ベイズ統計の漸近理論と SAS による実装

○玉乃井 崚太

（シミック株式会社　データサイエンス事業本部）

Asymptotic Theory of Bayesian Posteriors and Its Implementation in SAS

Ryota Tamanoi

Data Science Division, CMIC Co., Ltd.

要旨

近年、医療統計においてもベイズ流の解析が用いられる場面が増えている。頻度論的な統計学では、標本平均のような推定量の妥当性が大数の法則や中心極限定理によって保証されている。ベイズ流の解析にも同じ役割を果たす理論が必要であり、それは「観測データが十分に存在するとき、事後分布はどのようにふるまうか」という問いになる。素直な答えは「事後分布は正規分布に近づく」であり、最適なパラメータが一点に定まるような性質のよい正則なモデルでは実際にそのとおりになる。しかし説明変数に完全な共線性がある回帰のような非正則なモデルにおいてはこの結論は破綻する。本稿では、正則なモデルの漸近正規性とその破綻を PROC MCMC で観察したうえで、破綻した場合にも成立する漸近理論、渡辺澄夫ベイズ理論を紹介する。パラメータそのものを追うことをやめ、真の分布との KL 情報量の値だけを追うという見方をとると、正則・非正則を問わず同じ形の漸近分布が得られる。応用として、正則なモデルにおけるパラメータ数に代わってモデルの実質的な自由度を表す実対数閾値と、正則性を仮定しない情報量規準である WAIC を導入し、SAS による実装を与える。

キーワード：ベイズ統計、漸近理論、実対数閾値、WAIC、PROC MCMC、非正則モデル

1. はじめに

1.1. 背景

近年、医療統計分野においてベイズ統計の応用が活発に進展している。ベイズ統計では観測データに加え外部情報を「事前分布」として組み込むことができるほか、複雑な試験デザインにも柔軟に対応できるなど多くの利点がある。ベイズ流でない頻度論的な統計学においては、標本平均という推定量の正当性は大数の法則と中心極限定理によって保証されている。ベイズ推定にも同じ役割を果たす理論はあるだろうか。すなわち、「観測データが十分に存在するとき、事後分布はどのようにふるまうか」という問いである。

素直な答えは「事後分布は正規分布に近づく」であり、実際に多くのモデルでそのとおりになる。信用区間などのベイズ流でよく用いられる道具は、すべてこの事実の上に立っている。ところが、モデルの書き方をわずかに変えるだけでこの答えは完全に破綻する。破綻するのは「近似の精度が落ちる」という意味ではなく、近似の枠組みそのものが使えなくなるという意味である。

破綻するモデルは決して病的 (人為的に作成した反例) ではない。説明変数に完全な共線性がある回帰、混合分布、ニューラルネットワーク、行列分解はいずれもこちら側に属する[1][2]。

渡辺澄夫によって整備された特異学習理論（本稿では渡辺澄夫ベイズ理論と呼ぶ[2]）は、この破綻した世界でも成り立つ漸近論を与える。本稿の目的は、その骨格を SAS で実際に事後分布を描きながら追うことである。

1.2. 本稿の構成

2 章でベイズ推定の基礎と PROC MCMC による実装を確認する。3 章で正則なモデルの漸近正規性を定式化し、SAS で確認する。4 章で共線性のある回帰を用いてそれが破綻する様子を観察する。5 章が本稿の中心で、パラメータを追うのをやめて$K(\theta)$の値だけを追うという発想を導入し、実対数閾値λを定義して SAS で測定する。6 章では応用として WAIC を導き、事後標本から計算する実装を与える。

2. ベイズ統計の基礎と SAS による実装

2.1. 事後分布

観測データを$y_1,\cdots,y_n$とし、統計モデルを$f(Y|\theta)$、事前分布を$\varphi(\theta)$とする。ベイズ推定の出発点は事後分布である。

$$\varphi(\theta|y_{1:n}) = \frac{1}{Z_n}\varphi(\theta)\prod_{i=1}^{n} f(y_i|\theta)$$

$$Z_n = \int \varphi(\theta)\prod_{i=1}^{n} f(y_i|\theta)\,d\theta$$

例えば身長のデータ$(y_1,y_2,y_3,y_4,y_5) = (165,167,170,173,180)$を用いて平均 μ 及び分散 σ^2 を推定する場合モデルの確率密度関数は$f(Y|\mu,\sigma^2) = N(Y|\mu,\sigma^2)$、事前分布を$\varphi(\mu,\sigma^2) = N(\mu|170,100^2)\cdot HalfCauchy(\sigma|0,25)$などのようにすることができる。ここで事前分布の正規分布の平均や分散、半コーシー分布の scale はハイパーパラメータと呼ばれる。この例ではそれぞれ 170、10000、25 としており、観測データとは無関係に事前に決定される。この例では 170 が「成人の身長はこのくらい」という事前の知識を、標準偏差 100 cm が「その知識をほとんど信用しない (解析に組み入れない)」という意思を表している。

このとき事後分布はμについて事後平均 171.0、95%信用区間 [163.2, 179.7]、σについて事後平均 8.40 (右にゆがんだ分布、事後中央値は 7.25)、95% 信用区間 [3.67, 20.96] となる。

従来の解析が観測データから点推定量や信頼区間を求めるのに対し、ベイズ流の統計は観測データと事前分布を合わせて事後分布を求め、そこから点推定量や信用区間を得る。

2.2. PROC MCMC による実装

SAS では PROC MCMC で事後分布からのサンプルが得られる[3]。2.1 章の例では以下のコードになる。

```
data height;
   input subjid $ y;
   datalines;
001 165
002 167
003 170
004 173
005 180
```

```
;
run;
proc mcmc data=height seed=12345
         nbi=1000  /* MCMCサンプルの最初はバイアスが大きいため捨てる(Buin in) その回数 */
         nmc=10000  /* MCMCの回数 */
         thin=20    /* 隣接するMCMCサンプルは独立でないためthin回ごとに取得する。*/
                    /* nmc / thin 個のMCMCサンプルが作成される。 */
         outpost=post_out
         monitor=(mu sigma)
         statistics(alpha=0.05 percent=(2.5 25 50 75 97.5))=(summary interval)
         diagnostics=none plots=none;
   parms mu 170 sigma 6;                          /* MCMCサンプルの初期値 */
   beginnodata;
      lp_sigma = -log(1 + (sigma/25)**2);         /* コーシー分布 */
   endnodata;
   prior mu    ~ normal(170, var=10000);          /* muの事前分布 */
   prior sigma ~ general(lp_sigma, lower = 0); /* lower = 0によって正に限定 */
   model y ~ normal(mu, sd=sigma);                /* モデルの分布 */
   ods output   PostSummaries = PostSummaries;
run;
```

注意として、PROC MCMC が返すのは事後分布そのものではなく事後分布に従うサンプルである。OUTPOST= で指定したデータセットにそのサンプルが格納される。ヒストグラムを描けば事後分布の概形が分かり、昇順に並べて下から 2.5% から 97.5% までの範囲が 95% 信用区間になる。ODS OUTPUT PostSummaries = で数値としても取り出せる。

2.3. 信用区間と信頼区間

上記のサンプルコードでは分散の大きい事前分布を指定しているが、無情報事前分布を指定することも可能である。平均と分散をともに未知とし、Jeffreys 事前分布 (無情報事前分布としてよく用いられる) の場合、平均に対する 95% 信用区間は t 検定による 95% 信頼区間と一致する[1]。

3. 正則なモデルの漸近理論

3.1. KL 情報量と最適なパラメータ

真の分布を$f_{TRUE}(Y)$とする。これは頻度論的な解析と同様、設定したモデルなどに全く依らず決定されているものであり、我々が推定したい分布の正体である。モデル$f(Y|\theta)$が真の分布からどれだけずれているかを表す量として、Kullback-Leibler 情報量を次で定義する[4]。

$$K(\theta) = E_{Y_{TRUE}}\left[\log\frac{f_{TRUE}(Y_{TRUE})}{f(Y_{TRUE}|\theta)}\right] = \int f_{TRUE}(Y)\log\frac{f_{TRUE}(Y)}{f(Y|\theta)}dY$$

ここで$E_{Y_{TRUE}}$は真の分布に従う確率変数Y_{TRUE}による期待値を表す。この$K(\theta)$が小さいほど真の分布に近い「良い」θであるといえる。$K(\theta)$を最小にするθの全体をΘ_{BEST}とおく。このようなθを推定することがベイズ

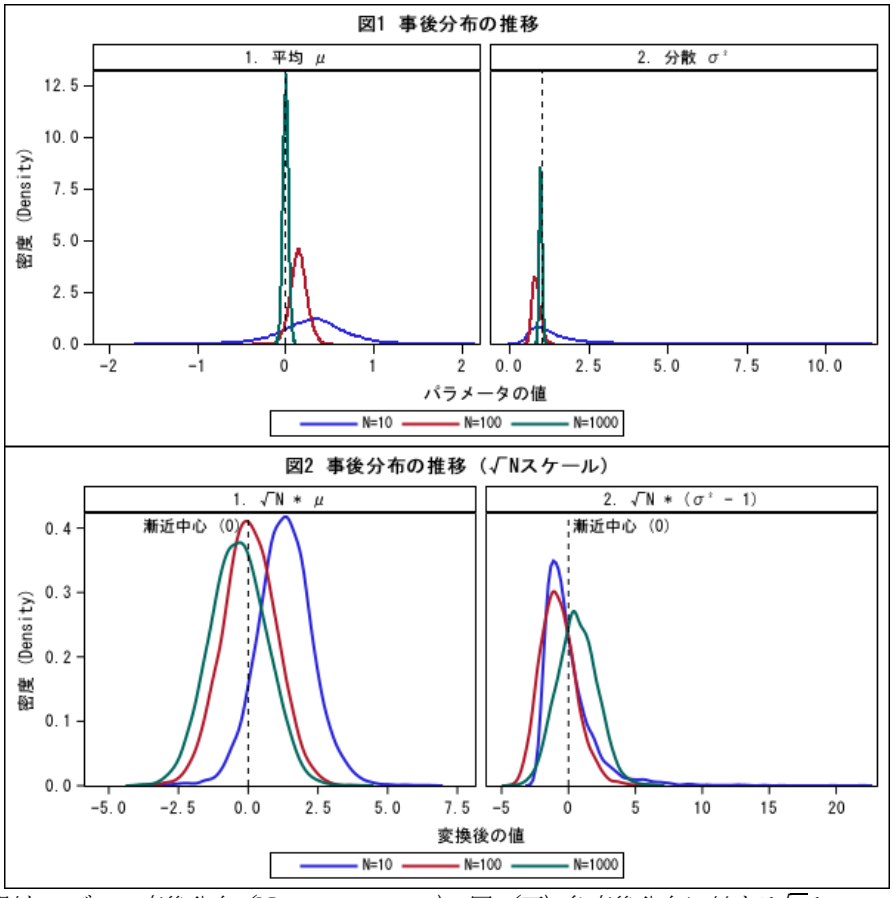


図 1(上) 正則なモデルの事後分布 (N=10, 100, 1000)。図 2(下) 各事後分布に対する$\sqrt{N}(\theta-\theta_{BEST})$の分布。

流の目標といえる。真の分布がモデルに含まれる場合、Θ_{BEST}上で$K(\theta)$は 0 になる。以下 3,4,5,6 章では簡単のため真の分布がモデルに含まれることを仮定する (実現可能性の仮定)[1]。

$$\Theta_{BEST} = \{\theta | K(\theta) \text{が最小値}\}$$
$$= \{\theta | K(\theta) = 0\}$$

多くの議論は実現可能性の仮定なしに成立するが、AIC など一部でこの仮定が必要になる。

3.2. 正則なモデルの定義

本稿では、次の 2 条件を満たすモデルを***正則なモデル***と呼ぶ。①Θ_{BEST}が 1 点$\{\theta_{BEST}\}$であり、θ_{BEST}はパラメータ空間の内点である(すなわち端の点などではない)こと。②その点で$K(\theta)$のヘッセ行列$J = \nabla^2 K(\theta)$が正定値であること。②は一見難しい数式のように見えるかもしれないが、これは単に関数$K(\theta)$が谷底θ_{BEST}のまわりで純粋な 2 次形式になり、いわば「きれいなお椀」のような形になることを意味する。すなわち次の近似がθ_{BEST}の周りで成立する。

$$K(\theta) \simeq \frac{1}{2}(\theta-\theta_{BEST})' J (\theta-\theta_{BEST})$$

この 2 条件のどちらが欠けても、3 章の以降の議論はすべて崩れる。4 章では①が破れる例、①が成り立っているが②が破れる例をそれぞれ紹介する。

3.3. 正則なモデルの例と漸近正規性

例として平均分散未知の正規分布を扱う$\theta = (\mu, \sigma^2)$, $f(Y|\mu,\sigma^2) = N(Y|\mu,\sigma^2)$。事前分布は Jeffreys 事前分布とする。また真の分布を平均 0 分散 1 の正規分布とする $f_{TRUE}(Y) = N(Y|0,1^2)$。最適なパラメータは$(\mu_{BEST}, \sigma^2_{BEST}) = (0,1^2)$であり、

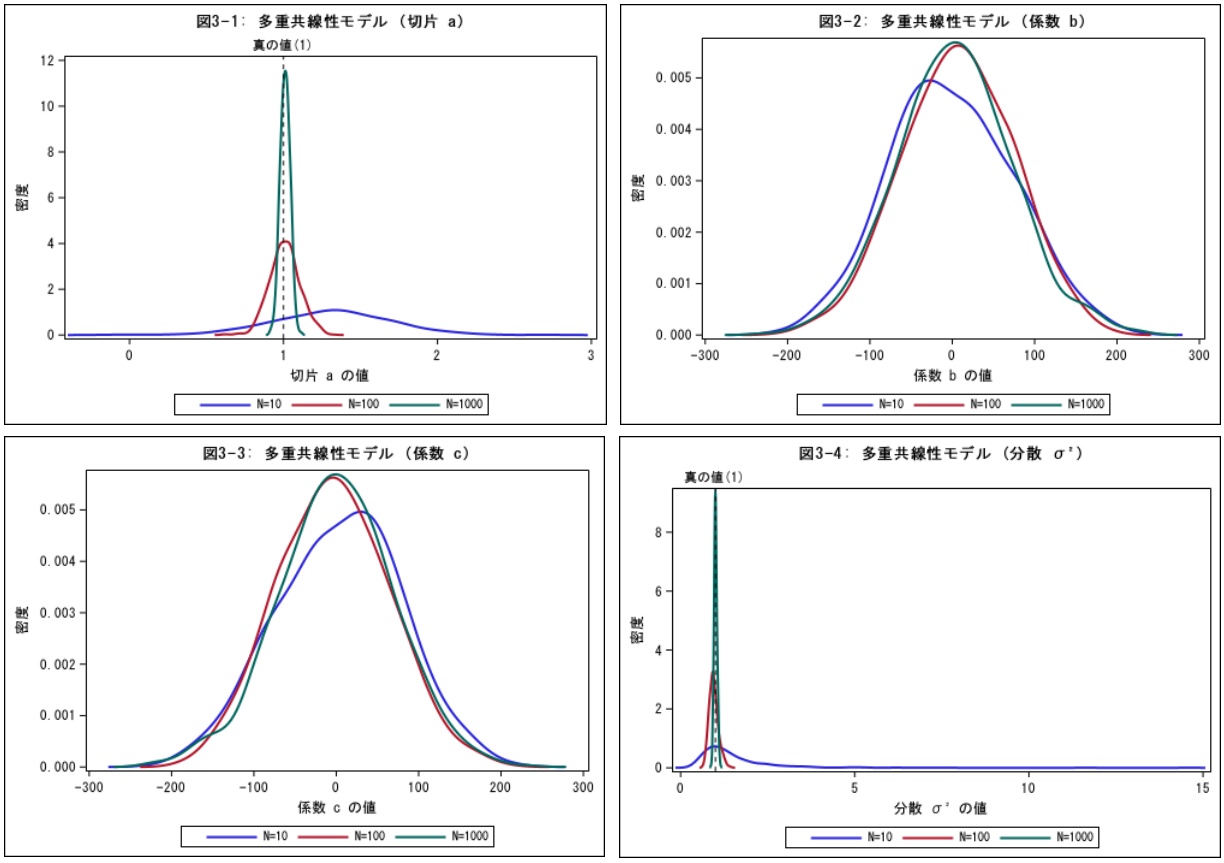


図 3-1, 図 3-2, 図 3-3, 図 3-4: 共線性を持つモデルの事後分布。a, σ^2 は観測データを増やすごとに分散が小さくなることがわかるが、b, c は観測データの個数の影響がない。

$$K(\mu, \sigma^2) = \frac{1}{2}\left(\log\sigma^2 + \frac{(1+\mu^2)}{\sigma^2} - 1\right)$$

$$J = \begin{pmatrix} 1/\sigma^2 & -\mu/\sigma^4 \\ -\mu/\sigma^4 & -\frac{1}{2\sigma^4} + \frac{1+\mu^2}{\sigma^6} \end{pmatrix}$$

になる。このとき$n = 10,100,1000$で事後分布は図 1 のようになる。例数が大きくなるにつれて真の平均、分散に収束していることがわかる。さらに収束の仕方の詳細を追うために、事後分布のθ_{BEST}からの離れ方を$\sqrt{n}$倍した確率変数を考える。

$$\sqrt{n}(\theta - \theta_{BEST})$$

$N = 10,100,1000$に対してこの確率変数は図 2 のようになる。$\sqrt{n}$倍することで一点に収束せず、正規分布に分布収束する（Bernstein–von Mises の定理)。

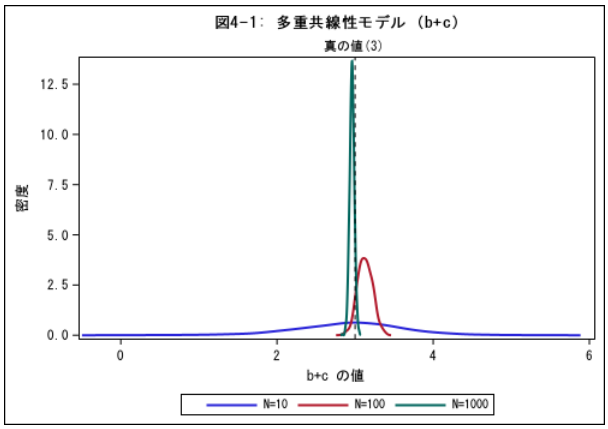
3.4. 事後分布の漸近正規性

3.3 章で見たように、正則なモデルでは、次の意味で事後分布は正規分布に近づく。

$$\sqrt{n}(\theta - \theta_{BEST} - \Delta_n) \to N(\theta|0, J^{-1})$$

ここでΔ_nはデータのゆらぎを表すベクトルで、経験過程η_nを用いて次で与えられるが本稿では深く立ち入らない。

$$\Delta_n = \frac{J^{-1}\nabla\eta_n(\theta_{BEST})}{\sqrt{n}}$$



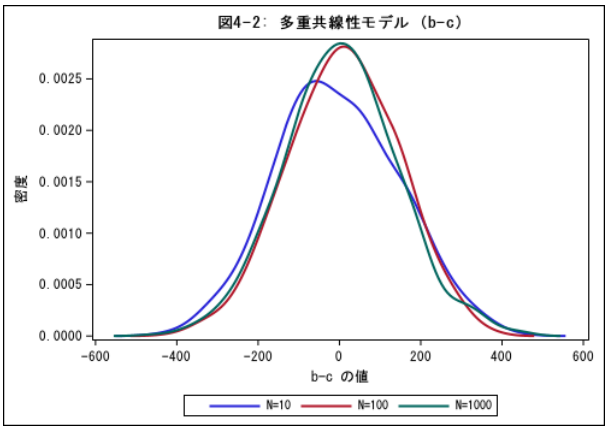


図 4-1,図 4-2: $(b+c)$の事後分布及び$(b-c)$の事後分布。$b+c$は観測データを増やすことで真の値 3 に収束する。

$$\eta_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ E_{Y_{TRUE}} \left[\log \frac{f(Y_{TRUE}|\theta_{BEST})}{f(Y_{TRUE}|\theta)} \right] - \log \frac{f(y_i|\theta_{BEST})}{f(y_i|\theta)} \right\}$$

重要な点は、Δ_n自体は 0 に収束するが、$\sqrt{n}\Delta_n$は 0 に収束しないことである。すなわち$\sqrt{n}(\theta - \theta_{BEST})$のバイアスである$\sqrt{n}\Delta_n$は消えず、$N(0, J^{-1})$に分布収束しない。

4. 非正則なモデルの例

3.2 章で挙げた 2 条件のうち、4.1 章では①が破れる例を、4.2 章では①は成り立つが②が破れる例を見る。

4.1. 共線性のある回帰

説明変数に完全な共線性がある回帰を考える。

$$Y = a + bx_1 + cx_2 + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

とし、真の分布を$x_1 = x_2 \sim N(0,1^2)$かつ$Y = 1 + 3x_1 + \epsilon$、$\epsilon \sim N(0,1)$とする。この設定は極端に見えるが、共線性の強い共変量を扱う場面の単純化である。このとき KL 情報量は次のようになる。

$$K(a,b,c,\sigma^2) = \frac{1}{2}\left\{ \log \sigma^2 - 1 + \frac{1 + (a-1)^2 + (b+c-3)^2}{\sigma^2} \right\}$$

よって$\Theta_{BEST} = \{(a,b,c,\sigma^2) | a = 1, b + c = 3, \sigma^2 = 1\}$となり一点ではなく直線となる。3 章の条件①が破れている。このとき$N = 10,100,1000$の事後分布は図 3 のようになる。b,cについて一点に収束しないことがわかる。原因は明らかで、このモデルではb,c個別の情報はなく、$b+c$について収束するのみである(図 4)。特に、情報の全くない$b-c$の分布は観測データによって変化しない。

4.2. 二乗平均のモデル

3 章の条件②が破れる例として次のモデルを考える

$$f(Y|\theta) = N(Y|\theta^2, 1^2)$$

真の分布を$f_{TRUE}(Y) = N(Y|0,1^2)$とする。このとき

$$K(\theta) = \theta^4/2$$

であり、最適なパラメータは$\theta_{BEST} = 0$の一点のみである。この一点におけるヘッセ行列は

$$J = 6\theta_{BEST}^2 = 0$$

であり、これは②を満たさない。このモデルにおいて事後分布は図 5 のようになる。N を増やしても正規分

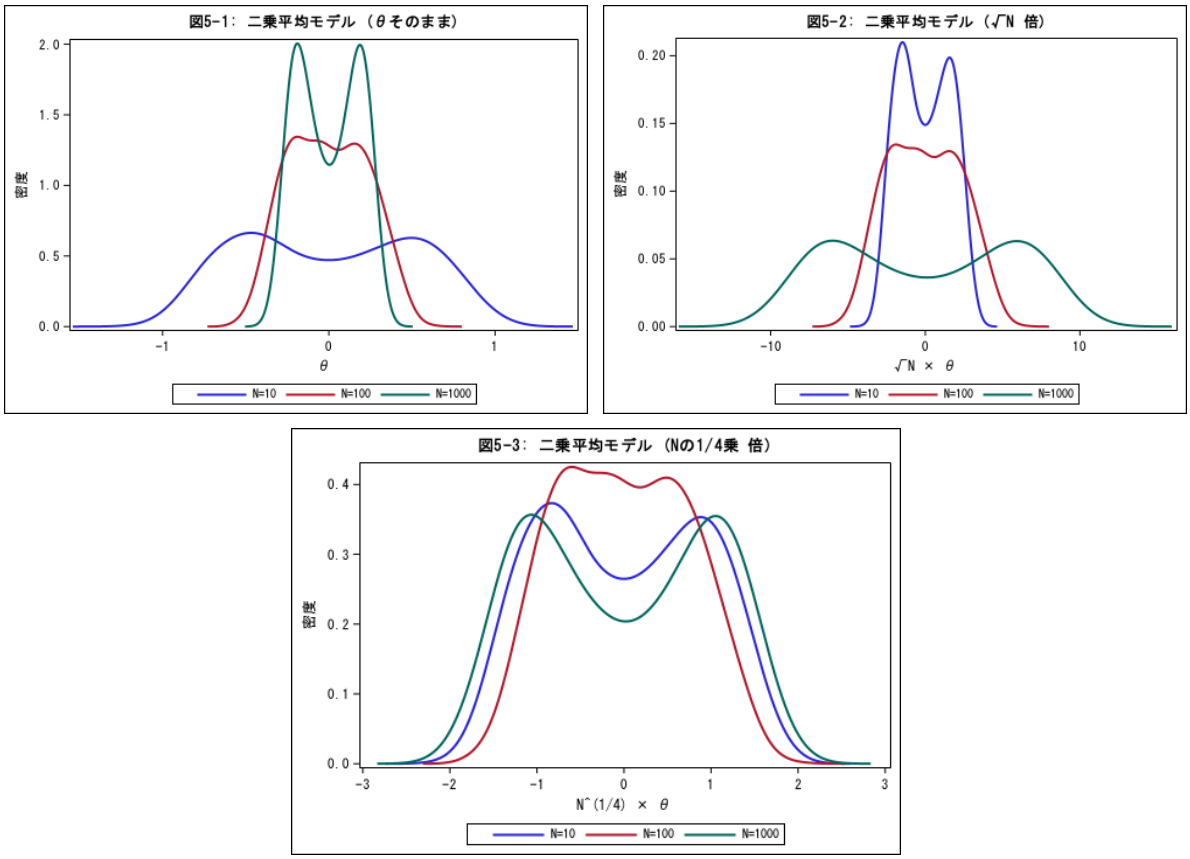


図 5-1, 図 5-2, 図 5-3:スケールを調整した二乗平均のモデルの事後分布。元の分布では観測データを増やすごとに分散が小さくなり、$\sqrt{N}$倍のスケールでは逆に大きくなる。$N^{1/4}$倍のスケールでは分散も 0 に収束もしない

布にはならず、また 3 章の例とは異なり$\sqrt{n}$倍すると発散していく。これは最適な点への収束が正則なモデルより遅いことを意味する（$\sqrt{n}$に速度で負けてしまう）。実際のスケールである$n^{1/4}$倍をかけると発散も一点への収束もしていないことが観察できる。しかし正規分布にはならない。

さらに図 5 では事後分布が二峰になっているがこれはデータに依存し、50%の確率で二峰に、50%の確率で単峰の分布となる。実際$u = n^{1/4}\theta$と置きなおすとuの事後分布は

$$\varphi(u|y_1,\cdots,y_n) \propto \exp(G \cdot u^2 - u^4/2)$$
$$G = \sqrt{n}\bar{y}$$

となり、観測データの平均$\bar{y}$が 0 より大きい場合は二峰、小さい場合は単峰である。正則なモデルでは観測データのゆらぎΔ_nによって分布の平均が変化していたが、このモデルでは分布の形も観測データによるゆらぎの影響をうけることがわかる。

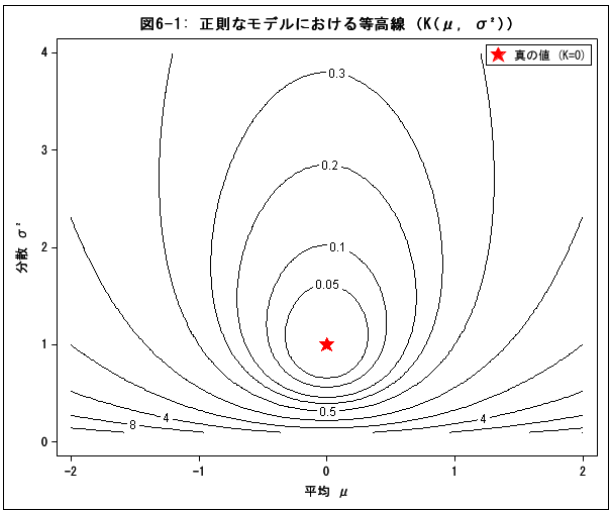
5. 非正則なモデルの漸近理論

5.1. 発想

4 章で見たように、非正則なモデルでは「Θ_{BEST}を中心にして$\theta - \theta_{BEST}$を測る」という測り方そのものが成立しない。Θ_{BEST}が 1 点でなければ中心が定まらず、Kのヘッセ行列が退化していれば$\sqrt{n}$という尺度も合わない。そこで測り方を変える。

θ がどこにあるかを追うのをやめて、K(θ) の値を追う。

θは d 次元のベクトルだが、$K(\theta)$は非負の実数である。d 次元の情報を 1 次元に潰す乱暴な操作に見えるが、



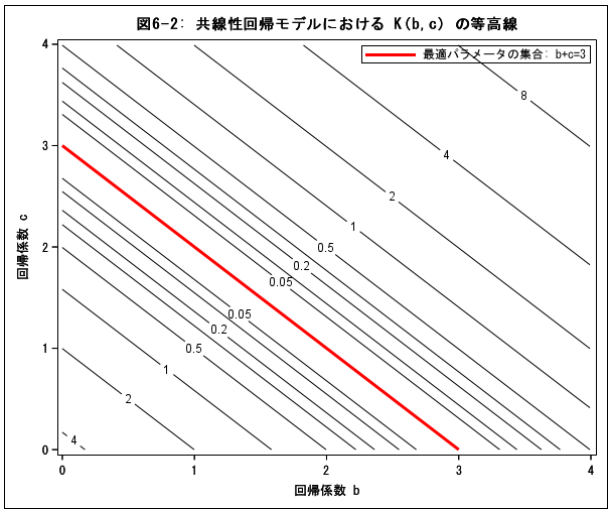
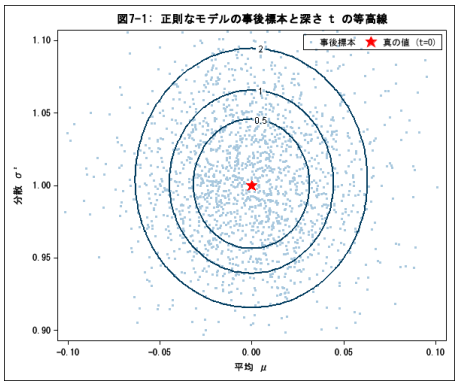
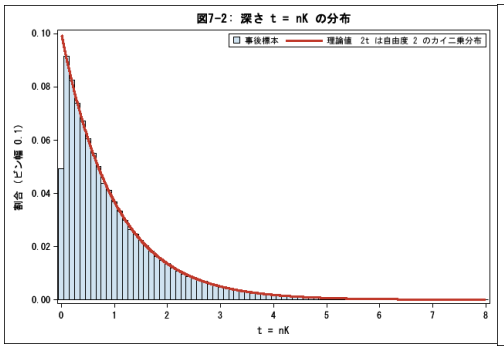


図 6-1:正則なモデルにおける$K(\mu,\sigma^2)$の値。図 6-2:共線性のある回帰モデルにおける$K(a,b,c,\sigma^2)$。ただし$a=1,\sigma^2=1$としてb,cに対してのみ図示している。





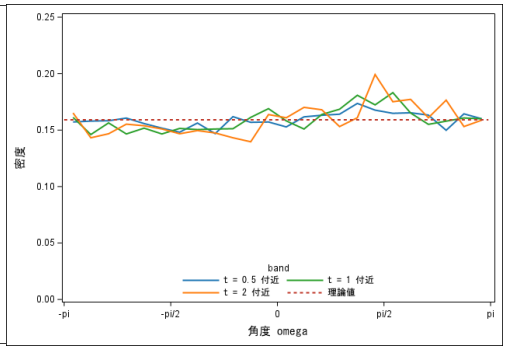


図 7-1: 事後分布の MCMC サンプルに t=0.5,1,2 の線をを重ねた図。(n=1000) 図 7-2: t で周辺化した分布。ヒストグラムは MCMC サンプルから作成され、赤線が理論値。図 7-3: t=0.5, 1, 2 の等高線の付近のωの分布。

$K(\theta)$はΘ_{BEST}が 1 点でなくとも定義でき、かつ真の分布との近さという、いちばん知りたい情報を持っている。

$\boldsymbol{t=nK(\theta)}$と置く。この$t$でパラメータ空間にいわば「等高線」を引き、各等高線の間と、各等高線内にわけて考える。図 6 は 3.3 章の正則モデルと 4.1 章のモデルである。ただし 4.1 章のモデルは簡単のため b,c についてのみ図示している。形はまったく違うが、「高さtでラベルを付ける」という操作自体はどちらでも同じようにできる。

一般論に入る前に、例として正則なモデルの場合のtの分布と各等高線内の分布を考える (図 7)。$K(\theta)\simeq$

$\frac{1}{2}(\theta-\theta_{BEST})^{\mathsf{T}}J(\theta-\theta_{BEST})$であったから

$$2t=\left(\sqrt{n}(\theta-\theta_{BEST})\right)^{\mathsf{T}}J\left(\sqrt{n}(\theta-\theta_{BEST})\right)$$

となる。3 章でみたように$\sqrt{n}(\theta-\theta_{BEST})$は観測データのブレを除いて正規分布であったから、**$2t$はおよそ自由度 d のカイ二乗分布になる。**($\varphi(t)\propto t^{\frac{d}{2}-1}\exp(-t)$)。また、

$$x=\mu,\ y=\frac{\sigma^2-1}{\sqrt{2}},\ x=r\cos\omega,\ y=r\sin\omega$$

とすると、ωの分布は観測データのブレを除いて一様分布となる。

5.2. 結論の形

前節で見たものは、実は正則なモデルに限った話ではない。同じ形が非正則なモデルでもそのまま成り立つ。先に答えを書く。θを「どの等高線か」を表すtと「その等高線の中のどこか」を表すuに分けて書き直すと、事後分布は次の形になる[1]。

$$\varphi(t,u|y_1,\cdots,y_n)\propto t^{\lambda-1}\exp\left(-t+\sqrt{t}\xi_n(u)\right)\cdot\varphi(u|t)$$

ここで各記号は次のとおりである。

$$t=nK(\theta)$$

$$\xi_n(u)=\frac{t-\sum_{i=1}^{n}\log\frac{f_{TRUE}(y_i)}{f(y_i|\theta)}}{\sqrt{t}}$$

$\xi_n(u)$はデータのゆらぎを表す量であり、$\varphi(u|t)$は各等高線に制限した時のuの事前分布である。

本稿ではこの式の導出は行なわず、この式を認めたうえでλなどについて考える。

5.3 章でこの漸近モデルを前提に、3.3 章、4.1 章、4.2 章の例に対して実際にλを天下り的に計算する。このときλの計算に当たっては$t\to 0$における挙動を考えればよい。$t\to 0$では$\exp\left(-t+\sqrt{t}\xi_n(u)\right)$は 1 に収束するため、原点における特異性はだけが担っており、λはこの因子から読み取れる。

5.4 章では一般のモデルにおいてλを事前分布から計算する方法を考える。

5.3. 計算例

前節の形が実際にどのように出てくるかを、3.3 章、4.1 章、4.2 章の例で確かめる。

5.3.1. 正則なモデル（3.3 章の例）

5.1 章で図 7 として見た「2t が自由度 d のカイ二乗分布になる」「各等高線の中で角度ω が一様になる」ことを、ここで導出する。真の分布を$f_{TRUE}(Y)=N(Y|0,1)$、モデルを$f(Y|\mu,\sigma^2)$とすると最適なパラメータは$(\mu_{BEST},\sigma^2_{BEST})=(0,1)$であった。3.3 章で記載したヘッセ行列 J を用いると、最適なパラメータの周りで$K(\mu,\sigma^2)$は次の二次形式で近似できる。

$$K(\mu,\sigma^2)\simeq\frac{1}{2}\mu^2+\frac{1}{4}(\sigma^2-1)^2$$

ここで高さsの等高線$K(\mu,\sigma^2)=s$を考えると、楕円

$$\mu^2+\frac{1}{2}(\sigma^2-1)^2=2s$$

で近似できることがわかる。この楕円を円にするために縦軸をスケーリングし、$x=\mu$、$y=\frac{1}{\sqrt{2}}(\sigma^2-1)$と座標変換する。すると$K(x,y)\simeq\frac{1}{2}(x^2+y^2)$となるため、中心からの距離$r=\sqrt{2s}$を用いて極座標

$$x=r\cos\omega,\ \ y=r\sin\omega$$

とできる。ここでsが「どの等高線か」を、角度ωが「その等高線の中のどこか」を表しており、これがまさに 5.2 章の(t,u)にほかならない。正則なモデルにおいて(t,u)は極座標表示によって実現できる。

このx,y,r,ωを用いてλを導出する。上述の変換によって(x,y)は独立な正規分布になる。

$$x,y\sim N(\Delta_n',1/n)$$

Δ_n'は観測データの揺らぎΔ_nを変換したものである。このときよく知られているように、$n(x^2+y^2)=nr^2=2ns=2nK(\theta)=2t$の分布は自由度 2 の非心カイ二乗分布となる。$t\to 0$においてこれは自由度 2 のカイ二乗分布で近似することができ、事後分布$\varphi(\theta|y_1,\cdots,y_n)$を$t$で周辺化すると

$$\varphi(t|y_1,\cdots,y_n)\propto t^{\frac{2}{2}-1}\exp(-t)$$

となる。したがって$\lambda=1$である。

以上の議論はパラメータが d 個ある正則なモデルにおいてほとんどそのまま成立する。θはd個の独立な正規分布$N(0,1/n)$に分解され、$2t$は自由度dのカイ二乗分布となる。したがって

$$\boldsymbol{\lambda=d/2}$$

となる。次の例から観察できるように、非正則なモデルにおいてλは$d/2$にはならない。実質的にモデルに関与しないパラメータが存在するためである。**2λは「実質的なモデルの自由度」**と考えることもできる。

5.3.2. 非正則なモデル (4.1 章の例)

説明変数$x_1=x_2$に対して真のモデルを$f_{TRUE}(Y)=N(Y|1+3x_1,1^2)$、モデルを$f(Y|a,b,c,\sigma^2)=N(Y|a+bx_1+cx_2,\sigma^2)$とすると最適なパラメータは直線$\Theta_{BEST}=\{(a,b,c,\sigma^2)|a=1,b+c=3,\sigma^2=1\}$であった。

$\alpha=a-1$、$\beta=b+c-3$、$\epsilon=\sigma^2-1$とおくとΘ_{BEST}の周りで$K(a,b,c,\sigma^2)$は次の二次形式で近似できる。

$$K(a,b,c,\sigma^2)\simeq\frac{1}{2}\left(\alpha^2+\beta^2+\frac{\epsilon^2}{2}\right)$$

このとき、$\gamma=b-c$はこの式に現れない。つまりこのモデルは、パラメータを 4 個持ちながら実質的には 3 個しか使っていない。

言い換えるとヘッセ行列 J はγの方向に 0 の固有値をもち、$\mathrm{rank}\,J=3$である。5.4 章で見るように$K(\theta)$が二次形式で近似できる場合$\lambda=\mathrm{rank}(J/2)$であるから、$\lambda=3/2$を得る。$d/2=4/2$にはならない。

5.3.3. 非正則なモデル (4.2 章の例)

最後に 4.2 章の$f(Y|\theta)=N(Y|\theta^2,1^2)$の例を見る。$K(\theta)=\theta^4/2$であった。

このモデルでは 5.3.1 章, 5.3.2 章と異なり$K(\theta)$は二次形式にならない。直接計算することでλが計算できる。

4.2 で紹介したように、$u=n^{1/4}\theta$と置きなおすとuの事後分布は

$$\varphi(u|y_1,\cdots,y_n)\propto\exp(G\cdot u^2-u^4/2)$$

である。$t=nK(\theta)=u^4/2$であるから$u=(2t)^{1/4}$であり、

$$\varphi(t)\propto t^{-\frac{3}{4}}\exp\left(-t+\sqrt{2}G\sqrt{t}\right)$$

となる。$t^{\lambda-1}\exp\left(-t+\sqrt{t}\xi_n(u)\right)$と比較すると、

$$\lambda=\frac{1}{4}$$

である。特筆すべき点として、このモデルは$2\lambda=1/2$であり整数値にはならない。λはパラメータの個数を数

えるだけでは得られず、一般の定義が必要になる。

5.4. 実対数閾値λ

本節で一般のλの定義を述べる。5.3 章の 3 つの例で共通して使ったのは、$K(\theta)$の形そのものではなく「高さsの等高線」に事前分布の重みがどれだけ乗っているかだけであった。これを**状態密度**と呼び、$v(s)$と書く。

$$v(s) = \int \delta\big(s - K(\theta)\big)\varphi(\theta)d\theta = \int_{s=K(\theta)} \frac{\varphi(\theta)}{|\nabla K(\theta)|} d\theta$$

ここでδはディラックのδ関数である。平たく言えば「$s = K(\theta)$のθにおいて事前分布$\varphi(\theta)$を積分する」ことを意味する。この$v(s)$の原点における振る舞い（極）によってλが定義できる

$$v(s) = v_0 s^{\lambda-1} \qquad (s \to 0)$$

このλを**実対数閾値**（real log canonical threshold）と呼ぶ。なお一般には、この式に重複度と呼ばれる自然数で定まる対数の因子が付く。本稿で扱う 3 つのモデルはいずれも重複度が 1 であり、その場合は上の形のままでよい。定義から、$K(\theta)$が二次形式で近似できる場合、

$$\lambda = \frac{\operatorname{rank} J}{2}$$

特に正則なモデルにおいては$\lambda = d/2$であり、一般に$\lambda \leq d/2$である[5]。ただし事前分布がΘ_{BEST}において 0 になる場合$\lambda > d/2$となりうる。

5.5. 実対数閾値λの推定量と実装

以上の議論ではすべて$K(\theta)$を用いており、これは真の分布$f_{TRUE}(Y)$がわかっている状況でしか厳密なλは導出できない。現実的には与えられた観測データからλに収束する何らかの統計量を作ることでλを推定することになる。詳細はここでは触れないが、以下の理論によってλの推定量を作ることができる[6]。

まず「逆温度βの事後分布」を以下で定義する

$$\varphi_\beta(\theta|y_1,\cdots,y_n) \propto \varphi(\theta) \times \prod_{i=1}^{n} f(y_i|\theta)^\beta$$

各観測データの尤度$f(y_i|\theta)$をβ乗した形であり、これは「各観測データがβ個得られた」と考えていることと等しい。$\beta = 2$であればすべての観測データを二回ずつかけており、$\beta = 0.5$であれば各観測データを 0.5 個分で計算している。特に$\beta = 0$の場合明らかに$\varphi_\beta(\theta|y_1,\cdots,y_n) = \varphi(\theta)$である。

このとき、次が成り立つ

$$E^\beta\left[\sum_{i=1}^{n} \log f(y_i|\theta)\right] \simeq C_n - \frac{\lambda}{\beta}$$

すなわち、$\varphi_\beta(\theta|y_1,\cdots,y_n)$で対数尤度の期待値をとり、それを$1/\beta$に対してプロットすると、傾きが$-\lambda$になる。

実際に本稿の 3 つのモデルに対して計算すると以下の値が得られた。(SAS コードは別添)

モデル	理論値	N=10	N=100	N=1000
正則$N(\mu,\sigma^2)$	1	1.35	1.10	0.998
共線性モデル	3/2	2.71	1.66	1.56
$N(\theta^2,1)$	1/4	0.188	0.249	0.288

ただしこの計算には逆温度βの事後分布を繰り返し計算する必要があり、特に観測データが多い場合時間がかかる。また非正則モデルでは特に観測データによる誤差が大きく、N=1000 程度では正しいλの算出は難しいモデルが存在する。

6. 応用：WAIC

5 章で導入した漸近近似の応用例として**渡辺・赤池情報量規準（WAIC）**を導出する。

6.1. 汎化損失と経験損失

与えられた観測データ$(y_1,\cdots,y_n)$から真の分布を予測するうえで、より予測精度の良いモデルを探索する手法を考える。例えば回帰分析において説明変数を選択するような状況である。

事後分布$\varphi(\theta|y_1,\cdots,y_n)$に関する期待値を$E_{post}[\cdot]$と書くことにする。このとき、**事後予測分布**が次で定義される。

$$r(Y|y_1,\cdots,y_n)=E_{post}[f(Y|\theta)]$$

すなわち、観測データ$y_1,\cdots,y_n$の下で新しいデータYの予測をする分布である。事後予測分布をもちいて、真の分布に対する予測の悪さである**汎化損失**、手元のデータの当てはまりの悪さである**経験損失**を次で定義する。

$$G_n=E_{Y_{TRUE}}[-\log r(Y_{TRUE}|y_1,\cdots,y_n)]$$

$$T_n=-\frac{1}{n}\sum_{i=1}^{n}\log r(y_i|y_1,\cdots,y_n)$$

汎化損失G_nは事後予測分布が真の分布$f_{TRUE}(Y)$に当てはまるほど小さくなる関数であり、G_nが小さいモデルは「良いモデル」であるといえる。一方でG_nは$f_{TRUE}(Y)$が既知でないと計算できないため、観測データ$(y_1,\cdots,y_n)$からG_nを推定することが必要になる。

経験損失T_nは一見するとG_nを推定する量になっているように見える。しかし、T_nはいわゆる過適合（over fitting）を起こすことが知られる。すなわち、事後分布の予測とG_nの予測両方に同じ観測データを用いているため、バイアスが発生する（$E[G_n]\neq E[T_n]$）。このずれは以下になることが知られる[1][7]。

$$G_n-T_n=(S-S_n)+\frac{E_{post}\left[\sqrt{t}\xi_n(u)\right]}{n}+o(1/n)$$

ここで、Sは真の分布のエントロピー、S_nは経験エントロピー：

$$S=-E_{Y_{TRUE}}[\log f_{TRUE}(Y_{TRUE})]$$

$$S_n=-\frac{1}{n}\sum_{i=1}^{n}\log f_{TRUE}(y_i)$$

$E_{Y_{TRUE}}[S-S_n]=0$であるから、バイアスの正体は$E_{post}\left[\sqrt{t}\xi_n(u)\right]/n$であることがわかる。

$$2\nu=E_{Y_{TRUE}}\left[E_{post}\left[\sqrt{t}\xi_n(u)\right]\right]$$

とおくと、

$$E_{Y_{TRUE}}[G_n]=E_{Y_{TRUE}}[T_n]+\frac{2\nu}{n}+o(1/n)$$

となる。もちろん2νの値は真の分布がわからないと計算できない。よってこの2νを観測データから推定することが必要になる。推定量を作ることができれば、それを罰則項として汎化損失のバイアスのない推定が可能になる。これが WAIC である。

正則かつ実現可能なモデルでは$2\nu=d$となる[1]。したがって次の**正則の情報量規準**が与えられる。

$$T_n+\frac{d}{n}$$

これは赤池情報量規準 AIC[8]のベイズ版に当たる。

一方で非正則なモデルでは$2\nu = d$とは限らず、正則の情報量規準はバイアスを含む。また、3.1 章で仮定した実現可能性（真のモデルが$f(Y|\theta)$に含まれている）を仮定しない場合も同様である。WAIC はいずれの場合にも成立する点で有用である。

6.2. WAIC

2νは観測データ$(y_1, \cdots, y_n)$から推測することができる。各観測の対数尤度の事後分布に関する分散を足し合わせた量を**汎関数分散**と呼ぶ。

$$V_n = \sum_{i=1}^{n} Var_{post}[\log f(y_i|\theta)]$$

これが2νの推定量になる[1]。WAIC は次で定義される。

$$WAIC = T_n + \frac{V_n}{n}$$

このとき

$$E_{Y_{TRUE}}[G_n] = E_{Y_{TRUE}}[WAIC] + O(1/n^2)$$

が成り立つことが知られる[1]。

6.3. SAS による実装

2.2 章のサンプルコードのデータセット height, post_out から WAIC を算出できる。

```
proc sql noprint;
   select count(*) into :M trimmed from post_out;
quit;
%put &=M : the number of posterior MCMC samples;
data loglik;
   set height;                              /* 被験者 subjid */
   do j = 1 to &M.;
      set post_out(keep=mu sigma) point=j;   /* 事後標本 j */
      ll  = logpdf('NORMAL', y, mu, sigma);
      lik = exp(ll);
      output;
   end;
run;
proc means data=loglik noprint;
   by subjid notsorted;
   var lik ll;
   output out=stat mean(lik)=Ebar var(ll)=v_i;
run;
data waic;
   set stat end=last;
   sumT + (-log(Ebar));           /* SUM -log E[f]  */
   Vn   + v_i;                   /* SUM Var[log f] */
```

```
  if last then do;
    Tn   = sumT / _n_;
    WAIC = Tn + Vn / _n_;
    output;
  end;
  keep Tn Vn WAIC;
run;
```

このコードでは$\log r(y_i|y_1,\cdots,y_n)$、$Var_{post}[\log f(y_i|\theta)]$を全ての$y_i$について計算するため「観測データの個数$n$ × 事後分布の MCMC サンプルの個数」行のデータセットを作成している。観測データを 5 個としているため問題ないが、観測データを増やすと過剰に重くなる。

Yが 1 変数で説明変数を持たない場合、Yの代表値をいくつか選び、その点を用いて事後予測分布及び$Var_{post}[\log f(Y|\theta)]$を推定することで軽量化できる。例えば 50 点の$y$に対して$r(y|y_1,\cdots,y_n)$、$Var_{post}[\log f(y|\theta)]$を求め、ラグランジュ 5 点内挿 (各$y_i$に近い 5 点を用いた 4 次近似) を用いて$r(y_i|y_1,\cdots,y_n)$、$Var_{post}[\log f(y_i|\theta)]$ $(i=1,\cdots,n)$ を求めることができる。

3.3 章、4.1 章、4.2 章の例に対してV_n、T_n、WAIC を計算すると以下になる。(SAS コードは別添) (4.2 章の例の$2v$の理論値は数値積分による。R コードを別添。)

モデル	n	$2v$（理論値）	V_n	T_n	WAIC	G_n
正則$N(\mu,\sigma^2)$	10	2	1.97	1.35	1.54	1.47
	100	2	2.04	1.36	1.38	1.42
	1000	2	1.98	1.41	1.41	1.42
共線性モデル	10	3	2.81	1.28	1.56	1.52
	100	3	3.13	1.46	1.49	1.42
	1000	3	2.92	1.40	1.41	1.42
$N(\theta^2,1)$	10	0.345…	0.471	1.33	1.38	1.45
	100	0.345…	0.293	1.36	1.36	1.42
	1000	0.345…	0.697	1.41	1.41	1.42

7. まとめ

本稿では「観測データが十分に存在するとき、事後分布はどのようにふるまうか」という問いを追ってきた。頻度論において標本平均の妥当性を中心極限定理が支えているのと、同じ役割を果たす定理である。答えはモデルが正則であるかどうかで分かれる。Θ_{BEST}が 1 点であり、その点で$K(\theta)$のヘッセ行列Jが正定値であれば、事後分布は$\sqrt{n}(\theta-\theta_{BEST})$の意味で正規分布に近づく。事後分布から作った信用区間が頻度論的にも妥当な被覆をもつことも、パラメータの個数を罰則とする AIC が使えることも、この事実の上に立っている。

ところがこの 2 条件は簡単に壊れる。そのときはθの位置を追うことをやめ、真の分布との KL 情報量の値$t=nK(\theta)$を追えばよい。事後分布は正則・非正則を通じて$t^{\lambda-1}\exp\left(-t+\sqrt{t}\xi_n(u)\right)\cdot\varphi(u|t)$という同じ形になり、モデルの違いは実対数閾値$\lambda$ただ一つに集約される。正則なモデルでは$2\lambda=d$となるが、非正則ではそうならない。応用として、汎化損失と経験損失のずれ$2v$を事後標本から推定する WAIC を導き、SAS による実装を与えた。

ここで強調しておきたいのは非正則なモデルは決して特殊なものではないということである。4.2 章の$N(\theta^2,1)$は、ごく普通の正規分布$N(\mu,1)$を変数変換しただけのモデルにすぎない。つまり非正則性は、モデルそのものというよりパラメータの付け方から生じている。パラメータを冗長に取ることや、意味のある量を非線形に書き直すことは、実務では自然に起こる。

応用として WAIC を導いた。汎化損失と経験損失のずれ2νは、正則かつ実現可能なモデルではdに一致するが、非正則ではそうならない。$N(\theta^2,1)$では$2\nu \simeq 0.35$であり、$2\lambda = 0.5$とも$d = 1$とも異なる値になる。それでも2νは各観測の対数尤度の事後分散を足し合わせるだけで推定できるため、モデルが正則かどうかを判定することなく汎化損失を見積もることができる。

本稿で扱わなかったことを 3 点挙げる。第一に、λを解析的に求めるには特異点解消と呼ばれる代数幾何の手続きが必要になる。実は正則なモデルにおいて用いた極座標は正に特異点解消の例である。一般の非正則なモデルでは「極座標の一般化」のような概念を導入する必要がある。第二に、理論が成り立つための条件である。本稿では 5.2 章の形を認めて話を進めたが、これは無条件に成り立つものではない。まずモデル$f(Y|\theta)$がθについて解析的であること。単に滑らかであるだけではなく、解析性が必要となる。次に、相対的に有限な分散を持つことが必要である。これは真の分布から離れたパラメータにおいても対数尤度比が$K(\theta)$に比べて暴れないことを要求している。このほかパラメータ空間が有界であることなどが必要である。第三に、WBIC である。5.5 章で用いた逆温度によるλの推定は、周辺尤度Z_nの対数を推定する WBIC と呼ばれる枠組みから得られたものだが、その導出には立ち入らなかった。WAIC が予測の良さを測るのに対し、WBIC はモデルと事前分布の適切さを測る規準であり、目的が異なる。

θの位置を追うことをやめて$K(\theta)$の値を追えば、正則・非正則を通じた漸近論が得られ、そこから正則性を仮定しない情報量規準 WAIC が、PROC MCMC の事後標本だけで計算できる形で導かれる。

8. 参考文献

[1] 渡辺澄夫. ベイズ統計の理論と方法. コロナ社, 2012

[2] 鈴木譲. 渡辺澄夫ベイズ理論 100 問 with R/Stan. 共立出版, 2023

[3] SAS Institute Inc. SAS/STAT® User's Guide. Chapter 80, "The MCMC Procedure". Cary, NC: SAS Institute Inc., 2026 https://documentation.sas.com/api/collections/pgmsascdc/v_077/docsets/statug/content/mcmc.pdf

[4] Kullback, S., Leibler, R. A. On information and sufficiency. Annals of Mathematical Statistics, 1951, 22(1): 79-86. https://doi.org/10.1214/aoms/1177729694

[5] Watanabe, S. Algebraic Analysis for Nonidentifiable Learning Machines. Neural Computation, 2001, 13(4): 899-933. https://doi.org/10.1162/089976601300014402

[6] Watanabe, S. A widely applicable Bayesian information criterion. J. Mach. Learn. Res., 2013, 14: 867-897 https://www.jmlr.org/papers/v14/watanabe13a.html

[7] Watanabe, S. Equations of States in Singular Statistical Estimation. Neural Networks, 2010, 23(1): 20-34. https://doi.org/10.1016/j.neunet.2009.08.002

[8] Akaike, H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 1974, 19(6): 716-723. https://doi.org/10.1109/TAC.1974.1100705