

Bayesian Data Analysis : Chapter4

Asymptotics and connections to non-Bayesian approaches

Andrew Gelman, John B.Carlin, Hal S.Stern,
David B.Dunson, Aki Vehtari, and Donald B.Rubin

山崎 遼也

情報学科 数理工学コース 4 回

2017/4/4～5

Table Contents

1 4.1 Normal approximations to the posterior distribution

2 4.2 Large-sample theory

3 4.3 Counterexamples to the theorems

4 4.4 Frequency evaluations of Bayesian inferences

5 4.5 Bayesian interpretations of other statistical methods

Key Words

- 漸近理論
- 一緻性
- 正規性
- 漸近正規性
- フラットな尤度
- エイリアシング
- 安定推定
- 漸近的不偏性
- 効率的
- 信頼区間
- 比率推定
- 統計的有意性
- ノンパラメトリック法
- ジャックナイフ
- ブートストラップ
- Wilcoxon 順位検定

無情報事前分布に基づく多くの単純なベイズ分析が、標準的な非ベイズ手法 (例えば, 未知の分散を伴う正規平均の事後 t 区間) に類似の結果を与えることを見出した. 無情報事前分布が客観的な仮定として正当化できるかどうかは, データで利用可能な情報の量によって決まる. 2, 3 章で議論された単純な場合では, サンプルサイズ n が増加すると, 事後分布に対する事前分布の影響が減少することが明らかになった. これらの考え方は, **漸近理論 (asymptotic theory)** と呼ばれ, n が大きくなるにつれて極限にある性質を参照するため, この章ではベイズ手法と非ベイズ手法との関係をより明示的に論じる. 大きなサンプルの結果は, ベイズデータ分析を実行するために実際には必要ではありませんが, しばしば近似として, そして理解のためのツールとして有用です.

この章では, 事後分布に対する正規近似のさまざまな使用法について説明します. 大きなサンプルの事後分布の **一貫性 (consistency)** と **正規性 (normality)** に関する定理は 4.2 節で概説され, 4.3 節でいくつかの反例が続く. 定理の証明は付録 B に示されている. 最後に, ベイズ推論の性質を評価するために頻度統計の方法をどのように使用することができるかを議論する.

4.1 事後分布の正規近似

結合事後分布の正規近似

事後分布 $p(\theta|y)$ が単峰性でおおよそ対称であれば、それを正規分布で近似することが便利である。すなわち、事後密度の対数は θ の二次関数で近似される。

ここでは、事後モードを中心とする対数事後密度 (一般に既存の最適化ルーチンを使用して計算が容易である) に対する二次近似を考える。13章では、単純なモードに基づく近似が失敗する設定で効果的な、より精巧な近似について説明します。

事後モード $\hat{\theta}$ を中心とする $\log p(\theta|y)$ のテイラー級数展開 (ここで、 θ はベクトルであり、 $\hat{\theta}$ はパラメータ空間の内部にあると仮定する) は

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots \quad (4.1)$$

を与える。

ここで, 対数事後密度はそのモードで微分が 0 になるから, 展開の中の線形項は 0 である. 4.2 節で議論するように, 残りの高次項は θ が $\hat{\theta}$ に近く, n が大きいときの二次項に対して重要である. θ の関数として (4.1) を考えれば, 第 1 項は定数で, 第 2 項は近似

$$p(\theta|y) \simeq N(\hat{\theta}, [I(\hat{\theta})]^{-1}) \quad (4.2)$$

となる, 正規密度に比例する. ここで, $I(\theta)$ は観測値の情報量である

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

である. モード $\hat{\theta}$ がパラメータ空間内にあるならば, 行列 $I(\hat{\theta})$ は正定値である.

例. 未知の平均, 分散に関する正規分布

サンプル定理の例に関する近似正規分布を示す. y_1, \dots, y_n を $N(\mu, \sigma^2)$ 分布に従う独立な観測値とし, 簡単のため, $(\mu, \log \sigma)$ に対して一様事前密度を仮定する. $(\mu, \log \sigma)$ の事後分布に正規近似を行う. これは, σ を正の値に制限するという利点がある. 近似を構成するには, 対数事後密度の二次導関数が必要であり,

$$\log p(\mu, \log \sigma | y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2} ((n-1)s^2 + n(\bar{y} - \mu)^2)$$

1 階微分は

$$\begin{aligned} \frac{d}{d\mu} \log p(\mu, \log \sigma | y) &= \frac{n(\bar{y} - \mu)}{\sigma^2} \\ \frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma | y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2} \end{aligned}$$

よって事後モードから,

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \log \left(\sqrt{\frac{n-1}{n}} s \right) \right)$$

が得られる.

対数事後密度の 2 階微分は

$$\begin{aligned}\frac{d^2}{d\mu^2} \log p(\mu \log \sigma | y) &= -\frac{n}{\sigma^2} \\ \frac{d^2}{d\mu d(\log \sigma)} \log p(\mu \log \sigma | y) &= -2n \frac{\bar{y} - \mu}{\sigma^2} \\ \frac{d^2}{d(\log \sigma)^2} \log p(\mu \log \sigma | y) &= -\frac{2}{\sigma^2} ((n-1)s^2 + n(\bar{y} - \mu)^2)\end{aligned}$$

モードでの 2 階微分の行列は $\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}$ である. (4.2) から, 事後分布は

$$p(\mu, \log \sigma | y) \approx N \left(\begin{pmatrix} \mu \\ \log \sigma \end{pmatrix} \middle| \begin{pmatrix} \hat{\mu} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix} \right)$$

と近似できる. もし $p(\mu, \sigma^2)$ に関して正規近似を行おうとするならば, 2 階微分行列は $\log \sigma$ から σ^2 への変換のヤコビアンをかけられ, モードはわずかに $\tilde{\sigma}^2 = \frac{n}{n+2} \hat{\sigma}^2$ と変化する. 2 つの要素 (μ, σ^2) は近時事後分布においても独立で, $p(\sigma^2 | y) \approx N(\sigma^2 | \tilde{\sigma}^2, 2\tilde{\sigma}^4/(n+2))$ である.

その最大値に関する事後密度関数の解釈

近似としての直接使用に加えて, 多変量正規分布は, 事後密度関数の解釈および等高線プロットのための基準を提供する. d 次元の正規分布では, 密度関数の対数は, $\text{constant} + \chi_d^2$ の分布を -2 で割ったものです. 例えば, χ_{10}^2 密度の 95 パーセンタイルは 18.31 であるので, 問題が $d = 10$ のパラメータを有する場合, 事後確率質量の約 95% は, $p(\theta|y)$ が 0 モードでの密度の $\exp(-18.31/2) = 1.1 \times 10^{-4}$ 倍未満である θ の値と関連する. 同様に, $d = 2$ のパラメータでは, 事後質量の約 95% が, モードでの密度に対して, $\exp(-5.99/2) = 0.05$ より高い密度に対応する. 事後密度の 2 次元等高線プロット (例えば, 図 3.3a) では, 0.05 等高線には, 確率質量の約 95% が含まれています.

点推定による事後分布の要約と標準誤差

4.2 節で概要を示した漸近定理は n が十分大きい場合、事後分布は正規分布によって近似されうるということを示す。応用の多くの分野で、標準的な推論要約は、最大尤度推定値 (一様事前密度の事後モードである)、 $\pm 2 \times$ 標準誤差などの点推定値 $\hat{\theta}$ を計算することによって得られる 95% 推定値 $I(\hat{\theta})$ の情報から推定された標準誤差である。異なる漸近議論は非ベイズ、この要約の頻度主義的解釈を正当化するが、多くの簡単な状況では、両方の解釈が成り立つ。ベイズの観点から、与えられた例での精度は事後分布を検査することによって直接的に決定することができる。

多くの場合で、パラメータ θ に対する事後分布の正規分布への収束は変換により劇的に向上する。もし ϕ が θ の連続変換ならば、 $p(\phi|y), p(\theta|y)$ は正規分布に近づくが、有限 n に対する近似の近さは、選択された変換によって実質的に変化し得る。

データ削減と要約統計量

正規近似の下で, 事後分布はそのモード $\hat{\theta}$ と事後密度の曲率 $I(\hat{\theta})$ で要約される. すなわち, 漸近的に, 十分統計量が存在する. 次の節の終わりの例で, それらの通常の理論によって十分統計量によって, 多くの情報源からの地方レベル, または個人レベルのデータを要約すると便利であることがわかるはずである. 要約統計量を使用するこのアプローチは, 階層的モデリング技術を比較的容易に適用して個々の推定を改善することを可能にする. 例えば, 5.5 節では, 一組の 8 つの実験の各々を, 以前の線形回帰分析から推定された点推定値および標準誤差によって要約する. 事後分布が正常に近い場合, 要約統計量を使用することは明らかに最も合理的です. 他の方法では重要な情報を破棄して誤った推論につなげることがある.

より低い次元の正規近似

有限サンプルサイズ n の場合, 通常近似は, θ の成分の条件付き分布および周辺分布について, 完全な結合分布よりも, より正確である. 例えば, 結合分布が多変量正規分布である場合, その周辺化はすべて正規ですが, 逆は真ではありません. θ の成分の限界分布を決定することは, θ の他の全ての成分で平均することと等価であり, 分布族の平均化は, 一般に, 中央極限定理の基礎となる同じ論理によって, それらを通常に近づける.

低次元 θ の事後分布の正規近似は, しばしば, 特に適切な変換後に, 完全に許容されることが多い. θ が高次元である場合, 一般に 2 つの状況が生じる. 第 1 に, θ の多くの個々の成分の周辺分布は, 近似的に正常であり得る. 個別に取られたこれらのパラメータのいずれか 1 つについての推論は, 点推定値および標準誤差によって十分に要約することができる. 第 2 に, θ を 2 つの部分ベクトル, $\theta = (\theta_1, \theta_2)$ に分割することが可能であり, $p(\theta_2|y)$ は必ずしも正規分布に近くないが, $p(\theta_1|\theta_2, y)$ はおそらく θ_2 の関数である平均と分散に関係している. 条件付き分布を用いた近似手法はしばしば有用であり, 13.5 節でそれをより体系的に考える. 低次元の近似は, 潜在的なガウスモデルの計算など, ますます普及している.

最後に, 正規分布に基づく近似は, コンピュータプログラムをデバッグしたり, 事後分布を近似するためのより精巧な方法を調べるのに役立つ.

例. バイオアッセイ実験 (続き)

3.7 節のバイオアッセイ実験のモデルとデータへの正規近似を示します. この実験でのサンプルサイズは比較的小さく, わずか 20 匹の動物であり, 正規近似は正確な事後分布に近いが, 重要な違いがあることがわかる.

(α, β) の結合事後分布への正規近似

初めに, (ロジスティック回帰問題を用いて) 事後分布のモードとモードで評価された正規近似 (4.2) を計算する. (α, β) の事後近似は, (α, β) に対して一様事前密度を仮定したので, 最大尤度推定量と同じである. 図 4.1 は, 二変量正規近似の等高線プロットと, この近似分布からの 1000 点の散布図を示しています. プロットは, 図 3.3 の実際の事後分布のプロットに似ていますが, 以前のプロットの右上隅に歪度はありません.

$(\alpha, \beta) = (0.8, 7.7)$ を実際の事後分布の平均 $(\alpha, \beta) = (1.4, 11.9)$ と比較したときの歪度の効果は明らかである. 図 3.3b に示されているシミュレーションから選択します.

(α, β) への正規近似を用いた LD50 に対する事後分布, 正規近似の流れ

通常の近似からの同じ 1000 回の抽出を使用して, β が正である確率を推定し, LD50 の事後分布 (条件付き β が正である) を推定することができる. 1000 シミュレーション抽出のうち, 950 は β の正の値を有し, 正確な分布 ($\Pr(\beta > 0) > 0.999$) とは異なる結果である推定値 $\Pr(\beta > 0) = 0.95$ を生じた. 引き続き通常の近似に基づく解析を行い, $\beta > 0$ の 950 回の抽出のそれぞれについて LD50 を $-\alpha/\beta$ として計算する. 図 4.2a に, LD50 値のヒストグラムを示します. 両尾の極値を除きます. (シミュレーションの全範囲が含まれていれば, 分布の形状はほとんど見えません). 分布の中心をよりよく把握するために, 図 4.2b に, LD50 の 950 シミュレーション抽出の中央 95% のヒストグラムを表示します. ヒストグラムは, 図 3.4 とほぼ同じ場所に配置されていますが, β がゼロに近い可能性があるため, 実質的により多くの変動があります.

要約すると, ここでの正規近似に基づく事後推論は, 正確な結果とほぼ同じですが, 小さなサンプルのために, 実際の結合事後分布は, 大きなサンプル近似よりも実質的に歪みがあり, 実際に LD50 の事後分布は結合正規近似を使用することによって暗示されるよりもはるかに短いテールを有する. これらの違いが, この例における実際の使用に通常の近似が不十分であることを示唆するかどうかは, 分析の最終的な目的に依存する.

4.2 大標本理論

なぜ正規近似がしばしば合理的なのかということを理解するために、いくつかの固定サンプリング分布からのデータ量が増加するにつれて事後分布がどのように挙動するかの理論を再検討する.

表記と数学的設定

大規模サンプルのベイズ推論の基本的なツールは、事後分布の**漸近正規性 (asymptotic normality)** です。より多くのデータが同じ潜在的なプロセスから来ると、データの真の分布が考慮中のパラメータ族内になくても、パラメータベクトルの事後分布は多変量正規性に近づく。数学的には、結果は同一分布 $f(y)$ からサンプリングされた独立な結果である観測 y_1, \dots, y_n に最も直接的に適用される。多くの状況で、データに対する真の潜在的な分布 $f(y)$ の概念は解釈が難しいが、漸近理論を発展させるために必要である。データがパラメータ族 $p(y|\theta)$ によってモデル化され、事前分布が $p(\theta)$ であると仮定する。一般に、データ点 y_i およびパラメータ θ はベクトルであってもよい。真のデータ分布がパラメータ族に含まれている場合、つまり、ある θ_0 について $f(y) = p(y|\theta_0)$ ならば、漸近正規性に加えて、一貫性の特性が成り立つ。事後分布は真のパラメータ値 θ_0 の点質量に $n \rightarrow \infty$ として収束する。真の分布がパラメータ族に含まれていない場合、真の値 θ_0 はなくなりますが、理論的結果におけるその役割は、モデル分布 $p(y|\theta)$ を付録 B で説明されているように、Kullback-Leibler ダイバージェンスを含む技術的意味での真の分布 $f(y)$ である。

事後分布の大標本特性の議論で、Jeffreys の事前分布の文脈で 2.8 節の (2.20) で導入したフィッシャー情報量 $J(\theta)$ の概念は重要な役割を果たす。

漸近正規性と一貫性

付録 B で与えられる基本的な数学的結果は、正規化条件 (特に、尤度は θ の連続関数であり、 θ_0 はパラメータ空間の境界上にない) の下で、 $n \rightarrow \infty$ としたとき、 θ の事後分布は平均 θ_0 、分散 $(nJ(\theta_0))^{-1}$ となる正規分布に近づく。最も単純なレベルでは、この結果は、事後モードを中心とする対数事後密度のテイラー級数展開 (4.1) によって理解することができる。予備的結果は、事後モードが θ_0 と一致することを示しているので、 $n \rightarrow \infty$ として、事後分布 $p(\theta|y)$ の質量は θ_0 のより小さい近傍に集中し、 $|\theta - \theta_0|$ は 0 に近づく。

従って、(4.1) の二次形式の係数を

$$\left[\frac{d^2}{d\theta^2} \log p(\theta|y) \right]_{\theta=\hat{\theta}} = \left[\frac{d^2}{d\theta^2} \log p(\theta) \right]_{\theta=\hat{\theta}} + \sum_{i=1}^n \left[\frac{d^2}{d\theta^2} \log p(y_i|\theta) \right]_{\theta=\hat{\theta}}$$

とかける。 θ の関数として考えると、この係数は定数と、それぞれが y_i の真のサンプリング分布、 $p(y|\theta_0)$ の下での期待値が、 $\hat{\theta}$ が θ_0 (いま、 θ_0 に対して $f(y) = p(y|\theta_0)$ と仮定している) に近い場合に限り、近似的に $-J(\theta_0)$ であるような n 項の和である。したがって、大きな n に対して、対数事後密度の曲率は、 θ または θ_0 (前者のみが実際に利用可能である) で評価されるフィッシャー情報量によって近似することができる。

要約すると, 大きな n の極限では, 特定のモデル群の文脈において, 事後モード θ が θ_0 に近づき, 曲率 (テイラー展開における第 2 項の係数の観測情報または負数) は $nJ(\theta)$ または $nJ(\theta_0)$ に近づく. また, $n \rightarrow \infty$ のとき, 尤度は事前分布を支配するので, 尤度のみを用いて正規近似のモードと曲率を得ることができる. 定理のより正確な記述と証明の概要は付録 B に示されている.

事前分布を支配する尤度

漸近的な結果は, 標本サイズが増加するにつれて, 事前分布の重要性が減少するという考えを形式化する. この結果の 1 つの結果は, 大きなサンプルサイズの問題では, すべての利用可能な情報を正確に反映する事前分布を定式化することを特に努力する必要はないということです. サンプルサイズが小さい場合, 事前分布はモデル仕様の重要な部分である.

4.3 定理の反例

大標本定理の結果の極限を理解するよい方法は定理が成り立たない場合を考えることである. 正規分布は通常, 最初の近似として役に立つが, 特に異常なパラメータ空間や分布の極端な部分では, 偏差を調べなければならない. 漸近定理への反例は一般的に, サンプルサイズの極限においても事前分布が事後分布へ大きな影響を持つという状況に対応する.

不確かなモデルと同定されていないパラメータ

尤度 $p(\theta|y)$ が θ の値の範囲で等しい場合, モデルはデータ y の下で不確かである. これは, **フラットな尤度 (flat likelihood)** とも呼ばれます (ただし, データによって弱く同定されるパラメータの尤度にも使用されるため, 尤度関数は厳密には値の範囲で等しくはありません). そのようなモデルでは, 事後分布が収束する単一点 θ_0 は存在しない.

例えば, モデル

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

を考える. ここで, 各対 (u, v) から u または v の一方のみが観測される. パラメータ ρ は同定されない. データは ρ について何も情報を与えないので, ρ の事後分布はデータセットがどんなに大きくても, 事前分布と同じである.

同定されていないか不明確なパラメータの問題に対する唯一の解決策は, 問題が存在することを認識し, これらのパラメータをより正確に推定したい場合は, パラメータを推定できるような情報を (将来のデータ収集や事前分布を知らせる外部情報から) 収集することです.

サンプルサイズとともに増加するパラメータ数

複雑な問題では、パラメータ数が大きくなることがあり、漸近理論の異なるタイプを区別する必要がある。 n が増加しモデルが変化してパラメータの数も増えると、固定モデルクラス $p(y_i|\theta)$ を仮定する 4.1, 4.2 節で説明した単純な結果は適用されない。例えば、研究中のサンプリング単位ごとにパラメータが割り当てられることがある。例えば、 $y_i \sim N(\theta_i, \sigma^2)$ である。パラメータ θ_i は、各サンプリングユニットから収集されたデータの量がユニットの数とともに増加しない限り、一般的に一貫して推定することができない。ガウス過程 (21 章を参照) などのノンパラメトリックモデルでは、各データポイントに対応する新しい潜在パラメータが存在する可能性があります。

不確実なパラメータの場合と同様に、新しいデータが θ_i についての十分な情報をもたらさない場合、 θ_i の事後分布は点質量に収束しない。ここで、事後分布は一般に拡大パラメータ空間の点に収束することはなく (θ の次元性が高まることを反映して)、固定された空間への投影 (例えば、特定の θ_i の周辺事後分布など) いずれかの点に収束する。

エイリアシング

エイリアシング (Aliasing) は, 同一の尤度関数が離散点集合で繰り返される, 同定されていないパラメータの特殊なケースである. 例えば, 独立同一分布のデータ y_1, \dots, y_n , パラメータベクトル $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$ である次の正規混合モデルを考える.

$$p(y_i | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(y_i - \mu_1)^2} + (1 - \lambda) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2\sigma_2^2}(y_i - \mu_2)^2}$$

もし, $(\mu_1, \mu_2), (\sigma_1^2, \sigma_2^2)$ のそれぞれを交換し, λ を $(1 - \lambda)$ で置き換えると, データの尤度は同じである. このモデルの事後分布は少なくとも 2 つのモードを持ち, 互いの鏡像である 2 つの分布の (50%, 50%) 混合物からなる. データセットの大きさにかかわらず, 単一点に収束することはない.

一般に, エイリアシングの問題は, 重複が現れないようにパラメータ空間を制限することによって排除される. 上記の例では, μ_1 を μ_2 以下に制限することで, エイリアシングを除去することができる.

非有界な尤度

尤度関数が非有界な場合, パラメータ空間内に事後モードがなく, 一致性結果と正規近似の両方が無効になる可能性があります. 例えば, 以前の正規混合モデルを考えてみましょう. 単純化のために, λ が既知である (そして 0 または 1 に等しくない) と仮定する. 任意の y_i に対して $\mu_1 = y_i$ とし, $\sigma_1^2 \rightarrow 0$ とすると, 尤度は無限に近づく. $n \rightarrow \infty$ とすると, 尤度のモードの数が増加する. 事前分布が 0 に近い領域の σ_1^2 と σ_2^2 で一様であれば, 同様に対応する正規近似を持たない事後モードの数が増加する. $\sigma_1^{-2}\sigma_2^{-2}$ に比例する事前分布は事態を悪化させます. なぜなら, これは 0 付近でより多くの確率を生じさせ, 事後分布が 0 でさらに速く発散するためである.

非有界な尤度の極は, モデルの非現実的な条件に対応するため, 一般に, この問題は実際にはめったに起こらないはずである. 問題は, もっともらしい分布の集合に限定することによって解決することができる. 0 に近い分散成分で問題が発生した場合, 境界で 0 になる事前分布を使用するか, 分散パラメータの比率に情報事前分布を割り当てるなど, さまざまな方法で解決できます.

不適切な事後分布

不適切な事前分布を表す正式な事前密度を尤度に掛け合わせることに
よって得られる非正規化事後密度が無限大に積分される場合, 1 に合計
する確率に依存する漸近的結果はついてこない. 不適当な事前分布を除
いて, 不適切な事後分布は起こり得ない.

単純な例は, 2 項の割合に $\text{Beta}(0, 0)$ 事前分布を, n 個の成功と 0 個の失敗
からなるデータと組み合わせることから生じる. 階層的な二項モデルと
正規モデルを用いたさらに微妙な例については, 5.3, 5.4 節で説明します.
この問題に対する解決策は明らかです. 不適当な事前分布は単なる近似
であり, 適切な事後分布をもたらさない場合, 求められる利便性は失われ
る. この場合, 適切な事前分布が必要であるか, または尤度と組み合わせ
られたとき有限積分を有する不適切な事前密度が少なくとも必要とされる.

収束点を除外した事前分布

離散パラメータ空間に対して $p(\theta_0) = 0$ か, 連続パラメータ空間に対して
 θ_0 についての近傍で $p(\theta) = 0$ ならば, 事前分布を支配する尤度に基づく
収束の結果は成り立たない. この解は, 離れて妥当な θ のすべての値に対
して, 事前分布における正の確率密度を与えることである.

パラメータ空間の端への収束

θ_0 がパラメータ空間の境界にあるならば、テイラー級数の展開はいくつかの方向で切り捨てられなければならない、正規分布は極限においても必ずしも適切ではありません。

例えば、 $\theta_i \geq 0$ のモデル $y_i \sim N(\theta, 1)$ を考える。モデルが正確であり、 $\theta = 0$ が真の値であると仮定する。 θ の事後分布は、 y を中心とする正規分布であり、正の値に切り捨てられる。 θ の事後分布の形状は、 $n \rightarrow \infty$ の極限で、正の値に切り捨てられた 0 を中心とする正規分布の半分である。

別の例として、仮定した同じモデルであるが、真の θ が仮定されたパラメータ空間外の値 -1 である場合を考えてみる。 θ の極限事後分布は、0 に鋭いスパイクを有し、正規分布と全く類似していない。実際の解決策は、パラメータ空間のエッジ付近のパラメータ値に関心がある場合には、正規近似を適用するのが困難であることを認識することです。さらに重要なことは、離れた取りうる θ のすべての値に対して、または離れた取りうる値の近傍に、正の事前確率密度を与える必要があるということです。

分布の尾

正規近似は事実上全ての事後分布の質量に対して成り立つが、尾では正確ではない。たとえば、 $p(\theta|y)$ は $|\theta| \rightarrow \infty$ としたとき、定数 c について $e^{-c|\theta|}$ に比例すると仮定する。比較すると、正規密度は $e^{-c\theta^2}$ に比例する。分布関数は正規分布に収束するが、任意の有限標本サイズ n に対して、近似は尾ではまったく成り立たない。別の例として、正であるように制約される任意のパラメータを考える。有限標本サイズの場合、近似は分布の尾のその点では適切ではないので、パラメータが負である可能性を認めるだろうが、 n が増加するにつれてその点は尾で遠くなる。

4.4 ベイズ推論の頻度主義的評価

単純な古典的手法を正当化するためにベイズのパラダイムが見られるのと同様に, 頻繁主義的な統計の手法は, ベイズ推論の性質 (その動作特性) を評価するため, これらが一連の繰り返しサンプルに埋め込まれていると見なし有用なアプローチを提供します. すでに一致性と漸近正常性の考え方を議論する際にこの概念を使用している. 固定されたモデルに対して, よりデータが集まるにしたがって事後分布が 1 点に近づくことを示す (極限において, 推論の確実性を導く), **安定推定 (stable estimation)** の概念は反復されるサンプリングの概念に基づく. 仮定された確率モデル族に真の分布が含まれている (また, それに非ゼロの事前密度を割り当てる) 場合, θ に関するより多くの情報が得られると, 事後分布は θ の真の値に収束することは確かに魅力的です.

大標本対応

θ の事後分布に対する正規近似 (4.2) が成立するとする. 標準多変量正規分布に変換できる.

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta})|y \sim N(0, I) \quad (4.3)$$

ここで, $\hat{\theta}$ は事後モードで, $[I(\hat{\theta})]^{1/2}$ は $I(\hat{\theta})$ の任意の行列平方根である. 加えて, $\hat{\theta} \rightarrow \theta_0$ で, ちょうど同様に $I(\theta_0)$ の近似を書くことができます. もし, ある θ に対して $f(y) \equiv p(y|\theta)$ となるように, 真のデータ分布がモデルのクラスに含まれているならば, 定数 θ に関する反復されるサンプリングで, 極限 $n \rightarrow \infty$ において,

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta})|\theta \sim N(0, I) \quad (4.4)$$

が示せる. 一般的に $\hat{\theta}$ に対して示される, 古典的な統計理論からの結果は最大尤度推定量に等しいが, $\hat{\theta}$ が事後モードに等しい場合に簡単に拡張できる. これらの結果は, $(\theta - \hat{\theta})$ の任意の関数に対して, (4.3) から導かれる事後分布は (4.4) から導かれる反復されるサンプリング分布と漸近的に等しくなるということを意味している. そして, 例えば, θ に対する 95% 中央事後区間は, 固定された真の θ に関する反復サンプリング分布の下での回数の真値の 95% をカバーする.

点推定, 一致性, 効率性

ベイズの枠組みで, θ の推定量を得ることは大きなサンプルで, 事後モード $\hat{\theta}$ が明らかに θ の事後分布の中心であり, $nI(\hat{\theta})$ による不確かさが非常に小さい場合に, 事実上重要ではない. さらに一般的に, しかしながら, 小さいサンプルで, 特に θ の事後分布がより多くの変数またはさらには非対称であるとき, θ について 1 つの値で要約することは不適切である. 正式には, 決定論的文脈 (9.1 節および 9.6 節を参照) に損失関数を組み込むことによって, 最適点推定を定義することができる. ベイズデータ解析の目的では, しかしながら, 完全な結合分布 (例えば, 50%および 95%の中央事後区間) はより役に立つ. 多くの問題, 特に大きなサンプルで, 点推定量とその推定標準誤差は事後推論を要約するのに適しているが, 推定量は, 決定問題の解法ではなく推論要約として解釈する. どのような場合でも, 推定量がベイズ解析から導出されたかどうかを考慮せずに, 任意の推定量の大きなサンプル周波数特性を評価することができる.

サンプルが大きくなるにつれて、それが推定するために主張されるパラメータの真の値に収束するならば、点推定量はサンプリング理論の意味において**一致している (consistent)**と言われる。そして、 $f(y) \equiv p(y|\theta_0)$ の場合、データサンプルサイズ n が増加するにしたがって、サンプリング分布が θ_0 の点質量に収束するならば、 θ の点推定量 $\hat{\theta}$ は一致する。(すなわち、 $\hat{\theta}$ を y の関数として考えると、 θ_0 の条件付き確率変数である)。似た概念は**漸近的不偏性 (asymptotic unbiasedness)** である。ここで、 $(E(\hat{\theta}|\theta_0) - \theta_0)/\text{sd}(\hat{\theta}|\theta_0)$ は 0 に収束する。(ここでもまた、 $\hat{\theta}(y)$ を $p(y|\theta_0)$ によって決定される確率変数として考える) 真値が適合するモデル族に含まれる場合、事後モード θ と事後平均および中央値は、穏やかな規則性条件の下で一致性があり、漸近的に不偏である。

点推定量 $\hat{\theta}$ は、 θ を平均二乗和誤差がより小さくなるよう、すなわち、最適値での $E((\hat{\theta} - \theta_0)^2|\theta_0)$ が最小値であるよう推定する y の関数がない場合、**効率的 (efficient)**と言われる。より一般的には、 θ の効率は、最適平均二乗誤差を θ の二乗平均誤差で割ったものである。推定量は、標本サイズ $n \rightarrow \infty$ のときに効率が 1 に近づくと漸的に効率的です。軽度の規則性条件の下では、事後分布の中心(事後平均, 中央値, またはモードによって定義される)は漸的に効率的である。

信頼区間

領域 $C(y)$ が反復サンプルの少なくとも $100(1 - \alpha)\%$ (θ_0 の任意の値を与えられた) の θ_0 を含む場合, $C(y)$ はパラメータ θ に対する $100(1 - \alpha)\%$ **信頼区間 (confidence region)** と呼ばれる. そのような区間を確率区間と区別し, 以下のふるまいの意味を伝えるために, 信頼という単語を慎重に選択します. α が十分に小さい (例えば 0.05 または 0.01) ように選択すると, 信頼区間は実際の少なくとも $1 - \alpha$ で真値をカバーするため, 各応用で真値が領域内にあることを確信し, そうであるかのように行動します. 以前に, 漸近的に $100(1 - \alpha)\%$ の中央事後区間は, y の反復サンプルにおいて, 間隔の $100(1 - \alpha)\%$ が値 θ_0 を含むという性質を有することを見出した.

4.5 他の統計手法のベイズ的解釈

ベイズ統計手法と他の手法を比較できる3つのレベルを検討します. 1つ目に,すでに述べたように,ベイズ手法は,固定確率モデルからの大きなサンプルを含む問題の他の統計的アプローチとよく似ています. 2つ目に,小さなサンプルであっても,多くの統計的手法は,特定の事前分布に基づくベイズ推論の近似値とみなすことができる. 統計的手法を理解する方法として,暗黙の潜在的な事前分布を決定することはしばしば有益である. 3つ目に,古典的統計からのいくつかの方法(特に仮説検定)は,ベイズ法によって与えられた結果とは大きく異なる結果を与えることができる. この節では,点推定と間隔推定,尤度推論,不偏推定,信頼区間の頻度カバレッジ,仮説検定,多重比較,ノンパラメトリック法,ジャックナイフとブートストラップといったいくつかの統計的概念を簡単に考察し,ベイズ手法との関係を議論する.

可能なモデルを開発する 1 つの方法は、特定のモデルのもとでのベイズ推論への近似として、粗いデータ分析手順の解釈を調べることです。例えば、サンプル調査で広く使用されている手法は、単純なランダムサンプルからのデータが与えられた場合、8 章の表記で $\bar{y}_{obs}/\bar{x}_{obs}$ で $R = \bar{y}/\bar{x}$ を推定するといった、**比率推定 (ratio estimation)** です。この推定値は独立した観測値 $y_i|x_i \sim N(Rx_i, \sigma^2 x_i)$ と無情報事前分布を仮定してベイズ事後推定の要約に相当することが示される。比率の推定は、このモデルが成り立たないさまざまなケースで有用であるが、データがこのモデルから大きく外れた場合、比率推定は一般に適切ではありません。

別の例として、**統計的有意性 (statistical significance)** に基づく回帰予測を選択する標準的方法は、14.6 節でさらに論じるように、各係数の事前分布がゼロのピークと範囲に広がる分布である係数に関する交換可能な事前分布の下でのベイズ分析である。この対応関係を理解することで、そのようなモデルをいつ有効に適用でき、どのように改善できるかが示唆されています。多くの場合、実際には、バッチで似ている可能性が高い回帰係数をクラスタリングすることによって、多数の予測を含む問題などの追加情報を含めることによって、このような手順を改善できます。

最大尤度推定量とほかの点推定量

ベイズデータ解析の観点から, 古典的な推定量を暗示的な完全な確率モデルに基づく正確または近似事後要約と解釈できる. 大きなサンプルサイズの極限で, 実は, 古典的な最尤推定に対する理論的なベイズ正当化を構築するため漸近理論を用いることができる. 極限 (正規化条件を仮定して) では, 最尤推定量, $\hat{\theta}$ は十分統計量で, 事後モード, 事後平均, 事後中央値も同様である. すなわち, 十分大きな n に対して, 最尤推定量 (もしくはほかの要約) は本質的にデータから利用可能な θ についての情報を提供する. 事前分布の漸近的無相関は, 便利な無情報事前モデルの使用を正当化するために取ることができる.

$\theta = \theta_0$ に関する反復サンプリングでは,

$$p(\hat{\theta}(y)|\theta = \theta_0) \approx N(\hat{\theta}(y)|\theta_0, (nJ(\theta_0))^{-1})$$

である. すなわち, $\hat{\theta}(y)$ のサンプリング分布は平均 θ_0 で精度 $nJ(\theta_0)$ となる漸近的正規である.

ここで, わかりやすさのため $\hat{\theta}$ は y の関数であることを強調する. 事前分布は真値 θ の近くで局所的ー様分布 (または連続で非ゼロ) であることを仮定して, 正規平均の簡単な解析 (3.5 節) は事後ベイズ推論は

$$p(\theta|\hat{\theta}) \equiv N(\theta|\hat{\theta}, (nJ(\hat{\theta}))^{-1})$$

であることを示す. この結果は漸近正規理論から直接現れるが, 与えられた $\hat{\theta}$ を用いて間接的にベイズ推論を導出することは, 点推定と標準誤差に基づく古典的漸近推論のベイズ論理に洞察を与える.

有限の n に対して, 上の手法は, θ が十分統計量ではない場合, 非効率的であるか, 無駄である. パラメータ数が大きいとき, 一致性の結果はしばしば有用ではなく, 無情報事前分布は正当化するのが難しい. 5 章で説明したように, 多数のパラメータを扱う場合, それらの共通の分布はデータから推定できるため, 階層モデルが好まれる. さらに, 尤度のみに基づく推論の任意の方法は, 尤度関数に含まれるものに実質的に寄与するのに十分に強い実際の事前情報が利用可能である場合に改善することができる.

不偏推定量

非ベイズ統計手法の中には、推定の望ましい原則として不偏性を重視するものがあり、反復サンプリングでは、パラメータ推定量の平均 (またはおそらく中央値) を真の値と等しくすることが直感的に魅力的です. 正式には、 θ の任意の値について $E(\hat{\theta}(y)|\theta) = \theta$ である場合、推定量 $\hat{\theta}(y)$ は不偏である呼ばれ、この期待値はデータ分布 $p(y|\theta)$ で取られる. ベイズの観点からは、不偏性の原則は大きなサンプルの極限で合理的ですが (92 ページ参照), それ以外の場合は誤解を招く可能性がある. 大きな困難は、推定すべき多くのパラメータが存在し、これらのパラメータのうちのいくつかの知識または部分的な知識が、他者の推定に明らかに関連している場合に生じる. 不偏推定量を必要とすると、しばしば関連する情報が無視される (5 章の階層モデルで議論します). サンプリング理論では、偏りを最小限に抑えることは、しばしば分散の逆効果的な増加をもたらす.

不偏性 (および一般的に点推定) に関する 1 つの一般的な問題は、偏っていなくても一度にいくつかのパラメータを推定することはできないことが多いことである. 例えば、 $\theta_1, \dots, \theta_J$ は、 θ_j の分散の上向きに偏った推定値を生成する (θ_j が正確に知られている些細な場合を除いて).

将来の観測可能な値を予測問題のパラメータとして扱うとき、不偏性の原理に関する別の問題が生じる.

例. 回帰を用いる予測

母親の身長は y が与えられた, 大人の娘の長さ θ の推定の問題を考える. 簡単のために, 母親と娘の長さは, 160 センチメートルの等平均, 等分散, および 0.5 の既知の相関で結合正規分布であると仮定する. 既知の y の値へ条件付けし (言い換えると, ベイズ推論を用い), θ の平均は

$$E(\theta|y) = 160 + 0.5(y - 160) \quad (4.5)$$

しかしながら, 定数 θ が与えられた y の反復サンプリングの意味で, 事後平均は θ の不偏推定量ではない. 娘の身長 θ が与えられたときの, 母の身長 y は平均 $E(y|\theta) = 160 + 0.5(\theta - 160)$ である. そして, 定数 θ が与えられた y の反復サンプリングの下で, 事後平均 (4.5) は期待値 $160 + 0.25(\theta - 160)$ をもち, 160 の総平均に偏っている対称的に, 推定量

$$\hat{\theta} = 160 + 2(y - 160)$$

は θ の条件で y の反復サンプリングの下で不偏である. 残念ながら, 推定量 $\hat{\theta}$ は, y の値が 160 に等しくない場合に意味をなさない. たとえば, 母親が平均よりも 10 センチ高い場合, 娘は平均よりも 20 センチ高いと推定される.

θ が許容される母集団分布を有するこの単純な例では, 合理的な非ベイズ統計学者は不偏推定値 θ を使用しない. 代わりに, この問題は推定ではなく予測に分類され, 手法は確率変数 θ に対して条件付きで評価されない. この例では, 一般的な原則としての不偏性の限界を示しています. 推定には異なる意味を持つが, 明確な実質的な区別はなく, 未知の量をパラメータまたは予測として特徴付ける必要がある. 5 章では, θ の母集団分布を特定の値で調整するのではなく, データから推定しなければならない同様の状況を考慮する.

この例で示される重要な原則は, 平均に対する回帰のものです. 任意の母親について, 娘の身長期待値は母親の身長と母集団平均の間にある. この原理は, 19 世紀後半の Galton によるこの種の分析のための回帰という用語の本来の使用にとって基本的なものであった. 多くの点で, ベイズ解析は, 回帰の原則を論理的に拡張して, さまざまな情報源からの情報を適切に重み付けすることを保証するものと見なすことができます.

信頼区間

小さなサンプルにおいても、ベイズ $(1 - \alpha)$ 事後区間はしばしば、 θ の条件での反復サンプリングの下で $(1 - \alpha)$ 信頼区間に近くなる。しかし、サンプリング理論の議論から純粹に導き出され、ベイズ確率区間とはかなり異なるいくつかの信頼区間が存在する。私たちの見解では、これらの区間は疑わしい値です。例えば、多くの著者は、無条件のふるまいに基づく一般的な理論は、明らかに直観に反する結果、例えばゼロまたは無限の長さの信頼区間の可能性につながることを示している。簡単な例は、5%が空であり、実線の 95%がすべて含まれている信頼区間です。これは、反復サンプルの 95%に (実際の値のパラメータの) 真の値を常に含んでいます。そのような例は、信頼区間の概念に価値がないことを意味するのではなく、区間のみが合理的な推論を形成するのに十分な基礎ではないことを示す。

仮説検定

この本の観点は、非ベイズの仮説検定概念、特に $\theta = \theta_0$ という形式の帰無仮説を指すことに関連するものにあまり役割を負わない。ベイズ解析に対して点帰無仮説に対して非ゼロの確率をもたらすためには、その仮説に対して非ゼロの事前確率で始まらなければならない。連続パラメータの場合、そのような事前分布 (別の場所で連続密度と混合された θ_0 での離散質量、例えば 0.5 を含む) は、通常、考案されたようである。実際、仮説検定の難しさの多くは、 $\theta = \theta_0$, $\theta \neq \theta_0$ の間で必要とされる人工的な二分法により生じる。この二分法に関する難しさは統計的推論のすべての観点から広く知られている。連続パラメータ θ を含む問題 (例えば、2つの平均の間の違い) では、 θ が 0 であるという仮説は、ほとんど合理的はなく、事後分布を推定すること、または対応する θ の区間推定量により興味がある。連続するパラメータ θ については、「 θ は 0 に等しいか?」という質問は、一般に、「 θ の事後分布とは何か?」というように、より有用に言い換えることができる。

様々な単純な片側仮説検定では、従来の p 値は、無情報事前分布の事後確率と対応することができる。例えば、モデル $y \sim N(\theta, 1)$ から $y = 1$ を観測し、 θ に一様な事前密度を仮定すると仮定する。 $\theta = 0$ であるという仮説を棄却することはできない。片側の p 値は 0.16 であり、両側の p 値は 0.32 であり、両方とも従来の統計的有意性 0.05 のカットオフ値より大きい。一方、 $\theta > 0$ の事後確率は 84% であり、これは二分検定の棄却または採用よりもより満足で有益な結論である。

特定のモデル内のパラメータについての推論を行う問題とは対照的に、確率モデルの適合度を評価する際に有用な仮説検定の一形式を見つける。ベイズの枠組みでは、6 章で詳しく説明するように、観察されたデータと予測可能な結果とを比較することによってモデルをチェックすることは有用である。

多重比較とマルチレベルモデリング

J 個のパラメータのそれぞれについて独立した測定値 $y_j \sim N(\theta_j, 1)$ の問題を考えてみる. ここで, 目標は連続パラメータ θ_j の差と順序付けを検出することです. いくつかの競合する複数の比較手続きが古典的な統計で導出されており, 様々な θ_j が著しく異なると宣言されるときについての規則がある. ベイズ手法では, パラメータは結合事後分布を有する. 各 J の事後確率を必要に応じてオーダーし, 計算できる. 順序付けに事後不確実性がある場合, いくつかの置換は実質的な確率を持つことになります. これは, 異なると宣言が可能な θ_j のリストを生成するよりも合理的な結論です (他の θ_j が正確に等しいという誤った意味合いで). J が大きい場合, 正確な順序付けは重要ではない可能性があり, 集団内の各 θ_j の分位数の事後中央値および区間推定値を与える方がより合理的かもしれない.

5.5 節の 8 つの学校 (5.3 節も参照) の取り扱い効果の比較で説明するように, 階層モデルを使用して多重比較問題を扱うことを好む. 階層的モデリングは, 実際の変化の証拠がほとんどない場合, 異なる θ_j の推定値を自動的に部分的にプールします. 結果として, このベイズ手法は, 古典的な多重比較分析の重要な関心事を自動的に解決します. これは, 非常に多くの可能性を探索する副産物として大きな違いを見つける可能性があります. 例えば, 教育テストの例では, 8 つの学校が $8 \cdot 7/2 = 28$ の可能な比較を行い, どれも統計的に有意義に近い (学校のペア間の効果の差異のすべてについて, 95%の区間内にゼロが含まれているという意味で) とは限らないとしている. これは学校間の差 (そのモデルのパラメータ τ) が小さいと推定されていることを意味する.

ノンパラメトリック法, 順列検定, ジャックナイフ, ブートストラップ

サンプリングレベルであっても完全な確率モデルを避ける多くの非ベイズ手法が開発されている. ベイズの観点からこれらの多くは評価することが困難である. 例えば, 順位に基づく中央値を比較するための仮説検定は, ベイズ推論に直接対応するものではない. 従って, ベイズの観点から結果の推定値および p 値を解釈することは困難です (例えば, パラメータや予測値, 区間, 確率など). 複雑な問題では, 使用される手順にある程度の恣意性があります. 例えば, 仮説的複製において **ジャックナイフ/ブートストラップ** に対して **ノンパラメトリック法** または推定量を構築するための明確な方法は一般にない. 特定の確率モデルがなければ, 特定のノンパラメトリック法の根底にある仮定をどのようにテストするかを見ることは困難です. このような問題では, 推定確率を構築し, その頻度特性を評価するよりも, 合同確率分布を構築し, それをデータと照らし合わせてチェックする方が満足できることが分かる (第 6 章のように). ノンパラメトリックな方法は, 私たちがモデルを構築したり, 推論を全く異なる視点から評価したりするのに役立つデータの要約と説明のツールとして私たちにとって有益である.

ベイズ手法は、モデルの任意の選択を伴い、評価が困難なモデルの重要な側面が常に存在するため、評価が困難である。この目的は、古典的なノンパラメトリック手法を却下したり、捨てたりするのではなく、これが可能な限りベイジアン文脈に入れることです。

実験についての順列検定や調査のための理論的推論のようないくつかのノンパラメトリック手法は、ベイジアンモデルがデータを合理的にうまく収めるために構築されていれば、単純な問題で同様の結果をもたらす。このような単純な問題には、データの欠落のないバランスの取れた設計や、単純なランダムサンプリングに基づく調査が含まれます。一度にいくつかのパラメータを推定したり、(共分散または回帰の分析などの方法を使用して) 分析の説明変数を含めると、パラメータに関する事前の情報が含まれる場合、置換/サンプリング理論の方法は直接的な答えを示さず、モデルベースのベイズ手法に移行する。

例. Wilcoxon 順位検定

他の関連はノンパラメトリック手法を明示的なモデルに関して解釈することではしうる. 例えば, 2 サンプル $(y_1, \dots, y_{n_y}), (z_1, \dots, z_{n_z})$ を比較するための **Wilcoxon 順位検定 (Wilcoxon rank test)** は, まず 1 から

$n = n_y + n_z$ へとら合成されたデータ内の点の各々をランク付けし, 次に y, z の平均ランクの間の違いを計算し, 最後に n 個のランクのランダム割当の仮定に基づいて計算された表形式の基準分布と比較することによって, この差の p 値を計算するという様に進む. これは各点を, 2 つの変換されたサンプルの平均の比較に従い結合データ内でのそのランクで置き換えるという非線形変換として定式化できる. さらに明確に, ランク $1, \dots, n$ を分位数 $\frac{1}{2n}, \dots, \frac{2n-1}{2n}$ に変換すると, 2 つの手段の間の差は, 結合された分布の分位数のスケールにおける平均距離として解釈され得る. 中心極限定理から, 平均差はほぼ正規分布で, 古典正規理論の信頼区間は, この節の冒頭で説明したベイズ事後確率表現として解釈できる.

近似ベイズ推論としてランクテストを表現するのには2つの大きな利点があります. 第1に, ベイズの枠組みは, 回帰予測などの追加情報や, 打ち切りや切り捨てられたデータなどの合併症から生じる合併症を処理するためのランクテストよりも柔軟性があります. 第2に, 非線形変換の問題を設定することは, モデルベースのアプローチの一般性を示しています. おそらく便利なデフォルト選択として結合された分位数を扱うかもしれませんが, 問題に適した変換を自由に使うことができます.

まとめ 4.1~4.

- 大標本理論とは, 固定のサンプリング分布からのデータ量が増えるにつれて, 事後分布がどのように挙動するかについて考える理論である.

