^{(m)}, \cdots, n_{\beta,j}^{(m)}$ を状態 $s_j(m)$ のもとでの各シンボル $0, 1, \cdots, \beta$ の生起回数とする. すなわち, $n = \sum_j n_j^{(m)}$ かつ $n_j^{(m)} = \sum_i n_{i,j}^{(m)}$ である. このとき, 尤度関数 $P(x^n|m, \theta^{k_m})$ は

$$P(x^n|m, \theta^{k_m}) = \prod_{j=0}^{|S(m)|-1} \prod_{i=0}^{\beta} (\theta_{i,j}^{k_m})^{n_{i,j}^{(m)}}, \tag{5.4}$$

ここで, $\theta_{\beta,j}^{k_m} = 1 - \sum_{i=0}^{\beta-1} \theta_{i,j}^{k_m}$ であり, この尤度関数は唯一の最大値をもつ.

θ^{k_m} の最尤推定量 $\hat{\theta}^{k_m}$ の各要素は

$$\hat{\theta}_{i,j}^{k_m} = \frac{n_{i,j}^{(m)}}{n_j^{(m)}}, \tag{5.5}$$

で与えられる. ここで, 情報行列 $J^*(\theta^{k_m}|m)$

$$J^*(\theta^{k_m}|m) = -\lim_{n\to\infty} \frac{1}{n} E^* \left[\frac{\partial^2 \log P(X^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} \right], \tag{5.6}$$

を定義する. ただし, E^* は真の分布 $P^*(\cdot)$ による期待値を表す. このとき, FSMX モデルの $J^*(\theta^{k_m}|m)$ は

$$J^*(\theta^{k_m}|m) = -\frac{\partial^2 \sum_{i,j} q^*_{s_j(m)} \theta_{i,j}^{k_m^*} \log \theta_{i,j}^{k_m}}{\partial \theta^{k_m} (\partial \theta^{k_m})^T}, \tag{5.7}$$

と与えられる. ただし, 最適パラメータ $\theta^{k_m^*}$ は

$$\theta^{k_m^*} = \arg \min_{\theta^{k_m} \in \Theta^{k_m}} \lim_{n\to\infty} E^* \left[-\frac{1}{n} \log P(X^n|m, \theta^{k_m}) \right], \tag{5.8}$$

で定義するが, FSMX モデルにおいては, $N_{i,j}^{(m)}$ を X^n 内の状態 $s_j(m)$ のもとでの各シンボル $0, 1, \cdots, \beta$ の生起回数を表す確率変数, $N_j^{(m)}$ を状態 j の生起回数を表す確率変数として,

$$\theta_{i,j}^{k_m^*} = \lim_{n\to\infty} E^* \left[\frac{N_{i,j}^{(m)}}{N_j^{(m)}} \right], \tag{5.9}$$

で与えられる.

もし $P^*(\cdot) = P(\cdot|m^*, \theta^{k_{m^*}^*})$ とすれば, $J^*(\theta^{k_{m^*}^*}|m^*)$ はフィッシャー情報行列となり, $\det J^*(\theta^{k_{m^*}^*}|m^*)$ は

$$\det J^*(\theta^{k_{m^*}^*}|m^*) = \prod_{j=0}^{|S(m^*)|-1} (q^*_{s_j(m^*)})^{\beta} \frac{1}{\theta_{0,j}^{k_{m^*}^*} \theta_{1,j}^{k_{m^*}^*} \cdots \theta_{\beta,j}^{k_{m^*}^*}}, \tag{5.10}$$

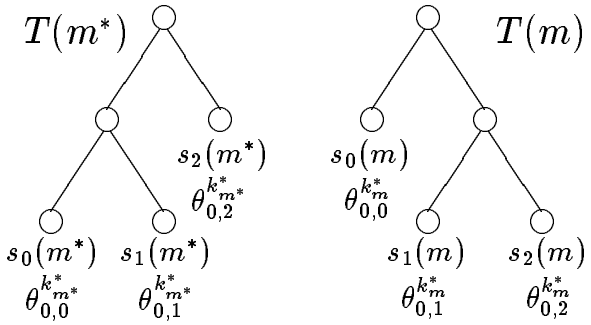

図 5.2: $\theta^{k^*_m}$ の例

と与えられる.

ここで, $m \neq m^*$ を満たすモデル m のパラメータ θ^{k_m} を x^n から推定した場合について考える. FSMX モデルにおいては, $m \neq m^*$ の最適パラメータ $\theta^{k^*_m}_{i,j}$ は次のように与えられることが容易にわかる. すなわち, $T(m)$ と $T(m^*)$ の構造を比較し, もし $T(m)$ の葉 $s_j(m)$ が $T(m^*)$ の葉 $s_l(m^*)$ と対応する, あるいは $T(m)$ の中間接点ノードが $T(m^*)$ の葉 $s_l(m^*)$ と対応する場合には,

$$\theta^{k^*_m}_{i,j} = \theta^{k^*_{m^*}}_{i,l}, \tag{5.11}$$

が成り立つ. 一方, $T(m)$ の葉 $s_j(m)$ が $T(m^*)$ の中間接点ノードに対応する場合には,

$$\theta^{k^*_m}_{i,j} = \sum_{s_l(m^*) \in \mathcal{S}(j,m,m^*)} \theta^{k^*_{m^*}}_{i,l} \frac{q^*_{s_l(m^*)}}{q^*_{\mathcal{S}(j,m,m^*)}}, \tag{5.12}$$

となる. ここで, $\mathcal{S}(i,m,m^*)$ は $T(m^*)$ の $s_j(m)$ に対応するノードの全ての子孫の葉ノードを表し, $q^*_{\mathcal{S}(j,m,m^*)}$ は

$$q^*_{\mathcal{S}(j,m,m^*)} = \sum_{s_l(m^*) \in \mathcal{S}(j,m,m^*)} q^*_{s_l(m^*)}, \tag{5.13}$$

で与えられる. $m \neq m^*$ に対する $\theta^{k^*_m}$ の例を図5.2に示す. 上述の関係から, $\theta^{k^*_m}_{0,0} = \frac{q^*_{s_0(m^*)}}{q^*_{s_0(m^*)}+q^*_{s_1(m^*)}} \theta^{k^*_{m^*}}_{0,0} + \frac{q^*_{s_1(m^*)}}{q^*_{s_0(m^*)}+q^*_{s_1(m^*)}} \theta^{k^*_{m^*}}_{0,1}$ かつ $\theta^{k^*_m}_{0,1} = \theta^{k^*_m}_{0,2} = \theta^{k^*_{m^*}}_{0,2}$ である. 多項分布族と同様の性質から, m を固定したとき, $\theta^{k^*_m}$ は x^n を生成する真のパラメータとみなして扱うことができる.

FSMX モデルは有限エルゴードマルコフモデルの一つであり, 次の性質が知られている.

補題 5.1 (重複対数の法則 [34]) $\forall m \in \mathcal{M}$ に対し，

$$\hat{\theta}^{k_m} = \theta^{k_m^*} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{5.14}$$

$$\frac{n_j^{(m)}}{n} = q^*_{s_j(m)} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{5.15}$$

□

5.2.2 FSMX モデル族に対するベイズ符号

前節までと同様に，$P(m)$ と $f(\theta^{k_m}|m)$ を事前分布とし，$P(x^n|m)$ と $f(\theta^{k_m}|x^n, m)$ を

$$P(x^n|m) = \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{5.16}$$

$$f(\theta^{k_m}|x^n, m) = \frac{P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m)}{P(x^n|m)}, \tag{5.17}$$

と記述する．本章で扱うベイズ符号は次で定義される．

定義 5.1 (FSMXモデル族に対するベイズ符号) ベイズ符号の符号長 $L_{Bayes}^{m,\theta^{k_m}}(x^n)$ は

$$\begin{aligned} L_{Bayes}^{m,\theta^{k_m}}(x^n) &= -\log \sum_{m\in\mathcal{M}} \int_{\theta^{k_m}\in\Theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m} \\ &= -\log \sum_{m\in\mathcal{M}} P(x^n|m) P(m), \end{aligned} \tag{5.18}$$

で与えられる．ここで，和と積分はそれぞれ $\mathcal{M}$ と Θ^{k_m} に対してとられる． □

5.3 主結果

5.3.1 前提条件

本章では，FSMXモデルと事前分布について，次の条件を仮定する．

条件 5.1

(1) $\forall m \in \mathcal{M}$ に対し $\Theta^{k_m} = (0,1)^{k_m}$，かつ $\theta^{k_{m^*}^*} \in \Theta^{k_{m^*}}$.

(2) $\forall m \in \mathcal{M}$ と $\forall \theta^{k_m} \in \Theta^{k_m}$ に対し，$f(\theta^{k_m}|m) > 0$ かつ $P(m^*) > 0$.

(3) $\forall m \in \mathcal{M}$ に対し，$f(\theta^{k_m}|m)$ は Θ^{k_m} 上で3階微分可能とする．

注意 5.1 条件5.1, (1)について考える．パラメータの定義域は，要素が 0 や 1 をとる境界を含まないが，有限長のデータ列 x^n に対する最尤推定量 $\hat{\theta}^{k_m}$ を考えると，この境界上の点となる場合が存在する．しかし，$\theta^{k^*_{m^*}} \in \Theta^{k_{m^*}}$ という条件のもとで，$n \to \infty$ のとき，$\hat{\theta}^{k_m} \in \Theta^{k_m}$, $a.s.$ であり，その意味で漸近的状況では $\Theta^{k_m} = (0,1)^{k_m}$ として議論して差し支えない．

次に条件5.1, (3)について考える．例えば，共役な事前分布である Dirichlet分布は，この条件を満たす．これは，ベイズ符号化アルゴリズム[81]で使われる事前分布である．

5.3.2 事後確率密度の漸近正規性

本章でも，パラメータの事後確率密度の漸近正規性が本質的な役割を演じる．しかし，前章までの議論では，事後確率最大推定量 $\tilde{\theta}^{k_m}$ のまわりで展開した漸近正規性を用いたが，本章では最尤推定量 $\hat{\theta}^{k_m}$ のまわりで展開した場合の漸近正規性を同時に導く．これにより最尤推定量の漸近正規性を用いて，平均符号長の解析が可能となる．

まず，FSMXモデル族が第2章で示した[16, 22]の条件(c.1),(c.2),(c.3)を満たすことを示し，$\tilde{\theta}^{k_m}$ のまわりの漸近正規性を示す．

補題 5.2 [16, 22] 無限系列 x^∞ を固定する．$\tilde{\theta}^{k_m}$ を $L_n(\theta^{k_m}|m) = \log f(\theta^{k_m}|x^n, m)$ を局所的唯一に最大化する点，すなわち

$$L'_n(\tilde{\theta}^{k_m}|m) = \mathbf{0}, \tag{5.19}$$

かつ

$$\Sigma_n = -\left(L''_n(\tilde{\theta}^{k_m}|m)\right)^{-1}, \tag{5.20}$$

が正定値であるとする．ここで，$L'_n(\tilde{\theta}^{k_m}|m)$ と $L''_n(\tilde{\theta}^{k_m}|m)$ を

$$L'_n(\tilde{\theta}^{k_m}|m) = \left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}}, \tag{5.21}$$

$$L''_n(\tilde{\theta}^{k_m}|m) = \left.\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}(\partial \theta^{k_m})^T}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}}, \tag{5.22}$$

と定義する．また，球 $B_\delta(\tilde{\theta}^{k_m}) = \{\theta^{k_m} \in \Theta^{k_m} | \|\theta^{k_m} - \tilde{\theta}^{k_m}\| < \delta\}$ を定義する．このとき，次の3つの条件のもとで漸近正規性が成り立つ．

(c.1) *"Steepness"* $\lim_{n\to\infty}\bar{\sigma}_n^2 \to 0$, ここで $\bar{\sigma}_n^2$ は Σ_n の最大固有値である.

(c.2) *"Smoothness"* 任意の正定数 $\epsilon > 0$ に対し, ある N と $\delta > 0$ が存在し, $\forall n > N$ と $\forall \theta^{k_m} \in B_\delta(\tilde{\theta}^{k_m})$ に対して, $L_n''(\theta^{k_m})$ は

$$I - A(\epsilon) \leq L_n''(\theta^{k_m}|m)\left\{L_n''(\tilde{\theta}^{k_m}|m)\right\}^{-1} \leq I + A(\epsilon), \tag{5.23}$$

を満たす. ただし, I は $k_m \times k_m$ 単位行列, $A(\epsilon)$ はその最大固有値が $\epsilon \to 0$ のとき 0 に収束する $k_m \times k_m$ 正定値行列とする.

(c.3) *"Concentration"* 任意の $\delta > 0$ に対し, N と $c, d > 0$ が存在し, $\forall n > N$ と $\forall \theta^{k_m} \notin B_\delta(\tilde{\theta}^{k_m})$ に対し,

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) < -c\left\{(\theta^{k_m} - \tilde{\theta}^{k_m})^T\Sigma_n^{-1}(\theta^{k_m} - \tilde{\theta}^{k_m})\right\}^d, \tag{5.24}$$

が成り立つ.

条件(c.1),(c.2),(c.3)のもとで[3]

$$f(\tilde{\theta}^{k_m}|x^n, m)\,(\det\Sigma_n)^{1/2} = (2\pi)^{-k_m/2} + o(1), \tag{5.27}$$

が成り立つ. □

上記の補題の条件 (c.1), (c.2), (c.3) は, x^∞ を固定したときの一つの分布列 $f(\theta^{k_m}|x^n, m), n = 1, 2, \cdots$ に対する条件であるが, x^n は $P^*(x^n)$ に従って生起する確率変数であるので, 以下ではこれらが $P^*(x^n)$ に関して確率1で(ほとんど確実に)成り立つことを示す.

まず $\frac{1}{n}L_n(\theta^{k_m}) = \frac{1}{n}\log f(\theta^{k_m}|x^n, m)$ は

$$\frac{1}{n}L_n(\theta^{k_m}|m) = \sum_{i,j}\frac{n_{i,j}^{(m)}}{n}\log\theta_{i,j}^{k_m} + \frac{1}{n}\log f(\theta^{k_m}|m) - \frac{1}{n}\log P(x^n|m), \tag{5.28}$$

[3]前章で示した (c.3) と本章の (c.3) は異なるが, 両者とも [16] で示されているものである. 前章で示したように, 条件 (c.1),(c.2),(c.3) が成り立てば, $\phi_n^{k_m} = \Sigma_n^{-1/2}(\theta^{k_m} - \tilde{\theta}^{k_m})$ の事後密度 $f_\phi(\phi_n^k|x^n, m)$ は

$$\int_{\phi_n^{k_m}\in R} f_\phi(\phi_n^{k_m}|x^n, m)d\phi_n^{k_m} \to \int_{\phi^{k_m}\in R}(2\pi)^{-k_m/2}\exp\left\{-\frac{1}{2}(\phi^{k_m})^T\phi^{k_m}\right\}d\phi^{k_m}, \tag{5.25}$$

も満たす. ただし, R は任意の直方体で, $f_\phi(\phi_n^{k_m}|x^n, m)$ は

$$f_\phi(\phi_n^{k_m}|x^n, m) = (\det\Sigma_n)^{1/2}\,f(\theta^{k_m}|x^n, m), \tag{5.26}$$

で与えられる.

で与えられ，$f(\theta^{k_m}|m)$ は n に依存しないから，$\forall x^n \in \mathcal{X}^n$ に対し $n \to \infty$ のとき

$$\frac{1}{n}L_n(\theta^{k_m}|m) \to \sum_{i,j}\frac{n_{i,j}^{(m)}}{n_j^{(m)}}\frac{n_j^{(m)}}{n}\log\theta_{i,j}^{k_m} - \frac{1}{n}\log P(x^n|m), \tag{5.29}$$

となる．ここで，$\log P(x^n|m)$ は θ^{k_m} に依存しないので，$\frac{1}{n}L_n(\theta^{k_m}|m)$ を最大化する $\tilde{\theta}^{k_m}$ は $\hat{\theta}^{k_m}$ に収束する．

条件5.1, (3) より，

$$\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = \frac{\partial^2\sum_{i,j}n_{i,j}^{(m)}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} + \frac{\partial^2\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \tag{5.30}$$

は θ^{k_m} に関して微分可能である．補題より $\frac{n_{i,j}^{(m)}}{n} \to q^*_{s_j(m)}\theta_{i,j}^{k_m^*}$, $a.s.$ が成り立つから，

$$\begin{aligned}\frac{1}{n}\frac{\partial^2\sum_{i,j}n_{i,j}^{(m)}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} &= \frac{\partial^2\sum_{i,j}\frac{n_{i,j}^{(m)}}{n}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} \\ &\to \frac{\partial^2\sum_{i,j}q^*_{s_j(m)}\theta_{i,j}^{k_m^*}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \quad a.s.\end{aligned} \tag{5.31}$$

が得られる．したがって，条件5.1, (3) より，$\frac{\partial^2\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$ は微分可能で n に依存しないので，(5.7)式より $-\frac{1}{n}\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$ は，$0 < \delta_0 < \inf_{i,j}\theta_{i,j}^{k_m}$ となる θ^{k_m} に対して一様に $J^*(\theta^{k_m}|m)$ に概収束する．ここで正定数 $\delta_0 > 0$ は任意である．よって，

$$L_n''(\theta^{k_m}|m)\left\{L_n''(\tilde{\theta}^{k_m}|m)\right\}^{-1} \to J^*(\theta^{k_m}|m)\left\{J^*(\tilde{\theta}^{k_m}|m)\right\}^{-1}, \quad a.s. \tag{5.32}$$

も $0 < \delta_0 < \inf_{i,j}\theta_{i,j}^{k_m}$ となる θ^{k_m} に対して一様に成り立つ．一方，$J^*(\theta^{k_m}|m)$ の連続性と $\tilde{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ から，$\exists\delta > 0$ が存在して，$\forall\theta^{k_m} \in B_\delta(\tilde{\theta}^{k_m})$ に対し，$n \to \infty$ のとき

$$I - A(\epsilon) \le J^*(\theta^{k_m}|m)\left\{J^*(\tilde{\theta}^{k_m}|m)\right\}^{-1} \le I + A(\epsilon), \ a.s. \tag{5.33}$$

が成り立つ．

よって，$\tilde{\theta}^{k_m} \to \hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ から，(c.2)が $n \to \infty$ のとき，ほとんど確実に成り立つことがわかる．

さらに $\lim_{n\to\infty} n_{i,j}^{(m)} \to \infty$, $a.s.$ より，$\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$ の最大固有値はほとんど確実に ∞ に発散する．これは，(c.1)がほとんど確実に成り立つことを意味する．

一方，$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m)$ は

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) = \sum_{i,j}n_{i,j}^{(m)}\log\frac{\theta_{i,j}^{k_m}}{\tilde{\theta}_{i,j}^{k_m}} + \log\frac{f(\theta^{k_m}|m)}{f(\tilde{\theta}^{k_m}|m)}, \tag{5.34}$$

で与えられる．ここで，$\tilde{\theta}^{k_m} = (\tilde{\theta}^{k_m}_{0,0}, \tilde{\theta}_{1,0}, \cdots, \tilde{\theta}^{k_m}_{\beta-1,|S(m)|-1})^T$ である．これをテーラー展開すれば

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) = (\theta^{k_m} - \tilde{\theta}^{k_m})^T \left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}, \tag{5.35}$$

ただし，θ^{k_m+} は θ^{k_m} と $\tilde{\theta}^{k_m}$ で結ばれる線分上の点である．ここで，$\frac{1}{n}\left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}$ は

$$\begin{aligned}\frac{1}{n}\left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}} &= \frac{1}{n}\frac{\partial \sum_{i,j} n^{(m)}_{i,j} \log \theta^{k_m}_{i,j}}{\partial \theta^{k_m}} + \frac{1}{n}\left.\frac{\partial \log f(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}} \\ &\to \left.\frac{\partial \sum_{i,j} q^*_{s_j(m)}\theta^{k^*_m}_{i,j} \log \theta^{k_m}_{i,j}}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}, \quad a.s. \end{aligned} \tag{5.36}$$

が成り立つ．これは，$\frac{n^{(m)}_{i,j}}{n} \to q^*_{s_j(m)}\theta^{k^*_m}_{i,j}$, $a.s.$ であることと，$f(\theta^{k_m}|m)$ が3階微分可能であることから明らかである．(5.36)式の最後の項は $\theta^{k_m+} = \theta^{k^*_m}$ のとき 0 に収束する．一方，$n \to \infty$ のとき，$\theta^{k^*_m} \in B_\delta(\tilde{\theta}^{k_m})$, $a.s.$ であるから，$\exists c_\delta > 0$ が存在して，$\forall \theta^{k_m} \notin B_\delta(\tilde{\theta}^{k_m})$ に対して $n \to \infty$ のとき，$c_\delta < \left\| \frac{1}{n}\left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}} \right\|$, $a.s.$ が成り立つ．以上より，$\forall \theta^{k_m} \notin B_\delta(\tilde{\theta}^{k_m})$ に対して $n \to \infty$ のとき[4]

$$\begin{aligned} L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) &\le -\|\theta^{k_m} - \tilde{\theta}^{k_m}\| \left\| L'_n(\theta^{k_m+}|m) \right\| \\ &< -nc_\delta \|\theta^{k_m} - \tilde{\theta}^{k_m}\|, \quad a.s. \end{aligned} \tag{5.37}$$

ここで $\frac{1}{n}\Sigma_n^{-1} \to J^*(\tilde{\theta}^{k_m}|m)$, $a.s.$ かつ $\tilde{\theta}^{k_m} \to \hat{\theta}^{k_m} \to \theta^{k^*_m}$, $a.s.$ であるから，$\exists c_\psi$ が存在して，$n \to \infty$ のとき $\bar{\psi} < c_\psi$, $a.s.$ が成り立つ．ただし，$\bar{\psi}$ は $\frac{1}{n}\Sigma_n^{-1}$ の最大固有値である．したがって，$\exists c_{\delta,\psi}$ が存在して $n \to \infty$ のとき

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) < -c_{\delta,\psi}\left\{(\theta^{k_m} - \tilde{\theta}^{k_m})^T \Sigma_n^{-1} (\theta^{k_m} - \tilde{\theta}^{k_m})\right\}^{1/2}, \quad a.s. \tag{5.38}$$

が成り立ち，これは(c.3)が概収束の意味で成り立つことを示している．

以上より，(c.1), (c.2), (c.3) が概収束の意味で成り立ち，$\lim_{n\to\infty} \frac{1}{n}\Sigma_n^{-1} = J^*(\tilde{\theta}^{k_m}|m)$, a.s. であることから，次の補題を得る．

補題 5.3 条件5.1のもとで，$\forall m \in \mathcal{M}$ に対して，

$$f(\tilde{\theta}^{k_m}|x^n, m) = \left(\frac{n}{2\pi}\right)^{k_m/2} \sqrt{\det J^*(\tilde{\theta}^{k_m}|m)} + o(n^{k_m/2}), \quad a.s. \tag{5.39}$$

が成り立つ． □

[4]この部分の同様の議論については [16],pp.293-294 を参照．

[16]での議論，および上の補題における漸近正規性は，事後確率最大推定量 $\tilde{\theta}^{k_m}$ のまわりで展開されている．しかし，よく知られている最尤推定量の漸近的性質とあわせて議論するためには，$\tilde{\theta}^{k_m}$ の代わりに $\hat{\theta}^{k_m}$ のまわりで展開した方が後の解析で有用となる．また，J.Rissanen が最尤符号の漸近符号長を最尤推定量を用いて展開しているように，最尤推定量を用いて展開することが望まれる．最尤推定量を用いた $-\log P(x^n|m,\hat{\theta}^{k_m})$ からの符号長の増加分はリグレット (regret) と呼ばれ，これに関連する性質について現在活発に研究されている [12, 138, 139]．そこで本稿でも符号長を最尤推定量を用いて展開するために，以下では漸近正規性を最尤推定量 $\hat{\theta}^{k_m}$ を用いて導く．

補題 5.4 条件5.1のもとで，$\forall m \in \mathcal{M}$ に対し，

$$f(\hat{\theta}^{k_m}|x^n,m) = \left(\frac{n}{2\pi}\right)^{k_m/2} \sqrt{\det J^*(\hat{\theta}^{k_m}|m)} + o(n^{k_m/2}), \quad a.s. \tag{5.40}$$

(証明) 5.5.1節参照． □

注意 5.2 情報行列 $I^*(\hat{\theta}^{k_m}|m)$ を

$$I^*(\hat{\theta}^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E^* \left[\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial \theta^{k_m}} \frac{\partial \log P(x^n|m,\theta^{k_m})}{(\partial \theta^{k_m})^T} \right]_{\theta^{k_m}=\hat{\theta}^{k_m}}, \tag{5.41}$$

と定義する．真の分布がモデル族に含まれていない場合，$J^*(\hat{\theta}^{k_m}|m) = I^*(\hat{\theta}^{k_m}|m)$ は一般に成り立たない．もし $P^*(\cdot) \in \mathcal{H}^{\theta^{k_m}}$ であれば，$J^*(\theta^{k_m^*}|m) = I^*(\theta^{k_m^*}|m)$ が成り立つ．

一方，最尤推定量についても漸近正規性が成り立つ．すなわち，$\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*})$ の分布は，$N(0, \{K^*(\theta^{k_m^*}|m)\}^{-1})$ に法則収束する．ただし，$\{K^*(\theta^{k_m}|m)\}^{-1}$ は

$$\{K^*(\theta^{k_m}|m)\}^{-1} = \{J^*(\theta^{k_m}|m)\}^{-1} I^*(\theta^{k_m}|m) \{J^*(\theta^{k_m}|m)\}^{-1}, \tag{5.42}$$

で与えられる情報行列である．しかし，離散分布である多項分布族や FSMX モデル族では，真の分布がモデル族に含まれていなくても，$J^*(\theta^{k_m^*}|m) = I^*(\theta^{k_m^*}|m)$ が成り立つので，$\{K^*(\theta^{k_m^*}|m)\}^{-1} = \{J^*(\theta^{k_m^*}|m)\}^{-1}$ となる．これは，これらの分布が $\theta^{k_m^*}$ を真のパラメータと考えて扱うことができる特殊性のためである．

しかし，一般の確率モデル族においては最尤推定量の漸近共分散は $\{K^*(\theta^{k_m^*}|m)\}^{-1}$ で与えられるが，この場合でもパラメータの事後確率密度の漸近共分散は $\{J^*(\hat{\theta}^{k_m}|m)\}^{-1} \sim \{J^*(\theta^{k_m^*}|m)\}^{-1}$ で与えられ，両者で分散が異なる点に注意すべきである． □

5.3.3 概収束の意味で成り立つ漸近的符号長

まず，以下の定理が成り立つ．

定理 5.1 $\forall x^n \in \mathcal{X}^n$ に対し

$$L_{Bayes}^{m,\theta^{k_m}}(x^n) = -\log P(x^n|m^*) - \log P(m^*) + \log P(m^*|x^n), \tag{5.43}$$

が成り立つ．
(証明) ベイズ規則より明らか． □

したがって，$-\log P(x^n|m^*)$ と $\log P(m^*|x^n)$ が解析できれば，ベイズ符号の符号長が与えられることがわかる．一方，前節で導いた漸近正規性により，個々の系列の符号長に関する次の定理を得る．

定理 5.2 条件5.1のもとで，$\forall m \in \mathcal{M}$ に対して

$$\begin{aligned} -\log P(x^n|m) &= -\log P(x^n|m,\tilde{\theta}^{k_m}) + \frac{k_m}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)} + o(1), \quad a.s. \\ &= -\log P(x^n|m,\hat{\theta}^{k_m}) + \frac{k_m}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)} + o(1), \quad a.s. \end{aligned} \tag{5.44}$$

が成り立つ．
(証明) (5.40)式より

$$\log f(\hat{\theta}^{k_m}|x^n,m) = \frac{k_m}{2}\log\frac{n}{2\pi} + \log\sqrt{\det J^*(\hat{\theta}^{k_m}|m)} + \log\Big(1+o(1)\Big), \quad a.s. \tag{5.45}$$

が成り立つ．これを，ベイズ規則から得られる

$$-\log P(x^n|m) = -\log\frac{P(x^n|m,\hat{\theta}^{k_m})f(\hat{\theta}^{k_m}|m)}{f(\hat{\theta}^{k_m}|x^n,m)}, \tag{5.46}$$

に代入して(5.44)式の右辺第2項を得る．同様に，(5.39)式から(5.44)式の右辺第1項が導かれる． □

定理5.2は，一つの FSMX モデルをパラメトリックモデル族と考え，パラメータで混合をとったベイズ符号の漸近符号長を与えている．

さらに，本章で考えている，モデル m でも混合をとったベイズ符号の符号長を議論するために，以下の補題を準備する．

補題 5.5 条件5.1のもとで，$\forall m \neq m^*, m \in \mathcal{M}$ に対し，

$$\frac{P(x^n|m)}{P(x^n|m^*)} = o^+(1), \quad a.s. \tag{5.47}$$

が成り立つ．ただし，$o^+(1)$ は正で $\lim_{n\to\infty} o^+(1) = +0$, $a.s.$ を満たす項である．
(証明) 5.5.2節参照． □

定理5.1, 定理5.2と補題5.5をあわせて，以下の主定理が得られる．

定理 5.3 条件5.1のもとで，ベイズ符号の漸近符号長は次式で与えられる．

$$\begin{aligned} L_{Bayes}^{m,\theta^{k_m}}(x^n) &= -\log P(x^n|m^*, \tilde{\theta}^{k_{m^*}}) - \log P(m^*) \\ &\quad + \frac{k_{m^*}}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\tilde{\theta}^{k_{m^*}}|m^*)}}{f(\tilde{\theta}^{k_{m^*}}|m^*)} + o(1), \quad a.s. \\ &= -\log P(x^n|m^*, \hat{\theta}^{k_{m^*}}) - \log P(m^*) \\ &\quad + \frac{k_{m^*}}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\hat{\theta}^{k_{m^*}}|m^*)}}{f(\hat{\theta}^{k_{m^*}}|m^*)} + o(1), \quad a.s. \end{aligned} \tag{5.48}$$

(証明) 補題5.5と $\sum_m P(m) = 1$, および $\mathcal{M}$ が有限集合であることより，

$$\begin{aligned} -\log\sum_m \frac{P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)} &= -\log\left\{1 + \sum_{m\neq m^*}\frac{P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}\right\} \\ &= -\log\left\{1 + o^+(1)\right\}, \quad a.s. \end{aligned} \tag{5.49}$$

したがって，

$$L_{Bayes}^{m,\theta^{k_m}}(x^n) = -\log P(x^n|m^*) - \log P(m^*) - o^+(1), \quad a.s. \tag{5.50}$$

定理5.2を適用して証明が完結する． □

5.3.4 平均符号長の漸近的解析

本節では，ベイズ符号の平均符号長の漸近式を与える．そのための準備として，以下の補題を用意する．

補題 5.6 条件5.1のもとで，$\forall m \in \mathcal{M}$ に対し

$$E^*\left[\log\frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m, \theta^{k_m^*})}\right] = \frac{k_m}{2} + o(1), \tag{5.51}$$

ここで，E^* は真の分布 $P^*(\cdot)$ による期待値を表す．
(証明) 5.5.3節参照． □

この補題により，次の定理が得られる．

定理 5.4 条件5.1のもとで，

$$E^*\left[L^{m,\theta^{k_m}}_{Bayes}(x^n)\right] = -E^*\left[\log P(x^n|m^*,\theta^{k^*_{m^*}})\right] - \log P(m^*)$$

$$+\frac{k_{m^*}}{2}\log\frac{n}{2\pi e} + \log\frac{\sqrt{\det J^*(\theta^{k^*_{m^*}}|m^*)}}{f(\theta^{k^*_{m^*}}|m^*)} + o(1), \quad a.s. \tag{5.52}$$

が成り立つ．ここで，$J^*(\theta^{k^*_{m^*}}|m^*)$ は $\theta^{k^*_{m^*}}$ におけるフィッシャー情報行列に一致する．

(証明) (5.48)式は概収束の意味の漸近式であるので，有界収束定理が適用できる．よって，(5.51)式と $J^*(\theta^{k_m}|m)$ と $f(\theta^{k_m}|m)$ の連続性から(5.52)式を得る． □

5.4 考察

本章ではまず，ほとんど確実に生起する系列に対して成り立つベイズ符号の漸近的符号長を解析した．確率的コンプレキシティ(stochastic complexity)の概念に基づけば，ベイズ符号の符号長は得られた系列 x^n の情報量を与えられた確率モデル族によって測っているものと解釈することができ，モデル族の相対的な良さを与えている．この符号長の漸近式は FSMXモデル族という一つの階層モデル族に対して与えているものであり，その点で従来の結果を拡張している．また，FSMX モデルが i.i.d.情報源ではない点も重要である．その式の形は，パラメトリックモデル族に対する最尤符号化の漸近的符号長[101]と類似の項が導かれており，パラメトリックモデルによる符号化において情報行列 $J^*(\theta^{k_m}|m)$ は本質的であるといえる．

本章では，概収束の意味で得られた漸近的符号長をもとに，ベイズ符号の平均符号長の漸近式を導いた．[25]において，Clarke と Barron は i.i.d.パラメトリックモデル族に対して平均符号長を導いているが，本研究では階層構造をもつマルコフモデルの1種である FSMXモデル族について漸近式を示した．FSMXモデルは実用的なベイズ符号化アルゴリズムが与えられている重要なモデル族であり，本研究の成果はその意味で重要と考えられる．

5.5 補題と定理の証明

5.5.1 補題 5.4 の証明

まず $L_n(\theta^{k_m}|m) = \log f(\theta^{k_m}|x^n, m)$ と $B_\delta(\hat{\theta}^k) = \left\{\theta^k \in \Theta^k \middle| \|\theta^k - \hat{\theta}^k\| < \delta\right\}$, および

$$\hat{\Sigma}_n = -\left(L_n''(\hat{\theta}^{k_m}|m)\right)^{-1}, \tag{5.53}$$

を定義しておく.

$\forall\theta^{k_m}$ に対し, テーラー展開より,

$$f(\theta^{k_m}|x^n, m) = f(\hat{\theta}^{k_m}|x^n, m)\exp\left\{(\theta^{k_m} - \hat{\theta}^{k_m})^T L'(\hat{\theta}^{k_m}|m)\right\}$$

$$\cdot \exp\left\{\frac{1}{2}(\theta^{k_m} - \hat{\theta}^{k_m})^T(I + R_n)L''(\hat{\theta}^{k_m}|m)(\theta^{k_m} - \hat{\theta}^{k_m})\right\}, \tag{5.54}$$

ここで, θ^{k_m+} を θ^{k_m} と $\hat{\theta}^{k_m}$ で結ばれる線分上の点とし, R_n は

$$R_n = L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} - I, \tag{5.55}$$

で与えられる. ここで, $\hat{\theta}^{k_m}$ の定義より, $L_n'(\hat{\theta}^{k_m}|m)$ は

$$L_n'(\hat{\theta}^{k_m}|m) = \left.\frac{\partial \log f(\theta^{k_m}|m)}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}, \tag{5.56}$$

で与えられる. 一方,

$$-\frac{1}{n}L_n''(\theta^{k_m}|m) \to J^*(\theta^{k_m}|m),\ a.s. \tag{5.57}$$

より,

$$L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \to J^*(\theta^{k_m+}|m)\{J^*(\hat{\theta}^{k_m}|m)\}^{-1},\ a.s. \tag{5.58}$$

が成り立つ. $J^*(\theta^{k_m}|m)$ の連続性と $\hat{\theta}^{k_m} \to \theta^{k_m^*}, a.s.$ より, $\exists\delta > 0$ が存在して, $\forall\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$ と $\forall\epsilon > 0$ に対し, $n \to \infty$ のとき

$$I - A(\epsilon) \le L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \le I + A(\epsilon),\ \ a.s. \tag{5.59}$$

が成り立つ. よって $n \to \infty$ のとき

$$P_n(\delta) = \int_{\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})} f(\theta^{k_m}|x^n, m)d\theta^{k_m}, \tag{5.60}$$

はほとんど確実に

$$f(\hat{\theta}^{k_m}|x^n,m)\exp\{\bar{c}_f\delta\}|\det\hat{\Sigma}_n|^{1/2}|\det(I-A(\epsilon))|^{-1/2}\int_{\|z^{k_m}\|<s_n}\exp\{-\frac{1}{2}(z^{k_m})^Tz^{k_m}\}dz^{k_m}, \tag{5.61}$$

で上界され，

$$f(\hat{\theta}^{k_m}|x^n,m)\exp\{-\bar{c}_f\delta\}|\det\hat{\Sigma}_n|^{1/2}|\det(I+A(\epsilon))|^{-1/2}\int_{\|z^{k_m}\|<t_n}\exp\{-\frac{1}{2}(z^{k_m})^Tz^{k_m}\}dz^{k_m}, \tag{5.62}$$

で下界される．ただし，z^{k_m} は k_m次元のベクトルである．ここで，$\bar{c}_f$ は $\frac{\partial\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}}$ の最大絶対値であり，s_n と t_n は，$\overline{a(\epsilon)}$ ($\underline{a(\epsilon)}$) と $\overline{l_n}$ ($\underline{l_n}$) をそれぞれ $A(\epsilon)$ と $\hat{\Sigma}_n$ の最大(最小)固有値として，$s_n=\delta(1-\overline{a(\epsilon)})^{1/2}/\underline{l_n}^{1/2}$，$t_n=\delta(1-\underline{a(\epsilon)})^{1/2}/\overline{l_n}^{1/2}$ で与えられる．

$$\frac{1}{n}(\hat{\Sigma}_n)^{-1}\to J^*(\theta^{k_m^*}|m),\quad a.s. \tag{5.63}$$

より，$\overline{l_n}\to 0,\ a.s.$ かつ $\underline{l_n}\to 0,\ a.s.$ が成り立ち，$s_n\to\infty,\ a.s.$ かつ $t_n\to\infty,\ a.s.$ が導かれる．よって $n\to\infty$ のとき，ほとんど確実に

$$\begin{aligned}|I-A(\epsilon)|^{1/2}\exp\{-\bar{c}_f\delta\}\lim_{n\to\infty}P_n(\delta)&\le\lim_{n\to\infty}f_\xi(\hat{\xi}^{k_m}|x^n,m)\left\{\det\left(\frac{1}{n}L''(\hat{\theta}^{k_m}|m)\right)\right\}^{-1/2}(2\pi)^{k_m/2}\\&\le|I+A(\epsilon)|^{1/2}\exp\{\bar{c}_f\delta\}\lim_{n\to\infty}P_n(\delta),\end{aligned} \tag{5.64}$$

が成り立つ．

よってもし $\forall\epsilon>0$ に対し，$\lim_{n\to\infty}P_n(\delta)=1,\ a.s.$ が成り立てば，

$$\lim_{n\to\infty}f_\xi(\hat{\xi}^{k_m}|x^n,m)\to\frac{\{\det J^*(\hat{\theta}^{k_m}|m)\}^{1/2}}{(2\pi)^{k_m/2}},\quad a.s. \tag{5.65}$$

が成り立つことがわかる．よって最後に $\forall\epsilon>0$ に対して $\lim_{n\to\infty}P_n(\delta)=1,\ a.s.$ となることを示そう．

再びテーラー展開より，

$$L_n(\theta^{k_m}|m)-L_n(\hat{\theta}^{k_m}|m)=(\theta^{k_m}-\hat{\theta}^{k_m})^T\left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}. \tag{5.66}$$

ただし，θ^{k_m+} は θ^{k_m} と $\hat{\theta}^{k_m}$ で結ばれる線分上の点である．(5.36)式~(5.37)式と同様の議論により，ある定数 $c_\delta>0$ が存在して，$\theta^{k_m}\in B_\delta(\hat{\theta}^{k_m})$ に対し，$n\to\infty$ のとき $c_\delta<\left\|\frac{1}{n}L_n'(\theta^{k_m+}|m)\right\|,\ a.s.$ が成り立つ．したがって，$\exists c_{\delta,\phi}>0$ と $\theta^{k_m}\in B_\delta(\hat{\theta}^{k_m})$ に対し

$$L_n(\theta^{k_m}|m)-L_n(\hat{\theta}^{k_m}|m)=\left(\theta^{k_m}-\hat{\theta}^{k_m}\right)^TL_n'(\theta^{k_m+}|m)$$

$$\leq -nc_{\delta} \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|$$

$$\leq -c_{\delta,\phi} \left\{ (\theta^{k_m} - \hat{\theta}^{k_m})^T L''(\hat{\theta}^{k_m}|m)(\theta^{k_m} - \hat{\theta}^{k_m}) \right\}^{1/2}, \quad a.s. \tag{5.67}$$

が成り立つ．この不等式より，

$$\int_{\theta^{k_m} \in \Theta^{k_m} - B_{\delta}(\hat{\theta}^k)} f(\theta^{k_m}|x^n, m) d\theta^{k_m}$$

$$\leq f(\hat{\theta}^{k_m}|x^n, m) \left(\det L''(\hat{\theta}^{k_m}|m) \right)^{-1/2} \int_{||z^{k_m}|| > \delta/\overline{l_n}^{1/2}} \exp\{ -c_{\delta,\phi} \left((z^{k_m})^T z^{k_m} \right)^{1/2} \} dz \tag{5.68}$$

(5.64)式と $\overline{l_n} \to 0$, $a.s.$ であることから，

$$\int_{\theta^{k_m} \in \Theta^{k_m} - B_{\delta}(\hat{\theta}^k)} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 0, \quad a.s. \tag{5.69}$$

したがって，$\lim_{n\to\infty} P_n(\delta) = 1$, $a.s.$ が成り立ち，(5.65)式が示された． □

5.5.2 補題 5.5 の証明

補題5.4より $\forall m \in \mathcal{M}$ に対し

$$\log \frac{P(x^n|m)}{P(x^n|m^*)} = \log \frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \left(\frac{k_m}{2} - \frac{k_{m^*}}{2} \right) \log \frac{n}{2\pi}$$

$$- \log \frac{f(\hat{\theta}^{k_{m^*}}|m^*) \sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m) \sqrt{\det J^*(\hat{\theta}^{k_{m^*}}|m^*)}} + o(1), \quad a.s. \tag{5.70}$$

が成り立つ．

$f(\theta^{k_m}|m)$ と $J^*(\theta^{k_m}|m)$ が連続微分可能で，$\hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ であるから，$f(\hat{\theta}^{k_m}|m) = O(1)$, $a.s.$ と $\det J^*(\hat{\theta}^{k_m}|m) = O(1)$, $a.s.$が成り立つ．よって，(5.70)式より，もし $\forall m \neq m^*$ に対し

$$\log \frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2} \log n \to -\infty, \quad a.s. \tag{5.71}$$

が示されれば，証明は完結する．

まず，$\log \frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m, \theta^{k_m^*})}$ を評価する．

$$-\log P(x^n|m, \theta^{k_m^*}) = -\sum_{j=0}^{|S(m)|-1} \sum_{i=0}^{\beta} n_{i,j}^{(m)} \log \theta_{i,j}^{k_m^*}, \tag{5.72}$$

と

$$\frac{n_{i,j}^{(m)}}{n} \to q^*_{s_j(m)}\theta_{i,j}^{k_m^*} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{5.73}$$

より，$-\log P(x^n|m,\theta^{k_m^*}) = O(n)$, $a.s.$ が成り立つ．よって，テーラー展開より

$$-\log P(x^n|m,\theta^{k_m^*}) = -\log P(x^n|m,\hat{\theta}^{k_m})$$

$$-\frac{1}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)$$

$$+ O\left(n\left\|\hat{\theta}^{k_m} - \theta^{k_m^*}\right\|^3\right), \quad a.s. \tag{5.74}$$

が得られる．

$$-\frac{1}{n}\left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} = -\frac{1}{n}\left.\frac{\partial^2 \sum_{j=0}^{|S(m)|-1}\sum_{i=0}^{\beta} n_{i,j}^{(m)} \log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}$$

$$\to -\left.\frac{\partial^2 \sum_{j=0}^{|S(m)|-1}\sum_{i=0}^{\beta} \theta_{i,j}^{k_m^*} q^*_{s_j(m)} \log\theta_{i,j}^{k_m}}{\partial\theta^{k_m T}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s.$$

$$= J^*(\hat{\theta}^{k_m}|m) + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{5.75}$$

が成り立ち，また $\hat{\theta}^{k_m}$ も重複対数の法則を満たすことから，

$$-\frac{1}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)$$

$$= \frac{n}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T J^*(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right) + O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \ a.s. \tag{5.76}$$

が成り立つ．

一方，

$$\left\|\hat{\theta}^{k_m} - \theta^{k_m^*}\right\| = O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{5.77}$$

より，

$$n\left\|\hat{\theta}^{k_m} - \theta^{k_m^*}\right\|^3 = O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \tag{5.78}$$

よって，(5.74)式より，

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})} = \frac{n}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T J^*(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)$$

$$+ O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \tag{5.79}$$

また，$\sqrt{n}\|\hat{\theta}^{k_m} - \theta^{k_m^*}\| = O((\log\log n)^{1/2})$, $a.s.$ と $J^*(\hat{\theta}^{k_m}|m) \to J^*(\theta^{k_m^*}|m)$, $a.s.$ より，

$$\frac{n}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T J^*(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right) = O(\log\log n), \quad a.s. \tag{5.80}$$

したがって，

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n$$

$$= \log\frac{P(x^n|m,\theta^{k_m^*})}{P(x^n|m^*,\theta^{k_{m^*}^*})} - \frac{k_m - k_{m^*}}{2}\log n + o(\log n), \quad a.s. \tag{5.81}$$

が得られる．

まず，$P^*(x^n) \notin \mathcal{H}^{k_m}$ である場合を考える．このとき，(5.12)式と(5.73)式より

$$\forall\theta^{k_m} \in \Theta^{k_m}, \ \log\frac{P(x^n|m,\theta^{k_m})}{P(x^n|m,\theta^{k_{m^*}^*})} = -Cn + o(n), \tag{5.82}$$

ただし，C はある正定数である．

したがって，$\forall m \in \mathcal{M}$, かつ $P^*(x^n) \notin \mathcal{H}^{k_m}$ に対して

$$\begin{aligned}\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n &= -Cn - \frac{k_m - k_{m^*}}{2}\log n + o(n) \\ &\to -\infty, \quad a.s.\end{aligned} \tag{5.83}$$

が成り立つ．

一方 $P^*(x^n) \in \mathcal{H}^{k_m}$ のとき，定義より

$$P(x^n|m,\theta^{k_m^*}) = P(x^n|m^*,\theta^{k_{m^*}^*}), \tag{5.84}$$

が成り立つから，$\forall m \in \mathcal{M}$, $P^*(x^n) \in \mathcal{H}^{k_m}$ に対して

$$\begin{aligned}\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n &= \frac{k_{m^*} - k_m}{2}\log n + o(\log n) \\ &\to -\infty, \quad a.s.\end{aligned} \tag{5.85}$$

が得られる．よって $\forall k_m \neq k_{m^*}$ に対し，(5.71)式が成り立つことが示され，証明が完結した． □

5.5.3 補題 5.6 の証明

重複対数の法則と (5.76) 式より，$J^*(\theta^{k_{m^*}}|m^*)$ が連続であることから，

$$\log \frac{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}{P(x^n|m^*,\theta^{k^*_{m^*}})} = \frac{n}{2}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)^T J^*(\theta^{k^*_{m^*}}|m^*)\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right) + O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \tag{5.86}$$

が成り立つ．

上式の収束が概収束であることから，優収束定理が適用できて，

$$E^*\left[\log \frac{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}{P(x^n|m^*,\theta^{k^*_{m^*}})}\right] = E^*\left[\frac{n}{2}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)^T J^*(\theta^{k^*_{m^*}}|m^*)\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)\right] + O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \tag{5.87}$$

が成り立つ．

$\xi^{k_{m^*}} = \sqrt{n}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)$ の分布は $N(0, \{J^*(\theta^{k^*_{m^*}}|m^*)\}^{-1})$ に法則収束する [34] ことから，

$$E^*\left[\frac{1}{2}\sqrt{n}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)^T J^*(\theta^{k^*_{m^*}}|m^*)\sqrt{n}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)\right] = \frac{k_{m^*}}{2} + o(1), \tag{5.88}$$

が得られ，証明が完結する． □

5.6 本章のまとめ

本章では，FSMXモデル族に対して，ほとんど全ての系列に対して成り立つベイズ符号の符号長の漸近式，及び平均符号長の漸近式を導出した．本章においてもその解析において本質的となるのは，パラメータの事後確率密度の漸近正規性である．この結果，FSMXモデル族に対するベイズ符号の漸近的な評価が精密な形で与えられた．これは，予測の観点からは，累積対数損失を評価関数とした場合のベイズ規準による予測を漸近的評価したことになる．すなわち，本章の結果は，対数損失を用いた場合のオンライン型の逐次予測-学習問題などにそのまま適用可能といえる．

さらにベイズリスクの解析と漸近的にミニマックスリスクを達成する事前分布を示すことは今後の課題である．漸近的にミニマックス符号と等価になる事前分布を与えられれば，事前分布が未知の場合にミニマックス規準という観点から符号化を構成することが可能になる．本章の結果より，Fisher情報行列が発散するパラメータ空間の境界を除けば，モデル m に対しては一様分布，パラメータ θ^{k_m} に対してはジェフリーズ事前分布が，漸近的にミニマックス規準を達成することが予想できる．しかし，これを厳密に示すためには，本章で示した平均符号長の漸近式，すなわちリスク関数の漸近式をパラメータ集合内で一様収束の意味で示す必要がある．このためには漸近正規性の一様性を示す必要があり，これは今後の課題である．また，本章の結果は最適パラメータの要素 $\theta_{i,j}^{k_m^*}$ が 0 や 1 をとるようなFisher情報行列が発散してしまう点を除いた解析であり，このようなパラメータ集合の境界点を含む漸近式を議論することも今後の課題である．

第 6 章

一時点の推定と予測に対するモデル選択と一致性

6.1 本章の目的

統計的モデル選択は，与えられたデータをもとに候補である確率モデルの中から1つのモデルを選択する問題であり，極めて広い応用範囲をもつ重要な課題の一つである．モデル選択の規準としては，AIC[1]とその拡張[108, 111]，BIC[102]とその拡張[92]，MDL[11, 95, 101, 146]，HQ[57]などがそれぞれ異なる視点から導かれている．その他にも様々な視点から研究がなされている[71, 76, 85, 84, 87, 115].

モデル選択のための情報量規準として基本的な AIC は予測の観点から導かれた規準であり，対数損失に対するリスク関数である KL-情報量を考え，その漸近的不偏推定量となっている．これを真の分布がモデルクラスに含まれない場合に拡張した竹内の情報量規準(TIC)[131]も KL-情報量を考えている．村田らは，その損失関数を拡張し，任意の損失関数を用いてモデル選択規準(NIC)を構成した[86]．一方，BICは漸近的に事後確率を最大化するモデルを選択する規準であり，モデル間の 0-1 損失を考えている．これはベイズの事後確率で測った選択誤り率を最小化している[14, 27]．D.S.Poskitt は BIC をパラメータの推定精度を考慮する形に拡張している[92]．また MDL はモデルの良さを記述長で測るという考え方に基づいており[56], ベイズ統計の観点からは BIC と本質的に等価である[101]．さらにベイズ統計に基づいた規準として，予測のためのモデル選択規準(predictive model selection criterion)[71, 76]が提案されている．

このように多くの従来研究がベイズ統計理論に基づいているが，これはモデル選択問題とベイズ統計理論の相性が良いという認識のためと考えられる．しかし，厳密にベイズ決定理論[14, 27, 35, 80]に基づくモデル選択については議論がなされていない．もし事前確率を与えるならば，それに対して最適なモデル

選択を構成することは，ベイズ規準という観点から一貫性があると考えられる．

本章ではまず，ベイズ決定理論に基づくモデル選択を定式化する．その際，モデル選択の目的，すなわち “未来データの予測” と “真のモデルの発見”，にあわせて様々な損失関数を設定することが可能である．さらに，定式化したベイズ最適なモデル選択の，階層モデル族に対する一致性を議論する．多くの従来研究においても一致性の問題は議論されてきた[89]．興味深いことに，AIC は階層モデル族に対して一致性をもたないが[111]，BIC や MDL はもつことが示されている[57, 125]．AIC は予測を目的としているので，階層モデル族に対して一致性がなくてもよいと考えられるが，一方で漸近的状況では予測に対しても BIC や MDL が優れるという議論がある[89]のも確かである．しかし，多くの現実問題で AIC の有効性が示されており[74]，また他の状況では AIC の優れた性質を示すこともできる[108, 111]．そのため当然ながら一致性だけでモデル選択の優劣を議論することはできないが，モデル選択の一つの性質を示すという意味で十分議論されるべき性質であると考えられる．本章では，ベイズ決定理論に基づくモデル選択が，その目的が “未来データの予測” であるか，“真のモデルの発見” であるかに関わらず一致性を持つことを示す．

6.2 準備と仮定

6.2.1 階層モデル族

長さ n のデータ系列 x^n は真の分布 $p^*(x^n) = p^*(X^n = x^n)$ に従い発生するものとする．$p^*(\cdot)$ がモデルクラスに含まれることは仮定しない．本章では階層モデル族に対するモデル選択問題を扱う．階層モデル族は第2章で定義したように，次のようなモデル族である．

$\mathcal{M}$ を離散有限集合とし，その要素 m は一つのモデルを表す．すなわち，$m \in \mathcal{M}$ かつ $|\mathcal{M}|$ は有限である．各モデル m は k_m 次元パラメータ θ^{k_m} をもち，パラメータ集合 Θ^{k_m} は $\mathcal{R}^{k_m}$ の部分集合であるとすると，m は一つのパラメトリックモデル族を構成し，これを $\mathcal{H}^{k_m}$ で表す．

$$\mathcal{H}^{k_m} = \{p(\cdot|m, \theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}\}. \tag{6.1}$$

階層モデル族 $\mathcal{H}$ は $\mathcal{H} = \cup_{m\in\mathcal{M}}\mathcal{H}^{k_m}$ で定義されるが，$m1, m2, \cdots \in \mathcal{M}$ かつ $k_{m1} < k_{m2} < \cdots$ に対して，入れ子構造

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots\cdots, \tag{6.2}$$

が成り立つものとする．この入れ子構造は $\mathcal{M}$ 内で全順序でも，半順序でもよいものとする[1]．

2つの情報行列 $I^*(\theta^{k_m}|m), J^*(\theta^{k_m}|m)$, および $H^*(m,\theta^{k_m})$ を次のように定義する．

$$I^*(\theta^{k_m}|m) = \lim_{n\to\infty}\frac{1}{n}E^*\left[\frac{\partial \log p(X^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\frac{\partial \log p(X^n|m,\theta^{k_m})}{(\partial\theta^{k_m})^T}\right], \tag{6.3}$$

$$J^*(\theta^{k_m}|m) = -\lim_{n\to\infty}\frac{1}{n}E^*\left[\frac{\partial^2 \log p(X^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right], \tag{6.4}$$

$$H^*(m,\theta^{k_m}) = -\lim_{n\to\infty}\frac{1}{n}E^*[\log p(X^n|m,\theta^{k_m})], \tag{6.5}$$

ここで E^* は真の分布 $p^*(\cdot)$ による期待値を表す.

また $D^*(p_{m1}^{\theta^{k_{m1}}} : p_{m2}^{\theta^{k_{m2}}})$ を

$$D^*(p_{m1}^{\theta^{k_{m1}}} : p_{m2}^{\theta^{k_{m2}}}) = \lim_{n\to\infty}\frac{1}{n}E^*\left[\log\frac{p(X^n|m1,\theta^{k_{m1}})}{p(X^n|m2,\theta^{k_{m2}})}\right], \tag{6.6}$$

と定義すると，

$$D^*(p_{m_1}^{\theta^{k_{m_1}}} : p_{m_2}^{\theta^{k_{m_2}}}) = H^*(m_2,\theta^{k_{m_2}}) - H^*(m_1,\theta^{k_{m_1}}), \tag{6.7}$$

が成り立つ．

$\forall m\in\mathcal{M}$ に対し，$\theta^{k_m^*}$ を

$$\theta^{k_m^*} = \arg\min_{\theta^{k_m}} H^*(m,\theta^{k_m}), \tag{6.8}$$

と定義する．$\theta^{k_m^*}$ をモデル m の最適パラメータとよぶ．

6.2.2 一致性

データ系列 x^n は真の分布 $p^*(X^n)$ に従って生起するものとする．本章においても，最適モデルを以下のように定義する．

$$m^* = \arg_m\min\left\{k_m\middle|H^*(m,\theta^{k_m^*}) = H^*\right\}, \tag{6.9}$$

ただし，H^* は

$$H^* = \min_{p(\cdot|m,\theta^{k_m})\in\mathcal{H}} H^*(m,\theta^{k_m}), \tag{6.10}$$

すなわち，$\mathcal{H}^{k_m}\subset\mathcal{H}^{k_{m^*}}$ となる m に対して

$$\forall\theta^{k_m}\in\Theta^{k_m},\ \ H^*(m,\theta^{k_m}) - H^*(m^*,\theta^{k_{m^*}^*}) > 0, \tag{6.11}$$

[1]以下の解析を見れば明らかなように，全く階層構造がない分離型のモデル族，階層構造と分離構造が混在するモデル族に関しても本章の結果はそのまま成り立つ．

が成り立つ．

ここで $p(\cdot|m_1,\theta^{k_{m_1}}) = p(\cdot|m_2,\theta^{k_{m_2}})$ であれば， $H^*(m_1,\theta^{k_{m_1}}) = H^*(m_2,\theta^{k_{m_2}})$ であるが，逆は必ずしも成り立たない．しかし，実際の応用で扱われる現実的な確率モデルにおいては $H^*(m_1,\theta^{k_{m_1}}) = H^*(m_2,\theta^{k_{m_2}})$ は $p(\cdot|m_1,\theta^{k_{m_1}}) = p(\cdot|m_2,\theta^{k_{m_2}})$ を意味する場合がほとんどである．したがって，本章では $H^*(m_1,\theta^{k_{m_1}}) = H^*(m_2,\theta^{k_{m_2}})$ は $p(\cdot|m_1,\theta^{k_{m_1}}) = p(\cdot|m_2,\theta^{k_{m_2}})$ と等価であるとする．よって，$k_{m^*} \leq k_m$ を満たす m に対して

$$p(\cdot|m,\theta^{k_m^*}) = p(\cdot|m^*,\theta^{k_{m^*}^*}), \tag{6.12}$$

が成り立つものとする．

つぎに，モデル選択の弱一致性(weak consistency)と強一致性(strong consistency)を以下のように定義する．

定義 6.1 (弱一致性) $\hat{m}$ を選択されたモデル，すなわち m の推定量とする．以下の性質が成り立つとき，モデル選択は弱一致性をもつという[57]:
$\forall\lambda > 0$ に対し，ある n_λ が存在して， $\forall n > n_\lambda$ に対して

$$P^*\Big\{\hat{m} = m^*\Big\} > 1 - \lambda, \tag{6.13}$$

が成り立つ．これは確率収束を意味し，$\hat{m} \to m^*$, *in probability* と記述する． □

定義 6.2 (強一致性) 以下の性質が成り立つとき，モデル選択は強一致性をもつという[57]:

$$P^*\Big\{\hat{m} \neq m^*, infinitely\ often\Big\} = 0, \tag{6.14}$$

ただし P^* は真の分布 $p^*(\cdot)$ に基づく確率を表す．これは，無限個の部分系列 $x^n \in \{x^1,x^2,\cdots\}$ に対して $\hat{m} \neq m^*$ が成り立つような無限系列 x^∞ の生起する確率は 0 であることを意味する．これは $P^*\{\hat{m} \to m^*\} = 1$ と等価であり，概収束を意味する．この概収束を $\hat{m} \to m^*$, *a.s.* と記述する． □

6.3 ベイズ決定理論に基づくモデル選択

本節では，ベイズ決定理論にもとづくベイズ最適なモデル選択を定式化する．もし事前確率が仮定できるのであれば，ベイズリスクを最小化するモデルがベイズ最適である．我々は，モデル選択の目的に応じて，様々な損失関数を設定することができる．

6.3.1 真のモデルの発見が目的の場合

$P(m)$ を $\mathcal{M}$ 上の事前確率，$f(\theta^{k_m}|m)$ を Θ^{k_m} 上の事前密度とする．モデル m で特徴づけられる x^n の確率分布 $p(x^n|m)$ を

$$p(x^n|m) = \int_{\theta^{k_m} \in \Theta^{k_m}} p(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{6.15}$$

とする．

$L(m : Am)$ を m と Am の損失関数とする．ここで $Am \in \mathcal{M}$ は決定であり，これは m が正しかったときに Am を選択した場合の損失を意味する．このとき，ベイズ最適なモデル選択は

$$BL_{DT}(Am|x^n) = \sum_{m \in \mathcal{M}} L(m : Am) P(m|x^n), \tag{6.16}$$

の Am に関する最小化で与えられる．ここで $P(m|x^n)$ は m の事後確率であり

$$P(m|x^n) = \frac{p(x^n|m)P(m)}{\sum_{m \in \mathcal{M}} p(x^n|m)P(m)}, \tag{6.17}$$

で与えられる．

損失関数としてはまず，0-1損失が考えられる．

$$L_1(m : Am) = \begin{cases} 0 & \text{if} \quad m = Am, \\ 1 & \text{if} \quad m \neq Am. \end{cases} \tag{6.18}$$

この損失関数のもとで，ベイズ決定は事後選択誤り確率 $P(m \neq m^*|x^n)$ を最小化する．ベイズ最適解は事後確率最大のモデルとなり，これは BIC の最適解である．

次にもし過小評価 (underestimation)，すなわち $k_{\hat{m}} < k_{m^*}$ よりも，過大評価 (overestimation)，すなわち $k_{\hat{m}} > k_{m^*}$ を避けたいという要求があったとしよう [106]．同様のアイデアは[84]にもある．ここで $k_{\hat{m}}$ は推定量 $\hat{m}$ の次元であり，k_m の推定量である．このとき次のような損失関数が考えられる [87, 156]．

$$L_2(m : Am) = \begin{cases} 0 & \text{if} \quad m = Am, \\ a_0 & \text{if} \quad m \neq Am, k_m = k_{Am}, \\ a_1 & \text{if} \quad k_m > k_{Am}, \\ a_2 & \text{if} \quad k_m < k_{Am}, \end{cases} \tag{6.19}$$

ただし $0 < a_0 < a_2 < a_1$ である．$0 < a_0 < a_1 < a_2$ とすれば，過小評価(underestimation)の確率を抑えることが可能である．さらにこの議論を一般化し，次の一

般化損失関数を定義しよう.

$$L_3(m:Am) = \begin{cases} C_a & \text{if} \quad m = Am, \\ a(m, Am) & \text{if} \quad m \neq Am, \end{cases} \tag{6.20}$$

ここで，$a(m, Am)$ は $C_a < a(m, Am) < \infty$ をみたす任意の有界関数であるとする. ここで，$C_a = 0$ としても一般性は失われない．よって以下の解析では $C_a = 0$ として取り扱う.

このとき，ベイズ最適なモデル選択は

$$\hat{m}^B_{DT} = \arg \min_{Am \in \mathcal{M}} BL_{DT}(Am|x^n), \tag{6.21}$$

で与えられる.

6.3.2 予測を目的とした場合

過去のデータ系列 x^n から次のデータ $x = x_{n+1}$ を予測することを目的としたモデル選択を考える．この場合第2章で議論したように，$x = x_{n+1}$ は $p^*(x|x^n)$ に従って生起するので，この確率を精度良く推定することが目的となる．ここで，$p^*(x|x^n)$ は真の確率分布における x^n のもとでの x の条件付き分布であり，もし i.i.d. モデルならば $p^*(x|x^n) = p^*(x)$ である.

$L(p(x|x^n, m, \theta^{k_m}) : Ap(x|x^n))$ を，x の予測を与える決定 $Ap(x|x^n)$ と $p(x|x^n, m, \theta^{k_m})$ の間の損失関数とする．x^n で条件付けられたリスク関数 $R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n))$ は，

$$R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) = \int_{x \in \mathcal{X}} L(p(x|x^n, m, \theta^{k_m}) : Ap(x|x^n)) dp(x|x^n, m, \theta^{k_m}), \tag{6.22}$$

で与えられる．本来ならこのリスクを最小化したいところであるが，m, θ^{k_m} は未知であるので，これらの事後分布 $P(m|x^n)$, $f(\theta^{k_m}|x^n, m)$ で平均化した条件付きベイズリスクを考える．x^n による条件付きベイズリスクは

$$BR_{PD}(Ap|x^n) = \sum_{m \in \mathcal{M}} \int_{\theta^{k_m} \in \Theta^{k_m}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|x^n, m) P(m|x^n) d\theta^{k_m}, \tag{6.23}$$

で与えられ，$\forall n \in \{0, 1, 2, \cdots\}$ と $\forall x^n \in \mathcal{X}^n$ に対して， $BR_{PD}(Ap|x^n) < \infty$ を満たす $Ap(x|x^n)$ が存在するものとする．もし $\forall Ap(x|x^n)$ に対して $BR_{PD}(Ap|x^n) = \infty$ であるならば，ベイズ最適解は存在しない.

もし損失関数として対数損失

$$L_4(p(x|x^n, m, \theta^{k_m}) : Ap(x|x^n)) = -\log Ap(x|x^n) + \log p(x|x^n, m, \theta^{k_m}), \tag{6.24}$$

をとればリスク関数は KL-情報量となり，AIC と同じリスクを考えていることになる．他にも二乗誤差損失

$$L_5(p(x|x^n,m,\theta^{k_m}):Ap(x|x^n))=(Ap(x|x^n)-p(x|x^n,m,\theta^{k_m}))^2, \tag{6.25}$$

をとることもできる．

損失関数 $L_4(\cdot:\cdot)$ や $L_5(\cdot:\cdot)$ に対するリスク関数 $R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))$ は，適当な正則条件を満たす現実的な確率モデル族と $R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))<\infty$ を満たす決定 $Ap(X|x^n)$, $x^n\in\mathcal{X}^n$ に対し一様連続である[27, 35]．

ベイズ最適解は(6.23)式の最小化で与えられる．しかし，もし決定空間に制限を設けない場合，ベイズ最適解は $Ap(x|x^n)=p_{pred}(x|x^n)$ で与えられる[14]．ただし，

$$p_{pred}(x|x^n)=\sum_{m\in\mathcal{M}}\int_{\theta^{k_m}\in\Theta^{k_m}}p(x|x^n,m,\theta^{k_m})f(\theta^{k_m}|x^n,m)P(m|x^n)d\theta^{k_m}, \tag{6.26}$$

である．しかし，本章ではモデル選択問題を扱っており，m の推定量を与えることを考えている．よって，ベイズリスクは m に関して最小化するという方針で定式化を進める．そのための方法は一意ではないが，本章では以下の2つの方法を提案する．

1つ目の方法は決定関数 $Ap(x|x^n)$ の定義域，すなわち決定空間を以下のように与える方法である．

$$Ap(x|x^n)\in\{p^m_{pred}(x|x^n)|m\in\mathcal{M},x^n\in\mathcal{X}^n\}, \tag{6.27}$$

ここで $p^m_{pred}(x)$ はパラメータ θ^{k_m} に対する混合分布であり，

$$p^m_{pred}(x|x^n)=\int_{\theta^{k_m}\in\Theta^{k_m}}p(x|x^n,m,\theta^{k_m})f(\theta^{k_m}|x^n,m)d\theta^{k_m}, \tag{6.28}$$

で与えられる．このとき，ベイズ最適解は

$$\hat{p}^B_{PD}(x|x^n)=\arg\min_{Ap(x|x^n)\in\{p^m_{pred}(x|x^n)\}}BR_{PD}(Ap|x^n), \tag{6.29}$$

$$\hat{m}^B_{PD}=\arg\min_{m|Ap(x|x^n)\in\{p^m_{pred}(x|x^n)\}}BR_{PD}(Ap|x^n). \tag{6.30}$$

で与えられる．

2つ目の方法は，パラメータに関しては最尤推定量を用いる方法である．これは，多くの情報量規準が最尤推定量に基づいて導出されていることを受けた考え方である．すなわち，決定空間を

$$Ap(x|x^n)\in\{p(x|x^n,m,\hat{\theta}^{k_m})|m\in\mathcal{M},x^n\in\mathcal{X}^n\}, \tag{6.31}$$

とする．ここで $\hat{\theta}^{k_m} = \hat{\theta}^{k_m}(x^n)$ は x^n により計算される最尤推定量を表す．このとき，ベイズ最適な推定は

$$\hat{p}^B_{ML}(x|x^n) = \arg \min_{Ap(x|x^n) \in \{p(x|x^n, m, \hat{\theta}^{k_m})\}} BR_{PD}(Ap|x^n), \tag{6.32}$$

$$\hat{m}^B_{ML} = \arg \min_{m | Ap(x|x^n) \in \{p(x|x^n, m, \hat{\theta}^{k_m})\}} BR_{PD}(Ap|x^n), \tag{6.33}$$

で与えられる．

注意 6.1 もしベイズリスクを m と θ^{k_m} の両方で最適化することを考えると，最適モデルは一意に決まらなくなる可能性がある．さらに(6.26)式の混合分布がモデル族内に存在しない場合には，常に最大次数のモデルが最適になるなど，モデル選択自体が意味をなさなくなる．またこの場合には，選ばれた θ^{k_m} の意味も不明確となってしまう． □

6.4 一致性に関する解析

6.4.1 モデル族に対する条件

まず，事前分布に対して次を仮定する．

条件 6.1 (事前分布の条件)

(1) $\forall m \in \mathcal{M}$ に対して $P(m) > 0$ である．

(2) ($f(\theta^{k_m}|m)$ の有界連続性) $\forall \theta^{k_m} \in \Theta^{k_m}$ と $\forall m \in \mathcal{M}$ に対して，$f(\theta^{k_m}|m) > 0$. さらに，$\forall m \in \mathcal{M}$ に対し，$f(\theta^{k_m}|m)$ の θ^{k_m} に関する導関数は

$$\sup_{\theta^{k_m} \in \Theta^{k_m}} \left\| \frac{\partial f(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < c_f, \tag{6.34}$$

を満たす．ただし c_f はある正定数である．

(3) ($f(\theta^{k_m}|m)$ の健全性) $\forall m \in \mathcal{M}$ に対し，事前密度 $f(\theta^{k_m}|m)$ は

$$\sup_{x \in \mathcal{X}, n \in \{1,2,\cdots\}, x^n \in \mathcal{X}^n} \int_{\theta^{k_m} \in \Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m} < c_{int}, \tag{6.35}$$

を満たす．ただし，c_{int} はある正定数である．さらに，$\forall m \in \mathcal{M}, \forall Ap(X|x^n) \in \{p^m_{pred}(x|x^n)\}$，$\forall Ap(X|x^n) \in \{p(x|x^n, m, \hat{\theta}^{k_m})\}$ に対し，事前密度 $f(\theta^{k_m}|m)$

は

$$\sup_{x\in\mathcal{X},n\in\{1,2,\cdots\},x^n\in\mathcal{X}^n}\int_{\theta^{k_m}\in\Theta^{k_m}} R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))f(\theta^{k_m}|m)d\theta^{k_m} < c'_{int}, \tag{6.36}$$

を満たす．ただし，c'_{int} はある正定数である． □

次にモデル族に関する条件を示す．ここでは，弱一致性のために条件6.2を，強一致性のために条件6.3を仮定する．当然ながら，強一致性を示すための条件6.3の方が強い条件となっている．これまでと同様に，$\forall\delta>0$ に対して球 $B_\delta(\theta^{k^*_m}) = \left\{\theta^{k_m}\in\Theta^{k_m}\middle|\|\theta^{k_m}-\theta^{k^*_m}\|<\delta\right\}$ を定義しておく．

条件 6.2 モデル族の条件 I

(1) ($\theta^{k^*_m}$ の存在性) $\forall m\in\mathcal{M}$ に対し，最適パラメータ $\theta^{k^*_m}$ は Θ^{k_m} の内部に唯一存在する．

(2) ($\hat{\theta}^{k_m}$ の存在性) $\forall x^n\in\mathcal{X}^n$，$\forall m\in\mathcal{M}$ に対し，尤度関数 $p(x^n|m,\theta^{k_m})$ は Θ^{k_m} 内で単峰形，あるいは単調関数とする．さらに，$\forall x^n\in\mathcal{X}^n$，$\forall m\in\mathcal{M}$ に対し，$\hat{\theta}^{k_m}=\hat{\theta}^{k_m}(x^n)$ は Θ^{k_m} で唯一に存在する．

(3) (有界性) $\forall m\in\mathcal{M}$ に対し，

$$\sup_{x\in\mathcal{X},n\in\{1,2,\cdots\},x^n\in\mathcal{X}^n} p(x|x^n,m,\theta^{k_m}) < c_p(m,\theta^{k_m}), \tag{6.37}$$

ただし $c_p(m,\theta^{k_m})$ は m と θ^{k_m} に依存する有限値である．さらに，$\forall\theta^{k_m}\in\Theta^{k_m}$, $\forall m\in\mathcal{M}$ に対し $0<\det I^*(\theta^{k_m}|m)<\infty$ かつ $0<\det J^*(\theta^{k_m}|m)<\infty$ とする．

(4) (モデル族の smoothness) $\forall m\in\mathcal{M}$ に対し，$I^*(\theta^{k_m}|m)$ と $J^*(\theta^{k_m}|m)$ は Θ^{k_m} 上で連続微分可能とする．さらに，$\forall m\in\mathcal{M}$, $\forall Ap(X|x^n)\in\{p^m_{pred}(x|x^n)\}$, $\forall Ap(X|x^n)\in\{p(x|x^n,m,\hat{\theta}^{k_m})\}$ に対し，リスク関数 $R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))$ は Θ^{k_m} 上で一様連続とする[14, 35].

(5) (健全性) $\forall m\in\mathcal{M}$ と $\forall\delta>0$ に対し

$$\int_{\theta^{k_m}\in B_\delta(\theta^{k^*_m})} f(\theta^{k_m}|x^n,m)d\theta^{k_m}\to 1,\quad in\ probability, \tag{6.38}$$

が成り立つ．

(6) ($\hat{\theta}^{k_m}$ の確率収束) $\forall m \in \mathcal{M}$ に対し，最尤推定量 $\hat{\theta}^{k_m} = \hat{\theta}^{k_m}(x^n)$ は

$$\hat{\theta}^{k_m} = \theta^{k_m^*} + O(1/\sqrt{n}), \quad in\ probability, \tag{6.39}$$

を満たす.

(7) (情報行列の一致性) $\forall m \in \mathcal{M}$ に対して $\delta > 0$ が存在し，$\theta^{k_m} \in B_\delta(\theta^{k_m^*})$ に対して一様に

$$-\frac{1}{n}\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} \to J^*(\theta^{k_m}|m), \quad in\ probability. \tag{6.40}$$

$$-\frac{1}{n}\log p(x^n|m,\theta^{k_m}) \to H^*(m,\theta^{k_m}), \quad in\ probability. \tag{6.41}$$

が成り立つ. □

条件 6.3 モデル族の条件 II

(1) ~ (4) 条件6.2, (1) ~ (4) と同じ.

(5) $\forall m \in \mathcal{M}$ と $\forall \delta > 0$ に対し

$$\int_{\theta^{k_m} \in B_\delta(\theta^{k_m^*})} f(\theta^{k_m}|x^n,m)d\theta^{k_m} \to 1, \quad a.s. \tag{6.42}$$

(6) (重複対数の法則 1) $\forall m \in \mathcal{M}$ と $\forall \theta^{k_m} \in \Theta^{k_m}$ に対し，

$$\hat{\theta}^{k_m} = \theta^{k_m^*} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{6.43}$$

(7) (重複対数の法則 2) $\forall m \in \mathcal{M}$ に対して $\delta > 0$ が存在し，$\forall \theta^{k_m} \in B_\delta(\theta^{k_m^*})$ に対して一様に

$$-\frac{1}{n}\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = J^*(\theta^{k_m}|m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{6.44}$$

$$-\frac{1}{n}\log p(x^n|m,\theta^{k_m}) = H^*(m,\theta^{k_m}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{6.45}$$

が成り立つ. □

注意 6.2 条件6.3は 条件6.2よりも強い仮定になっている．現実的なほとんどのモデル族は条件6.3を満たすので，同時に条件6.2も満たす．実際の問題に適用される例で，条件6.2は満たすが条件6.3をみたさない例はほとんどない．したがっ

て，実際には条件6.3を仮定した議論のみで十分であり，情報量規準HQなどの強一致性を扱った解析では同様の仮定をおいている．しかし，強一致性と弱一致性という結果の差に本質的に効いてくるのは，重複対数の法則であることを示すために，これらの条件を設定する．以下では，条件6.3のもとで強一致性を示し，条件6.2のもとで弱一致性を示すが，これにより重複対数の法則が本質的であることが確かめられる． □

条件6.2のもとでは，(3),(7)から明らかなように， $\forall m \in \mathcal{M}, \forall \theta^{k_m} \in B_\delta(\theta^{k^*_m})$ に対し，$n \to \infty$ のとき

$$\frac{1}{n}\left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial^3 \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right| < c'_p(m,\theta^{k_m}),\ \ in\ probability, \tag{6.46}$$

が成り立つ．ここで，$c'_p(m,\theta^{k_m}) < \infty$ は m と θ^{k_m} に依存する有限値である．一方，条件6.3を仮定すれば，$\forall m \in \mathcal{M}, \forall \theta^{k_m} \in B_\delta(\theta^{k^*_m})$ に対し，$n \to \infty$ のとき

$$\frac{1}{n}\left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial^3 \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right| < c'_p(m,\theta^{k_m}),\ \ a.s. \tag{6.47}$$

が成り立っている．

条件6.2, (5) と条件6.3, (3) は，事後確率密度は最適パラメータの近傍に集中してくることを述べている．同様の仮定は[16, 25, 51, 145]などにある．このための十分条件については[16]で議論されている．

条件6.2, (7)のもとで，m^*, $\theta^{k^*_{m^*}} \in \Theta^{k^*_m}$, $\forall m \in \mathcal{M}$, $\forall \theta^{k_{m^*}} \in \Theta^{k_{m^*}}$, $\forall \theta^{k_m} \in \Theta^{k_m}$ に対し，

$$\frac{1}{n}\log\frac{p(x^n|m^*,\theta^{k^*_{m^*}})}{p(x^n|m,\theta^{k_m})} \to D^*(p_{m^*}^{\theta^{k^*_{m^*}}} : p_m^{\theta^{k_m}}),\ in\ probability. \tag{6.48}$$

が成り立つ．一方，条件6.3, (7), のもとでは

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{p(x^n|m^*,\theta^{k^*_{m^*}})}{p(x^n|m,\theta^{k_m})} = D^*(p_{m^*}^{\theta^{k^*_{m^*}}} : p_m^{\theta^{k_m}}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right),\ \ a.s. \tag{6.49}$$

例 6.1 (正規分布族) i.i.d.の正規分布を考える．モデル m_1 を正規分布 $N(\mu,(\sigma_0)^2)$, ただし σ_0 は固定値であり，μ を自由パラメータとする．モデル m_2 を正規分布 $N(\mu,\sigma^2)$, ただし μ と σ^2 の両方が自由パラメータであるとする．すなわち，$\theta^{k_{m_1}} = \theta^1 = \mu$, $\theta^{k_{m_2}} = \theta^2 = (\mu,\sigma)^T$ である．

このモデル族では，条件6.3, (2), (3), (4) が成り立っていることがわかる[74, 131]．最尤推定量はそれぞれ $\hat\theta^{k_{m1}} = \hat\mu = \frac{1}{n}\sum_{i=1}^n x_i$, $\hat\theta^{k_{m2}} = (\hat\mu,\hat\sigma)^T = (\hat\mu, \sqrt{\frac{1}{n}\sum(x_i-\hat\mu)^2})^T$ と与えられる．正規分布では重複対数の法則[34]が成り立ち，$\hat\mu \to \mu^* + O(\frac{(\log\log n)^{1/2}}{\sqrt{n}})$

$a.s.$ かつ $\hat{\sigma} \to \sigma^* + O(\frac{(\log\log n)^{1/2}}{\sqrt{n}})$ $a.s.$ が得られる．すなわち，条件6.3, (6)は成り立つ．ここで μ^* と σ^* は真の平均値と分散とする．m_1, m_2 に対し $\frac{1}{n}\log p(x^n|m,\hat{\theta}^{k_m})$ は

$$\frac{1}{n}\log p(x^n|m_1,\hat{\theta}^{k_{m_1}}) = -\frac{1}{2}\log 2\pi - \log\sigma_0 - \frac{\hat{\sigma}^2}{2(\sigma_0)^2}, \tag{6.50}$$

$$\frac{1}{n}\log p(x^n|m_2,\hat{\theta}^{k_{m_2}}) = -\frac{1}{2}\log 2\pi - \log\hat{\sigma} - \frac{1}{2}, \tag{6.51}$$

と与えられる．$\hat{\mu} \to \mu^* + O(\frac{(\log\log n)^{1/2}}{\sqrt{n}})$ $a.s.$ かつ $\hat{\sigma} \to \sigma^* + O(\frac{(\log\log n)^{1/2}}{\sqrt{n}})$ $a.s.$ であるから，m_1 と m_2 に対して(6.45)式が成り立つことがわかる．

さらに，m_1 と m_2 に対する $\frac{1}{n}\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\Big|_{\theta^{k_m}=\hat{\theta}^{k_m}}$ は

$$\frac{1}{n}\frac{\partial^2 \log p(x^n|m_1,\theta^{k_{m_1}})}{\partial\theta^{k_{m_1}}(\partial\theta^{k_{m_1}})^T}\Big|_{\theta^{k_{m_1}}=\hat{\theta}^{k_{m_1}}} = -\frac{1}{(\sigma_0)^2}, \tag{6.52}$$

$$\frac{1}{n}\frac{\partial^2 \log p(x^n|m_2,\theta^{k_{m_2}})}{\partial\theta^{k_{m_2}}(\partial\theta^{k_{m_2}})^T}\Big|_{\theta^{k_{m_2}}=\hat{\theta}^{k_{m_2}}} = \begin{pmatrix} -1/\hat{\sigma}^2 & 0 \\ 0 & -2/\hat{\sigma}^2 \end{pmatrix}, \tag{6.53}$$

となる．$\hat{\sigma} = \sigma^* + O(\frac{(\log\log n)^{1/2}}{\sqrt{n}})$ $a.s.$ であり，$J^*(\theta^{k^*_{m_1}}|m_1) = (1/(\sigma^*)^2)$,

$$J^*(\theta^{k^*_{m_2}}|m_2) = \begin{pmatrix} 1/(\sigma^*)^2 & 0 \\ 0 & 2/(\sigma^*)^2 \end{pmatrix}, \tag{6.54}$$

であるから，m_1 と m_2 に対して(6.44)式が得られる．よって, 条件6.3, (7) が成り立つことがわかる．さらに，

$$\begin{aligned}\frac{1}{n}\log f(\theta^{k_{m_1}}|x^n,m_1) &\propto \frac{1}{n}\log p(x^n|m_1,\theta^{k_{m_1}})f(\theta^{k_{m_1}}|m_1) \\ &\to \log\frac{1}{\sqrt{2\pi}\sigma_0} - \frac{1}{2(\sigma_0)^2}\{(\mu-\mu^*)^2 + (\sigma^*)^2\}, \quad a.s.\end{aligned} \tag{6.55}$$

$$\begin{aligned}\frac{1}{n}\log f(\theta^{k_{m_2}}|x^n,m_2) &\propto \frac{1}{n}\log p(x^n|m_2,\theta^{k_{m_2}})f(\theta^{k_{m_2}}|m_2) \\ &\to \log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2(\sigma)^2}\{(\mu-\mu^*)^2 + (\sigma^*)^2\}, \quad a.s.\end{aligned} \tag{6.56}$$

これらの式の右辺はそれぞれ唯一の点 $\theta^{k^*_{m_1}} = \mu^*$ と $\theta^{k^*_{m_2}} = (\mu^*,\sigma^*)^T$ で最大化される．よって，条件6.3，(5)が成り立つ．したがって，このモデル族は条件6.3をみたし，明らかに条件6.2も満たす.

同様の議論から，正規雑音をもつ線形回帰モデル[74]においても，条件6.3が成り立つことが示される． □

6.4.2 事後確率密度の漸近正規性

本章の議論においても，事後確率密度の漸近正規性 [16, 22, 39, 51, 60, 145] が本質的な役割を演ずる．ただし，これまでは概収束のみの議論であったが，ここでは確率収束の意味で成り立つ漸近正規性についても述べる．

補題 6.1 条件6.1，条件6.2のもとで， $\forall m \in \mathcal{M}$ に対し $\xi^{k_m} = \sqrt{n}(\theta^{k_m} - \hat{\theta}^{k_m})$ の事後確率密度は平均ゼロ，共分散 $\{J^*(\hat{\theta}^{k_m}|m)\}^{-1}$ の k_m-次元正規分布に確率収束する．すなわち R_ξ を任意の直方体として

$$\int_{\xi^{k_m} \in R_\xi} f_\xi(\xi^{k_m}|x^n, m) d\xi^{k_m} \to \frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{(2\pi)^{k_m/2}} \int_{\xi^{k_m} \in R_\xi} \exp\left\{ -\frac{1}{2} \left\| \xi^{k_m} \right\|^2_{J^*(\hat{\theta}^{k_m}|m)} \right\} d\xi^{k_m}, \quad in\ probability, \tag{6.57}$$

が成り立つ．ここで $f_\xi(\xi^{k_m}|x^n, m)$ は ξ^{k_m}-空間上の事後密度で

$$f_\xi(\xi^{k_m}|x^n, m) = \frac{1}{\sqrt{n}^{k_m}} f(\theta^{k_m}|x^n, m). \tag{6.58}$$

で与えられる．さらに $f_\xi(\xi^{k_m}|x^n, m)$ は

$$f_\xi(\hat{\xi}^{k_m}|x^n, m) \to \frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{(2\pi)^{k_m/2}}, \quad in\ probability, \tag{6.59}$$

も満たす．ただし，$\hat{\xi}^{k_m} = \mathbf{0}$ である．

もし条件6.1，条件6.3を仮定すれば，(6.57)式と(6.59)式の確率収束は概収束で置き換えることができる．

(証明) 6.5.1節を参照． □

注意 6.3 補題6.1は重要な事実を述べている．すなわち，最尤推定量を規準化した $\zeta^{k_m} = \sqrt{n}(\hat{\theta}^{k_m}(x^n) - \theta^{k^*_m})$ は漸近的に $N(0, \{K(\theta^{k^*_m}|m)\}^{-1})$ に法則収束するのに対し， $\xi^{k_m} = \sqrt{n}(\theta^{k_m} - \hat{\theta}^{k_m})$ の事後密度は $N\left(0, \{J^*(\hat{\theta}^{k_m}|m)\}^{-1}\right)$ に収束する．ただし，$\{K(\theta^{k_m}|m)\}^{-1}$ は

$$\{K(\theta^{k_m}|m)\}^{-1} = \{J^*(\theta^{k_m}|m)\}^{-1} I^*(\theta^{k_m}|m) \{J^*(\theta^{k_m}|m)\}^{-1} \tag{6.60}$$

で与えられる．この $J^*(\hat{\theta}^{k_m}|m)$ と $K(\theta^{k^*_m}|m)$ の違いは本質的である．すなわち，最尤推定量は x^n の経験分布からのモデル族への写像であり，真の分布に従う x^n によって分布するが，事後密度は与えられた1本の x^n に対する確率モデルの相対的な当てはまりの良さを表しているために生じる差違である． □

6.4.3 モデルの事後確率の収束

まず最初に $-\log p(x^n|m)$ の漸近式を与えた次の補題を示そう.

補題 6.2 条件6.1, 条件6.2のもとで

$$-\log p(x^n|m) = -\log p(x^n|m,\hat{\theta}^{k_m}) + \frac{k_m}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)} + o_p(1), \quad (6.61)$$

が成り立つ. ここで, $o_p(1)$ は $\lim_{n\to\infty} o_p(1) = 0$, *in probability* となる項である.

一方, 条件6.1, 条件6.3のもとでは

$$-\log p(x^n|m) = -\log p(x^n|m,\hat{\theta}^{k_m}) + \frac{k_m}{2}\log\frac{n}{2\pi} + \log\frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)} + o(1), \quad a.s. \quad (6.62)$$

が成り立つ.

(証明) 6.5.2節を参照. □

X^n が離散確率変数であれば, $-\log P(x^n|m)$ はパラメトリックモデル族に対するベイズ符号の理想符号長を表す [80]. m に関する(6.61)式の最小化は MDL原理 [101]であり, これは一様事前分布 $P(m) = 1/|\mathcal{M}|$ に対する事後確率最大規準を与える. (6.61)式は BIC の精密な漸近式を与えている.

補題6.2を用いることにより, 一致性の議論で重要となる以下の補題が導かれる.

補題 6.3 条件6.1, 条件6.2のもとで, $\forall m \neq m^*$, $m \in \mathcal{M}$ に対し,

$$\frac{p(x^n|m)}{p(x^n|m^*)} \to 0, \quad \textit{in probability}. \quad (6.63)$$

が成り立つ. 一方, 条件6.1, 条件6.3のもとで

$$\frac{p(x^n|m)}{p(x^n|m^*)} \to 0, \quad a.s. \quad (6.64)$$

が成り立つ.

(証明) 6.5.3節を参照. □

よって, 以下の定理が成り立つ.

定理 6.1 条件6.1, 条件6.2のもとで

$$P(m^*|x^n) \to 1, \quad in\ probability, \tag{6.65}$$

が成り立つ．一方，条件6.1, 条件6.3が成り立てば

$$P(m^*|x^n) \to 1, \quad a.s. \tag{6.66}$$

が成り立つ．

(証明) ベイズ規則より

$$\frac{p(x^n|m)}{p(x^n|m^*)} = \frac{P(m|x^n)P(m^*)}{P(m^*|x^n)P(m)}, \tag{6.67}$$

である．条件6.1, 条件6.2のもとでは，(6.63)式と条件6.2, (4),より， $\forall m \neq m^*$ に対し

$$\frac{P(m|x^n)}{P(m^*|x^n)} \to 0, \quad in\ probability, \tag{6.68}$$

が成り立つ．$\sum_{m\in\mathcal{M}} P(m|x^n) = 1$ と $\mathcal{M}$ が有限集合であることから，

$$\forall m \neq m^*, P(m|x^n) \to 0, \quad in\ probability, \tag{6.69}$$

かつ

$$P(m^*|x^n) \to 1, \quad in\ probability, \tag{6.70}$$

が成り立つ．(6.66)式についても同様である． □

定理6.1は，最適モデルの事後確率が1に収束することを述べている．これは，事後確率最大のモデル選択，すなわち BIC が弱一致性，あるいは強一致性をもつことを示している．

6.4.4 主定理: モデル選択の一致性

(1) 真のモデルの発見を目的とする場合

定理6.1より，次の定理を導くことができる．

定理 6.2 条件6.1, 条件6.2のもとで，損失関数 $L_1(\cdot:\cdot)$, $L_2(\cdot:\cdot)$, $L_3(\cdot:\cdot)$ に対する(6.21)式のモデル選択は弱一致性

$$\hat{m}^B_{DT} \to m^*, \quad in\ probability, \tag{6.71}$$

をもつ．

もし 条件6.1, 条件6.3が成り立てば, 強一致性

$$\hat{m}_{DT}^{B} \to m^{*}, \quad a.s. \tag{6.72}$$

が成り立つ.

(証明) 6.5.4節参照. □

(2) 未来データの予測を目的とする場合

次のデータの予測を目的とした場合のモデル選択についても, 次のように一致性が導かれる. まず, (6.30)式, (6.33)式によるモデル選択の一致性も示すために,

$$BR(Ap|m,x^{n}) = \int_{\theta^{k_m}\in\Theta^{k_m}} R(p(X|x^{n},m,\theta^{k_m}) : Ap(X|x^{n}))f(\theta^{k_m}|x^{n},m)d\theta^{k_m}, \tag{6.73}$$

と定義し, 以下の2つの補題を示そう.

補題 6.4 条件6.1, 条件6.2のもとで, $\forall m \in \mathcal{M}$ に対し,

$$p_{pred}^{m}(x|x^{n}) = p(x|x^{n},m,\hat{\theta}^{k_m}) + o_p(1), \tag{6.74}$$

を満たす.

もし, 条件6.1, 条件6.3が成り立てば, 上式の確率収束は概収束で置き換えることができ,

$$p_{pred}^{m}(x|x^{n}) = p(x|x^{n},m,\hat{\theta}^{k_m}) + o(1), \quad a.s. \tag{6.75}$$

となる.

(証明) 6.5.5節参照. □

補題 6.5 条件6.1, 条件6.2のもとで, 損失関数 $L_4(\cdot:\cdot)$, $L_5(\cdot:\cdot)$ に対し $BR(Ap|m,x^{n})$ は, $\forall m \in \mathcal{M}$ と $\forall Ap(\cdot|x^{n}) \in \{p_{pred}^{m}(\cdot|x^{n})|m \in \mathcal{M}\}$, $\forall Ap(\cdot|x^{n}) \in \{p(\cdot|x^{n},m,\hat{\theta}^{k_m})|m \in \mathcal{M}\}$ に対して一様に,

$$BR(Ap|m,x^{n}) = R(p(X|x^{n},m,\hat{\theta}^{k_m}) : Ap(X|x^{n})) + o_p(1), \tag{6.76}$$

を満たす.

もし, 条件6.1, 条件6.3が成り立てば, 上式の確率収束は概収束で置き換えることができ,

$$BR(Ap|m,x^{n}) = R(p(X|x^{n},m,\hat{\theta}^{k_m}) : Ap(X|x^{n})) + o(1), \quad a.s. \tag{6.77}$$

となる.

(証明) 6.5.6節参照. □

これらの補題より，以下の定理が導かれる．

定理 6.3 条件6.1，条件6.2のもとで，損失関数 $L_4(\cdot:\cdot), L_5(\cdot:\cdot)$ に対する(6.30)式のモデル選択は弱一致性

$$\hat{m}^B_{PD} \to m^*, \quad in\ probability, \tag{6.78}$$

をもつ．

もし 条件6.1，条件6.2が成り立てば，強一致性

$$\hat{m}^B_{PD} \to m^*, \quad a.s. \tag{6.79}$$

が成り立つ．

(証明) 6.5.7節参照． □

補題6.4，補題6.5と定理6.3の証明より，$BR_{ML}(Af|x^n)$ の最小化は $BR_{PD}(Af|x^n)$ の最小化と漸近的に等価であることから，ただちに以下の定理が導かれる．

定理 6.4 条件6.1，条件6.2のもとで，損失関数 $L_4(\cdot:\cdot), L_5(\cdot:\cdot)$ に対する(6.33)式のモデル選択は弱一致性

$$\hat{m}^B_{ML} \to m^*, \quad in\ probability, \tag{6.80}$$

をもつ．

もし 条件6.1，条件6.2が成り立てば，強一致性

$$\hat{m}^B_{ML} \to m^*, \quad a.s. \tag{6.81}$$

が成り立つ． □

6.4.5 考察

定理6.1は階層モデル族に対する事後確率最大規準の一致性を示している．BIC や MDL の一致性は E.J.Hannan と B.G.Quinn の議論[57]からも示されているが，これは線形回帰モデルに対する結果である．より広いモデル族を扱った結果としては，指数型分布族に対する解析が[50]にある．一方，MDL 規準の密度関数の推定問題 (density estimation) における強一致性については A.R.Barron と T.M.Cover[11]にある．

本章では，ベイズ決定理論に基づくモデル選択を定式化し，これが一致性をもつことを示した．一致性は最適モデルを発見することが目的の場合には望ま

しい性質と考えられる．しかしその場合だけでなく，次のデータの予測が目的の場合においても一致性が示され，ベイズ規準では漸近的に最適モデルで予測するのが良いという結論を得る．AIC が階層モデル族に対して一致性をもたないのはリスクの推定量を考えているためであるが[56, 89]，逆に他の観点からのAICの優位性も示されている[108, 111]．しかし，対象を階層モデル族に限定した場合には，最適なモデル m^* で予測することが最適であるから，一致性を持つことは予測の面からも望ましいことであるといえる[89]．

条件6.2では概収束と重複対数の法則を仮定していないので確率収束しか示せないが，重複対数の法則を仮定した条件6.3では概収束を示すことができる．情報量規準 HQ のペナルティ項 $c\log\log n$ は重複対数の法則のもとで導かれており[57]，BIC が強一致性をもつという議論も重複対数の法則を前提にしていることになる．すなわち，現実的ではないかも知れないが，重複対数の法則が成り立たないようなモデル族に対しては，BIC でも強一致性は成り立たないことになる．どんな規準に対しても，一致性を満足しないような特殊なモデル族を作ることができることには注意するべきである[117]．一致性だけがモデル選択基準の有効性を決定する尺度ではないので，様々な観点からの成果を体系的にとらえる必要がある．

6.5 補題と定理の証明

6.5.1 補題 6.1 の証明

1つの分布列の漸近正規性については[16],pp.285-297, [22] にある．また，[16, 22] では事後確率最大推定量 $\tilde{\theta}^{k_m} = \arg_{\theta^{k_m}} \max \log f(\theta^{k_m}|x^n, m)$ のまわりで展開した議論を行っている．ここでは第5章の議論と同様に，最尤推定量 $\hat{\theta}^{k_m}$ を中心にした漸近正規性[39]を示す．最初に定理の前半である確率収束について証明を与える．

第5章と同様に $L_n(\theta^{k_m}|m) = \log f(\theta^{k_m}|x^n, m)$ と球 $B_\delta(\hat{\theta}^{k_m}) = \left\{\theta^{k_m} \in \Theta^{k_m} | \|\theta^{k_m} - \hat{\theta}^{k_m}\| < \delta\right\}$, および

$$\Sigma_n = -\left(L_n''(\hat{\theta}^{k_m}|m)\right)^{-1} = -\left(\left.\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}\right)^{-1}. \tag{6.82}$$

を定義しておく．テーラー展開により

$$f(\theta^{k_m}|x^n, m) = f(\hat{\theta}^{k_m}|x^n, m)\exp\left\{(\theta^{k_m} - \hat{\theta}^{k_m})^T L'(\hat{\theta}^{k_m}|m)\right\}$$

$$\cdot \exp\left\{-\frac{1}{2}(\theta^{k_m}-\hat{\theta}^{k_m})^T(I+R_n)L''(\hat{\theta}^{k_m}|m)(\theta^{k_m}-\hat{\theta}^{k_m})\right\}, \tag{6.83}$$

が得られる．ここで R_n は

$$R_n = L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} - I, \tag{6.84}$$

であり，θ^{k_m+} は θ^{k_m} と $\hat{\theta}^{k_m}$ を結ぶ線分上の点である．$L_n'(\hat{\theta}^{k_m}|m)$ は定義より，

$$L_n'(\hat{\theta}^{k_m}|m) = \left.\frac{\partial \log f(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}, \tag{6.85}$$

で与えられる．一方，(6.39)式より

$$\hat{\theta}^{k_m} \to \theta^{k_m^*}, \quad in\ probability, \tag{6.86}$$

が成り立っている．

よって，(6.34)式, (6.40)式，及び

$$L_n''(\theta^{k_m}|m) = \frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} + \frac{\partial^2 \log f(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \tag{6.87}$$

より，$\theta^{k_m} \in B_\delta(\theta^{k_m^*})$ に対して一様に

$$-\frac{1}{n}L_n''(\theta^{k_m}|m) \to J^*(\theta^{k_m}|m),\ in\ probability, \tag{6.88}$$

が成り立つ．一方，$\hat{\theta}^{k_m} \to \theta^{k_m^*}$ $in\ probability$ より，$B_\delta(\hat{\theta}^{k_m}) \to B_\delta(\theta^{k_m^*})$ $in\ probability$ が成り立つので，$\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$ に対して一様に

$$L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \to J^*(\theta^{k_m+}|m)\{J^*(\hat{\theta}^{k_m}|m)\}^{-1},\ \ in\ probability, \tag{6.89}$$

が成り立つ．よって，条件6.2,(4) と $\hat{\theta}^{k_m} \to \theta^{k_m^*}, in\ probability$ より，$\delta > 0$ が存在して，$\forall\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$ と $\forall\epsilon > 0$ に対して $n \to \infty$ のとき

$$I - A(\epsilon) \le L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \le I + A(\epsilon),\ \ in\ probability, \tag{6.90}$$

が成り立つ[2]．ただし，I は $k_m \times k_m$ 基本行列，$A(\epsilon)$ は $\epsilon \to 0$ のときその最大固有値が 0 に収束する $k_m \times k_m$ 正定値行列である．$J^*(\theta^{k_m}|m)$ の連続性から，(6.90)式において $\epsilon \to 0$ のとき $\delta \to 0$ とできることに注意しておく．

[2]事象 $\mathcal{A}$ に対して $\lim_{n\to\infty} P^*(\mathcal{A}) \to 1$ となるとき，“$n \to \infty$ のとき $\mathcal{A}$, $in\ probability$” と記述する．

以上より,

$$P_n(\delta) = \int_{\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})} f(\theta^{k_m}|x^n, m) d\theta^{k_m}, \tag{6.91}$$

は $n \to \infty$ のとき確率収束の意味で

$$f(\hat{\theta}^{k_m}|x^n, m) \exp\{\bar{c}_f \delta\} |\det \Sigma_n|^{1/2} |\det(I - A(\epsilon))|^{-1/2} \int_{\|z^{k_m}\| < s_n} \exp\{-\frac{1}{2} z^T z\} dz, \tag{6.92}$$

で上界され,

$$f(\hat{\theta}^{k_m}|x^n, m) \exp\{-\bar{c}_f \delta\} |\det \Sigma_n|^{1/2} |\det(I + A(\epsilon))|^{-1/2} \int_{\|z^{k_m}\| < t_n} \exp\{-\frac{1}{2} z^T z\} dz, \tag{6.93}$$

で下界される. これは, $P_n(\delta)$ が上の上下界を越える確率が $n \to \infty$ で 0 に収束するという意味である. ここで $\bar{c}_f$ は $\frac{\partial \log f(\theta^{k_m}|m)}{\partial \theta^{k_m}}$ の最大絶対値, s_n と t_n は $\overline{a(\epsilon)}$ ($\underline{a(\epsilon)}$) と $\overline{l_n}$ ($\underline{l_n}$) を $A(\epsilon)$ と Σ_n の最大(最小)固有値として, $s_n = \delta(1 - \overline{a(\epsilon)})^{1/2} / \underline{l_n}^{1/2}$, $t_n = \delta(1 - \underline{a(\epsilon)})^{1/2} / \overline{l_n}^{1/2}$ で定義される.

(6.86)式と(6.88)式より, $J^*(\theta^{k_m}|m)$ は連続であるので,

$$n\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \to \{J^*(\theta^{k_m^*}|m)\}^{-1}, \quad in\ probability. \tag{6.94}$$

となり, $\overline{l_n} \to 0$, *in probability* かつ $\underline{l_n} \to 0$, *in probability* が得られる. よって, $s_n \to \infty$, *in probability* と $t_n \to \infty$, *in probability* が成り立つ. したがって $n \to \infty$ のとき

$$\begin{aligned} &|\det(I - A(\epsilon))|^{1/2} \exp\{-\bar{c}_f \delta\} \lim_{n\to\infty} P_n(\delta) \\ &\quad \le \lim_{n\to\infty} f_\xi(\hat{\xi}^{k_m}|x^n, m)(2\pi)^{k_m/2} \left\{ \det\left(\frac{1}{n} L''(\hat{\theta}^{k_m}|m) \right) \right\}^{-1/2} \\ &\quad \le |\det(I + A(\epsilon))|^{1/2} \exp\{\bar{c}_f \delta\} \lim_{n\to\infty} P_n(\delta), \end{aligned} \tag{6.95}$$

が確率収束の意味で成り立つ.

以上より, もし $\forall \delta > 0$ に対して $\lim_{n\to\infty} P_n(\delta) = 1$, *in probability*, が成り立てば

$$\lim_{n\to\infty} f_\xi(\hat{\xi}^{k_m}|x^n, m) \to \frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{(2\pi)^{k_m/2}}, \quad in\ probability, \tag{6.96}$$

が明らかに成り立つ. 一方, $B_\delta(\hat{\theta}^{k_m}) \to B_\delta(\theta^{k_m^*})$ であるので, 条件6.2, (3) より, $\forall \delta > 0$ に対して $\lim_{n\to\infty} P_n(\delta) = 1$, *in probability* が成り立ち, (6.96)式が示された.

条件6.1, 条件6.3のもとでは, 上の議論の確率収束は全て概収束で置き換えることができる. すなわち, 同様の議論により

$$f_\xi(\hat{\xi}^{k_m}|x^n, m) \to \frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{(2\pi)^{k_m/2}}, \ a.s. \tag{6.97}$$

が成り立ち, 証明が完結する.

6.5.2 補題 6.2 の証明

条件6.1，条件6.2のもとで，補題6.1と

$$f(\theta^{k_m}|x^n,m)=\sqrt{n}^{k_m}f_\xi(\xi^{k_m}|x^n,m), \tag{6.98}$$

より

$$f(\hat{\theta}^{k_m}|x^n,m)=\frac{\sqrt{n}^{k_m}\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{\sqrt{2\pi}^{k_m}}+o_p(\sqrt{n}^{k_m}), \tag{6.99}$$

が成り立つ．ここで $o_p(\sqrt{n}^{k_m})$ は $\lim_{n\to\infty}\frac{o_p\left(\sqrt{n}^{k_m}\right)}{\sqrt{n}^{k_m}}=0$, *in probability* を満たす項である．よって，

$$-\log f(\hat{\theta}^{k_m}|x^n,m)=\frac{k_m}{2}\log\frac{n}{2\pi}+\log\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}+\log(1+o_p(1)). \tag{6.100}$$

が得られる．一方，ベイズの定理より

$$-\log f\left(x^n|m\right)=-\log\frac{f(x^n|m,\hat{\theta}^{k_m})f(\hat{\theta}^{k_m}|m)}{f(\hat{\theta}^{k_m}|x^n,m)}. \tag{6.101}$$

である．(6.100)式を(6.101)式に代入して，補題の前半，すなわち確率収束の意味での漸近式が得られる．

条件6.1，条件6.3が仮定されれば，上の議論の確率収束は概収束で置き換えることができ，定理の後半が示される． □

6.5.3 補題 6.3 の証明

まず，条件6.1，条件6.2のもとで，定理の前半である確率収束を示そう．

条件6.2, (6)より，$\forall\epsilon,\delta>0$ に対して $n_{\epsilon,\delta}$ が存在し， $n>n_{\epsilon,\delta}$ のとき

$$P^*\left\{\left\|\hat{\theta}^{k_m}-\theta^{k_m^*}\right\|<\frac{\epsilon(g(n))^{1/3}}{\sqrt{n}}\right\}>1-\delta, \tag{6.102}$$

が成り立つ．ここで $g(n)$ は $\lim_{n\to\infty}g(n)=\infty$ を満たす任意の関数である．以下記述の簡便性のため，確率収束の意味でのオーダー $O_p(\cdot)$, $o_p(\cdot)$ を用いて議論する．(6.102)式は $\|\hat{\theta}^{k_m}-\theta^{k_m^*}\|\le O_p((g(n))^{1/3}/\sqrt{n})$ を意味する．これは $\forall m\in\mathcal{M}$ に対し $\hat{\theta}^{k_m}$ は $\theta^{k_m^*}$ の弱一致推定量，すなわち $\hat{\theta}^{k_m}\to\theta^{k_m^*}$, *in probability* であることを意味する．これを $\|\hat{\theta}^{k_m}-\theta^{k_m^*}\|=o_p(1)$ と記述する．

補題6.2より， $\forall m\in\mathcal{M}$ に対し

$$\log\frac{p(x^n|m)}{p(x^n|m^*)}=\log\frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\left(\frac{k_m}{2}-\frac{k_{m^*}}{2}\right)\log\frac{n}{2\pi}$$

$$-\log \frac{p(\hat{\theta}^{k_{m^*}}|m^*)\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)\sqrt{\det J^*(\hat{\theta}^{k_{m^*}}|m^*)}} + o_p(1), \tag{6.103}$$

が成り立つ．ここで $\hat{\theta}^{k_m} \to \theta^{k_m^*}$, *in probability* より，$f(\hat{\theta}^{k_m}|m) = O_p(1)$ と $\det J^*(\hat{\theta}^{k_m}|m) = O_p(1)$ も成り立つ．$O_p(1)$ は $\exists C > 0$ が存在して $n \to \infty$ のとき $|O_p(1)| \leq C$ *in probability* が成り立つような項である．

(6.102)式と(6.103)式より，もし $\forall m \neq m^*$ に対して

$$\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n \to -\infty, \quad \textit{in probability}, \tag{6.104}$$

が成り立てば証明が完結する．

まず $\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k_m^*})}$ を評価しよう．テーラー展開より

$$-\log p(x^n|m, \theta^{k_m^*}) =$$

$$-\log p(x^n|m, \hat{\theta}^{k_m}) - \frac{1}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)$$

$$+ O_p\Big(\Big\|\hat{\theta}^{k_m} - \theta^{k_m^*}\Big\|^3 \left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m} \frac{\partial \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}\Big), \tag{6.105}$$

が得られる．

条件6.1より，

$$-\frac{1}{n}\left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \to J^*(\hat{\theta}^{k_m}|m), \quad \textit{in probability}, \tag{6.106}$$

であり，$\sqrt{n}\|\hat{\theta}^{k_m} - \theta^{k_m^*}\| \leq O_p((g(n))^{1/3})$ も成り立つ．$g(n)$ は $\lim_{n\to\infty} g(n) = \infty$ であれば任意だったので，$g(n)$ の発散速度を十分ゆっくりとれば

$$-(\hat{\theta}^{k_m} - \theta^{k_m^*})^T \left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} (\hat{\theta}^{k_m} - \theta^{k_m^*})$$

$$= \sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*})^T J^*(\hat{\theta}^{k_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*}). \quad a.s. \tag{6.107}$$

一方，

$$\Big\|\hat{\theta}^{k_m} - \theta^{k_m^*}\Big\| \leq O_p\Big(\frac{(g(n))^{1/3}}{\sqrt{n}}\Big), \tag{6.108}$$

と(6.46)式より，

$$\Big\|\hat{\theta}^{k_m} - \theta^{k_m^*}\Big\|^3 \left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m} \frac{\partial \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}} \leq O_p\Big(\frac{g(n)}{\sqrt{n}}\Big). \tag{6.109}$$

が成り立つ．したがって，$g(n) = o(n^{1/2})$ と定義すれば

$$-\log p(x^n|m,\theta^{k_m^*}) = -\log p(x^n|m,\hat{\theta}^{k_m}) + \frac{1}{2}\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*})^T J^*(\hat{\theta}^{k_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k_m^*}) + o_p(1). \tag{6.110}$$

条件6.2, (3) と $\|\hat{\theta}^{k_m} - \theta^{k_m^*}\| = o_p(1)$ より，$J^*(\hat{\theta}^{k_m}|m) \to J^*(\theta^{k_m^*}|m)$, *in probability* も成り立つ．よって，条件6.2, (6)より $\hat{\theta}^{k_m} = \theta^{k_m^*} + O_p(1/\sqrt{n})$ であるので，$\forall c > 0$ に対し

$$\lim_{n\to\infty} P^*\left\{|\log p(x^n|m,\theta^{k_m^*}) - \log p(x^n|m,\hat{\theta}^{k_m})| \le c\log n\right\} = 1, \tag{6.111}$$

が成り立つ．すなわち

$$\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k_m^*})} = o_p(\log n), \tag{6.112}$$

であり，$o_p(\log n)$ は $\frac{o_p(\log n)}{\log n} \to 0$, *in probability* となる項である．よって

$$\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n = \log \frac{p(x^n|m,\theta^{k_m^*})}{p(x^n|m^*,\theta^{k_{m^*}^*})} - \frac{k_m - k_{m^*}}{2}\log n + o_p(\log n), \tag{6.113}$$

が成り立つ．

まず $k_{m^*} > k_m$ の場合を考える．(6.41)式より，$\forall \theta^{k_m} \in \Theta^{k_m}$ に対し

$$\log \frac{p(x^n|m,\theta^{k_m})}{p(x^n|m,\theta^{k_{m^*}^*})} = -Cn + o_p(n), \tag{6.114}$$

が成り立つ．ただし C はある正定数であり，$o_p(n)$ は $\frac{o_p(n)}{n} \to 0$, *in probability* となる項である．したがって，$\forall m \in \mathcal{M}, k_{m^*} > k_m$ に対し

$$\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n = -Cn - \frac{k_m - k_{m^*}}{2}\log n + o_p(n) \to -\infty, \quad \textit{in probability}, \tag{6.115}$$

が得られる．

一方，$k_{m^*} < k_m$ のときは(6.12)式より

$$p(x^n|m,\theta^{k_m^*}) = p(x^n|m^*,\theta^{k_{m^*}^*}), \tag{6.116}$$

が成り立っている．よって，(6.113)式より，$\forall m \in \mathcal{M}$，$k_{m^*} < k_m$ に対し

$$\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n = -\frac{k_m - k_{m^*}}{2}\log n + o_p(\log n) \to -\infty, \quad in\ probability, \tag{6.117}$$

が成り立つ．以上より，$\forall k_m \neq k_{m^*}$ に対し(6.104)式が示され，題意の確率収束が証明された．

次に条件6.1，条件6.3のもとで，定理の後半の概収束を示す．

条件6.3, (7)より，$\forall m \in \mathcal{M}$ に対し

$$\left\| \hat{\theta}^{k_m} - \theta^{k_m^*} \right\| = O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right), \ a.s. \tag{6.118}$$

であり，これは $\hat{\theta}^{k_m}$ が $\theta^{k_m^*}$ の強一致推定量であることを示している．

補題6.2より，$\forall m \in \mathcal{M}$ に対し

$$\log \frac{p(x^n|m)}{p(x^n|m^*)} = \log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \left(\frac{k_m}{2} - \frac{k_{m^*}}{2}\right)\log\frac{n}{2\pi} - \log \frac{f(\hat{\theta}^{k_{m^*}}|m^*)\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)\sqrt{\det J^*(\hat{\theta}^{k_{m^*}}|m^*)}} + o(1), \ a.s. \tag{6.119}$$

となる．$\hat{\theta}^{k_m} \to \theta^{k_m^*} \ \ a.s.$ より，$f(\hat{\theta}^{k_m}|m) = O(1) \ \ a.s.$ かつ $\det J^*(\hat{\theta}^{k_m}|m) = O(1)$ $a.s.$ であるから，(6.119)式より，もし $\forall m \neq m^*$ に対して

$$\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n \to -\infty, \quad a.s. \tag{6.120}$$

が成り立てば証明は完結する．

確率収束のときの議論と同様に $\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k_m^*})}$ を評価しよう．テーラー展開より，

$$-\log p(x^n|m,\theta^{k_m^*}) = -\log p(x^n|m,\hat{\theta}^{k_m}) - \frac{1}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right) + O\Big(\left\|\hat{\theta}^{k_m} - \theta^{k_m^*}\right\|^3 \left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m} \frac{\partial \log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}\Big), \tag{6.121}$$

が成り立つ．

条件6.3, (7)より,

$$-(\hat{\theta}^{k_m}-\theta^{k_m^*})^T\left.\frac{\partial^2\log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}(\hat{\theta}^{k_m}-\theta^{k_m^*})$$

$$=\sqrt{n}(\hat{\theta}^{k_m}-\theta^{k_m^*})^T J^*(\hat{\theta}^{k_m}|m)\sqrt{n}(\hat{\theta}^{k_m}-\theta^{k_m^*})+O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \tag{6.122}$$

一方,

$$\left\|\hat{\theta}^{k_m}-\theta^{k_m^*}\right\|=O\Big(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\Big),\ a.s. \tag{6.123}$$

と(6.46)式より

$$\left\|\hat{\theta}^{k_m}-\theta^{k_m^*}\right\|^3\left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial\log p(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m+}}=O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \tag{6.124}$$

が成り立つ. したがって,

$$-\log p(x^n|m,\theta^{k_m^*})=-\log p(x^n|m,\hat{\theta}^{k_m})$$

$$+\frac{1}{2}\sqrt{n}(\hat{\theta}^{k_m}-\theta^{k_m^*})^T J^*(\hat{\theta}^{k_m}|m)\sqrt{n}(\hat{\theta}^{k_m}-\theta^{k_m^*})+O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right),\quad a.s. \tag{6.125}$$

が得られる.

$\|\hat{\theta}^{k_m}-\theta^{k_m^*}\|=o(1)\ \ a.s.$ より, $J^*(\hat{\theta}^{k_m}|m)\to J^*(\theta^{k_m^*}|m)\ \ a.s.$ が得られるから,

$$\log\frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k_m^*})}=O(\log\log n),\quad a.s. \tag{6.126}$$

が成り立ち,

$$\log\frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n$$

$$=\log\frac{p(x^n|m,\theta^{k_m^*})}{p(x^n|m^*,\theta^{k_{m^*}^*})}-\frac{k_m-k_{m^*}}{2}\log n+O(\log\log n),\quad a.s. \tag{6.127}$$

が得られる.

まず $k_{m^*}>k_m$ の場合を考えると, (6.49)式より, $\forall\theta^{k_m}\in\Theta^{k_m}$ に対し

$$\log\frac{p(x^n|m,\theta^{k_m})}{p(x^n|m,\theta^{k_{m^*}^*})}=-Dn+o(n),\quad a.s. \tag{6.128}$$

が成り立つ. ただし, D はある正定数である. よって $\forall m\in\mathcal{M}$ かつ $k_{m^*}>k_m$ に対して

$$\log\frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n=-Cn-\frac{k_m-k_{m^*}}{2}\log n+o(n)$$

$$\to-\infty,\quad a.s. \tag{6.129}$$

が成り立つ.

$k_{m^*} < k_m$ のときは(6.12)式より

$$p(x^n|m, \theta^{k_m^*}) = p(x^n|m^*, \theta^{k_{m^*}^*}), \tag{6.130}$$

であるから, (6.127)式より, $\forall m \in \mathcal{M}$ かつ $k_{m^*} < k_m$ に対して

$$\begin{aligned} \log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2} \log n &= \frac{k_{m^*} - k_m}{2} \log n + O(\log n \log n) \\ &\to -\infty, \quad a.s. \end{aligned} \tag{6.131}$$

が成り立つ. よって $\forall k_m \neq k_{m^*}$ に対して(6.120)式が示され, 証明が完結した. □

6.5.4 定理 6.2 の証明

まず, 条件6.1, 条件6.2のもとで弱一致性を示す. $L_3(m^* : Am)$ は $L_1(m^* : Am)$, $L_2(m^* : Am)$ を含むので, $L_3(m^* : Am)$ のみを考えれば十分である.

定理6.1と $L_3(m^* : Am)$ の有界性より,

$$BL_{DT}(Am|x^n) = \sum_{m \in \mathcal{M}} L_3(m : Am) P(m|x^n) = L_3(m^* : Am) + o_p(1), \tag{6.132}$$

したがって,

$$\min_{Am \in \mathcal{M}} BL_{DT}(Am|x^n) \to \min_{Am \in \mathcal{M}} L_3(m^* : Am), \quad in\ probability. \tag{6.133}$$

が成り立つ. よって

$$\arg \min_{Am \in \mathcal{M}} L_3(m^* : Am) = m^*, \tag{6.134}$$

より弱一致性が成り立つことが示された.

もし 条件6.1, 条件6.3が成り立てば, 定理6.1より, 上の議論の確率収束を概収束で置き換えることができる. したがって, 条件6.1, 条件6.2のもとで強一致性が成り立つことがわかる. □

6.5.5 補題 6.4 の証明

まず, 条件6.1, 条件6.2のもとで確率収束を示す.

$p(x|x^n, m, \theta^{k_m})$ は θ^{k_m}. に対して一様連続である. ここで, $\forall D > \forall C > 0$ に対して

$$E_{n,C} = \left\{ \theta^{k_m} \middle| \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} \leq \frac{C}{\sqrt{n}} \right\}, \tag{6.135}$$

$$F_{n,C} = \left\{ \theta^{k_m} \left| \frac{C}{\sqrt{n}} < \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} \leq D \right. \right\}, \tag{6.136}$$

$$G = \left\{ \theta^{k_m} \left| \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} > D \right. \right\}, \tag{6.137}$$

を定義すると，$p^m_{pred}(x|x^n)$ は

$$\int_{\theta^{k_m} \in \Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m} = I_1 + I_2 + I_3, \tag{6.138}$$

とかける．ここで，

$$I_1 = \int_{E_{n,C}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m}, \tag{6.139}$$

$$I_2 = \int_{F_{n,C}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m}, \tag{6.140}$$

$$I_3 = \int_{G} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m}. \tag{6.141}$$

である．

条件6.1,(3)と条件6.2,(3)より，

$$\sup_{x \in \mathcal{X}} \int_{\theta^{k_m} \in \Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m} < C', \tag{6.142}$$

となる正定数 C' が存在する．

まず，$p(x|x^n, m, \theta^{k_m})$ の θ^{k_m} に対する連続性と条件6.2,(3)の有界性より，$\forall \theta^{k_m} \in E_{n,c}$ と $\forall x \in \mathcal{X}$ に対して一様に

$$p(x|x^n, m, \theta^{k_m}) \to p(x|x^n, m, \hat{\theta}^{k_m}), \quad in\ probability, \tag{6.143}$$

が成り立つ．したがって，$\forall \theta^{k_m} \in E_{n,c}$ と $\forall x \in \mathcal{X}$ に対して一様に

$$I_1 = p(x|x^n, m, \hat{\theta}^{k_m}) \int_{E_{n,C}} f(\theta^{k_m}|m, x^n) d\theta^{k_m} + o_p(1). \tag{6.144}$$

次に，$\hat{\theta}^{k_m} \to \theta^{k^*_m}, in\ probability$ と $p(x|x^n, m, \theta^{k_m})$ の連続性から，正定数 $C_R > 0$ が存在して $n \to \infty$ のとき，

$$\sup_{x \in \mathcal{X}, \theta^{k_m} \in F_{n,C}} p(x|x^n, m, \theta^{k_m}) < C_R, \quad in\ probability, \tag{6.145}$$

が成り立つ．一方，事後密度の漸近正規性より，

$$\int_{F_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 1 - \int_{E_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m}, \ in\ probability, \tag{6.146}$$

であるから，(6.145)式と(6.146)式より，$\forall x \in \mathcal{X}$ に対して一様に

$$I_2 \to C_R \left\{ 1 - \int_{E_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \right\}, \quad in\ probability, \tag{6.147}$$

を得る．一方，条件6.1,(3)より

$$\int_G p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m} < c_{int}, \tag{6.148}$$

であり，かつ尤度関数は確率収束の意味で漸近的に単峰形で

$$\int_G f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 0, \quad in\ probability, \tag{6.149}$$

が成り立つから，$\forall x \in \mathcal{X}$ に対して一様に

$$I_3 \to 0, \quad in\ probability. \tag{6.150}$$

が得られる．

C, D を十分大きくとれば，$\int_{\theta^{k_m} \in \Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m} = I_1 + o_p(1)$ となり，すなわち $\forall m \in \mathcal{M}$ と $\forall x \in \mathcal{X}$ に対して一様に

$$\int_{\theta^{k_m} \in \Theta^{k_m}} p(x|x^n, m, \theta^{k_m}) f(\theta^{k_m}|m, x^n) d\theta^{k_m} = p(x|x^n, m, \hat{\theta}^{k_m}) + o_p(1), \tag{6.151}$$

が成り立つことがわかる．

条件6.1，条件6.3を仮定すれば，上の議論の確率収束は概収束で置き換えられる．

6.5.6 補題 6.5 の証明

まず，条件6.1，条件6.2のもとで確率収束を示す．

リスク関数 $R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n))$ は θ^{k_m} に対して一様連続である．補題6.4の証明と同様に，$\forall D > \forall C > 0$ に対して

$$E_{n,C} = \left\{ \theta^{k_m} \left| \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} \leq \frac{C}{\sqrt{n}} \right. \right\}, \tag{6.152}$$

$$F_{n,C} = \left\{ \theta^{k_m} \left| \frac{C}{\sqrt{n}} < \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} \leq D \right. \right\}, \tag{6.153}$$

$$G = \left\{ \theta^{k_m} \left| \left\| \theta^{k_m} - \hat{\theta}^{k_m} \right\|_{J(\hat{\theta}^{k_m}|m)} > D \right. \right\}, \tag{6.154}$$

を定義すると，条件付きベイズリスクは

$$\int_{\theta^{k_m} \in \Theta^{k_m}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|m, x^n) d\theta^{k_m} = I_1 + I_2 + I_3, \tag{6.155}$$

とかける．ここで，

$$I_1 = \int_{E_{n,C}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|m, x^n) d\theta^{k_m}, \tag{6.156}$$

$$I_2 = \int_{F_{n,C}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|m, x^n) d\theta^{k_m}, \tag{6.157}$$

$$I_3 = \int_{G} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|m, x^n) d\theta^{k_m}, \tag{6.158}$$

である．

条件6.1,(3)と条件6.2,(4)より，$\forall Ap(X|x^n) \in \{p^m_{pred}(X|x^n)\}$, $\forall Ap(X|x^n) \in \{p(X|x^n, m, \hat{\theta}^{k_m})\}$ に対し一様に

$$\int_{\theta^{k_m} \in \Theta^{k_m}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) f(\theta^{k_m}|m, x^n) d\theta^{k_m} < C', \tag{6.159}$$

となる正定数 C' が存在する．

まず，$R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n))$ の連続性より，$\forall \theta^{k_m} \in E_{n,c}$ と $\forall Ap(X|x^n) \in \{p^m_{pred}(X|x^n)\}$, $\forall Ap(X|x^n) \in \{p(X|x^n, m, \hat{\theta}^{k_m})\}$ に対して一様に

$$R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) \to R(p(X|x^n, m, \hat{\theta}^{k_m}) : Ap(X|x^n)), \tag{6.160}$$

が成り立つ．したがって，$\forall \theta^{k_m} \in E_{n,c}$ と $\forall Ap(X|x^n) \in \{p^m_{pred}(X|x^n)\}$, $\forall Ap(X|x^n) \in \{p(X|x^n, m, \hat{\theta}^{k_m})\}$ に対して一様に

$$I_1 = R(p(X|x^n, m, \hat{\theta}^{k_m}) : Ap(X|x^n)) \int_{E_{n,C}} f(\theta^{k_m}|m, x^n) d\theta^{k_m} + o_p(1). \tag{6.161}$$

次に，$\hat{\theta}^{k_m} \to \theta^{k^*_m}$, *in probability* と $R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n))$ の連続性から，正定数 $C_R > 0$ が存在して $n \to \infty$ のとき，$\forall Ap(X|x^n) \in \{p^m_{pred}(X|x^n)\}$, $\forall Ap(X|x^n) \in \{p(X|x^n, m, \hat{\theta}^{k_m})\}$ に対して一様に

$$\sup_{\theta^{k_m} \in F_{n,C}} R(p(X|x^n, m, \theta^{k_m}) : Ap(X|x^n)) < C_R, \quad in\ probability, \tag{6.162}$$

が成り立つ．一方，事後密度の漸近正規性より，

$$\int_{F_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 1 - \int_{E_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m}, \ in\ probability, \tag{6.163}$$

であるから，(6.162)式と(6.163)式より，

$$I_2 \to C_R \left\{ 1 - \int_{E_{n,C}} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \right\}, \quad in\ probability, \tag{6.164}$$

を得る．一方，条件6.1,(3)より

$$\int_G R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))f(\theta^{k_m}|m)d\theta^{k_m}<\infty, \tag{6.165}$$

であり，かつ尤度関数は確率収束の意味で漸近的に単峰形で

$$\int_G f(\theta^{k_m}|x^n,m)d\theta^{k_m}\to 0,\ \ in\ probability, \tag{6.166}$$

が成り立つから，$\forall Ap(X|x^n)\in\{p^m_{pred}(X|x^n)\}$，$\forall Ap(X|x^n)\in\{p(X|x^n,m,\hat{\theta}^{k_m})\}$ に対して一様に

$$I_3\to 0,\ \ in\ probability, \tag{6.167}$$

が成り立つ．

$C\to\infty, C'\to\infty$ とすれば，$\int_{\theta^{k_m}\in\Theta^{k_m}} R(p(X|x^n,m,\theta^{k_m}):Ap(X|x^n))f(\theta^{k_m}|m,x^n)d\theta^{k_m}$ $=I_1+o_p(1)$ となり，すなわち $\forall m\in\mathcal{M}$, $\forall Ap(X|x^n)\in\{p^m_{pred}(X|x^n)\}$, $\forall Ap(X|x^n)$ $\in\{p(X|x^n,m,\hat{\theta}^{k_m})\}$ に対して一様に

$$\begin{aligned}&\int_{\theta^{k_m}\in\Theta^{k_m}} R(p(X|x^n,m,\theta^{k_m}):Ap^*(X))f(\theta^{k_m}|m,x^n)d\theta^{k_m}\\ &\qquad=R(p(X|x^n,m,\hat{\theta}^{k_m}):Ap^*(X))+o_p(1),\end{aligned} \tag{6.168}$$

が成り立つ．

条件6.1，条件6.3を仮定すれば，上の議論の確率収束は概収束で置き換えられる．

6.5.7 定理 6.3 の証明

まず，条件6.1，条件6.2のもとで弱一致性を示す．

補題6.5と $|\mathcal{M}|$ が有限であることから，ある定数 C_R が存在して，$n\to\infty$ のとき

$$\sup_{Ap(\cdot|x^n),m} BR(Ap|m,x^n)<C_R,\ \ in\ probability, \tag{6.169}$$

が成り立つ．したがって，$P(m^*|x^n)\to 1, in\ probability$ より，

$$\begin{aligned}BR_{PD}(Ap|x^n)&=\sum_{m\in\mathcal{M}} BR(Ap|m,x^n)P(m|x^n)\\ &=BR(Ap|m^*,x^n)+o_p(1).\end{aligned} \tag{6.170}$$

よって，

$$\min_{Ap(\cdot|x^n)\in\{p^m_{pred}(\cdot|x^n)\}} BR_{PD}(Ap|x^n)\to \min_{Ap(\cdot|x^n)\in\{p^m_{pred}(\cdot|x^n)\}} BR(Ap|m^*,x^n),\ \ in\ probability. \tag{6.171}$$

したがって,

$$\arg \min_{Ap \in \{p^m_{pred}(x|x^n)\}} = p^{m^*}_{pred}(x|x^n), \tag{6.172}$$

より, 弱一致性が示された.

もし 条件6.1, 条件6.3が成り立てば, 定理6.1より, 上の議論の確率収束を概収束で置き換えることができる. したがって, 条件6.1, 条件6.2のもとで強一致性が成り立つことがわかる. □

6.6 本章の結論

本章ではベイズ決定理論に基づくモデル選択を定式化した．さらに，このベイズ規準によるモデル選択の一致性を示した．これにより，真のモデル(最適モデル)の発見が目的の場合，および未来のデータの予測が目的の場合の両者で一致性が成り立つことが明らかとなった．すなわち，未来のデータの予測が目的の場合においても，漸近的には最適モデルによる予測がベイズ最適であるといえる．ただし，ここでもベイズ最適解を求めるための計算量削減は残された課題である．

一致性は，統計的モデル選択問題において多くの議論がなされてきた性質であるが，この一致性の議論の難しさは本質的にモデル族の階層構造に起因する．すなわち，最適なパラメータを用いた場合には，最適モデルよりも高次のモデルを用いても予測に対して最適であるので，一致性は意味がないとも考えられる．しかし，各モデル $m \in \mathcal{M}$ でパラメータの推定量を用いて予測する場合には，最適モデル m^* の予測精度が最も優れている．予測を目的とした場合のベイズ決定理論に基づくモデル選択は，漸近的にこの最適モデルを選択でき，これは予測に対しても最適なモデルを選択することを意味している．勿論，一致性が統計的モデル選択基準の唯一の評価ではない．どのようなモデル選択規準に対しても，一致性をもたないようなモデル族を構成することは可能であることに注意すべきである．ただし，実際の統計的モデル選択問題は本章で議論した階層構造をもつモデル族を対象とすることがほとんどであり，その意味で本章の結果は意味があるものと考えられる．

本章ではモデル選択の一つの性質として一致性を議論したが，他の観点からの評価も行われるべきである．例えば，ノンパラメトリックなモデル族に対する収束速度，真の確率分布やモデル族をデータ数と共に変化させる場合の議論などは今後の課題である．

第 7 章

モデル推定を目的としたモデル選択の選択誤り率に関する漸近解析

7.1 本章の目的

統計的モデル選択問題[74]は，理論的に興味深いテーマであるだけでなく，非常に広範囲の応用を持つ極めて重要な問題の一つである．とくに候補であるモデル族が階層的な構造を持っている場合，次数の低いモデルはより次数の高い(表現能力の高い)モデルに含まれるという性質が問題(あるいは問題設定)を難しくしている．モデル選択に関しては，AIC [1], BIC [102], MDL [95] といった情報量規準を代表として，様々な観点からの研究 [92, 111, 131] がなされている．一方，第6章で述べた通り，もし事前分布が仮定できれば，ベイズ決定理論[16, 27]に基づくモデル選択規準が定式化できる[40]．このベイズ規準では様々な損失関数に対し，有限長のデータに対してベイズ最適なモデル選択とパラメータ推定が可能となる．ベイズ決定理論に基づくモデル選択は，ベイズ統計の枠組みで統一した議論が行われるという点で一貫性があり，BIC や MDL でも事前分布が仮定されているように，モデル選択問題に対して非常に相性のよい解法の一つである．

本章では，真の構造(真のモデル: m^* で表す)の発見を目的とする統計的モデル選択問題に対し，ベイズ決定理論によって最適な推定を行うことを考える．ここで考慮に入れるべきことは，実際の応用ではデータからモデルを推定した後，この結果をもとにアクションをとることである．従って，損失関数は現実的応用を見据えたものであり，かつ一般性があることが望まれる．そこで本稿ではまず，0-1損失を含み，問題に応じて設定できる一般化された損失関数を定義し，ベイズ決定理論に基づくモデル選択を定式化する．

一方，真のモデル m^* の発見を目的としたモデル選択基準の一つの合理的な

評価基準は，

$$P_e^* = P^*\{\hat{m} \neq m^*\}, \tag{7.1}$$

で与えられる選択誤り率であろう．ここで，$\hat{m}$ はモデル選択によって選ばれたモデル(推定値)であり，P^* はデータを発生する真の確率分布による確率を表す．情報量規準によるモデル選択に対しては，この選択誤り率の漸近的特性を解析した結果が示されており[39, 111, 125]，モデル選択基準の特性として選択誤り率を議論することは意味があると思われる．

そこで本章ではまず，定式化したベイズ規準によるモデル選択に関して，一般的な条件のもとでこの選択誤り率の漸近的上界を解析することにより，その漸近的性能について考察を与える．また，このようなモデル選択問題の適用例として，確率モデルのパラメータ変化点検出問題を取り上げ，具体的な計算式を与えてる．さらにシミュレーション実験により損失関数を変えたときの性能を調べ，現実問題での有効性を検証する．

7.2 問題設定

母集団(情報源)からの長さ n のデータ系列を x^n とし，確率変数 X^n の1つの実現値を表すものとする．また，確率モデルのクラスを $\mathcal{P} = \{p(\cdot|m, \theta^{k_m})\}$ とする．$p(\cdot|m, \theta^{k_m})$ は m と θ^{k_m} で特定される1つの確率分布を表す．m は1つのパラメトリッククラスを特定する離散ラベルで，モデルと呼び，可算有限集合 $\mathcal{M}$ の要素とする．すなわち，$m \in \mathcal{M}$ かつ $|\mathcal{M}| < \infty$ である．$\theta^{k_m} = (\theta_1^{k_m}, \theta_2^{k_m}, \cdots, \theta_{k_m}^{k_m})^T$ はモデル m の k_m次元パラメータであり，$\theta^{k_m} \in \Theta^{k_m}$, $\Theta^{k_m} \subseteq \mathcal{R}^{k_m}$ とする．データ系列 x^n は真の分布 $p^*(\cdot) = p(\cdot|m^*, \theta^{k_{m^*}^*})$ に従って得られるものとする．m^* は真のモデル，$\theta^{k_{m^*}^*}$ は真のパラメータと呼ぶ．モデル選択の目的は，x^n に基づき $\mathcal{M}$ から m を選ぶこと，すなわち m の推定量 $\hat{m}$ を得ることである．

ここで，モデル選択が議論される現実的問題のほとんどは，全順序の階層構造，あるいは半順序の階層構造を持ったモデル族であることがわかる．例えば，ARモデルは全順序を持つ階層モデル族，重回帰モデルは半順序の階層モデル族に属する．以下にこれらを包含する広義の階層モデル族を定義する．

定義 7.1 (階層モデル族) モデル m で構成される確率分布のクラスを $\mathcal{H}^{k_m}$ とする．すなわち，

$$\mathcal{H}^{k_m} = \left\{ p(\cdot|m, \theta^{k_m}) \middle| \theta^{k_m} \in \Theta^{k_m} \right\}, \tag{7.2}$$

とする．さらにそれぞれのモデルに関し，$k_{m1} < k_{m2}$ となる $m1, m2 \in \mathcal{M}$ に対し，

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}}, \tag{7.3}$$

という入れ子構造が存在してもかまわないものとする．この階層構造は，$\mathcal{M}$ の中で全順序構造でも半順序構造でもよい．階層モデル族を $\mathcal{H}$

$$\mathcal{H} = \cup_{m \in \mathcal{M}} \mathcal{H}^{k_m}, \tag{7.4}$$

と定義する． □

本稿の定義した階層モデル族は，半順序の階層構造のモデル族を含む極めて一般的なモデル族を指している[1]．$H^*(m, \theta^{k_m})$ を $-\log p(\cdot|m, \theta^{k_m})$ の平均レート，すなわち

$$H^*(m, \theta^{k_m}) = -\lim_{n \to \infty} \frac{1}{n} E^*[\log p(X^n|m, \theta^{k_m})], \tag{7.5}$$

と定義する．ここで，$E^*[\cdot]$ は真の確率測度 $p^*(\cdot)$ に関する期待値である．また $D^*(p^* \| p_m^{\theta^{k_m}})$ を，

$$D^*(p^* \| p_m^{\theta^{k_m}}) = \lim_{n \to \infty} \frac{1}{n} E^* \left[\log \frac{p^*(X^n)}{p(X^n|m, \theta^{k_m})} \right], \tag{7.6}$$

で定義される $p(\cdot|m, \theta^{k_m})$ の $p^*(\cdot)$ に対する KL情報量とする．$-\infty < H^*(p^*) < \infty$ かつ $-\infty < H^*(p_m^{\theta^{k_m}}) < \infty$ であり，確率測度 $p(\cdot|m, \theta^{k_m})$ のゼロ集合(確率測度 0 を持つ集合)が $p^*(\cdot)$ のそれと等しければ，

$$D^*(p^* \| p_m^{\theta^{k_m}}) = H^*(m, \theta^{k_m}) - H^*(p^*), \tag{7.7}$$

が成り立つ．ただし，

$$H^*(p^*) = -\lim_{n \to \infty} \frac{1}{n} E^*[\log p^*(X^n)], \tag{7.8}$$

である．さらに，$\forall m \in \mathcal{M}$ に対して，最適パラメータ $\theta^{k_m^*}$ を

$$\theta^{k_m^*} = \arg\min_{\theta^{k_m}} H^*(m, \theta^{k_m}), \tag{7.9}$$

と定義する．データ系列 x^n は真の分布 $p^*(x^n)$ に従って発生するものとするが，ここでは選択誤り率を考えるため，真の分布は確率モデル族に含まれるものと

[1]本章の解析結果は，入れ子構造を全く持たない分離的なモデル族，すなわち $\forall m1, \forall m2 \in \mathcal{M}$ に対して $\mathcal{H}^{k_{m1}} \cap \mathcal{H}^{k_{m2}} = \phi$ (空集合) の場合，さらには階層構造と分離構造が混在するモデル族に対しても成り立つ．

する．ただし，真の分布がモデル族に含まれなくても，前章までの議論と同様に最適モデルを考え，最適モデルの選択できない誤り確率を議論すれば以下と全く同様の議論が成り立つ．階層モデル族では m^* を次のように定義する必要がある．すなわち，$\mathcal{H}^{k_{m^*}} \neq \mathcal{H}$ の場合に $p(x^n|m,\theta^{k_m}) = p(x^n|m^*,\theta^{k^*_{m^*}})$ を満たす $m(\neq m^*)$ と θ^{k_m} が存在してしまうので，これまでの研究と同様に次のように定義する [56, 57].

定義 7.2 (真のモデル m^*) 真のモデル m^* は

$$m^* = \arg_m \min\Big\{k_m \Big| \exists \theta^{k_m}, D^*(p^* \| p_m^{\theta^{k_m}}) = 0\Big\}, \tag{7.10}$$

で定義する． □

その結果，$\mathcal{H}^{k_m} \subset \mathcal{H}^{k_{m^*}}$ に対しては

$$\forall \theta^{k_m} \in \Theta^{k_m}, \quad D^*(p^* \| p_m^{\theta^{k_m}}) > 0, \tag{7.11}$$

が成り立つ．

一般的には，$D^*(p^* \| p_m^{\theta^{k_m}}) = 0$ であるからといって，$\forall x^n, n = 1, 2, \cdots$ に対して必ずしも $p(x^n|m,\theta^{k_m}) = p^*(x^n)$ とは限らない．逆に，もし $\forall x^n, n = 1, 2, \cdots$ に対して $p(x^n|m,\theta^{k_m}) = p^*(x^n)$ であれば $D^*(p^* \| p_m^{\theta^{k_m}}) = 0$ は成り立つ．しかし，現実的応用で必要となってくるモデル族に関しては，$D^*(p^* \| p_m^{\theta^{k_m}}) = 0$ は $p(x^n|m,\theta^{k_m}) = p^*(x^n)$ と等価であるといえる．したがって，本章では $D^*(p^*_{m^*} \| p_m^{\theta^{k_m}}) = 0$ は $\forall x^n, n = 1, 2, \cdots$ に対して $p(x^n|m,\theta^{k_m}) = p^*(x^n)$ を意味するものとする．このとき，(7.3)式, (7.10)式 より，$\mathcal{H}^{k_{m^*}} \subset \mathcal{H}^{k_m}$ を満たす m に対し，

$$p(x^n|m,\theta^{k^*_m}) = p(x^n|m^*,\theta^{k^*_{m^*}}), \tag{7.12}$$

が成り立つ．

具体的なモデル選択問題の一例として，品質管理の分野で特によく研究されている母集団(情報源)パラメータの変化位置検出問題を示す[156][2].

例 7.1 (確率モデルのパラメータ変化点検出問題) ある特性値 (説明変数)が何水準(カテゴリ)かに分かれており，目的変数 X の確率分布がそれぞれのカテゴリで条件付けられているモデル族を考える[40]．そのカテゴリ間で分布のパラメー

[2]このような問題の例として，薬効問題がある．ある薬品の投入量に対して，その薬の効果がどのように変化するかを調べために，投入量を何段階かに分けて実験を行い解析する問題である．

タが変化している可能性があり，どのカテゴリ間でパラメータが変化しているかをデータから推定するものとする．この問題でパラメータに変化があるカテゴリ間の位置とそれぞれのパラメータの組は，一つの確率モデルととらえることができ，半順序を持つ階層モデル族からのモデル選択問題として定式化することができる．

r 水準のカテゴリを $C_1, C_2, \ldots, C_r$とする．X の確率分布は，クラス $\{p(x|\theta^k)|\theta^k \in \Theta^k\}$ に属するものとする．ここで，θ^k は k 次元パラメータであり，このパラメータは，カテゴリ間で変化する可能性がある．問題は，全ての水準から得られたデータに基づいて，パラメータの変化位置とそれぞれのパラメータ値を推定することである．

変化の数が $s-1$ の場合のパラメータの変化位置を $\tau_1, \tau_2, \cdots, \tau_{s-1}$ で表す．ただし，$s \in \{1, \cdots, r\}$である．全ての変化位置をモデルと呼び，インデックス m で表す．$s \geq 2$ に対し

$$m = \left\{\tau_1, \tau_2, \cdots, \tau_{s-1}\right\}, \tag{7.13}$$

とし，$s = 1$ のとき $m = \phi$(変化なし)とする．変化回数 $s-1$ のモデルのパラメータ数は $k_m = sk$ である．m の集合を $\mathcal{M}$， s の集合を $S = \{0, 1, \cdots, r\}$，s を固定したときの m の集合を $T(s)$，モデル m のパラメータ集合を Θ_m，モデル m の次元(パラメータ数)を $k_m = sk$ とする．このとき，全モデルクラスは，

$$\mathcal{H} = \left\{p(\cdot|m, \theta^{sk}) \middle| s \in S, m \in T(s), \theta^{sk} \in \Theta_m\right\}, \tag{7.14}$$

で与えられる．変化位置検出問題は，クラス $\mathcal{M}$ から一つの確率モデル m を選択する問題と等価である．ただし，$\mathcal{M} = \{m|s \in S, m \in T(s)\}$ 図7.1に $\theta = \theta^1$ の場合の，カテゴリ C_2 と C_3 の間でパラメータが変化しているモデルの例を示す．このモデルではパラメータ数は 2 となり，パラメータ空間は $\mathcal{R}^2$ で与えられる．
□

7.3 ベイズ決定理論による定式化

これまでモデル選択問題に対して主に適用されている 0-1損失は，選択したモデルが真のモデルであるとき損失 0，それ以外のモデルであるとき損失 1 を与えたものであり，ベイズ的に選択誤り率を最小化する規準を導くという点で一つの客観的な基準を与えている．しかし，適用場面によっては誤り率による評価だけでなく，モデル間に何らかの擬距離構造を考えることができる場合があり，なるべく真のモデルに近いモデルを選択するための規準も有用と考えられる [87].

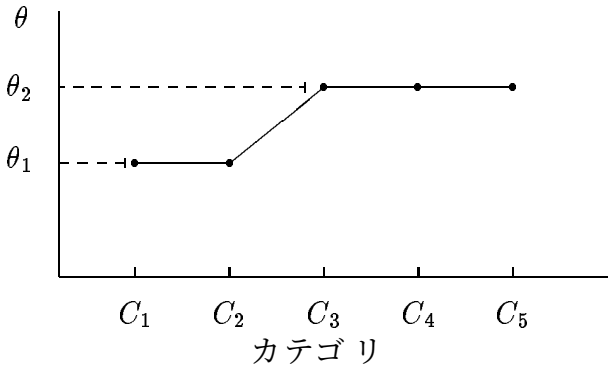

図 7.1: パラメータ変化の例

また，真のモデルよりも次数を少なく見積もる誤り(過小評価: underestimation)より，多く見積もる誤り(過大評価: overestimation)を避けたいという状況や，この逆の状況も考えられる [156][3]．そこで本稿では，このような要求に応じたベイズ規準を検討し，これらのモデル間の距離構造も表現できる一般的な損失関数を提案，この損失関数に対するベイズ最適なモデル選択を定式化する．

0-1損失はモデル間の距離として，同じモデルには 0，そうでないときは 1 を考えていることになるが，これを含み，より一般的なモデル間の距離を表現できる形として，次の損失関数を定義する．

$$L(m:Am) = \begin{cases} C_a & \text{if } \ Am = m \\ a(m,Am) & \text{if } \ Am \neq m \end{cases} \tag{7.15}$$

ここで，$a(m,Am)$ は $a(m,Am) > C_a$, かつ $\max_{m,Am} a(m,Am) < \infty$ を満たす関数，Am は決定関数を表す．$a(m,Am)$ は C_a より大の値域をとる有界関数であれば任意であるので，様々な形をとることができる．ここで，$C_a = 0$ としても一般性を失わない．よって，以下では $C_a = 0$ として取り扱う．例えば，真のモデルからパラメータ数が離れる程大きな損失を与えることを考えるならば，$C_a = 0$, $a(m,Am) = |k_m - k_{Am}|^2$ などと定義することができる．また，$C_a = 0, m \neq Am$ のとき $a(m,Am) = 1$ とすれば，0-1損失となる．その他の具体的に損失を与える方法については [87, 156]と7.5.2節を参照．

本研究では，モデル m の選択のみを議論しているので，パラメータ推定に関する損失を議論していない．しかし，現実的応用を考えると次のような m と

[3]過小評価と過大評価は，分離の誤り，非分離の誤りと呼ぶこともある [123].

θ^{k_m} の推定に対する損失関数の設定が可能である．このとき，以下の議論で示すように m の決定はパラメータの決定とは独立して行うことができ，m のみを議論することの正当性が示される．パラメータ推定については，問題に応じて二乗誤差損失など適切な損失関数を設定して決定することができ，これは本章の議論に影響を与えない．

まず，$p(\cdot|m,\theta^{k_m})$ に対して，決定 $Am, A\theta^{k_m}$ を与えたときの損失関数を

$$L(m,\theta^{k_m}:Am,A\theta^{k_m})=\begin{cases} a_\theta(\theta^{k_m},A\theta^{k_m}) & \text{if } \ Am=m, \\ a_m(m,Am) & \text{if } \ Am\neq m, \end{cases} \tag{7.16}$$

と定義する．ここで $a_\theta(\theta^{k_m},A\theta^{k_m})$ と $a_m(m,Am)$ は $\sup a_\theta(\cdot,\cdot)\leq a_m(\cdot,\cdot)<\infty$ を満たす任意の関数とし，$Am\in\mathcal{M}$ と $A\theta^{k_m}=(A\theta_1,A\theta_2,\cdots,A\theta_{k_m})^T$ は決定関数である[14]. $a_\theta(\theta^{k_m},A\theta^{k_m})$, $a_m(m,Am)$ は任意であるので，現実の応用を考えて適切な損失関数を設定することができる．これらの損失関数は応用を考慮して決められるべきである．

前節で述べたように，ベイズ最適解は次式の条件付きベイズリスクを，Am と $A\theta^{k_m}$ に関して最小化することによって与えられる[14].

$$BL(Am,A\theta^{k_m})=\sum_{m\in\mathcal{M}}\int_{\theta^{k_m}\in\Theta^{k_m}}L(m,\theta^{k_m}:Am,A\theta^{k_m})f(\theta^{k_m}|m,x^n)P(m|x^n)d\theta^{k_m}, \tag{7.17}$$

ここで，$L(m:Am)$, $BL(Am)$, $BL(A\theta^{k_m}|m)$ を

$$L(m:Am)=\begin{cases} 0 & \text{if } \ Am=m, \\ a_m(m,Am) & \text{if } \ Am\neq m, \end{cases} \tag{7.18}$$

$$BL(Am)=\sum_{m\in\mathcal{M}}L(m,Am)P(m|x^n), \tag{7.19}$$

$$BL(A\theta^{k_m}|m)=\int_{\theta^{k_m}\in\Theta^{k_m}}a_\theta(\theta^{k_m},A\theta^{k_m})f(\theta^{k_m}|x^n,m)d\theta^{k_m}, \tag{7.20}$$

と定義しよう．このとき，次の定理が成り立つ．

定理 7.1 $BL(Am,A\theta^{k_m})$ を最小化するベイズ最適なモデル $\hat{m}^B_{DT}$ とパラメータ $\hat{\theta}^{k_m}_{DT}$ は

$$\hat{m}^B_{DT}=\arg_{Am}\min BL(Am), \tag{7.21}$$

$$\hat{\theta}^{k_m}_{DT}=\arg_{A\theta^{k_m}}\min BL(A\theta^{k_m}|m)\Big|_{m=\hat{m}^B_{DT}}, \tag{7.22}$$

で与えられる．

(証明) $0 \leq a_\theta(\cdot,\cdot) < a_m(\cdot,\cdot)$ という定義より，(7.21)式は明らか．$\hat{m}^B_{DT}$ が与えられたもとでは，ベイズ最適なパラメータ推定は(7.22)式で与えられる． □

この定理より，モデル m の推定をパラメータ推定と独立に考え，モデル m の推定のみを考えれば，(7.15)式の損失に関して モデル m の選択を定式化すればよいことがわかる．(7.16)式のパラメータ推定の損失に対し，二乗誤差損失を仮定した場合について以下の系を与えておく．

系 7.1 (7.16)式において，二乗誤差損失 $a_\theta(\theta^{k_m}, A\theta^{k_m}) = \frac{1}{C}\sum_j^{sk}(A\theta_j - \theta_j)^2$ を与えた場合，ベイズ最適な推定量 $\hat{\theta}^{k_m}_{DT}$ は $\hat{m}^B_{DT}$ が与えられたもとで

$$\hat{\theta}^{k_m}_{DT} = \int_{\theta^{k_m}\in\Theta^{k_m}} \theta^{k_m} f(\theta^{k_m}|x^n, m) d\theta^{k_m}\Big|_{m=\hat{m}^B_{DT}}, \tag{7.23}$$

となる．

(証明) 文献[14]を参照．

(7.15)式の損失関数のもとで，データ x^n に条件付けられたベイズ最適な決定は，

$$BL(Am) = \sum_{m\in\mathcal{M}} L(m:Am)P(m|x^n), \tag{7.24}$$

の最小化によって与えられ，$BL(Am)$ を最小化するベイズ最適なモデル $\hat{m}^B_{DT}$ は，

$$\hat{m}^B_{DT} = \arg\min_{Am\in\mathcal{M}} \sum_m L(m:Am)P(m|x^n), \tag{7.25}$$

となる．ただし，$P(m|x^n)$ はモデル m の事後確率であり，

$$P(m|x^n) = \frac{p(x^n|m)P(m)}{\sum_{m\in\mathcal{M}} p(x^n|m)P(m)}, \tag{7.26}$$

$$p(x^n|m) = \int_{\theta^{k_m}\in\Theta^{k_m}} p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{7.27}$$

で与えられる．$P(m)$ はモデル m の事前確率，$f(\theta^{k_m}|m)$ は θ^{k_m} の事前確率密度である．$m \neq Am$ のとき $a(m,Am) = 1$ とした0-1損失を適用した場合には，$\hat{m}^B_{DT}$ は

$$\hat{m}^B_{DT} = \arg\max_{m\in\mathcal{M}} \Big\{\log \int_{\theta^{k_m}\in\Theta^{k_m}} p(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m} + \log P(m)\Big\}, \tag{7.28}$$

で与えられる．

7.4 漸近的性質の解析

本章では，広いクラスのモデル族に関して，前章で定式化したベイズ最適なモデル選択の漸近的な特性を解析する．そのために，まず2つの情報行列 $I^*(\theta^{k_m}|m), J^*(\theta^{k_m}|m)$ を定義する[4]

$$I^*(\theta^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E^* \left[\frac{\partial \log p(X^n|m,\theta^{k_m})}{\partial \theta^{k_m}} \frac{\partial \log p(X^n|m,\theta^{k_m})}{(\partial \theta^{k_m})^T} \right], \tag{7.29}$$

$$J^*(\theta^{k_m}|m) = -\lim_{n\to\infty} \frac{1}{n} E^* \left[\frac{\partial^2 \log p(X^n|m,\theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} \right]. \tag{7.30}$$

7.4.1 仮定

まず，ベイズ理論で設定される事前分布に関して以下を仮定する．

条件 7.1

(1) $\forall m \in \mathcal{M}$ に対して，$P(m) > 0$.

(2) $\forall m \in \mathcal{M}, \forall \theta^{k_m} \in \Theta^{k_m}$ に対し，$f(\theta^{k_m}|m) > 0$，かつ $f(\theta^{k_m}|m)$ は θ^{k_m} に関して3階微分可能とする．

また対象とする階層モデル族に関して次を仮定する．

条件 7.2

(1) (有界性) $\forall \theta^{k_m} \in \Theta^{k_m}, \forall m \in \mathcal{M}$ に対して，$0 < \det I^*(\theta^{k_m}|m) < \infty$ かつ $0 < \det J^*(\theta^{k_m}|m) < \infty$ とする．また，$\forall m \in \mathcal{M}, \forall \theta^{k_m} \in \Theta^{k_m}$ に対して $|H^*(m,\theta^{k_m})| < \infty$ であるとする．

(2) (連続性) $\forall \theta^{k_m} \in \Theta^{k_m}, \forall m \in \mathcal{M}$ に対して，$I^*(\theta^{k_m}|m), J^*(\theta^{k_m}|m)$ は Θ^{k_m} 上で θ^{k_m} について微分可能とする．

(3) (健全性) パラメータ空間上に半径 $\delta > 0$ の球 $B_\delta(\theta^{k_m^*}) = \{\theta^{k_m} \in \Theta^{k_m} | \|\theta^{k_m} - \theta^{k_m^*}\| < \delta\}$ を定義する．このとき，$\forall m \in \mathcal{M}$ と $\forall \delta > 0$，条件7.1を満たす事前密度関数 $f(\theta^{k_m}|m)$ に対して，事後密度関数は

$$\int_{\theta^{k_m} \in B_\delta(\theta^{k_m^*})} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 1, \quad a.s. \tag{7.31}$$

を満たす．

[4] $p^*(\cdot) \in \mathcal{H}^{k_m}$ のとき $J^*(\theta^{k_m^*}|m) = I^*(\theta^{k_m^*}|m)$ となる．

(4) (漸近正規性) $\forall m \in \mathcal{M}$ に対して，最尤推定量 $\hat{\theta}^{k_m} = \hat{\theta}^{k_m}(x^n)$ は漸近正規性を満たす [74, 131]．すなわち，$\zeta^{k_m} = \sqrt{n}(\hat{\theta}^{k_m}(x^n) - \theta^{k_m^*})$ の分布は平均 $\mathbf{0}$，共分散 $\{K^*(\theta^{k_m^*}|m)\}^{-1}$ の k_m-次元正規分布に分布収束する[5]．ここで，$\{K^*(\theta^{k_m}|m)\}^{-1}$ は

$$\{K^*(\theta^{k_m}|m)\}^{-1} = \{J^*(\theta^{k_m}|m)\}^{-1} I^*(\theta^{k_m}|m) \{J^*(\theta^{k_m}|m)\}^{-1}. \tag{7.32}$$

で与えられる行列とする．

(5) (重複対数の法則) $\forall m \in \mathcal{M}$, $\forall \theta^{k_m} \in \Theta^{k_m}$ に対して，次の3つの推定量は重複対数の法則を満たす:

$$\hat{\theta}^{k_m} = \theta^{k_m^*} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{7.33}$$

$$-\frac{1}{n}\log p(x^n|m,\theta^{k_m}) = H^*(p_m^{\theta^{k_m}}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{7.34}$$

$$-\frac{1}{n}\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = J^*(\theta^{k_m}|m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right), \quad a.s. \tag{7.35}$$

□

条件7.2,(3)の健全性は，n が大きくなると，事後確率は真のパラメータの近傍に集中してくることを述べており，例えば[16, 22]において同様の仮定が示されている．Θ^{k_m} が有界で Fisher 情報行列が退化していなければ健全性は満足される．しかし，有界でない場合でも，正規分布族と共役事前分布などを当てはめればすぐ明らかなように，通常有用な分布族はこの条件を満たすものとなっている．これらの分布族を仮定した例7.1のモデルが条件7.2,(3)を満たすことも明らかである．条件7.2,(3)の必要条件に関しては[16]参照．

条件7.2,(4)の最尤推定量の漸近正規性と条件7.2,(5)の重複対数の法則については，最も基本的な真のパラメータがパラメータ空間内に存在する場合について[34]で詳しく議論されている．モデル選択で必要となってくるように，真の分布を表現するパラメータが集合内にない場合(真のモデルより次数の低いモデルでは真の分布を表現できないので，このようなケースが生じる)については，

[5]厳密には，R_ζ を任意の直方体として，測度 $p^*(\zeta^{k_m})$ にもとづく確率が [74, 131]

$$P^*\left\{\zeta^{k_m} \in R_\zeta\right\} = \int_{\zeta^{k_m} \in R_\zeta^n} p^*\left\{d\zeta^{k_m}\right\} \to \frac{\sqrt{\det K(\theta^{k_m^*}|m)}}{(2\pi)^{k_m/2}} \int_{\zeta^{k_m} \in R_\zeta} e^{-\frac{1}{2}||\zeta^{k_m}||^2_{K^*(\theta^{k_m^*}|m)}} d\zeta^{k_m},$$

を満たすことを仮定している．

[74, 131]を参照．これらの仮定は，実務的に必要となる多くのモデル族で満足されることがわかる．

例えば，二項分布族は，そのパラメータ(生起確率)の範囲を $(0,1)$ とすれば，上の条件7.2を満たすことがわかり[34, 74]，例7.1に対して二項分布を仮定した場合のパラメータの変化時点検出問題は，この二項分布の組み合わせで与えられることから，明らかに条件7.2を満たす．正規分布に対するパラメータ変化時点検出問題においても，同様に条件7.2はみたされる．また，正規雑音を仮定した重回帰モデル，ARモデルも条件7.2は成り立つ[34, 74]．

条件7.2では，第6章までと異なり，$\theta^{k_m^*}$ が Θ^{k_m} の内部に一意に存在することを陽に仮定していない．しかし，これは (4)の漸近正規性の仮定に含まれている．すなわち，$\theta^{k_m^*}$ が Θ^{k_m} の内部に唯一存在しなければ，最尤推定量の漸近正規性は成り立たない．

7.4.2 事後確率の漸近正規性

本節では，ベイズ規準に基づくモデル選択の漸近解析に本質的となるパラメータの事後確率密度の漸近正規性[16]を示す．具体的には条件7.1と条件7.2のもとで，以下の定理が成り立つ．

補題 7.1 条件7.1, 条件7.2のもとで，$\forall m \in \mathcal{M}$ に対し，以下の式が成り立つ．$\xi^{k_m} = \sqrt{n}(\theta^{k_m} - \hat{\theta}^{k_m})$ の事後確率密度は次式を満たす．

$$f_\xi(\hat{\xi}^{k_m}|x^n, m) \to \frac{\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{(2\pi)^{k_m/2}}, \quad a.s. \tag{7.36}$$

ただし，$\hat{\xi}^{k_m} = 0$ は $\hat{\theta}^{k_m}$ に対応する ξ 空間上の原点，$f_\xi(\xi^{k_m}|x^n, m)$ は θ^{k_m} を変換した ξ^{k_m} 空間上の事後確率密度であり，

$$f_\xi(\xi^{k_m}|x^n, m) = \frac{1}{\sqrt{n}^{k_m}} f(\theta^{k_m}|x^n, m). \tag{7.37}$$

で与えられる．

(証明) 第6章の補題6.1と同様に証明できる． □

[16],pp.287 - 294 は，事後確率最大推定量を用いた漸近正規性の議論をしているが，補題7.1は最尤推定量を用いた漸近展開となっている．最尤推定量を用いた展開をすることにより，モデル選択問題の漸近的性質の解析において，すでによく知られている最尤推定量の漸近性質を用いた議論が可能となる．また，

定理7.1は事後確率密度 $f_\xi(\hat{\xi}^{k_m}|x^n,m)$ の漸近式のみを示しているが，分布の収束についても同様に示すことが可能である．本章の以下の議論では，$f_\xi(\hat{\xi}^{k_m}|x^n,m)$ の漸近式のみが使われる．

7.4.3 モデル選択の誤り率の上界

ここでは，条件7.1, 条件7.2のもとで，$BL(Am, A\theta^{k_{Am}})$ を最小化するモデル選択の m に関する選択誤り率の上界を導出する．類似した上界式は，木構造と多項分布で表現される確率モデル族に対し情報量規準型のモデル選択を行った場合について，J.Suzuki [125]により解析されている．

定理 7.2 条件7.1, 7.2のもとで，真のモデルの選択誤り確率 P_e^* は，次のように与えられる．

$$P_e^* \leq O\left(\frac{1}{(1-\log n)^2}\right), \tag{7.38}$$

(証明) 7.6.1節参照. □

本章での結果は Suzukiの解析 [125]を拡張したものとなっており，結果は当然ながら，情報量規準のペナルティ項を $\log n$ とした場合と同じ上界式になっている．

定理7.2は

$$P_e^* \leq O\left(\frac{1}{(\log n)^2}\right), \tag{7.39}$$

であることを述べており，これは $BL(Am)$ を最小化するモデル選択は漸近弱一致性を持つこと，すなわち，

$$\hat{m}_{DT}^B \to m^*, \ in\ probability, \tag{7.40}$$

を示している．本章では重複対数の法則が成り立つクラスを考えたが，第6章，および[57]では同様の条件のもとで，BIC や MDL が強一致性を持つことを示しており，本稿ではこのより強い結果を導くには至っていない．したがって，より厳しい上界式が存在する可能性があり，これは今後の課題である．

7.5 確率モデルのパラメータ変化点検出問題への適用

7.5.1 事後確率の計算式

本節では，例7.1のパラメータの変化時点検出問題において，値 0 を確率 θ で，値 1 を確率 $1-\theta$ で出力する二項分布族を考え $(k=1)$，ベイズ解の計算式を示

す．具体的応用では，例えば 0 は正常，1 は異常，などと 0, 1 を何らかの事象と対応づけることが多い．

n_i をカテゴリ C_i のサンプル数，y_i をそのうちの 0 の個数とする．$x_i = (y_i, n_i - y_i)$ である．このとき s 次元パラメータを持つモデル $m = \{\tau_1, \tau_2, \cdots, \tau_{s-1}\}$ の尤度関数は，

$$P(x^n|m, \theta^s) = \prod_{j=1}^{s} \theta_j^{\sum_{i=\tau_{j-1}+1}^{\tau_j} y_i} (1-\theta_j)^{\sum_{i=\tau_{j-1}+1}^{\tau_j} (n_i - y_i)}, \tag{7.41}$$

で与えられる．ただし，$\tau_{-1} = 0$ である．ここで，θ_j の事前分布として共役事前分布であるベータ分布

$$f(\theta_j) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}, \tag{7.42}$$

を仮定すれば，

$$\log \int_{\theta^{sk} \in \Theta^{sk}} P(x^n|m, \theta^{sk}) f(\theta^{sk}|m) d\theta^{sk} + \log P(m)$$

$$= \sum_{j=1}^{s} \log \frac{\Gamma(\sum_{i=\tau_{j-1}+1}^{\tau_j} y_i + \alpha)\Gamma(\sum_{i=\tau_{j-1}+1}^{\tau_j} (n_i - y_i) + \beta)}{\Gamma(\sum_{i=\tau_{j-1}+1}^{\tau_j} n_i + \alpha + \beta)} + s \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} + \log P(m), \tag{7.43}$$

となる．したがって，(7.43)式を使って $BL(Am, A\theta_{Am})$ を計算すれば，ベイズ最適なモデル選択が可能となる．0-1損失の場合は，(7.43)式を最小化するモデルがベイズ最適である．

7.5.2 選択誤りのトレードオフを考慮したモデル推定

本節では有用な損失関数の一つとして，

$$a_\theta(A\theta^{k_m}, \theta^{k_m}) = \frac{1}{C} \sum_{j}^{sk} (A\theta_j - \theta_j)^2, \tag{7.44}$$

$$a_m(Am, m) = \begin{cases} a_0 & \text{if } \ Am \neq m, k_{Am} = k_m, \\ a_1 |k_m - k_{Am}| & \text{if } \ k_{Am} < k_m, \\ a_2 |k_{Am} - k_m| & \text{if } \ k_{Am} > k_m, \end{cases} \tag{7.45}$$

を考える．ここで C は $C \geq \sup \sum_j (A\theta_j - \theta_j)^2$ を満たす定数，a_0, a_1, a_2 は $1 \leq a_0, a_1, a_2 < \infty$ を満たす任意の重み(定数)である．パラメータに対する二乗誤差損失は一般に広く適用されている有用な損失の一つである．(7.45)式はモデル間の次元の絶対誤差損失 $a_m(Am, m) = |k_{Am} - k_m|$ に重みを考慮した一般化に

なっている．重み a_1 と a_2 はそれぞれ，2つの方向の誤り，すなわち過大評価(overestimation)と過小評価(underestimation)に対する損失の大きさを表している．重み a_1 と a_2 を目的に合わせて設定することにより，過大評価と過小評価の一方の誤り率を小さくすることが可能である．このような要求は，実際の問題において，例えば変化がある場合には必ず検出したいという場合や，逆に変化がないのにあると誤判断してしまうことは避けたいというような場合に生じる[6]．

7.5.3 シミュレーション実験

この損失関数を適用した場合のモデル選択の挙動を認識するために，7.5.1節のモデルに対してシミュレーション実験を行った．モデルの選択のみを議論し，損失としては(7.45)式を考える．本章の結果より，モデル選択の誤り率は 0 に収束することが予想されるが，シミュレーションによりその速度がどの程度かを把握することができる．

m,θ_j の事前分布はともに一様分布とし，$r = 5$, $m^* = \{\tau_1, \tau_2\} = \{2, 3\}$ とした．真のパラメータは $\theta(S_1) = \theta(S_2) = 0.2$, $\theta(S_3) = 0.5$, $\theta(S_4) = \theta(S_5) = 0.8$ とする．$a_0 = 1$ とし，以下の 5つのケースを実験した．

1) $a_1 = 2$, $a_2 = 1$,

2) $a_1 = 1.5$, $a_2 = 1$,

3) $a_1 = a_2 = 1$,

4) $a_1 = 1$, $a_2 = 1.5$,

5) $a_1 = 1$, $a_2 = 2$.

サンプル数を変え，それぞれのサンプル数で 10000回の繰り返しを行い，モデル選択の選択誤り率を計算した．

$\hat{m}^B_{DT} \neq m^*$ となる選択誤り率の推移を図7.2に示す．この図より，誤り率は 0 に収束していく様子が伺え，(7.40)式の弱一致性が確認された．n が小さいときには，選択誤り率の逆転現象が起こっているが，n が大きくなるに従って規則的に 0 に収束していくことが分かる．

2種類の方向の誤りの確率の推移を示したのが，図7.3と図7.4である．これらの図より，よりシンプルなモデルを選ぶ確率を小さくすると複雑なモデルを選

[6]これは仮説検定において 2 方向の誤りを考え，それらのトレードオフの関係を議論することと同様の議論である．

図 7.2: モデル選択の選択誤り率の推移

図 7.3: 真のモデルより次元数の小さいモデルを選択した確率

択する確率が大きくなってしまい，逆も成り立つことがわかる．すなわち，重み a_1, a_2 を変えることによって，誤りの方向のトレードオフを取り扱うことができることがわかる．しかし，よりシンプルなモデルを選択してしまう確率はサンプル数の増加と共に急速に減少することがわかる．これは，モデル間で次元数に差があるという，統計的モデル選択特有の問題設定の本質的な性質である．図7.2において，n が小さいときに逆転現象が起きるが，n が大きくなると規則的になるのは，このようにシンプルなモデルへの誤り率が速やかに 0 に収束していくためである．

図 7.4: 真のモデルより次元数の高いモデルを選択した確率

7.6 補題と定理の証明

7.6.1 定理 7.1 の証明

$BL(Am) = \sum_m L(m:Am)P(m|x^n)$ と定義すると，m^* が真でもあるにかかわらず，他のモデルを選択してしまう選択誤り確率 P_e^* は，

$$P_e^* = P^*\Big[x^n\Big|\exists m \neq m^* \;\; BV(m^*) \geq BV(m)\Big], \tag{7.46}$$

である．従って，$BL(m^*) \geq BL(m)$ となる事象の確率を評価して，ユニオンバウンドをとればよい．

ここで，$L_{max} = \max_{m,m'} L(m,m') = \max_{m,m'} a(m,m')$, $L_{min} = \min_{m,m'} a(m,m')$ と定義する．損失に関する前提より，$0 < L_{min} \leq L_{max} < \infty$ であることに注意．

まず，$L(m,m) = 0$ であることから

$$\begin{aligned} BL(m^*) &= \sum_{m' \neq m^*} L(m':m^*)P(m'|x^n) \\ &= \sum_{m' \neq m^*, m} L(m':m^*)P(m'|x^n) + L(m:m^*)P(m|x^n) \end{aligned} \tag{7.47}$$

$$\begin{aligned} BL(m) &= \sum_{m' \neq m} L(m':m)P(m'|x^n) \\ &= \sum_{m' \neq m, m^*} L(m':m)P(m'|x^n) + L(m^*:m)P(m^*|x^n) \end{aligned} \tag{7.48}$$

が成り立つことが分かる．従って，

$$BL(m^*) - BL(m)$$

$$
\begin{aligned}
&= L(m:m^*)P(m|x^n) - L(m^*:m)P(m^*|x^n) \\
&\geq L_{min}P(m|x^n) - L_{max}P(m^*|x^n) \qquad (7.49)
\end{aligned}
$$

が導かれるので，$C = \frac{L_{min}}{L_{max}}$ として

$$
CP(m|x^n) \geq P(m^*|x^n) \Rightarrow BL(m^*) \geq BL(m) \qquad (7.50)
$$

である．

従って，

$$
P^*\left[BL(m^*) \geq BL(m)\right] \leq P^*\left[CP(m|x^n) \geq P(m^*|x^n)\right] \qquad (7.51)
$$

と上界できるので，ユニオンバウンドをとって

$$
\begin{aligned}
P_e^* &\leq P^*\left[\exists m \neq m^*,\ \ CP(m|x^n) \geq P(m^*|x^n)\right] \\
&\leq \sum_{m \neq m^*} P^*\left[\log \frac{P(m^*|x^n)}{P(m|x^n)} \leq \log C\right] \qquad (7.52)
\end{aligned}
$$

という形で誤り確率が上界できる．

一方，補題7.1より，

$$
f(\hat{\theta}^{k_m}|x^n, m) = \frac{\sqrt{n}^{k_m}\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}{\sqrt{2\pi}^{k_m}} + o\left(\sqrt{n}^{k_m}\right),\ a.s. \qquad (7.53)
$$

が得られる．ベイズ規則

$$
p(x^n|m) = \frac{p(x^n|m, \hat{\theta}^{k_m})f(\hat{\theta}^{k_m}|m)}{f(\hat{\theta}^{k_m}|x^n, m)}, \qquad (7.54)
$$

より，(7.53)式を(7.54)式に代入すると，$\forall m \in \mathcal{M}$ に対して，

$$
\begin{aligned}
P(m|x^n) &\propto p(x^n|m)P(m) \\
&= \frac{p(x^n|m, \hat{\theta}^{k_m})f(\hat{\theta}^{k_m}|m)P(m)\sqrt{2\pi}^{k_m}}{\sqrt{n}^{k_m}\sqrt{\det J^*(\hat{\theta}^{k_m}|m)}}\{1 + o(1)\},\ \ a.s. \qquad (7.55)
\end{aligned}
$$

が得られる．(7.33)式より，$f(\hat{\theta}^{k_m}|m) = O(1),\ a.s.,\ \sqrt{\det J^*(\hat{\theta}^{k_m}|m)} = O(1),\ a.s.$ であるから，(7.52)式より，ある定数 C' が存在して，

$$
P_e^* \leq \sum_{m \neq m^*} P^*\left[\log \frac{P(x^n|m, \hat{\theta}^{k_m})n^{k_{m^*}/2}}{P(x^n|m^*, \hat{\theta}^{k_{m^*}})n^{k_m/2}} \geq C'\right], \qquad (7.56)
$$

が得られる．さらに，チェビシェフの不等式[8]より，

$$
P_e^* \leq \sum_{m \neq m^*} \frac{V^*\left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m^*, \hat{\theta}^{k_{m^*}})} + \frac{k_{m^*} - k_m}{2}\log n + C'\right]}{\left\{E^*\left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m^*, \hat{\theta}^{k_{m^*}})} + \frac{k_{m^*} - k_m}{2}\log n + C'\right]\right\}^2}, \qquad (7.57)
$$

$E^*[\cdot]$ と $V^*[\cdot]$ は $p(\cdot|m^*, \theta^{k^*_{m^*}})$ による期待値と分散である．

まず，$m \in \mathcal{M}$ に対する $\log \frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k^*_m})}$ を評価することから始める．テーラー展開により [56],

$$-\log p(x^n|m, \theta^{k^*_m}) = -\log p(x^n|m, \hat{\theta}^{k_m})$$

$$-\frac{1}{2}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T \left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} (\hat{\theta}^{k_m} - \theta^{k^*_m})$$

$$+ O\left(\left\|\hat{\theta}^{k_m} - \theta^{k^*_m}\right\|^3 \log p(x^n|m, \theta^{k_m+})\right), \tag{7.58}$$

が得られる．ここで，θ^{k_m+} は $\theta^{k^*_m}$ と $\hat{\theta}^{k_m}$ の間にある点である．(7.33)式より，ある定数 $C_1 > 0$ が存在して，$n \to \infty$ のとき，

$$\sqrt{n}\left\|\hat{\theta}^{k_m} - \theta^{k^*_m}\right\| \le C_1(\log\log n)^{1/2}, \quad a.s. \tag{7.59}$$

が成り立つ．この式と (7.35) 式より，ある定数 $C_2 > 0$ が存在して，$n \to \infty$ のとき，

$$\left|\frac{1}{2}\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})\right.$$
$$\left.+\frac{1}{2}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T \left.\frac{\partial^2 \log p(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}(\hat{\theta}^{k_m} - \theta^{k^*_m})\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}\right.$$
$$\le \frac{(\log\log n)^{3/2}}{\sqrt{n}}, \quad a.s. \tag{7.60}$$

が成り立つ．一方，(7.34)式と，$\theta^{k_m+} \to \theta^{k^*_m}, a.s.$ であることから，$\forall\epsilon > 0$ に対して正定数 $C_3 > 0$ が存在して $n \to \infty$ のとき，

$$\left|\log p(x^n|m, \theta^{k_m+})\right| \le nC_3 + n\epsilon, \quad a.s. \tag{7.61}$$

となり，したがって正定数 $C_4 > 0$ が存在して $n \to \infty$ のとき，

$$\left|\left\|\hat{\theta}^{k_m} - \theta^{k^*_m}\right\|^3 \log p(x^n|m, \theta^{k_m+})\right| \le C_4\frac{(\log\log n)^{3/2}}{\sqrt{n}}, \quad a.s. \tag{7.62}$$

が成り立つ．(7.58)式 と (7.60)式, 及び (7.62)式より，ある正定数 $C_5 > 0$ が存在して，$n \to \infty$ のとき，

$$\left|\log\frac{p(x^n|m,\hat{\theta}^{k_m})}{p(x^n|m,\theta^{k^*_m})} - \frac{1}{2}\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})\right|$$
$$\le C_5\frac{(\log\log n)^{3/2}}{\sqrt{n}}, \quad a.s. \tag{7.63}$$

が成り立つ．上の展開より，ある正定数 $C_6 > 0$ が存在して，$n \to \infty$ のとき，

$$\left| \log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \log \frac{p(x^n|m, \theta^{k^*_m})}{p(x^n|m^*, \theta^{k^*_{m^*}})} \right.$$

$$\left. - \frac{1}{2}\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m}) \right.$$
$$\left. + \frac{1}{2}\sqrt{n}(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}})^T J^*(\theta^{k^*_{m^*}}|m)\sqrt{n}(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}) \right|$$

$$\leq C_6 \frac{(\log\log n)^{3/2}}{\sqrt{n}}, \quad a.s. \tag{7.64}$$

が成り立つことがわかる．

まず，$p(\cdot|m, \theta^{k^*_m}) \neq p^*(\cdot)$ の場合を考える．この場合には，(7.35)式と(7.11)式から，正定数 $D_1 > 0, C_7 > 0$ が存在して，$n \to \infty$ のとき，

$$\left| \log \frac{p(x^n|m^*, \theta^{k^*_{m^*}})}{p(x^n|m, \theta^{k^*_m})} - nD_1 \right| \leq C_7\sqrt{n}(\log\log n)^{1/2}, \quad a.s. \tag{7.65}$$

が成り立つ．また(7.33)式,より，定数 $C_8 > 0$ が存在して，$n \to \infty$ のとき，

$$\left| \sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m)\sqrt{n}(\hat{\theta}^{k_m} - \theta^{k^*_m}) \right| \leq C_8 \log\log n, \quad a.s. \tag{7.66}$$

が成り立つ．(7.65)式と(7.66)式より，正定数 $D_2 > 0, C_9 > 0$ が存在して，$n \to \infty$ のとき，

$$nD_2 + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n - C_9\sqrt{n}(\log\log n)^{1/2}$$

$$\leq \log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n + C'$$
$$\leq nD_2 + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n + C_9\sqrt{n}(\log\log n)^{1/2}, \quad a.s. \tag{7.67}$$

が成り立つことがわかる．これまでの議論が，確率1での収束を扱っていることから，これらの期待値を取ることができて，$D_2 > 0$ に対して，$n \to \infty$ のとき，

$$nD_2 + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n - C_9\sqrt{n}(\log\log n)^{1/2}$$

$$\leq E^*\left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n + C'\right]$$
$$\leq nD_2 + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2}\right)\log n + C_9\sqrt{n}(\log\log n)^{1/2}, \tag{7.68}$$

が成り立つ．(7.67)式より，正定数 $C_{10} > 0$ が存在して，$\forall \epsilon > 0$ に対して $n \to \infty$ のとき

$$V^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n + C' \right] \leq C_{10} n \log \log n. \tag{7.69}$$

が成り立つ．(7.57)式, (7.68)式, (7.69)式より，$p(\cdot|m, \theta^{k^*_m}) \neq p^*(\cdot)$ を満たす m を選択してしまう誤り確率 $P^*_{e1}[m]$ は，

$$P^*_{e1}[m] \leq O\left(\frac{\log \log n}{n} \right), \tag{7.70}$$

とバウンドされる．

次に，$p(\cdot|m, \theta^{k^*_m}) = p^*(\cdot)$ の場合を考える．このときは

$$\log \frac{p(x^n|m, \theta^{k^*_m})}{p(x^n|m, \theta^{k^*_{m^*}})} = 0, \tag{7.71}$$

が成り立つので，ある正定数 $C_{11} > 0$ が存在して，$n \to \infty$ のとき，

$$\begin{aligned} &\left| \log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \frac{1}{2} \sqrt{n} (\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}})^T J^*(\theta^{k^*_{m^*}}|m) \sqrt{n} (\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}) \right. \\ &\left. - \frac{1}{2} \sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m) \sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m}) \right| \\ &\leq C_{11} \frac{(\log \log n)^{3/2}}{\sqrt{n}}, \quad a.s. \end{aligned} \tag{7.72}$$

ここで，$\lim_{n \to \infty} \frac{(\log \log)^{3/2}}{\sqrt{n}} = 0$ である．一方，$\hat{\theta}^{k_m}$ の漸近正規性より，

$$\lim_{n \to \infty} E^* \left[\sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m) \sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m}) \right] = \mathrm{Tr} \{ J^*(\theta^{k^*_m}|m) \}^{-1} I^*(\theta^{k^*_m}|m), \tag{7.73}$$

が成り立つ．また，$p(\cdot|m, \theta^{k^*_m}) = p^*(\cdot)$ のときは，$J^*(\theta^{k^*_m}|m) = I^*(\theta^{k^*_m}|m)$ が成り立つので，$\mathrm{Tr} \{ J^*(\theta^{k^*_m}|m) \}^{-1} I^*(\theta^{k^*_m}|m) = k_m$ となり，

$$\lim_{n \to \infty} E^* \left[\sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m) \sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m}) \right] = k_m, \tag{7.74}$$

が得られる．同様に，

$$\lim_{n \to \infty} V^* \left[\sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m})^T J^*(\theta^{k^*_m}|m) \sqrt{n} (\hat{\theta}^{k_m} - \theta^{k^*_m}) \right] = k_m, \tag{7.75}$$

となる．(7.72) 式は確率 1 で成り立つので，ある正定数 $C_{12} > 0$ が存在して，$n \to \infty$ のとき，

$$\left| E^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n + C' \right] \right.$$

$$\left. + \frac{k_{m^*}}{2} - \frac{k_m}{2} - \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n \right|$$

$$\leq C_{12} \frac{(\log \log n)^{3/2}}{\sqrt{n}}, \tag{7.76}$$

が成り立つ．このことは，$n \to \infty$ のときに

$$\left\{ \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) (1 - \log n) - C_{12} \frac{(\log \log n)^{3/2}}{\sqrt{n}} \right\}^2$$

$$\leq \left\{ E^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n + C' \right] \right\}^2$$

$$\leq \left\{ \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) (1 - \log n) + C_{12} \frac{(\log \log n)^{3/2}}{\sqrt{n}} \right\}^2, \quad a.s. \tag{7.77}$$

であることを意味している．一方，

$$V^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n + C' \right] = V^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} \right]. \tag{7.78}$$

であることから，$C_{13} > 0$ が存在して $n \to \infty$ のとき，

$$V^* \left[\log \frac{p(x^n|m, \hat{\theta}^{k_m})}{p(x^n|m, \hat{\theta}^{k_{m^*}})} + \left(\frac{k_{m^*}}{2} - \frac{k_m}{2} \right) \log n + C' \right]$$

$$\leq k_m + k_{m^*} + 2k_m k_{m^*} + C_{13} \frac{(\log \log n)^{3/2}}{\sqrt{n}}, \tag{7.79}$$

が成り立つ．これは，

$$V[X - Y] \leq V[X] + V[Y] + 2V[X]V[Y], \tag{7.80}$$

から導かれる．(7.57) 式, (7.77) 式，(7.79) 式より，$p(\cdot|m, \theta^{k_m^*}) = p^*(\cdot)$ を満たす m^* 以外のある m を選択してしまう誤り確率 $P_{e2}^*[m]$ は

$$P_{e2}^*[m] \leq O\left(\frac{1}{(1 - \log n)^2} \right), \tag{7.81}$$

が導かれる．(7.70) 式と (7.81) 式より，(7.57) 式のようにユニオンバウンドをとれば定理が得られる． □

7.7 本章の結論

本章では，第6章で定式化したベイズ決定理論に基づく統計的モデル選択において，真のモデルを発見するという目的に対する一般的な損失関数を定義してベイズ最適な決定方法を示し，一般的な条件のもとで，この選択方法の選択誤り率の上界を導出した．その誤り率の上界は J.Suzuki が木構造のモデル族の情報量規準によるモデル選択に対して導出した結果を，ベイズ規準とより広いモデル族へ拡張したものとなっている．勿論，情報量規準によるモデル選択についても，木構造に限定しない階層モデル族に対して J.Suzuki の結果と同じ形の上界を導くことができる[41].

さらに，本章では現実問題への適用例として，確率モデルのパラメータ変化点検出問題を取り上げ，2方向の選択誤りを考慮した損失関数を設定し，シミュレーション実験を通じてその特性を評価した．これにより，損失関数の重みを変化させることにより，2方向の誤り率をトレードオフを制御できることが明らかとなった．また，真のモデルより次元の低いモデルへの選択誤り率は速やかに 0 に近づき，より次元の高いモデルへの誤り率は収束が遅いことが示されたが，これは階層構造のモデル族の本質的な性質である．このため，より次元の高いモデルへの誤りを抑えるように重みを設定すれば，全体として選択誤り率が減少することになる．実際の問題では解析の目的を考慮した上で，本章で示した特性も考慮して損失関数を適切に設定する必要がある．

本章の選択誤り率の上界の導出において議論した階層モデル族は，現実的応用で考えられる広いクラスの分布族を含むものとなっている．しかし，その手法はチェビシェフの不等式に基づくものであり，さらに厳しい上界が存在することが予想される．この点を踏まえ，さらに厳しい上界を示すことが今後の課題である．

第 8 章

一般の累積損失に対する拡張確率的コンプレキシティの漸近解析

8.1 本章の目的

J.Rissanen は統計的モデルの良さを測る尺度として 確率的コンプレキシティ (stochastic complexity: SC) [101] という概念を提案し，現在も広く研究が行われている．これは，与えられたデータ系列の情報量を，ある仮定した確率モデル族を用い，最小記述長の概念[95]で測ろうとしたものである．これは，あるパラメトリックなモデル族が与えられたときに，このモデルでデータ系列を符号化したときの最小記述長を考えることになる．もし，パラメータ空間上に事前確率が与えられているならば，この事前分布に対して平均的な意味で，平均符号長を最小化する符号化を考えればよい．これは J.Rissanen が 1987年に提案した SC の一つの形であり，いわゆるベイズ符号の符号長で与えられる．

SC では符号長を考えているので，対数損失を設定していることになる．最近，K.Yamanishi は学習問題などへの応用を踏まえ，SC の対数損失をより一般の損失関数へ拡張し，拡張確率的コンプレキシティ(extended stochastic complexity: ESC) という規準を提案した[153]．さらに，ESC の漸近評価を与え，ESC をバッチ的な学習問題とオンライン的逐次予測-学習問題に適用して，それらの学習の漸近性能を示している．しかし，K.Yamanishi の漸近解析では個々の全ての系列に対して一様に成り立つ ESCの漸近式の上界を与えているが，$o(\log n)$ 以下の項は評価できていない．ESCの平均値に関しても上界を与えているが，これらの下界が示されていない．しかし，ESC を適用した学習問題の評価を精密な形で与えるためにも，より詳細な項までの解析が必要であることが[154]においても指摘されている．

本研究では，データ系列に対して一様に成り立つ式を議論する代わりに，真

の情報源から出現する系列に関して成り立つ性質を考える．まず，ESC においても成り立つベイズ統計に似た興味深い性質，すなわちベイズ規則の一般化とオンライン型の ESC で定義されるパラメータ上の分布列に関する漸近正規性，という2つの性質を示す．さらに，この結果を用い，$o(1)$ の範囲内まで詳細な ESC の漸近式を，概収束と平均収束の意味で与える．

8.2 準備

8.2.1 確率的コンプレキシティ: SC

本章では K.Yamanishi の定式化に準じて，二つの確率変数の関係を考える問題を扱う．例えば回帰モデルの説明変数と目的変数がこの二つの変数にあたるが，説明変数を常に固定させれば，これは一つの確率変数列を考えていることになる．

$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ 上の確率変数を X, Y, Z とする．ここで，$Z = (X,Y)$ である．$z^n = z_1 z_2 \cdots z_n$ を長さ n の i.i.d. 系列とし，$z_i = (x_i, y_i)$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $z_i \in \mathcal{Z}$, $i = 1,2,\cdots,n$ とする．確率モデル族はパラメトリックな形で $\mathcal{P}^k = \{p_{\theta^k}(Y|X) : \theta^k \in \Theta^k\}$ と与えられ，θ^k は k次元パラメータ，Θ^k は k次元ユークリッド空間 $\mathcal{R}^k$ の部分集合とする．p は確率，あるいは確率密度を表す.

ここで，パラメータ空間上に事前密度 $\pi(\theta^k)$ [1] が与えられているものとする．

SC の形としてはいくつか提案されているが，ここでは次のものを考える[100].

$$SC(y^n|x^n) = -\log q(y^n|x^n), \tag{8.1}$$

$$q(y^n|x^n) = \int_{\theta^k \in \Theta^k} p(y^n|x^n,\theta^k)\pi(\theta^k)d\theta^k. \tag{8.2}$$

$-\log q(y^n|x^n)$ は x^n を受け取ったもとでの y^n の理想符号長を表している．ここで，

$$-\log q(y^n|x^n) = -\sum_{t=1}^{n} \log q(y_t|x_t, z^{t-1}), \tag{8.3}$$

$$q(y_t|x_t, z^{t-1}) = \int_{\theta^k \in \Theta^k} p(y_t|x_t,\theta^k)\pi(\theta^k|z^{t-1})d\theta^k, \tag{8.4}$$

と分解できる．ただし，ここでは $\pi(\theta^k|z^{t-1})$ はベイズの事後確率密度を表して

[1]前章まででは事前密度を $f(\theta^k)$ と与えていたが，ESC では後でデータをもとに (8.13) 式で事前分布を更新したとき，もはや $\pi(\theta^k|z^{t-1})$ はベイズの事後確率 $f(\theta^k|z^{t-1})$ とは異なるのでこれを明確にするため $\pi(\theta^k)$ と定義した．

おり，

$$\pi(\theta^k|z^{t-1}) = \frac{\prod_{i=1}^{t-1} p(y_i|x_i,\theta^k)\pi(\theta^k)}{\int_{\theta^k\in\Theta^k}\prod_{i=1}^{t-1} p(y_i|x_i,\theta^k)\pi(\theta^k)d\theta^k}, \tag{8.5}$$

で与えられる．このように，符号化の問題では，z^n を一括で符号化したときの符号長と逐次的に符号化した場合の全符号長が一致し，符号長の面からは両者を別々に取り扱う必要がない．

8.2.2 拡張確率的コンプレキシティ: ESC

まず最初に決定関数のクラス $\mathcal{D}$ を定義する．これは $\mathcal{D} = \{d_{\theta^k}(X)|\theta^k \in \Theta^k\}$ で表される関数 $d(\cdot)$ の集合，あるいは $\mathcal{D} = \{p(Y|X,\theta^k)|\theta^k \in \Theta^k\}$ で表される条件付き分布 $p(\cdot|\cdot,\theta^k)$ の集合であり，仮説集合と呼ぶ．θ^k は決定関数を特徴づける k 次元パラメータであり，k次元ユークリッド空間 $\mathcal{R}^k$ のコンパクト部分集合とする．

L を損失関数，$L(Y,D)$ を Y をある決定 D で予測したときの損失値とする．ここで，D は有限集合や $\mathcal{R}^1$ の部分集合，あるいは $\mathcal{Y}$ 上の確率や確率密度の集合の要素である．

仮説集合が実数値をとる関数の集合 $\mathcal{D} = \{d(X)\}$ で表されるとき，損失関数は $L(Y,d(X))$ で与えられる．例えば，α 損失 (α-loss function) は

$$L(Y,d(X)) = |Y - d(X)|^\alpha, \quad (\alpha \geq 1), \tag{8.6}$$

で与えられ，0-1 損失 (0-1 loss function) は

$$L(Y,d(X)) = \begin{cases} 0 & \text{if } Y = d(X) \\ 1 & \text{if } Y \neq d(X) \end{cases}, \tag{8.7}$$

で与えられる．

仮説集合が条件付き確率分布の集合 $\mathcal{D} = \{p(Y|X)\}$ で与えられるとき，損失関数は $L(Y,p(\cdot|X))$ で与えられる．例えば，対数損失 (logarithmic loss function) は

$$L(Y,p(\cdot|X)) = -\log p(Y|X), \tag{8.8}$$

で与えられる．$p(Y|X)$ が Y の条件付き確率，すなわち $P(Y|X)$ である場合の α 損失 (α-loss function) は

$$L(Y,P(\cdot|X)) = (1 - P(Y|X))^\alpha, \tag{8.9}$$

で与えられる．他の損失関数の例については [153] を参照．以後，記述の簡便性のため，$L(Y, d(X))$ も $L(Y, p(\cdot|X))$ も，$L(Z : d)$ と記述する．

$\lambda > 0$ をある定められた値とし，z^n の ESC は以下で定義される [153, 154].

$$ESC(z^n) = -\frac{1}{\lambda} \log \int_{\theta^k \in \Theta^k} \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : d_{\theta^k}) \right) \pi(\theta^k) d\theta^k. \tag{8.10}$$

ただし，d_{θ^k} は決定関数が連続パラメータ $\theta^k \in \Theta^k$ でラベル付けされていることを意味する．ESC に関しては，以下の性質が明らかにされている [153].

補題 8.1 $(\mathcal{P}, \mathcal{D}, \pi, L)$ が与えられたとき， 任意の z^n に対して ESC は次式を満たす.

$$ESC(z^n) = \sum_{t=1}^{n} \overline{L}(y_t|x_t, z^{t-1}), \tag{8.11}$$

ここで，

$$\overline{L}(y_t|x_t, z^{t-1}) = -\frac{1}{\lambda} \log \int_{\theta^k \in \Theta^k} \exp(-\lambda L(z_t : d_{\theta^k})) \pi(\theta^k|z^{t-1}) d\theta^k, \tag{8.12}$$

$$\pi(\theta^k|z^{t-1}) = \frac{\exp\left(-\lambda \sum_{j=1}^{t-1} L(z_j : d_{\theta^k}) \right) \pi(\theta^k)}{\int_{\theta^k \in \Theta^k} \exp\left(-\lambda \sum_{j=1}^{t-1} L(z_j : d_{\theta^k}) \right) \pi(\theta^k) d\theta^k}, \tag{8.13}$$

であり，また, $\pi(\theta^k|z^0) = \pi(\theta^k)$，$L(z_0 : d_{\theta^k}) = 0$ である． □

この補題により，ESC は逐次的に予測・学習を進めていくタイプの逐次型(オンライン型)学習-予測問題の累積損失として適用可能となる．しかし，$\pi(\theta^k|z^{t-1})$ は Θ^k 上の確率分布を与えるが，一般にベイズの事後密度確率ではない．設定した損失を平均的に小さくする θ^k に対して，相対的に高い密度が与えられる関数になっている．対数損失と $\lambda = 1$ を仮定したときは，

$$\overline{L}(y_t|z^{t-1}, x_t) = -\log \int_{\theta^k \in \Theta^k} p_{\theta^k}(y_t|x_t) \pi(\theta^k|z^{t-1}) d\theta^k, \tag{8.14}$$

$$\pi(\theta^k|z^{t-1}) = \frac{\prod_{j=1}^{t-1} p_{\theta^k}(y_j|x_j) \pi(\theta^k)}{\int_{\theta^k \in \Theta^k} \prod_{j=1}^{t-1} p_{\theta^k}(y_j|x_j) \pi(\theta^k) d\theta^k}, \tag{8.15}$$

となり，(8.15)式はベイズの事後確率密度となる．

8.3 主結果: ESC の解析

まず最初に，$h(x^n)$ と $h(x^n|\theta^k)$ を以下のように定義する．

$$h(z^n) = \int_{\theta^k \in \Theta^k} \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : d_{\theta^k})\right) \pi(\theta^k) d\theta^k, \tag{8.16}$$

$$h(z^n|\theta^k) = \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : d_{\theta^k})\right), \tag{8.17}$$

このとき，

$$\sum_{t=1}^{n} L(z_t : d_{\theta^k}) = -\frac{1}{\lambda} \log h(z^n|\theta^k), \tag{8.18}$$

という関係式が成り立っている．

このとき，以下の補題が容易に導かれる．

補題 8.2 (ベイズ規則の一般化) $h(z_n)$, $h(\theta^k|z^n)$ に対し，次式が成り立つ．

$$h(z_n) = \frac{h(z^n|\theta^k)\pi(\theta^k)}{\pi(\theta^k|z^n)}, \tag{8.19}$$

(証明) 定義より自明である． □

(8.19)式は，ベイズ規則

$$f(\theta^k|z^n) = \frac{p(z^n|\theta^k)f(\theta^k)}{p(z^n)}, \tag{8.20}$$

から導かれる式

$$p(z^n) = \frac{p(z^n|\theta^k)f(\theta^k)}{f(\theta^k|z^n)}, \tag{8.21}$$

の拡張になっている．ただし，$f(\theta^k) = \pi(\theta^k)$ である．(8.19)式より，

$$ESC(z^n) = -\frac{1}{\lambda} \log h(z^n|\theta^k)\pi(\theta^k) + \frac{1}{\lambda} \log \pi(\theta^k|z^n), \tag{8.22}$$

が成り立つ．したがって，$\log \pi(\theta^k|z^n)$ が解析できれば $ESC(z^n)$ の漸近式が与えられることになる．

つぎに，確率分布列 $\pi(\theta^k|z^n)$ に関する重要な性質を示す．多くの有用な確率モデル族に対して，ベイズの事後確率 $f(\theta^k|z^n)$ は漸近的に正規分布に収束することが知られている[16, 22]．実は次に示すように，ベイズの事後確率密度を一般化した $\pi(\theta^k|z^n)$ についても漸近正規性が成り立つことがわかる．

まず，情報行列 $Q^*(\theta^k)$，$R^*(\theta^k)$ を

$$\begin{aligned} Q^*(\theta^k) &= -E^*\left[\frac{\partial^2 \log h(Z|\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right] \\ &= \lambda E^*\left[\frac{\partial^2 L(Z:d_{\theta^k})}{\partial\theta^k(\partial\theta^k)^T}\right], \qquad (8.23) \\ R^*(\theta^k) &= E^*\left[\frac{\partial \log h(Z|\theta^k)}{\partial\theta^k}\frac{\partial \log h(Z|\theta^k)}{(\partial\theta^k)^T}\right] \\ &= \lambda^2 E^*\left[\frac{\partial L(Z:d_{\theta^k})}{\partial\theta^k}\frac{\partial L(Z:d_{\theta^k})}{(\partial\theta^k)^T}\right], \qquad (8.24) \end{aligned}$$

と定義しておく．ここで，E^* は真の分布による期待値であり，T は転値を表す．

損失最小推定量 $\hat{\theta}^k$，$\pi(\theta^k|z^n)$ を最大値 $\tilde{\theta}^k$, 最適パラメータ θ^{k^*} を次のように定義する．

$$\hat{\theta}^k = \arg\min_{\theta^k} \frac{1}{n}\sum_{t=1}^{n} L(z_t : d_{\theta^k}), \qquad (8.25)$$

$$\tilde{\theta}^k = \arg\max_{\theta^k} \pi(\theta^k|z^n), \qquad (8.26)$$

$$\theta^{k*} = \arg\min_{\theta^k} E^*\left[L(Z:d_{\theta^k})\right]. \qquad (8.27)$$

$\forall\delta > 0$ とし，Θ^k 上に球 $B_\delta(\theta^{k*}) = \{\theta^k \in \Theta^k | \|\theta^k - \theta^{k*}\| < \delta\}$ を定義する．同様に，$B_\delta(\tilde{\theta}^k) = \{\theta^k \in \Theta^k | \|\theta^k - \tilde{\theta}^k\| < \delta\}$ である．

まず，事前分布 $\pi(\theta^k)$ について，以下を仮定する．

条件 8.1 $\pi(\theta^k)$ は Θ^k 上で有界で3階微分可能とする．

次に，確率モデル族に対して次の条件を仮定する．これらは，損失最小推定量の漸近正規性のために [74], p.238 において仮定されている条件である．

条件 8.2

(1) Θ^k はコンパクト集合である．

(2) θ^{k*} は Θ^k の内部に唯一存在する．

(3) $Q^*(\theta^k)$ は正定値行列で，θ^k に対して連続微分可能である．

(4) $\frac{1}{n}\sum_{t=1}^{n} L(z_t : d_{\theta^k})$ は，$\forall n \in \{1,2,\cdots\}$ に対し，ほとんど確実に Θ^k 上で連続である．

(5) Θ^k 上で一様に，$\frac{1}{n}\sum_{t=1}^{n} L(z_t : d_{\theta^k}) \to E^*\left[L(Z:d_{\theta^k})\right]$, $a.s.$

(6) $E^*\left[L(Z:d_{\theta^k})\right]$ は Θ^k 上で 2 階微分可能，かつ $\frac{1}{n}\sum_{t=1}^n L(z_t:d_{\theta^k})$ も $\forall n \in \{1,2,\cdots\}$ に対して，ほとんど確実に Θ^k 上で2階微分可能である．

(7) Θ^k 上で一様に，

$$\frac{1}{n}\frac{\partial}{\partial\theta^k}\sum_{t=1}^n L(z_t:d_{\theta^k}) \to \frac{\partial}{\partial\theta^k}E^*\left[L(Z:d_{\theta^k})\right], \quad a.s. \tag{8.28}$$

$$\frac{1}{n}\frac{\partial^2}{\partial\theta^k(\partial\theta^k)^T}\sum_{t=1}^n L(z_t:d_{\theta^k}) \to E^*\left[\frac{\partial^2 L(Z:d_{\theta^k})}{\partial\theta^k(\partial\theta^k)^T}\right] = \frac{1}{\lambda}Q^*(\theta^k), \quad a.s. \tag{8.29}$$

(8) **(中心極限定理)** $\frac{1}{\sqrt{n}}\sum_{t=1}^n L(Z_t:d_{\theta^{k*}})$ は正規分布 $N(\mathbf{0},\frac{1}{\lambda^2}R^*(\theta^{k*}))$ に法則収束する． □

例 8.1 [153] $\mathcal{X}=\{X=(X_1,X_2,\cdots,X_k)^T\in[0,1]^k : X_1^2+X_2^2+\cdots X_k^n\le 1\}$ かつ $\mathcal{Y}=[0,1]$ とする．パラメータ集合を $\Theta^k=\{\theta^k=(\theta_1,\theta_2,\cdots,\theta_k)^T\in[0,1]^k : \theta_1^2+\theta_2^2+\cdots\theta_k^2\le 1\}$ とし，仮説集合を $\mathcal{H}=\{f_{\theta^k}(X)=(\theta^k)^TX : X\in\mathcal{X},\theta^k\in\Theta^k\}$ とする．この場合，条件8.2,(1)は明らかに成り立つ．ここで，二乗誤差損失 $L(Y,f_{\theta^k}(X))=(Y-f_{\theta^k}(X))^2$ を適用しよう．このとき条件8.2,(4)が成り立つ．また，

$$I^*(\theta^k)=\lambda E^*\left[\frac{\partial^2(Y-(\theta^k)^TX)^2}{\partial\theta^k(\partial\theta^k)^T}\right], \tag{8.30}$$

から，

$$I^*_{i,j}=2\lambda E^*\left[X_iX_j\right], \tag{8.31}$$

が成り立つ．ここで，$I^*_{i,j}$ は $I^*(\theta^k)$ の i-j 要素である．すなわち，$I^*(\theta^k)=2\lambda E^*\left[XX^T\right]$ となる．もし，$\forall i\in\{1,2,\cdots,k\}$ に対し $0<E^*\left[(X_i)^2\right]$ であれば，$I^*(\theta^k)$ は正定値行列であり，条件8.2,(3)が成り立つ．さらに，

$$\begin{aligned}E^*\left[L(Y,f_{\theta^k}(X))\right] &= E^*\left[(Y-(\theta^k)^TX)^2\right]\\ &= E^*\left[Y^2\right]-2\sum_{i=1}^k\theta_iE^*\left[YX_i\right]+\sum_{i=1}^k\sum_{j=1}^k\theta_i\theta_jE^*\left[X_iX_j\right].\end{aligned} \tag{8.32}$$

であるので，$i=1,2,\cdots,k$ に対し

$$\frac{\partial E^*\left[L(Y,f_{\theta^k}(X))\right]}{\partial\theta_i}=-2E^*\left[YX_i\right]+2\sum_{j=1}^k\theta_jE^*\left[X_iX_j\right], \tag{8.33}$$

が成り立つ．すなわち，

$$\frac{\partial E^*\left[L(Y,f_{\theta^k}(X))\right]}{\partial\theta^k}=-2E^*[YX]+2E^*\left[XX^T\right]\theta^k, \tag{8.34}$$

が成り立つ．したがって，もし θ^{k*} が Θ^k の内部に存在するならば，θ^{k*} は正規方程式

$$E^*[XX^T]\theta^{k*} = E^*[YX]. \tag{8.35}$$

をみたす．もし $\{E^*[XX^T]\}^{-1}$ が存在すれば，$\theta^{k*} = \{E^*[XX^T]\}^{-1}E^*[YX]$ が成り立つ．勿論，$\{E^*[XX^T]\}^{-1}E^*[YX] \in \Theta^k$ は常に存在するとは限らないが，$\{E^*[XX^T]\}^{-1}E^*[YX] \notin \Theta^k$ であるということは仮説集合 $\mathcal{H} = \{f_{\theta^k}(X) = (\theta^k)^T X : X \in \mathcal{X}, \theta^k \in \Theta^k\}$ が不適切であることを示している．そこで $0 < \left\|\{E^*[XX^T]\}^{-1}E^*[YX]\right\|^2 < 1$ を仮定しよう．これは，仮説集合が真の確率分布 (ターゲット) に対して適切に構成されていることを意味する．この場合には θ^{k*} は Θ^k の内部に一意に存在し，条件8.2,(2)が成り立つ．

確率変数 X と Y は有限の分散を持つので，$\frac{1}{n}\sum_{t=1}^n x_t(x_t)^T \to E^*[XX^T]$, $a.s.$, $\frac{1}{n}\sum_{t=1}^n y_t x_t \to E^*[YX]$, $a.s.$, $\frac{1}{n}\sum_{t=1}^n (y_t)^2 \to E^*[Y^2]$, $a.s.$ が成り立つ [34]．一方で $\frac{1}{n}\sum_{t=1}^n L(z_t : f_{\theta^k})$ は

$$\frac{1}{n}\sum_{t=1}^n L(z_t : f_{\theta^k}) = \frac{1}{n}\sum_{t=1}^n \left(y_t - f_{\theta^k}(x_t)\right)^2. \tag{8.36}$$

で与えられる．したがって，条件 8.2,(5)$\sim$ (7) が明らかに成り立つ．さらに，$E^*\left[\frac{1}{\sqrt{n}}\sum_{t=1}^n L(Z_t : f_{\theta^{k*}})\right] = 0$ かつ $L(Z_t : f_{\theta^{k*}})$ の分散は有限であるから，中心極点定理条件8.2,(8)が成り立つ [34].

以上により，条件8.2が成り立つことが分かる． □

例 8.2 $\mathcal{X} = \{X = (X_1, X_2, \cdots, X_k)^T \in \mathcal{R}^k\}$ かつ $\mathcal{Y} = \mathcal{R}^1$ とする．パラメータ集合を $\Theta^k = \{\theta^k = (\theta_1, \theta_2, \cdots, \theta_k)^T \in \mathcal{R}^k : \theta_1^2 + \theta_2^2 + \cdots \theta_k^2 \leq C_\theta\}$, 仮説集合を $\mathcal{H} = \{f_{\theta^k}(X) = (\theta^k)^T X : X \in \mathcal{X}, \theta^k \in \Theta^k\}$ とする．再び二乗誤差損失 $L(Y, f_{\theta^k}(X)) = (Y - f_{\theta^k}(X))^2$ を考える．

例8.1と同様の議論により，$\{E^*[XX^T]\}^{-1}$ が存在すれば $\theta^{k*} = \{E^*[XX^T]\}^{-1}E^*[YX]$ が成り立つ．C_θ を適当に大きくとることにより，θ^{k*} は Θ^k の内部で一意に存在する．したがって，例8.1と同様の議論から条件8.2が成り立つことが確かめられる． □

条件8.2を満たす他のモデル族に関しては，[74]などを参照．多くの実用的モデル族において，条件8.2はみたされる．2項分布などのようにパラメータ集合の境界で情報行列が発散するモデルでは，その境界において一様収束が成り立たなくなるが，この場合に議論を拡張することは難しくない．この条件のもとで，以下の性質が示されている[74].

補題 8.3 [74],p.238　条件8.2のもとで，$\hat{\theta}^k$ は強一致推定量となる．すなわち，

$$\hat{\theta}^k \to \theta^{k*}, \quad a.s. \tag{8.37}$$

□

次に $\hat{\theta}^k$ の分布を考える．よく知られているように $\hat{\theta}^k$ は正規分布に法則収束する．例えば，[63, 74]を参照．

補題 8.4 [74],p.239　条件8.2のもとで，$\sqrt{n}(\hat{\theta}^k - \theta^{k*})$ の分布は正規分布

$$N\left(0, \{Q^*(\theta^{k*})\}^{-1} R^*(\theta^{k*}) \{Q^*(\theta^{k*})\}^{-1}\right)$$

に法則収束する．□

これらの準備のもとに，$\pi(\theta^k|z^n)$ の漸近正規性を示す．

定理 8.1 条件8.1,条件8.2を仮定する．$\xi^k = \sqrt{n}(\theta^k - \hat{\theta}^k)$，および

$$\pi_\xi(\xi^k|z^n) = \frac{1}{\sqrt{n}^k} \pi(\theta^k|z^n), \tag{8.38}$$

と定義すると，次の漸近式が成り立つ．

$$\pi_\xi(\hat{\xi}^k|z^n) \to \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{\det Q^*(\hat{\theta}^k)} \quad a.s. \tag{8.39}$$

ここで，$\hat{\xi}^k = \mathbf{0}$ である．さらに，分布列 $\pi_\xi(\xi^k|z^n)$, $n = 1, 2, \cdots$ は正規分布に概収束する．すなわち，E を ξ^k-空間上の任意の直方体とすると，

$$\int_{\xi^k \in E} \pi_\xi(\xi^k|z^n) d\xi^k \to \frac{\sqrt{\det Q^*(\hat{\theta}^k)}}{(2\pi)^{k/2}} \int_{\xi^k \in E} \exp\left\{-\frac{1}{2}\|\xi^k\|^2_{Q^*(\hat{\theta}^k)}\right\} d\xi^k, \quad a.s. \tag{8.40}$$

が成り立つ．
(証明) 8.5.1節参照．□

さらに，$\hat{\theta}^k$ は $\tilde{\theta}^k$ で置き換えることができる．すなわち，次の定理が成り立つ．

定理 8.2 条件8.1,条件8.2を仮定する．$\eta^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$，および

$$\pi_\eta(\eta^k|z^n) = \frac{1}{\sqrt{n}^k} \pi(\theta^k|z^n), \tag{8.41}$$

と定義すると，次の漸近式が成り立つ．

$$\pi_\eta(\tilde{\eta}^k|z^n) \to \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{\det Q^*(\tilde{\theta}^k)} \quad a.s. \tag{8.42}$$

ただし，$\tilde{\eta}^k = \mathbf{0}$ である．さらに E を η^k-空間上の任意の直方体とすると，

$$\int_{\eta^k \in E} \pi_\eta(\eta^k|z^n) d\eta^k \to \frac{\sqrt{\det Q^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \int_{\eta^k \in E} \exp\left\{-\frac{1}{2}\|\eta^k\|^2_{Q^*(\tilde{\theta}^k)}\right\} d\eta^k, \quad a.s. \tag{8.43}$$

が成り立つ．

(証明) 8.5.2節参照． □

以上より，ベイズの事後確率を拡張した分布列 $\pi(\theta^k|z^n)$, $n = 1, 2, \cdots$ も漸近正規性を満たすことがわかる．その正規分布の共分散行列は $\{Q^*(\hat{\theta}^k)\}^{-1}$ で与えられる．本研究では，真の分布がモデル族に含まれることを仮定していない．当然ながら，一般に $Q^*(\hat{\theta}^k)$ はFisher情報行列と一致しない．この場合，損失最小推定量の漸近共分散が $\{Q^*(\theta^{k*})\}^{-1} R^*(\theta^{k*}) \{Q^*(\theta^{k*})\}^{-1}$ で与えられるのに対し，$\pi(\theta^k|z^n)$ の漸近分散には $\{Q^*(\hat{\theta}^k)\}^{-1}$ のみが現れる点は本質的である．これもベイズの事後確率の性質と同様である．

この漸近正規性を用いて，ESC の漸近式について，以下の定理が得られる．

定理 8.3 条件8.1,条件8.2のもとで，$ESC(z^n)$ の漸近式は

$$ESC(z^n) = \sum_{t=1}^n L(z_t : d_{\hat{\theta}^k}) + \frac{k}{2\lambda} \log \frac{n}{2\pi} - \frac{1}{\lambda} \log \frac{\sqrt{\det Q^*(\hat{\theta}^k)}}{\pi(\hat{\theta}^k)} + o(1), \quad a.s. \tag{8.44}$$

$$= \sum_{t=1}^n L(z_t : d_{\tilde{\theta}^k}) + \frac{k}{2\lambda} \log \frac{n}{2\pi} - \frac{1}{\lambda} \log \frac{\sqrt{\det Q^*(\tilde{\theta}^k)}}{\pi(\tilde{\theta}^k)} + o(1), \quad a.s. \tag{8.45}$$

で与えられる．

(証明) (8.22)式, (8.39)式, (8.42)式より明らかである． □

上で与えた定理では，真の分布に従って発生する個々の系列に対して，ESC の概収束の意味での漸近式を与えた．ここで，$\hat{\theta}^k$ は確率変数であるので，その期待値をとるために以下の補題を引用する．

補題 8.5 [74],pp.240 -241. 条件8.2のもとで，

$$E^* \left[\log \frac{h(Z^n|\hat{\theta}^k)}{h(Z^n|\theta^{k*})}\right] \to \frac{\mathrm{Tr} R^*(\theta^{k*}) \{Q^*(\theta^{k*})\}^{-1}}{2}. \tag{8.46}$$

が成り立つ．ただし，Trはトレースを表す. □

この補題を用いることにより，$ESC(z^n)$ の期待値 $E^*\left[ESC(Z^n)\right]$ は以下のように与えられる．

定理 8.4 条件8.1,条件8.2のもとで，$E^*\left[ESC(Z^n)\right]$ は

$$E^*\left[ESC(Z^n)\right] = \sum_{t=1}^{n} E^*\left[L(Z_t : d_{\theta^{k*}})\right] + \frac{k}{2\lambda}\log\frac{n}{2\pi}$$

$$- \frac{\mathrm{Tr}R^*(\theta^{k*})\{Q^*(\theta^{k*})\}^{-1}}{2\lambda} - \frac{1}{\lambda}\log\frac{\sqrt{\det Q^*(\theta^{k*})}}{\pi(\theta^{k*})} + o(1). \tag{8.47}$$

(証明) (8.44)式に有界収束定理を適用し，(8.46)式を用いて，(8.47)式が得られる．
□

8.4 考察

本章ではまず，ESC の個々の系列に対する漸近式を与えた．Yamanishi は Z^n に対して一様に成り立つ上界を与えているのに対し，本研究では概収束を考えている．よって，真の分布での期待値操作が含まれる情報行列 $R^*(\hat{\theta}^k)$，$Q^*(\hat{\theta}^k)$ が現れる．

もし真の確率分布がモデルクラス $\mathcal{P}^k$ に含まれるならば，多くの損失関数に対して $R^*(\theta^{k*}) = cQ^*(\theta^{k*})$ が成り立つ[74]．ここで，c はある定数である．よって，$TrR^*(\theta^{k*})Q^*(\theta^{k*})^{-1} = c^k k$ となり，$E^*\left[ESC(Z^n)\right]$ は

$$E^*\left[ESC(Z^n)\right] = -E^*\left[\log h(Z^n|\theta^{k*})\right] + \frac{k}{2\lambda}\log\frac{n}{2\pi e^C} - \frac{1}{\lambda}\log\frac{\sqrt{\det Q^*(\theta^{k*})}}{\pi(\theta^{k*})} + o(1), \tag{8.48}$$

と与えられる．ただし，$C = c^k$ とおいた．

この形をみると，Clarke と Barron のベイズ符号の漸近式の一般化となっている．もし対数損失を考え，$\lambda = 1$ とすれば，$c = 1$ となる．このとき，(8.48)式は Clarke と Barron の漸近式と等価となる．

8.5 定理と補題の証明

8.5.1 定理 8.1 の証明

ここでの漸近正規性も [16],pp.285-297 の議論を用いて，事後確率密度の場合と同様の手順で示すことができる．

まず，定理の前半，(8.39)式を示そう．

$$K_n(\theta^k) = \log \pi(\theta^k|z^n), \tag{8.49}$$

$$K_n''(\theta^k) = \frac{\partial^2 K_n(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}, \tag{8.50}$$

$$\Sigma_n = -\left(K_n''(\hat{\theta}^k)\right)^{-1}. \tag{8.51}$$

と定義し，テーラー展開により

$$\begin{aligned}\pi(\theta^k|z^n) = \pi(\hat{\theta}^k|z^n)\exp\left\{(\theta^k-\hat{\theta}^k)^T K'(\hat{\theta}^k)\right\} \\ \cdot\exp\left\{\frac{1}{2}(\theta^k-\hat{\theta}^k)^T(I+R_n)K''(\hat{\theta}^k)(\theta^k-\hat{\theta}^k)\right\},\end{aligned} \tag{8.52}$$

を得る．ここで，R_n は

$$R_n = K_n''(\theta^{k+})\{K_n''(\hat{\theta}^k)\}^{-1} - I, \tag{8.53}$$

で与えられ，θ^{k+} は θ^k と $\hat{\theta}^k$ を結ぶ線分上の点である．

$K_n'(\hat{\theta}^k)$ はその定義より，

$$K_n'(\hat{\theta}^k) = \left.\frac{\partial \log\pi(\theta^k)}{\partial\theta^k}\right|_{\theta^k=\hat{\theta}^k}. \tag{8.54}$$

一方，

$$K_n''(\theta^k) = \frac{\partial^2 \log h(z^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} + \frac{\partial^2\log\pi(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}, \tag{8.55}$$

と条件8.2,(7)より，$\forall\theta^k\in\Theta^k$ に対して一様に

$$\frac{1}{n}K_n''(\theta^k) \to Q^*(\theta^k),\ a.s. \tag{8.56}$$

が成り立つ．ここで

$$\hat{\theta}^k \to \theta^{k*},\quad a.s. \tag{8.57}$$

かつ $Q^*(\theta^k)$ は θ^k に関して連続微分可能であるので，$\forall\delta>0$ に対し，$\forall\theta^k\in B_\delta(\hat{\theta}^{k_m})$ に対して一様に

$$K_n''(\theta^{k+})\{K_n''(\hat{\theta}^k)\}^{-1} \to Q^*(\theta^{k+})\{Q^*(\hat{\theta}^k)\}^{-1},\quad a.s. \tag{8.58}$$

が成り立つ．

$Q^*(\theta^k)$ の連続性と $\hat{\theta}^k$ の一致性から，$\forall\epsilon>0$ に対し ある正定数 $\delta>0$ が存在して，$\forall\theta^k\in B_\delta(\hat{\theta}^k)$ に対し $n\to\infty$ のとき

$$I - A(\epsilon) \le Q^*(\theta^k)\{Q^*(\hat{\theta}^k)\}^{-1} \le I + A(\epsilon), \tag{8.59}$$

が成り立つ．ここで I は $k \times k$ 単位行列，$A(\epsilon)$ は $\epsilon \to 0$ のときその最大固有値が 0 に収束する $k \times k$ 対称正定値行列 0 である．ここで，$Q^*(\theta^k)$ は連続なので，$\epsilon \to 0$ のとき $\delta \to 0$ とできることに注意する．よって，(8.58)式と(8.59)式より，$\forall \epsilon > 0$ に対し $\delta > 0$ が存在し， $n \to \infty$ のとき $\forall \theta^k \in B_\delta(\hat{\theta}^k)$ に対して一様に

$$I - A(\epsilon) \leq K_n''(\theta^k)\{K_n''(\hat{\theta}^k)\}^{-1} \leq I + A(\epsilon), \ \ a.s. \tag{8.60}$$

ここで，$\epsilon \to 0$ のとき δ は任意に小さくとれる．

以上より，$n \to \infty$ のとき

$$P_n(\delta) = \int_{\theta^k \in B_\delta(\hat{\theta}^k)} \pi(\theta^k | z^n) d\theta^k, \tag{8.61}$$

はほとんど確実に

$$\pi(\hat{\theta}^k | z^n) \exp\{\bar{c}_f \delta\} |\det \Sigma_n|^{1/2} |\det(I - A(\epsilon))|^{-1/2} \int_{||u^k|| < s_n} \exp\{-\frac{1}{2}(u^k)^T u^k\} du^k, \tag{8.62}$$

で上界され，

$$\pi(\hat{\theta}^k | z^n) \exp\{-\bar{c}_f \delta\} |\det \Sigma_n|^{1/2} |\det(I + A(\epsilon))|^{-1/2} \int_{||u^k|| < t_n} \exp\{-\frac{1}{2}(u^k)^T u^k\} du^k, \tag{8.63}$$

で下界される．ここで，$\bar{c}_f$ は $\frac{\partial \log \pi(\theta^k)}{\partial \theta^k}$ の最大絶対値，s_n と t_n は $s_n = \delta(1 - \overline{a(\epsilon)})^{1/2} / \underline{l_n}^{1/2}$, $t_n = \delta(1 - \underline{a(\epsilon)})^{1/2} / \overline{l_n}^{1/2}$ と与えられる．ただし，$\overline{a(\epsilon)}$ ($\underline{a(\epsilon)}$) と $\overline{l_n}$ ($\underline{l_n}$) はそれぞれ，$A(\epsilon)$ と Σ_n の最大(最小)固有値とする．

(8.58)式と同様の議論により，

$$n\Sigma_n \to \{Q^*(\hat{\theta}^k)\}^{-1}, \ \ a.s. \tag{8.64}$$

であるので，(8.57)式より，$\overline{l_n} \to 0$, $a.s.$ かつ $\underline{l_n} \to 0$, $a.s.$ となる．したがって，$s_n \to \infty, a.s.$ と $t_n \to \infty, a.s.$ が成り立つ．したがって， $n \to \infty$ のとき

$$|\det(I - A(\epsilon))|^{1/2} \exp\{-\bar{c}_f \delta\} \lim_{n \to \infty} P_n(\delta) \leq \lim_{n \to \infty} \frac{\pi_\xi(\hat{\xi}^k | z^n)(2\pi)^{k/2}}{\left\{Q^*(\hat{\theta}^k)\right\}^{-1/2}}$$

$$\leq |\det(I + A(\epsilon))|^{1/2} \exp\{\bar{c}_f \delta\} \lim_{n \to \infty} P_n(\delta), \ a.s. \tag{8.65}$$

が成り立つ．

以上より，$\epsilon \to 0$ のとき $\delta \to 0$ とできるので，もし $\forall \delta > 0$ に対して $\lim_{n \to \infty} P_n(\delta) = 1$, $a.s.$ が成り立てば，

$$\lim_{n \to \infty} \pi_\xi(\hat{\xi}^k | z^n) \to \frac{\{\det Q^*(\hat{\theta}^k)\}^{1/2}}{(2\pi)^{k/2}}, \ \ a.s. \tag{8.66}$$

が示される．よって，最後に $\forall\delta > 0$ に対して $\lim_{n\to\infty} P_n(\delta) = 1,\ a.s.$ であることを示そう．

まず，条件8.1と条件8.1,(5)より，$\forall\delta_\epsilon > 0$ に対し，$\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ に対して一様に

$$\begin{aligned}\frac{1}{n}\{K_n(\theta^k) - K_n(\theta^{k*})\} &= \frac{1}{n}\log\frac{h(z^n|\theta^k)}{h(z^n|\theta^{k*})} + \frac{1}{n}\log\frac{\pi(\theta^k)}{\pi(\theta^{k*})} \\ &= -\frac{\lambda}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k}) + \frac{\lambda}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^{k*}}) + \frac{1}{n}\log\frac{\pi(\theta^k)}{\pi(\theta^{k*})} \\ &\to -\lambda E^*\left[L(Z : f_{\theta^k})\right] + \lambda E^*\left[L(Z : f_{\theta^{k*}})\right],\ a.s. \end{aligned} \tag{8.67}$$

が成り立つ．条件8.2,(2)より，$E^*\left[L(Z : f_{\theta^{k*}})\right]$ は唯一の最大値 $E^*\left[L(Z : f_{\theta^k})\right]$ を持つから，$\forall\delta_\epsilon > 0$ に対し，ある定数 $\exists C_{\delta_\epsilon} > 0$ が存在して，$n \to \infty$ のとき $\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ に対して一様に

$$\frac{1}{n}\{K_n(\theta^k) - K_n(\theta^{k*})\} < -C_{\delta_\epsilon},\ a.s. \tag{8.68}$$

が成り立つ．従って，$\forall\delta_\epsilon > 0$ に対し，$n \to \infty$ のとき，$\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ に対して一様に

$$\frac{\pi(\theta^k|z^n)}{\pi(\theta^{k*}|z^n)} < \exp\left\{-nC_{\delta_\epsilon}\right\},\ a.s. \tag{8.69}$$

が成り立つ．

一方，$n \to \infty$ のとき $\forall\theta^k \in B_{\delta_\epsilon}(\theta^{k*})$ に対して，

$$\pi_\xi(\xi^{k*}|z^n) < \frac{\{\det I^*(\hat{\theta}^k)\}^{1/2}}{(2\pi)^{k/2}},\quad a.s. \tag{8.70}$$

が成り立つ．ただし $\xi^{k*} = \sqrt{n}(\theta^{k*} - \hat{\theta}^k)$ である．

ここで条件8.2,(3)より，$\det I^*(\hat{\theta}^k) \to \det I^*(\theta^{k*}),\ a.s.$ が成り立つので，ある正定数 $C^* > 0$ が存在して $n \to \infty$ のとき

$$\pi(\theta^{k*}|z^n) < C^*\sqrt{n}^k,\quad a.s. \tag{8.71}$$

が成り立つ．

よって，2つの不等式 (8.69),(8.71) より，$\forall\delta_\epsilon > 0$ に対し，$\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ に対して一様に

$$\begin{aligned}\pi(\theta^k|z^n) &< \pi(\theta^{k*}|z^n)\exp\{-nC_{\delta_\epsilon}\},\quad a.s. \\ &< C^*\sqrt{n}^k\exp\{-nC_{\delta_\epsilon}\} \to 0,\quad a.s. \end{aligned} \tag{8.72}$$

が成り立つ．条件8.2,(1)より，Θ^k はコンパクト集合であるので，$\forall\delta_\epsilon > 0$ に対して

$$\int_{\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})} \pi(\theta^k|z^n)d\theta^k \to 0, \quad a.s. \tag{8.73}$$

が成り立つ．これは，$\forall\delta_\epsilon > 0$ に対して

$$\int_{\theta^k \in B_{\delta_\epsilon}(\theta^{k*})} \pi(\theta^k|z^n)d\theta^k \to 1, \quad a.s. \tag{8.74}$$

が成り立つことを意味している．

ここで，

$$\frac{1}{n}\log\pi(\theta^k|z^n) \to \frac{1}{n}\sum_{t=1}^{n} L(x_t : d_{\theta^k}), \tag{8.75}$$

より，$\hat{\theta}^k \to \tilde{\theta}^k \to \theta^{k*}$ $a.s.$ である．従って，$0 < \forall\delta' < \forall\delta''$ に対して $n \to \infty$ のとき，$B_{\delta'}(\theta^{k*}) \subset B_{\delta''}(\hat{\theta}^k)$, $a.s.$ であるから，$\lim_{n\to\infty} P_n(\delta_\epsilon) = 1$, $a.s.$ が成り立ち，(8.39)式が示された．

定理の後半，(8.40)式も同様の議論により，任意の直方体 $\forall E$ に対し

$$\int_{\xi^k \in E} \pi_\xi(\xi^k|x^n)d\xi^k, \tag{8.76}$$

をバウンドして得られる． □

8.5.2 定理 8.2 の証明

$\tilde{\theta}^k$ のまわりでのテーラー展開により，

$$\pi(\theta^k|z^n) = \pi(\tilde{\theta}^k|z^n)\exp\left\{\frac{1}{2}(\theta^k - \tilde{\theta}^k)^T(I + \tilde{R}_n)K''(\tilde{\theta}^k)(\theta^k - \tilde{\theta}^k)\right\}, \tag{8.77}$$

ここで $\tilde{R}_n$ は

$$\tilde{R}_n = K_n''(\theta^{k+})\{K_n''(\tilde{\theta}^k)\}^{-1} - I, \tag{8.78}$$

で与えられ，θ^{k+} は θ^k と $\tilde{\theta}^k$ を結ぶ線分上の点である．したがって，あとは定理8.1の証明と同様の議論により定理8.2が得られる． □

8.6 本章の結論

本章では，K.Yamanishi により提案された ESC の漸近式を定数オーダの項まで詳細に解析した．この結果，ESC を用いた学習アルゴリズムのより厳密な理論的評価が可能となった．この解析においても，本質的であるのはパラメータ集合上に定義される分布列の漸近正規性である．得られた漸近式はベイズ符号の漸近式と類似し，対数損失を仮定したときにはベイズ符号の漸近式となる．すなわち，Clarke と Barron のベイズ符号の漸近式の一般形となっている．

ESC はパラメトリックモデル族に対して Yamanishi により提案されたものであるが，仮説空間が階層モデル族になっている場合に拡張することができる．この場合の漸近式を導出できれば，階層構造を有する仮説空間に対する学習アルゴリズムを評価することが可能である．この場合への拡張は今後の課題である．また，ESC は逐次予測の累積損失とバッチ的に与えたデータの損失が一致するという性質を持っている．このため，オンライン的な逐次学習問題とバッチ的な学習問題の両方に適用可能であるが，事前分布を仮定していながらベイズ最適ではない．一般にベイズ最適解は逐次予測の累積損失とバッチ的に与えたデータの損失が異なってしまい，アルゴリズム的にも一般的な構成が難しい．例外が対数損失であり，その取り扱いの容易さから良く用いられる損失関数であるが，目的によっては他の損失関数に対するベイズ最適を考えたい状況も存在する．ESCはその近似解を提示しており，しかもアルゴリズム的に取り扱い易い性質を持っている．この観点からすれば ESC がベイズ最適に対してどの程度の性能をもつのかについて解析する必要があり，今後の課題である．

第 9 章

結論

9.1 本研究の成果

本研究では，ベイズ統計理論に基づくモデルの推定と確率分布の予測の漸近的評価をおこなった．

第3章，第4章においては，MDL規準とベイズ規準の差を累積対数損失の面から解析することにより，系列の予測問題においてモデルを1つ選択する方法と混合モデルを用いる方法の差を定量的に示した．その結果，MDL規準に対するベイズ規準の累積対数損失の面からの優位性が示された．これは，情報源符号化の立場で解釈すると，MDL符号とベイズ符号の符号長の差を解析したことと等価である．

第5章では，B.S.Clarke と A.R.Barron の成果を拡張し，応用上極めて重要な FSMX 情報源に対して，ベイズ規準による逐次予測の漸近的な累積対数損失，及びKL情報量をリスク関数としたときの漸近式を導出した．これにより，KL情報量という一般に広く使われている尺度を用いて，ベイズ規準に基づき混合分布を用いる方法の漸近的評価式が得られた．その漸近式は Clarke と Barron の結果を階層モデル族に拡張したものであり，応用例としてベイズ符号の最適な圧縮率への収束速度が従来よりも精密な形で与えられたことになる．

第6章においては，統計的モデル選択問題をベイズ決定理論に基づいて定式化し，これがある損失関数のクラスに対しては強一致性(漸近的に真のモデルに収束する性質)を持つことを数学的に証明した．この結果，目的が真のモデルの発見にある場合でも，未来のデータの予測にある場合でも，ベイズ決定理論に基づくモデル選択は漸近的に最適なモデルを選択することができる．一致性は統計的モデル選択問題で従来から議論されてきた一つの評価規準であり，ベイズ決定理論に基づくモデル選択の一つの漸近的性能の評価が得られたことになる．

第7章では，ベイズ決定理論に基づくモデル選択について，真のモデルを選択

できない選択誤り確率の上界を導出した．これにより，真のモデルを特定するという目的に対して，その選択誤り率という基準での評価が与えられたことになる．さらに応用例として情報源の変化時点検出問題を取り上げ，ベイズ最適解を導出し，現実問題におけるベイズ的モデル選択の有効性を示した．

第8章では，K.Yamanishi による ESC の漸近式を導出した．ESC の漸近式の導出はこれまでなされていなかったが，本論文でその解が示されたことになる．その結果，損失関数を限定しない拡張型確率的コンプレキシティを統計的推論に適用した際の，累積損失の漸近的評価を与えることが可能となった．

本研究の成果はベイズ統計に基づく推定と予測に関して従来明らかとされていなかった性質を理論的に示したものであり，その解析の切り口にも特徴があると考えている．また，情報技術が発達した現在，膨大な数のデータを解析する必要性も増えてきており，漸近的な状況を考えても差し支えないような場面が多くなってきている．この点からも本研究で行った漸近的評価は価値があると考えられる．

9.2 展望と今後の課題

本研究で得られた結果は，それぞれがベイズ統計に基づく推定理論と予測理論について，部分的ではあるがその漸近的性質を明らかとしたものであり，理論的興味のみならず，今後の応用に対しても有用な知見を得られたものと考えられる．

今後の課題としては以下が考えられる．

1. **条件の緩和:** 本論文ではパラメトリックモデル族と階層モデル族において，最適パラメータがパラメータ集合の内部に唯一存在することを条件として解析を行った．さらに，Fisher情報行列の行列式が ∞ とならないことを仮定している．しかし，例えば二項分布で二項確率が 0 や 1 の場合など，パラメータ集合の境界に位置するパラメータ，またはFisher情報行列の行列式が ∞ となるパラメータがあることは多くの確率モデルで生じる．最適パラメータがパラメータ集合の境界に位置する場合やFisher情報行列が発散する点である場合には，事後確率密度の漸近正規性は成り立たない．したがって，このような点も含めた解析を行うためには，新たな観点から解析を進める必要がある．ユニバーサル符号化の分野では，このような特異点を含めた解析が行われている．このような議論を押し進め，解析のための条件を緩和していくことは今後の課題である．

一方，統計的モデル選択の問題では，例えばニューラルネットワークモデルの中間層素子数の決定などの非線形モデルの選択問題[48]，混合正規分布を用いたHMMの混合数決定[129]などでは漸近正規性が成り立たない．このような非線形モデルでは一般に探索手法を用いてパラメータを推定するが，その探索終了条件によっても未知データに対する予測精度が変化してしまう[38]．このような非線形モデルに対する統計的モデル選択では，解析が非常に難しいこともあり，体系的な成果としてまとまっていない．しかし，多くの研究者により基礎理論の構築が試みられている最中である[37, 147]．計算機の発達によって非線形モデルの利用が可能となっている現在においては，このような非線形モデルの推定問題に関する基礎研究は不可欠であり，今後の課題である．

さらに，現実問題ではしばしばデータに欠測が含まれる場合がある．これはデータ採取の際に欠測が生じる場合と本質的に観測ができない変数を含むモデルを扱う場合とがある．これらの場合には，適当な条件のもとで漸近正規性が成り立つ場合があるが[116]，次元落ちや漸近分布の共分散行列の固有値が増大することが考えられる．このようなケースについて解析することも今後の課題である．

2. **ベイズ符号化アルゴリズムへの反映:** 本研究では，ベイズ符号の理想符号長を議論したが，現実的応用を考えると，実際の符号化アルゴリズムと理論的成果をうまく融合し，計算量，圧縮効率の両面で実用的な符号化法を構成していく必要がある．本研究で示したように，今後は MDL原理の研究も，ベイズ符号やミニマックス符号を中心に理論が展開されていくと考えられる．

   現在使われているユニバーサル符号化法は主にZiv-Lempel(ZL)法[159, 160]の改良型であり，理論的な観点から漸近的最良性が示され，有限長のデータ列に関しては実際のデータの圧縮性能を実験することでその有効性が認識されている．今後これに取って代わる方法としてベイズ符号を考えた場合，これは設定する確率モデル族に依存する符号化法であるので，現実のデータにより合った(圧縮効率の良い)確率モデルを設定し，少ない計算量で符号化できるアルゴリズムを構成しなくてはならない．本研究では FSMX モデル族に対してベイズ符号の漸近符号長を示したが，これはあくまで FSMX モデルで本来圧縮可能な限界からの冗長分の解析であり，実際のデータがより広いクラスの分布から生起するものとすれば，その FSMX

モデルがどこまで効率の良いものとなっているかを議論する必要がある．例えば，C言語のプログラムなどのデータファイルに対しては，過去の系列から次の 1シンボルの生起確率が決まるモデルよりも，いくつかのシンボルをまとめた単語単位で生起確率が決まってくるモデルの方が良く当てはまることが直感的に理解できる．このようなモデルに対するアルゴリズムの構成，及び理論的解析は今後の課題である[65].

3. **統計的モデル選択問題の諸問題:** 本研究では，ベイズ決定理論に基づくモデル選択を中心に議論を行った．その結果，ベイズ最適性と一致性という観点から一応の成果を得ている．しかし，統計的モデル選択問題が現在も広く研究されている理由の一つは，選択基準の評価が定まっていないことであり，今後も様々な観点から研究を進める必要がある．また，ベイズ最適なモデル選択を実際に適用する際，計算量を考慮する必要がある．また，一般に漸近正規性を仮定できないクラス，例えばノンパラメトリックなヒストグラム推定量，あるいは Fisher情報行列が退化するモデル族などにおける性質を議論する必要がある．とくにFisher情報行列が退化するモデルの一つであるニューラルネットワークモデルは，解析的に非常に難しいモデルであり，今後の理論的成果が期待されている．

   また，確率的コンプレキシティの概念を統計的モデル化の問題に適用する際には，全てのデータ系列に対して一様に成り立つ漸近式を導く必要がある．本研究では主に確率収束，概収束，平均収束を議論したが，真の母集団(情報源)を意識せずに，データ系列のみから計算できる漸近式が望まれる．このような観点から，データ系列に対して一様に成り立つ漸近式を導き，そこから得られる性質を議論することが必要であろう[77].

4. **損失関数の一般化に関連して:** 本研究では，最後に SC の対数損失を一般の損失へ拡張した ESC の漸近的性質について議論した．ここでの議論は，概収束と平均収束の意味であり，K.Yamanishi が議論している任意のデータ系列に対して漸近式を与えた訳ではない．この点はさらに議論を進める余地があると思われる．

   また，ESC は損失の一般化という意味で一つの形を与えているが，SC がもつベイズ最適性を保存した拡張にはなっていない．これは，バッチ型予測問題の損失とオンライン型の逐次予測問題の累積損失が同じ式で与えられるという性質を保存した形で拡張したためである．しかし，ベイズ統計理論の観点から考えると，事前分布が仮定されている以上，これに最適

な予測法を議論するべきであり，もし計算量などの面でこれを満たさない場合には，ベイズ最適からの劣化の程度を解析する必要があると考えられる．ESC がベイズ最適と比べてどの程度の性能を与えるかについての議論は今後の課題といえる．

以上，個別に今後の展望と課題について述べた．本研究では一般的な問題設定のもとで解析を行ったので，この結果をさらに精密化すると共に，より現実的応用に反映させるための研究も望まれる．とくに，ベイズ統計理論を現実問題に適用する際の計算量の問題を議論する必要がある．また，本研究の成果を実証的にも確認することを合わせて今後の課題とする．

謝辞

本論文は，著者が早稲田大学大学院に在学中における研究成果をまとめたものです．しかしその成果も決して自分一人によるものではなく，多くの支援があって初めてまとめることができました．本研究をまとめるに当たり，たくさんのご支援を戴きました全ての皆様に感謝の意を表します．

早稲田大学 理工学部 経営システム工学科 平澤茂一教授には，著者が平澤研究室の門を叩いたその日から，その広い懐で熱心にご指導を戴きました．その親身なご指導は研究に対してだけでなく，日常生活や趣味の面でも本当にお世話になりました．著者が本論文をまとめることができたのも平澤教授の熱心なご指導の賜物であり，ここに深く感謝し厚く御礼申し上げます．

また，本研究をまとめるにあたり有益なご助言と熱心なご指導を賜りました早稲田大学 理工学部 経営システム工学科 東 基衞教授，石渡徳彌教授，逆瀬川浩孝教授，松嶋敏泰教授に心より感謝申し上げます．先生方には本当に長期間にわたり貴重な時間を割いてご指導を戴きました．とくに，松嶋敏泰教授には研究室の先輩として私生活でもお世話になり，また論文連名者として議論をして戴きました．

武蔵工業大学 経営工学科 の俵 信彦教授には，著者が研究の道に進むきっかけを与えて戴き，研究指導だけでなく研究姿勢や生活習慣などについても一つ一つ丁寧にご指導戴きました．また博士過程進学後も日頃から激励下さり，常に前向きに研究に取り組むことが出来ました．その温かいご指導に深く感謝申し上げます．

早稲田大学 大学院 在学時，及び理工学部助手を務めさせて戴いた期間を含め，早稲田大学 理工学部 経営システム工学科の先生方には，教務と研究の両面で色々とご指導を戴きました．また，著者が武蔵工業大学に在学していた当時より，武蔵工業大学 経営工学科の先生方にも多くのご指導を戴きました．とくに，増井忠幸教授，横山真一郎教授には，常に励ましの言葉をかけて戴きました．また，俵研究室の先輩である東京都立短期大学の開沼泰隆先生には，初期の共同研究等を通じてご指導を戴き，研究の姿勢等についても教えて戴きました．先生方には心から感謝申し上げます．

また，平澤研究室の先輩であり日頃から適切な助言を戴きました法政大学 西島敏尚先生，大阪大学 鈴木譲先生，神奈川工科大学 新家稔央先生，大阪電気通信大学 鴻巣敏之先生に感謝致します．現在も勉学を共にする早稲田大学 平澤研究室 の鈴木誠氏，李相協助手，中澤真氏，二神常爾助手，小林学助手，松嶋研究室の浮田善文助手，野村亮氏，吉田隆弘氏には，日頃のゼミなどで熱心に議論を戴き，適切な助言を多数戴きました．早稲田大学 平澤研究室，松嶋研究室，武蔵工業大学 俵研究室の先輩，後輩の多くの方々にも大変お世話になりました．この場を借りまして心より感謝申し上げます．

最後に，著者の研究に理解を示し，研究生活を暖かく支えてくれた家族，親戚，友人の方々に心から感謝致します．

2000年2月

参考文献

[1] H. Akaike: "A New Look at the Statistical Model Identification", *IEEE Trans. Automatic Control*, Vol.AC-19, No.6, pp.716-722,(1974)

[2] 赤池弘次: "情報量規準 AIC とは何か", 数理科学, No.153, pp.5-11, (1976)

[3] 赤池弘次: "エントロピーとモデルの尤度", 日本物理学会誌, Vol.35, No.7, pp.608-614, (1980)

[4] 赤池弘次: "AIC と MDL と BIC", オペレーションズ・リサーチ, Vol.41, No.7, pp.375-378, (1996)

[5] P.H.Algoet: "The Strong Law of Large Numbers for Sequential Decision under Uncertainty", *IEEE Trans. Information Theory*, Vol.40, pp.609-633, (1994)

[6] 甘利俊一: 情報理論, ダイヤモンド社, (1970)

[7] 甘利俊一, 長岡浩司: 情報幾何の方法 (岩波講座 応用数学 12), 岩波書店, (1993)

[8] 有本 卓: 現代情報理論, 電子情報通信学会, (1978)

[9] 有本 卓: 確率・情報・エントロピー, 森北出版, (1980)

[10] A.C.Atkinson: "A Method for Discriminating between Models", *Journal of Royal Statistical Stociety*, Ser.B, Vol.32, pp.323-353, (1970)

[11] A.R.Barron and T.M.Cover: "Minimum Complexity Density Estimation", *IEEE Trans. Information Theory*, Vol.37, pp.1034-1054, (1991)

[12] A.Barron, J.Rissanen, and B.Yu: "The Minimum Description Length Principle in Coding and Modeling", *IEEE Trans. Information Theory*, Vol.44, No.6, pp.2743 - 2760, (1998)

[13] R.R.Bhattacharya and R.R.Rao: *Normal Approximation and Asymptotic Expansions*, Robert E.Krieger Publishing Company, (1986)

[14] J.O.Berger: *Statistical Decision Theory and Bayesian Analysis (Second Edition)*, Springer - Verlag, (1985)

[15] J.M.Bernardo: " Expected information as expected utility", *The Annals of Statistics*, Vol.7, pp.686-690, (1979)

[16] J.M.Bernardo and A.F.M.Smith: *Bayesian Theory*, John Wiley & Sons, (1994)

[17] R.N.Bhattacharya and R.Ranga Rao: *Normal Approximation and Asymptotic Expansions*, John Wiley & Sons, (1976)

[18] P.Billingsley: *Convergence of Probability Measures*, John Wiley & Sons, (1968)

[19] P.Billingsley: *Probability and Measures, Third Edition*, John Wiley & Sons, (1995)

[20] R.E.Blahut: *Principles and Practice of Information Theory*, Addison Weley, (1987)

[21] S.Chatterjee and B.Price: *Regression Analysis by Example*, John Wiley & Sons, (1977)

[22] Chan-Fu Chen: "On Asymptotic Normality of Limiting Density Function with Bayesian Implications", *Journal of Royal Statistical Society, Ser.B*, 47,No.3, pp.540-546, (1985)

[23] K.L.Chung: *A Course in Probability Theory (Second Edition)*, Academic Press, (1974)

[24] B.S.Clarke : "Asymptotic Normality of the Posterior in Relative Entropy", *IEEE Trans. Information Theory*, Vol.45, No.1, pp.165-176, (1999)

[25] B.S.Clarke and A.R.Barron : "Information - Theoretic Asymptotics of Bayes Methods ", *IEEE Trans. Information Theory*, Vol.36, No.3, pp.453 - 471, (1990)

[26] B.S.Clarke and A.R.Barron : "Jeffreys' Prior is Asymptotically Least Favorable under Entropy Risk", *Journal of Statistical Planning and Inference*, Vol.41, pp.37-60, (1994)

[27] G.E.P.Box and G.C.Tiao: *Bayesian Inference in Statistical Analysis*, John Wiley & Sons, (1992)

[28] L.D.Davisson: "Universal Noiseless Coding", *IEEE Trans. Information Theory*, Vol.19, No.6, pp.783-795, (1973)

[29] L.D.Davisson: "Minimax Noiseless Universal Coding for Markov Sources", *IEEE Trans. Information Theory*, Vol.29, No.2, pp.211-215, (1983)

[30] 情報理論とその応用学会 編: 情報源符号化, 培風館, (1994)

[31] B.Efron: "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, Vol.7, pp.1-26, (1979)

[32] P.Elias: "Minimax Optimal Universal Codeword Sets", *IEEE Trans. Information Theory*, Vol.29, No.4, pp.491-502, (1983)

[33] R.Estrada and R.P.Kanwal: *Asymptotic Analysis – A Distributed Approach*, Birkhauser, (1994)

[34] W.Feller: *An Introduction to Probability and Its Applications*, Vol.I and II, John Wiley & Sons, (1957,66)

[35] T.S.Ferguson: *Mathematical Statistics, A Decision Theoretic Approach*, Academic Press, (1967)

[36] M.Feder and N.Merhav: "Hierarchical Universal Coding", *IEEE Trans. Information Theory*, Vol.42, pp.1354-1364, (1996)

[37] 福水健次: "特異な Fisher 情報行列を持つ線形ニューラルネットワークの汎化誤差", 電子情報通信学会技術研究報告 NC96-2, (1996)

[38] 後藤正幸, 橋川弘紀, 俵　信彦: "学習アルゴリズムが与える汎化への影響について", 日本経営工学会秋季全国大会予稿集, (1994)

[39] 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, No.10, pp.2123 - 2132, (1998)

[40] M.Gotoh and S.Hirasawa: "Statistical Model Selection based on Bayes Decision Theory and Its Application to Change Detection Problems", *International Journal of Production Economics*, Vol.60-61, pp.629-638, (1999)

[41] M.Gotoh, T.Matsushima, and S.Hirasawa: "On Error Rates of Statistical Model Selection Based on Information Criteria", *The Proceedings of IEEE 1998 International Symposium on Information Theory*, (1998)

[42] M.Gotoh, T.Matsushima, and S.Hirasawa: "An Analysis on Difference of Codelengths between Codes Based on MDL Principle and Bayes Codes", Submitted to *IEEE Trans. Information Theory*, (1999)

[43] 後藤正幸, 松嶋敏泰, 平澤茂一: "事前分布が異なる場合の MDL 原理に基づく符号とベイズ符号の符号長に関する解析", 電子情報通信学会誌 A, Vol.J-82-A, No.5, pp.629-638,(1999)

[44] M.Gotoh, T.Matsushima, and S.Hirasawa: "Almost Sure and Mean Convergence of Extended Stochastic Complexity", *IEICE Trans. Fundamentals*, Vol.82-A, No.10, pp.2129-2137, (1999)

[45] M.Gotoh, T.Matsushima, and S.Hirasawa: "Statistical Model Selection based on Bayes Decision Theory and Its Consistency for Hierarchical Model Class", Submitted to *IEICE Trans. Fundamentals*, (1999)

[46] 後藤正幸, 松嶋敏泰, 平澤茂一: "ベイズ決定理論に基づく統計的モデル選択の選択誤り率に関する解析", 日本経営工学会誌, Vol.51, No.1, to appear,(1999)

[47] Robert M.Gray: *Entropy and Information Theory*, Springer-Verlag, (1990)

[48] 萩原克幸, 戸田尚宏, 臼井支朗: "階層型ニューラルネットワークにおける結合重みの非一意性と AIC", 電子情報通信学会論文誌 D-II, Vol.J76-D-II, No.9, pp.2058-2065, (1993)

[49] 橋川弘紀, 後藤正幸, 俵 信彦: "階層型ニューラルネットワークの混合モデルによるベイズ最適な予測について", 電子情報通信学会論文誌 D-II, Vol.J80-D-II, No.7, pp.1919 - 1928, (1997)

[50] D.M.A.Haughton: "On the Choice of a Model to Fit Data from an Exponential Family", *The Annals of Statistics*, Vol.16, No.1, pp.342-355, (1988)

[51] J.A.Hartigan: *Bayes Theory*, Springer-Verlag, (1983)

[52] P.Hall and E.J.Hannan: "On Stochastic Complexity and Nonparametric Density Estimation", *Biometrika*, Vol.75, No.4, pp.705-714, (1988)

[53] James D.Hamilton: *Time Series Analysis*, Princeton University Press, (1994)

[54] 韓 太舜："情報圧縮とはなにか", 数理科学, Vol.290, pp.5-15, (1987)

[55] 韓 太舜：”情報理論の最近の展開: エントロピーからコンプレキシティへ”, システム/制御/情報, Vol.34, No.2, pp.71-80, (1990)

[56] 韓 太舜, 小林欣吾: 情報と符号化の数理 (岩波講座 応用数学 13), 岩波書店, (1994)

[57] E.J.Hannan and B.G.Quinn: ”The Determination of the Order of an Autoregression”, *Journal of Royal Statistical Society, Ser.B*, Vol.41, No.2, pp.190-195, (1979)

[58] 平澤茂一: 情報理論, 培風館, (1997)

[59] 広津千尋: 離散データ解析, 教育出版, (1982)

[60] C.C.Heyde and I.M.Johnstone: ”On Asymptotic Posterior Normality for Stochastic Process”, *Journal of Royal Statistical Society, Ser.B*, Vol.41, No.2, pp.184-189, (1979)

[61] 伊庭幸人: ”重回帰分析の変数選択におけるベイズ的方法 -平均場近似に基づくアプローチ-, 第 19 回情報理論とその応用シンポジウム予稿集, pp.533-536, (1996)

[62] 伊庭幸人: ”モデル選択とその周辺” Proceedings of IBIS’98 (1998 情報理論的学習理論ワークショップ予稿集), pp.79-86, (1998)

[63] 池田和司: ”追加学習の漸近論的解析”, 電子情報通信学会論文誌 D-II, Vol.J80-D-II, No.7, pp.1913 - 1918, (1997)

[64] 伊藤秀一: ”MDL のパターン認識への応用”, 人工知能学会誌, Vol.7, No.4, pp.608 - 614, (1992)

[65] 石田 崇, 後藤正幸, 平澤茂一: ”単語単位で出現する系列に対するベイズ符号について”, 電子情報通信学会 技術研究報告 IT99-27, (1999)

[66] 伊藤秀一, 土屋英亮: ”MDL の画像処理への応用” *Proceedings of IBIS’98* (1998 情報理論的学習理論ワークショップ予稿集), pp.17-24, (1998)

[67] 今井秀樹: ”情報理論”, 昭晃堂, (1984)

[68] 金谷健一: ”情報量基準による幾何学的モデルの選択”, 情報処理学会論文誌, Vol.37, No.6, pp.1073-1080, (1996)

[69] 金谷健一: ”幾何学的推定における精度の理論的限界と最適アルゴリズム” Proceedings of IBIS’98 (1998 情報理論的学習理論ワークショップ予稿集), pp.25-32, (1998)

[70] 川端 勉: ”ベイズ符号と文脈木重み付け法”, 電子情報通信学会技術研究報告 IT93-121,(1994)

[71] P.W.Laud and J.G.Ibrahim: ”Predictive Model Selection”, *Journal of Royal Statistical Society, Ser.B*, Vol.57, No.1, pp.247-262, (1995)

[72] T.Leonard: ”Comment on a Simple Predictive Density Function”, *Journal of the American Statistical Association*, Vol.77, No.379, pp.657-658, (1982)

[73] Roderick J.R.Little and Donald B.Bubin: *Statistical Analysis with Missing Data*, John Wiley & Sons, (1987)

[74] H.Linhart and W.Zucchini: *Model Selection*, John Wiley & Sons, (1986)

[75] A.Y.Lo: "On a Class of Bayesian Nonparametric Estimators: I. Density Estimators", *The Annals of Statistics*, Vol.12, No.1, pp.351-357, (1984)

[76] A.S.Martini and F.Spezzaferri : " A Predictive Model Selection Criterion ", *Journal of Royal Stasistical Society, Ser.B*, Vol.46, No.2, pp.296-303, (1984)

[77] 丸山英昭, 後藤正幸, 平澤茂一: "事後確率密度の漸近正規性に関する一考察", 電子情報通信学会 技術研究報告 IT99-28, (1999)

[78] 松嶋敏泰: "情報理論に基づく知識情報処理に関する研究", 早稲田大学博士論文, (1991)

[79] 松嶋敏泰: "統計モデル選択の概要", オペレーションズ・リサーチ, Vol.41, No.7, pp.369-374, (1996)

[80] T.Matsushima, H.Inazumi, and S.Hirasawa: "A Class of Distortionless Codes Designed by Bayes Decision Theory", *IEEE Trans. Information Theory*, Vol.37, No.5, pp.1288-1293, (1991)

[81] T.Matsushima, and S.Hirasawa: "A Bayes Coding Algorithm for Markov Models", 電子情報通信学会技術研究報告 IT95-1, (1995)

[82] N.Merhav: "Estimating with Partial Statistics the Parameter of Ergodic Finite Markov Sources", *IEEE Trans. Information Theory*, Vol.35, No.2, pp.326-333, (1989)

[83] N.Merhav and M.Feder: "Universal Prediction", *IEEE Trans. Information Theory*, Vol.44, No.6, pp.2124 - 2147, (1998)

[84] N.Merhav, M.Gutman, and J.Ziv: "On the Estimation of the Order of a Markov Chain and Universal Data Compression", *IEEE Trans. Information Theory*, Vol.35, No.5, (1989)

[85] N.Merhav and J.Ziv: "Estimating with Partial Statistics the Parameter of Ergodic Finite Markov Sources", *IEEE Trans. Information Theory*, Vol.35, No.2, pp.326-333, (1989)

[86] N.Murata, S.Yoshizawa and S.Amari: "Network Information criterion - Determining the number of Hidden Units for an Artificial Neural Network Model", *IEEE Trans. Neural Networks*, Vol.5, No.6, pp.865-872, (1994)

[87] 中尾 峰, 後藤正幸, 松嶋敏泰, 平澤茂一: "統計的モデル選択問題における選択誤り率に関する一考察", 電子情報通信学会技術研究報告 IT96, (1996)

[88] 中川聖一: 確率モデルによる音声認識, 電子情報通信学会, (1988)

[89] 長岡浩司, 菊池 靖, 池田浩二: "モデル選択における一致性と予測誤差について: AIC と MDL の比較", 第 15 回情報理論とその応用シンポジウム予稿集, pp.369-372, (1992)

[90] 尾崎 統: 時系列論, 日本放送出版協会, (1988)

[91] D.S.Poskitt: "A Bayes Procedure for the Identification of Univariate Time Series Models" , *The Annals of Statistics*, Vol.44, No.2, pp.502-516, (1987)

[92] D.S.Poskitt: "Precision, Complexity and Bayesian Model Determination" , *Journal of Royal Statistical Society, Ser.B*, Vol.49, No.2, pp.199-208, (1987)

[93] G.Qian, G.Gabor, and R.P.Gupta: " On Stochastic Complexity Estimation : A Decision-Theoretic Approach", *IEEE Trans. Information Theory*, Vol.40, No.4, pp.1181-1191, (1994)

[94] J.R.Quinlan and R.Rivest: "Inferring Decision Trees Using Minimum Description Length Principle", *Information and Computation*, Vol.80, No.3, pp.227-248, (1989)

[95] J.Rissanen: "Modeling By Shortest Data Description", *Automatica*, Vol.46, pp.465-471, (1978)

[96] J.Rissanen: "Universal Modeling and Coding", *IEEE Trans. Information Theory*, Vol.27, No.1, pp.12-23, (1981)

[97] J.Rissanen: "A Universal Prior for Integers and Estimation by Minimum Description Length", *The Annals of Statistics*, Vol.11, No.2, pp.416-431, (1983)

[98] J.Rissanen: "Universal Coding, Information, Prediction, and Estimation", *IEEE Trans. Information Theory*, IT-30, No.4, (1984)

[99] J.Rissanen: "Stochastic Complexity and Modeling", *The Annals of Statistics*, Vol.14, No.3, pp.1080-1100, (1986)

[100] J.Rissanen: "Stochastic Complexity", *Journal of Royal Statistical Society*, Ser.B, Vol.49, pp.223-265, (1987)

[101] J.Rissanen: "Fisher Information and Stochastic Complexity", *IEEE Trans. Information Theory*, Vol.42, No.1, pp.40 - 47, (1996)

[102] C.Schwarz: "Estimating the Dimension of a Model", *The Annals of Statistics*, Vol.6, pp.461-464, (1978)

[103] 坂元慶行: カテゴリカルデータのモデル分析, 共立出版, (1985)

[104] 坂元慶行, 石黒真木夫, 北川源四郎: 情報量統計学, 共立出版, (1983)

[105] 佐藤 担: 測度から確率へ, 共立出版, (1994)

[106] 関 庸一, 橋本 巧: "MDL 基準に基づく正規母集団変化時点検出に関する研究", 日本経営工学会誌, Vol.47, No.3, pp.192 - 198, (1996)

[107] 関 庸一: "経営工学におけるモデル選択", オペレーションズ・リサーチ, Vol.41, No.7, pp.387-391, (1996)

[108] R.Shibata: "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion", *Biometrika*, Vol.63, No.1, pp.117-126, (1976)

[109] R.Shibata: "Asymptotically Efficient Selection of The Order of The Model for Estimating Parameters of a Linear Process", *Annals of Statistics*, Vol.8, pp.147-164, (1980)

[110] R.Shibata: "An Optimal Selection of Regression Variables", *Biometrika*, Vol.68, pp.45-54, (1981)

[111] R.Shibata: "Consistency of Model Selection and Parameter Estimation", *Applied Probability Trust*, pp.127-141, (1986)

[112] 柴田里程: "変数選択理論の現状", 数学, Vol.36, pp.344-352, (1988)

[113] 繁桝算男: ベイズ統計入門, 東京大学出版会, (1985)

[114] 島 幸夫, 後藤正幸, 松嶋敏泰, 平澤茂一: "モデル族の部分集合に基づく予測方法について", 第21回情報理論とその応用シンポジウム予稿集, pp.149-152, (1998)

[115] R.Shimizu: "Entropy Maximization Principle and Selection of the Order of an Autoregressive Gaussian Process", *The Annals of Statistics and Mathematics*, Vol.30, Part A, pp.263-270, (1978)

[116] 下平英寿: "不完全対数線形モデルにおける新しい情報量規準", http://www.ism.ac.jp/ shimo/, (1992)

[117] 下平英寿: "統計的モデル選択の研究 -モデルの信頼集合の構成-", 東京大学博士論文, (1995)

[118] 下平英寿: "モデル選択理論の新展開", http://www.ism.ac.jp/ shimo/, (1997)

[119] M.Stone: "Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of Royal Statistical Society, Ser.B*, Vol.36, pp.111-133, (1974)

[120] M.Stone: "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion", *Journal of Royal Statistical Society, Ser.B*, Vol.39, pp.44-47, (1977)

[121] W.Stout: *Almost Sure Convergence*, Academic Press, (1974)

[122] T.P.Speed and B.Yu: "Model Selection and Prediction: Normal Regression", *The Annals of Inst. Statistics and Mathematics*, Vol.45, pp.35-54, (1993)

[123] 鈴木 譲: "知識情報処理における規則の学習および事実の認識に関する研究", 早稲田大学博士論文, (1993)

[124] J.Suzuki: "A Universal Coding Scheme Based on Minimizing Minimax Redundancy with an Unknown Model", *IEICE Trans. Fundamentals*, Vol.E76-A, No.7, pp.1234-1239, (1993)

[125] J.Suzuki: "Evaluations for Estimation of an Information Source Based on State Decomposition", *IEICE Trans. Fundamentals*, Vol.E76-A, No.7, pp.1240-1251, (1993)

[126] J.Suzuki: "Some Notes on Universal Noiseless Coding", *IEICE Trans. Fundamentals*, Vol.E78-A, No.12, (1995)

[127] J.Suzuki: "Learning Bayesian Belief Networks based on the MDL Principle", *Proceedings of IBIS'98* (1998 情報理論的学習理論ワークショップ予稿集), pp.131-138, (1998)

[128] 鈴木雪夫, 国友直人 編: ベイズ統計学とその応用, 東京大学出版会, (1989)

[129] 高井 望, 浮田善文, 松嶋敏泰: "音声認識における Hidden Markov Model のパラメータ推定に関する研究", 第21回情報理論とその応用シンポジウム予稿集, pp.643-646, (1998)

[130] 竹内 啓: 統計的推定の漸近理論, 教育出版, (1974)

[131] 竹内 啓: "情報統計量の分布とモデルの適切さの規準", 数理科学, No.153, pp.12-18, (1976)

[132] 竹内 啓: "AIC 基準による統計的モデルの選択をめぐって", 計測と制御, Vol.22, No.5, pp.445-453, (1983)

[133] 竹内 啓ほか編: 統計学事典, 第 III 部, 13 章, モデル選択, pp.459-465, 東洋経済新報, (1989)

[134] 竹内 啓: 統計的方法 (岩波講座 応用数学 11), 岩波書店, (1994)

[135] 竹内純一: "MDL 推定量の Kullback - Leibler 情報量に関する収束速度について", 第 16 回情報理論とその応用シンポジウム予稿集, pp.161 -164, (1993)

[136] 竹内純一: "パラメータ推定問題における MDL 原理とベイズ符号について", 電子情報通信学会技術研究報告 IT93-122, pp.13-18, (1994)

[137] J.Takeuchi: "Characterization of the Bayes Estimator and the MDL Estimator for Exponential Families", *IEEE Trans. Information Theory*, Vol.43, No.4, pp.1165-1174, (1996)

[138] 竹内純一: "確率的コンプレキシティと Jeffreys 混合予測戦略" Proceedings of IBIS'98 (1998 情報理論的学習理論ワークショップ予稿集), pp.9-16, (1998)

[139] J.Takeuchi and A.R.Barron: "Asymptotically Minimax Regret for Exponential and Curved Exponential Families", *The Proceedings of IEEE 1998 International Symposium on Information Theory*, (1998)

[140] J.Takeuchi and A.R.Barron: "Robustly Minimax Codes for Universal Data Compression", 第 21 回情報理論とその応用シンポジウム予稿集, pp.213-316, (1998)

[141] 高橋伸夫: 組織の中の決定理論, 朝倉書店, (1993)

[142] L.Tierney and J.B.Kadane: "Accurate Approximation for Posterior Moments and Marginal Densities", *Journal of American Statistics Association*, Vol.81, pp.82-86, (1986)

[143] H.Tsuchiya, S.Itoh, and T.Hashimoto: "An Algorithm for Designing a Pattern Classifier by Using MDL Criterion", *IEICE Trans. Fundamentals*, Vol.E79-A, No.6, pp.910-920,(1996)

[144] 植松友彦: 現代シャノン理論, 培風館, (1998)

[145] A.M.Walker: "On the Asymptotic Behaviour of Posterior Distributions", *Journal of Royal Statistical Society, Ser.B*, Vol.31, pp.80-88, (1969)

[146] C.S.Wallace and P.R.Freeman: "Estimation and Inference by Compact Coding", *Journal of Royal Statistical Society*, Ser.B, Vol.49, No.3, pp.240-265, (1987)

[147] 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", 電子情報通信学会論文誌 A, Vol.J81-A, No.10, pp.1442-1452, (1998)

[148] F.M.J.Willems, Y.M.Shtarkov, and T.J.Tjalkens: "The Context - Tree Weighting Method: Basic Properties", *IEEE Trans. Information Theory*, Vol.41, No.3, pp.653 - 664, (1995)

[149] Q.Xie and A.R.Barron: "Minimax Redundancy for the Class of Memoryless Souece", *IEEE Trans. Information Theory*, Vol.43, No.2, pp.646 - 657, (1997)

[150] K.Yamanishi: " A Learning Criterion for Stochastic Rules", *Machine Learning*, Vol.9, pp.165-203, (1992)

[151] K.Yamanishi: " A Loss Bound Model for On-Line Stochastic Prediction Algorithms ", *Information and Computation*, 119, pp.39-54, (1995)

[152] 山西健司: "確率的コンプレキシティと学習理論", オペレーションズ・リサーチ, Vol.41, No.7, pp379-386, (1996)

[153] K.Yamanishi: "A Decision-Theoretic Extension of Stochastic Complexity and Its Applications to Learning", *IEEE Trans. Information Theory*, Vol.44, No.4, pp.1424-1439,(1998)

[154] 山西健司: "拡張型確率的コンプレキシティと学習理論", Proceedings of IBIS'98 (1998情報理論的学習理論ワークショップ予稿集), pp.33 - 40, (1998)

[155] 山西健司, 韓 太舜: "MDL 入門:情報理論の立場から", 人工知能学会誌, Vol.7, No.3, pp.427-434, (1992)

[156] 山下公子, 後藤正幸, 俵 信彦: "ベイズ的アプローチによるパラメータ変化時点の検出に関する一考察", 日本経営工学会 秋期大会予稿集, pp.105 -106, (1996)

[157] B.Yu: "Minimum Description Length Principle: A Review", *Proceeding of International Symposium of Information Theory and Its Application*, pp.432-435, (1996)

[158] Y.Yang and A.R.Barron: "An Asymptotic Property of Model Selection Criteria", *IEEE Trans. Information Theory*, Vol.44, No.1, pp.95-116, (1998)

[159] J.Ziv and A.Lempel: "A Universal Algorithm for Sequential Data Compression", *IEEE Trans. Information Theory*, Vol.23, No.3, pp.337-343, (1977)

[160] J.Ziv and A.Lempel: "Compression of Individual Sequences Via Variable-Rate Coding", *IEEE Trans. Information Theory*, Vol.24, No.5, pp.530-536, (1978)

[161] J.Ziv: "On The Classification with Empirically Observed Statistics and Universal Data Compression", *IEEE Trans. Information Theory*, Vol.IT-34, No.2, pp.278-286, (1988)

研究業績

種別	題名	発表年月	論文誌名	共著者
論文				
○1	共役勾配法を導入した BP 学習における安定化に関する研究	1995,4	日本経営工学会誌 Vol.46, No.1	俵　信彦
○2	変傾共役勾配法による BP 学習の安定化と高速化	1995,6	日本経営工学会誌 Vol.46, No.2	開沼泰隆 俵　信彦
3	最適レギュレータに基づく定期発注システムに関する研究	1996,2	日本経営工学会誌 Vol.46, No.6	田村嘉浩 俵　信彦
○4	FK 型発注システムによる定期発注システムの統一的考察	1996,2	日本経営工学会誌 Vol.46, No.6	内園みどり 俵　信彦
○5	A Formulation by Minimization of Differential Entropy for Optimal Control System	1996,4	IEICE Trans. on Fundamentals, Vol.E79-A, No.4	平澤茂一 俵　信彦
○6	有色雑音を持つ線形システムの最適制御則と定期発注システムへの適用	1996,6	日本経営工学会誌 Vol.47, No.2	内園みどり 俵　信彦
7	階層型ニューラルネットワークの混合モデルによるベイズ最適な予測について	1997,7	電子情報通信学会論文誌 D-II, Vol.80-D-II, No.7	橋川弘紀 俵　信彦
8	共役勾配法の探索効率向上法に関する一考察	1997,12	日本経営工学会誌 Vol.48, No.5	吉田隆弘 俵　信彦
9	発注サイクル期間を考慮した Push 型生産システムと Pull 型生産システムの特性解析	1998,9	日本経営工学会誌 Vol.49, No.3	加藤宏幸 俵　信彦
◎10	A Generalization of B.S.Clarke and A.R.Barron's Asymptotics of Bayes Codes for FSMX Sources	1998,10	IEICE Trans. on Fundamentals, Vol.E81-A, No.10	松嶋敏泰 平澤茂一
◎11	Statistical Model Selection Based on Bayes Decision Theory and Its Application to Change Detection Problem	1999,4	International Journal of Production Economics, Vol.60-61	平澤茂一

種別	題名	発表年月	論文誌名	共著者
論文				
◎ 12	事前分布が異なる場合の MDL 原理に基づく符号とベイズ符号の符号長に関する解析	1999,5	電子情報通信学会論文誌 A, Vol.J-82, No.5	松嶋敏泰 平澤茂一
13	マルコフ情報源に対し誤り伝搬を抑える VF 符号に関する一考察	1999,5	電子情報通信学会誌 A (研究速報), Vol.J-82, No.5	木村　勝 松嶋敏泰 平澤茂一
14	順序カテゴリカルデータ解析における母数推定に関する研究	1999,8	日本経営工学会誌, Vol.50, No.3	菊池淑子 俵　信彦
◎ 15	Almost Sure and Mean Convergence of Extended Stochastic Complexity	1999,10	IEICE Trans. on Fundamentals, Vol.E82-A, No.10	松嶋敏泰 平澤茂一
16	在庫量・発注量変動により発生するコストを制御する定期発注方式に関する研究	1999,10	日本経営工学会誌, Vol.50, No.4	西島　淳 俵　信彦
◎ 17	ベイズ決定理論に基づく統計的モデル選択の選択誤り率に関する解析	2000,2	日本経営工学会誌, Vol.50, No.6	松嶋敏泰 平澤茂一
18	線形回帰モデルのベイズ最適な予測法に関する研究	掲載決定	日本経営工学会誌, Vol.51, No.1	鈴木友彦 俵　信彦
○ 19	損失関数を考慮した拡張事後密度の漸近正規性	条件付採録	電子情報通信学会論文誌 A	松嶋敏泰 平澤茂一
◎ 20	An Analysis on Difference of Codelengths between Codes based on MDL Principle and Bayes Codes	条件付採録	IEEE Trans. Information Theory	松嶋敏泰 平澤茂一
◎ 21	Statistical Model Selection based on Bayes Decision Theory and Its Consistency for Hierarchical Model Class	投稿中	IEICE Trans. on Fundamentals	松嶋敏泰 平澤茂一

○は筆頭著者としての論文，◎は学位論文に関係する筆頭著者論文

種別	題名	発表年月	論文誌名	共著者
国際会議				
1	A Fast Learning Algorithm for Multilayer Neural Networks	1993,8	ICPR Production Research 93 (Finland)	開沼泰隆 俵 信彦
○ 2	An Analysis on Difference between the Code based on MDL Principle and the Bayes Code	1996,9	International Symposium on Information Theory and Its Application 96 (Canada)	松嶋敏泰 平澤茂一
○ 3	A Study on Difference of Codelengths between MDL Codes and Bayes Codes on Case Different Priors Are Assumed	1997,6	IEEE International Symposium on Information Theory 1997 (Germany)	松嶋敏泰 平澤茂一
○ 4	On Theory and Application of Statistical Model Selection Based on Bayes Decision Theory	1997,8	ICPR Production Research 97 (Japan)	松嶋敏泰 平澤茂一
○ 5	On Error Rates of Statistical Madel Selection Based on Information Criteria	1998,8	IEEE International Symposium on Information Theory 1998 (USA)	松嶋敏泰 平澤茂一
○ 6	Consistency of Bayesian Model Selection	1998,10	International Symposium on Information Theory and Its Applications 1998 (Mexico)	松嶋敏泰 平澤茂一

○ は筆頭著者としての発表

種別	題名	発表年月	論文誌名	共著者
講演				
1	バックプロパゲーション学習アルゴリズムについて	1993,5	日本経営工学会春期大会 (東京)	開沼泰隆 俵　信彦
○ 2	共役勾配法による BP 学習について	1993,11	日本経営工学会 秋季大会 (大分)	開沼泰隆 俵　信彦
3	階層型ニューラルネットワークにおける入力変数の指定法	1994,5	日本経営工学会 春期大会 (神奈川)	橋川弘紀 俵　信彦
○ 4	二次近似範囲を調整する共役勾配法と BP 学習への適用	1994,10	電子情報通信学会 技術研究報告 NC94-37 (仙台)	俵　信彦
○ 5	学習アルゴリズムが与える汎化への影響について	1994,11	日本経営工学会 秋季大会 (岡山)	橋川弘紀 俵　信彦
○ 6	エントロピー最小化によるフィードバック制御の定式化	1994,12	第 17 回情報理論とその応用シンポジウム (広島)	平澤茂一 俵　信彦
7	2 値分類問題へのニューラルネットワークの適用に関する一考察	1994,12	第 17 回情報理論とその応用シンポジウム (広島)	松原徳明 松嶋敏泰 平澤茂一
○ 8	有色雑音を持つ確率システムのフィードバック制御と生産-在庫システムへの適用	1995,5	日本経営工学会 春季大会 (東京)	内園みどり 俵　信彦
9	情報量基準による階層型ニューラルネットワークのモデル構造決定方法	1995,5	日本経営工学会 春季大会 (東京)	開沼泰隆 橋川弘紀 俵　信彦
10	情報源モデルのパラメータ推定と Ziv-Lempel 符号について	1995,7	電子情報通信学会 技術研究報告 IT95- 21 (東京)	木村　勝 松嶋敏泰 平澤茂一
○ 11	汎化を考慮した確率モデルの学習について	1995,9	情報処理学会全国大会 (富山)	松嶋敏泰 平澤茂一
○ 12	ベイズ決定理論に基づくデータ解析に関する一考察	1995,10	第 18 回情報理論とその応用シンポジウム (花巻)	松嶋敏泰 平澤茂一
13	ユニバーサル情報源符号化アルゴリズムに基づく事後確率推定アルゴリズム	1995,10	第 18 回情報理論とその応用シンポジウム (花巻)	阿部真士 松嶋敏泰 平澤茂一
14	観測雑音を考慮した不確実な知識の推論法について	1995,10	第 18 回情報理論とその応用シンポジウム (花巻)	川又英紀 松嶋敏泰 平澤茂一
○ 15	確率モデルの学習に関する一考察	1995,11	日本経営工学会 秋季大会 (大阪)	松嶋敏泰 平澤茂一 俵　信彦
16	最適制御を用いた発注方式の生産－在庫システムへの適用	1995,11	日本経営工学会 秋季大会 (大阪)	内園みどり 俵　信彦

種別	題名	発表年月	論文誌名	共著者
講演				
17	階層型ニューラルネットワークの混合モデルによるベイズ最適な予測について	1996,3	電子情報通信学会 技術研究報告 NC95 -121 (神奈川)	橋川弘紀 俵　信彦
○ 18	情報源の変化時点検出における MDL 原理とベイズ理論について	1996,5	日本経営工学会 春季大会 (東京)	松嶋敏泰 平澤茂一 俵　信彦
19	共役勾配法における探索効率向上法について	1996,5	日本経営工学会 春季大会 (東京)	吉田隆弘 俵　信彦
○ 20	MDL 基準に基づく符号とベイズ符号の符号長に関する解析	1996,5	電子情報通信学会技術研究報告 IT96 - 3 (東京)	松嶋敏泰 平澤茂一
21	統計的モデル選択問題における選択誤り確率に関する一考察	1996,7	電子情報通信学会 技術研究報告 IT96 (神奈川)	中尾　峰 松嶋敏泰 平澤茂一
○ 22	符号長を評価とした確率モデルの学習に関する一考察	1996,10	日本経営工学会 秋季大会 (名古屋)	松嶋敏泰 平澤茂一 俵　信彦
23	母集団のパラメータ変化時点検出のベイズアプローチに関する一考察	1996,10	日本経営工学会 秋季大会 (名古屋)	山下公子 俵　信彦
○ 24	異なる事前分布を持つ場合の MDL 原理に基づく符号とベイズ符号の符号長に関する一考察	1996,12	第 19 回情報理論とその応用シンポジウム (神奈川)	松嶋敏泰 平澤茂一
25	マルコフ情報源に対する VF 符号に関する一考察	1996,12	第 19 回情報理論とその応用シンポジウム (神奈川)	木村　勝 松嶋敏泰 平澤茂一
○ 26	Clarke と Barron の Bayesian Asymptotics について	1997,1	電子情報通信学会 技術研究報告 IT96-59 (金沢)	松嶋敏泰 平澤茂一
27	String Matching Algorithm による有歪み圧縮について	1997,1	電子情報通信学会 技術研究報告 IT96-60 (金沢)	小幡洋昭 平澤茂一
○ 28	ベイズ決定理論に基づく統計的モデル選択について	1997,7	電子情報通信学会 技術研究報告 IT97-21 (京都)	松嶋敏泰 平澤茂一
29	木構造のモデル族の学習・予測アルゴリズムに関する一考察	1997,7	電子情報通信学会 技術研究報告 IT97-36 (東京)	島　幸夫 松嶋敏泰 平澤茂一
30	線形回帰モデルのベイズ最適な予測法に関する研究	1997,11	日本経営工学会 秋季大会	鈴木友彦 俵　信彦

種別	題名	発表年月	論文誌名	共著者
講演				
31	分割表における母数推定に関する研究	1997,11	日本経営工学会 秋季大会	菊池淑子 俵　信彦
32	在庫量・発注量変動によるコスト増を考慮した定期発注方式に関する研究	1997,11	日本経営工学会 秋季大会	西嶋　淳 俵　信彦
33	指数分布に対する事前分布について ～ 寿命分布に対するベイズ推定量に関する研究 ～	1997,11	日本経営工学会 秋季大会	青木美穂 俵　信彦
34	木構造モデル族のモデル選択法に関する一考察	1997,12	第 20 回情報理論とその応用シンポジウム (愛媛)	中尾　峰 松嶋敏泰 平澤茂一
○ 35	階層モデル族のモデル選択における選択誤り率について	1998,1	電子情報通信学会 技術研究報告 IT97-64 (東京)	松嶋敏泰 平澤茂一
○ 36	事後確率密度の漸近正規性について	1998,5	日本経営工学会春季大会 (東京)	松嶋敏泰 平澤茂一
37	符号語コストを考慮した情報源符号化について	1998,5	日本経営工学会春季大会 (東京)	吉田隆弘 俵　信彦
38	部分的な復号を可能にする LZ78 符号の修正符号	1998,7	電子情報通信学会 技術研究報告 IT98-28 (東京)	対馬克也 平澤茂一
39	String Matching に基づく有歪み圧縮に関する研究	1998,7	電子情報通信学会 技術研究報告 IT98-29 (東京)	前田浩二 平澤茂一
○ 40	統計的モデル選択の規準について	1998,11	日本経営工学会秋期大会	松嶋敏泰 平澤茂一 俵　信彦
○ 41	Extended Stochastic Complexity の漸近式について	1998,12	第 21 回情報理論とその応用シンポジウム (名古屋)	松嶋敏泰 平澤茂一
42	モデル族の部分集合に基づく予測について	1998,12	第 21 回情報理論とその応用シンポジウム (名古屋)	島　幸夫 松嶋敏泰 平澤茂一
43	事後確率密度の漸近正規性に関する一考察	1999,7	電子情報通信学会 技術研究報告 IT99-27	丸山英昭 平澤茂一
44	単語単位で出現する系列に対するベイズ符号について	1999,7	電子情報通信学会 技術研究報告 IT99-28	石田　崇 平澤茂一
○ 45	On Asymptotic Normality of Density Functions and Its Applications	1999,8	1999 年情報論的学習理論ワークショップ予稿集 (静岡)	松嶋敏泰 平澤茂一
46	一対比較行列が不完全な場合の AHP に関する研究	1999,10	日本経営工学会秋期大会	本川　亨 俵　信彦

種別	題名	発表年月	論文誌名	共著者
講演				
47	線形回帰モデルにおけるベイズ最適な予測法に関する研究	1999,10	日本経営工学会秋期大会	肥土雅生 俵　信彦
48	業務プロセスにおけるアクティビティの関連性の定量化手法に関する一研究	1999,10	日本経営工学会秋期大会	平山宗一郎 俵　信彦
○ 49	単語単位で系列を出力する情報源	1999,12	第 22 回情報理論とその応用シンポジウム	松嶋敏泰 平澤茂一
50	統計的推定の立場からみた Ziv-Lempel 符号に関する一考察	1999,12	第 22 回情報理論とその応用シンポジウム	対馬克也 平澤茂一

○ は筆頭著者としての発表