一般事後分布に基づくベイズ推論とその応用

橋本 真太郎　(広島大学大学院理学研究科)*

概　　要

近年，仮定された統計モデルの中にデータを生成する真のモデルが含まれていないような場合，つまりモデルが誤特定されている場合のベイズ推論に関する研究が盛んに行われてきている．このような場合，通常のベイズの定理に基づくベイズ更新は意味をなさず，一つの解決法として一般的なベイズ更新に基づく一般事後分布を用いる方法がある．本論では，その枠組みを概観し，新たなロバストベイズ推定の方法を提案する．

1. はじめに

パラメータ θ を与えたもとでの X の密度関数を $f(x \mid \theta)$ とし，θ の事前密度を $\pi(\theta)$ とする．また，データ X を生成する真の分布を G とし，この真の分布は仮定されたモデル $\{f(x \mid \theta) \mid \theta \in \Theta\}$ に含まれているとする．このとき，データ $X = x$ を与えたもとでの θ の事後分布はベイズの定理により

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta')\pi(\theta')\mathrm{d}\theta'}$$

と表現される．ここで，θ の事前分布は $f(x \mid \theta)$ における θ の事前の不確かさを表していることに注意する．しかしながら，データ生成過程が想定したモデルの中に入っていない場合は，$\pi(\theta)$ は何を表しているのか，という疑問が生じる．例えば，頻度論の枠組みでモデルが誤特定されている場合の最尤法では

$$\theta^* = \arg\min_{\theta \in \Theta} \left\{ -\int \log f(x \mid \theta)\mathrm{d}G(x) \right\}$$

をターゲットとしこれを推定することを考え，推定量の漸近分布の共分散行列がサンドイッチ型になることはよく知られている事実である．Bissiri et al. (2016) では，この $-\log f(x \mid \theta)$ をより一般の損失関数 $\ell(\theta, x)$ に置き換え，$\int \ell(\theta, x)\mathrm{d}G(x)$ を最小にするパラメータに関する合理的 (coherent) かつ妥当な (valid) ベイズ更新を考えている．これを一般ベイズ更新 (general Bayesian updating) と呼び，対応する事後分布を一般事後分布 (general posterior) と呼ぶ．このように，データ x とパラメータ θ を尤度関数ではなく一般の損失関数により結びつけることによるベイズ的方法は実際にはもう少し前から用いられてきたのであるが (例えば，Chernozhukov and Hong (2003) 等)，この考え方を統一的に扱ったのが Bissiri et al. (2016) であり，この論文を起点に関連研究が盛んに行われてきている．

2. General posterior distribution

X を分布 P に従う確率変数とし，その n 個の独立な複製を $X^n = (X_1, \ldots, X_n)$ とおく．いま，$\theta = \theta(P)$ $(\theta \in \Theta)$ に関心があるとし，データと θ を結びつける関数として損失関

本研究は JSPS 科研費 若手研究 (B) (17K14233, 橋本真太郎) の助成を受けたものである。
キーワード：ベイズ統計，一般事後分布，ロバスト推定，縮小事前分布，マルコフ連鎖モンテカルロ法

* 〒 739-8521 広島県東広島市鏡山 1-7-1 総合科学部内 理学研究科
e-mail: s-hashimoto@hiroshima-u.ac.jp

数 $\ell(\theta, x)$ を準備する．また，ターゲットとなる母数を $\theta^* = \arg\min_{\theta\in\Theta} \mathrm{E}_{X\sim P}\ell(\theta, X)$ により定義する．ここで，$\mathrm{E}_{X\sim P}\ell(\theta, X)$ を θ の関数としてみたものを discrepancy function と呼び $D(\theta)$ とかくことにする．$D(\theta)$ をデータを用いて経験推定したものを $D_n(\theta)$ とかき，

$$D_n(\theta) = n^{-1}\sum_{i=1}^{n} \ell(\theta, X_i)$$

により表現する．後に述べる一般ベイズ更新の合理性を保証するには，$D_n(\theta)$ がデータ X_i に関して加法性を持っていることは重要である．θ の事前分布を Π とするとき，Π-可測集合 A に対する一般事後分布は次で計算される：

$$\Pi_{n,\omega}(A) = \frac{\int_A e^{-\omega n D_n(\theta)}\Pi(\mathrm{d}\theta)}{\int_\Theta e^{-\omega n D_n(\theta')}\Pi(\mathrm{d}\theta')}. \tag{1}$$

ここで，$\omega > 0$ は事前に決めておかないといけない scale parameter であり，learning rate と呼ばれ，事後分布の広がりをコントロールする役割を果たすが一般に選択は難しい．また，この一般事後分布は Gibbs posterior と呼ばれることもある．損失関数 $\ell(\theta, x)$ は分析の目的に応じて決めればよく，例えばメディアンの推定の興味があれば $\ell(\theta, x) = |x - \theta|$ と選べば，$D(\theta) = \mathrm{E}_{X\sim P}|X - \theta|$ の最小値としてのメディアンの推定をデータが従う統計モデルを仮定することなく行うことが可能である．

上記のことからわかるように，M-推定のときと同じように，データと母数は尤度の代わりに discrepancy function によりリンクしており，それによりモデルの誤特定を回避できる．さらに，ベイズ的なアプローチは母数の分布の推定を与えるため，事後分布をマルコフ連鎖モンテカルロ (MCMC) 法などにより適切に近似できれば，信用区間 (credible interval) の形で母数の不確かさの定量化 (uncertainly quantification) が容易にできる．

ここで，一般ベイズ更新の妥当性と合理性について少し触れる．

定理 1 (Zhang (2006)). ν を確率測度とし，損失関数 $\ell(\theta, x)$ はデータ x_i について加法的であると仮定する．このとき，(1) に対する密度関数 $\pi_{n,\omega}(\mathrm{d}\theta)$ は次の最小化問題の解となる：

$$\min_{\nu}\left[\int\sum_{i=1}^{n}\ell(\theta, x_i)\nu(\mathrm{d}\theta) + \frac{1}{\omega}D_{\mathrm{KL}}(\nu, \Pi)\right].$$

ただし，$\omega > 0$ は learning rate であり，$D_{\mathrm{KL}}(\nu, \Pi)$ は ν と Π の間のカルバック・ライブラー (KL) ダイバージェンスとする．

証明の概要は以下の通りである．目的関数を書き直すと

$$\begin{aligned}&\int\sum_{i=1}^{n}\ell(\theta, x_i)\nu(\mathrm{d}\theta) + \frac{1}{\omega}\int\log\left(\frac{\nu(\theta)}{\pi(\theta)}\right)\nu(\mathrm{d}\theta)\\ &= \frac{1}{\omega}\int\log\left[\frac{\nu(\theta)}{\exp\{-\omega\sum_{i=1}^{n}\ell(\theta, x_i)\pi(\theta)\}}\right]\nu(\mathrm{d}\theta)\end{aligned}$$

となり，これを最小にするのは $\nu(\theta) \propto \exp\{-\omega\sum_{i=1}^{n}\ell(\theta, x_i)\pi(\theta)\}$ のときであることから定理の主張を得る．定理 1 の目的関数の第二項の事前分布の罰則の尺度に関しては，Bissiri et al. (2016) により，合理性を保証するには KL-ダイバージェンスでなければいけないことが示されている．

次に，合理性についてであるが，これはデータを n 個持っているときに，これらすべてを使って更新した事後分布と最初の j 個を用いてベイズ更新し，それを事前分布としてみてさらに残りの $n-j$ 個のデータを用いて更新した事後分布が同じであるということである．もちろん対数尤度はデータに対する加法性を持つので通常のベイズ事後分布では成り立っている性質であるが，一般的な損失関数に基づく事後分布の場合は自明ではないことに注意する．実際，すぐ後に出てくる，γ-divergence はそのままでは加法性を持たない．

3. ロバスト統計への応用

本研究では，一般ベイズ更新の考え方を用いてデータ生成過程に外れ値等の何らかの異常がある場合に，その異常に対して頑健なベイズ推定について扱う． 前節でも述べたとおり，本来一般事後分布は model free な方法であることが長所であるが，本節では特に統計モデル $\{f(x \mid \theta) \mid \theta \in \Theta\}$ を一つ想定し，損失関数として $\ell(x, f(x \mid \theta))$ を考えることにする．すると，推定のターゲットとなるパラメータはダイバージェンスに対する相互エントロピー (cross entropy) を最小化するものであるとも考えることができる (この場合，損失関数がモデル $f(x \mid \theta)$ に依存する，つまり scoring rule となっていることを除けば，第 2 節の general Bayesian updating の枠組みに乗る．)．ダイバージェンスに基づくロバスト推定は近年，統計学・機械学習で非常によく用いられており density power divergence や γ-divergence など様々な良い性質をもつダイバージェンスが提案されてきている．特に，後者は外れ値の割合が小さくなくても安定した推定値を与えることから近年様々なモデルに対する手法が提案されてきている．

ベイズ的な観点で，γ-divergence を考えることの利点として，パラメータの不確かさに関する解析ができること，予測分布の精度が高いことなどが挙げられる．また，例えばスパースモデリングを行う際には多くの場合，調整パラメータの選択が必要になるが，ベイズ統計学ではそのパラメータにも不確かさを考え，調整パラメータもデータから推定できる良さも持つ．

確率密度 g, f_θ に対して，γ-divergence は次で定義される：

$$\begin{aligned} d_\gamma(g, f_\theta) =& \frac{1}{\gamma(\gamma+1)} \log \int g(x)^{1+\gamma} \mathrm{d}x \\ & - \frac{1}{\gamma} \log \int g(x) f(x \mid \theta)^\gamma \mathrm{d}x + \frac{1}{\gamma+1} \log \int f(x \mid \theta)^{1+\gamma} \mathrm{d}x. \end{aligned} \tag{2}$$

ここで，$\gamma > 0$ はあらかじめ分析者が指定する定数である．(2) 式の第 2，3 項目を合わせて γ-cross entropy とよび，この損失関数を最小にするような θ をターゲットにする (これを θ^*_γ とする)．ロバスト統計の文脈では，データ生成分布として，$g(x) = (1-\varepsilon) f_0(x) + \varepsilon w(x)$ の様な汚染分布を想定することが多く，ある条件を仮定すると，γ-divergence により g と f_θ の距離を測ったときに，ターゲット母数 θ^*_γ は ε の大きさによらず θ_0 近いことが Fujisawa and Eguchi (2008) により示されている．ここで，θ_0 は $f_0 = f_{\theta_0}$ を満たすようなターゲットとなる分布の母数であるとする．しかし，(2) をそのまま使おうとするとデータに関する損失関数の加法性が成り立たず，一般事後分布の合理性が失われるため，γ-cross entropy を単調変換して加法性をもたせる工夫をする：

$$-\exp\{-\gamma d_\gamma(g(x), f(x;\theta))\} \propto -\frac{\int f(x;\theta)^\gamma \mathrm{d}G(x)}{\{\int f(x;\theta)^{1+\gamma} \mathrm{d}x\}^{\gamma/(1+\gamma)}} =: q^{(\gamma)}(\theta, x).$$

ここで，単調変換に関してはターゲットとなる θ^*_γ の値は変わらないことに注意する．これらの準備のもとで Nakagawa and Hashimoto (2019) は γ-divergence に基づく θ の事後分布を

$$\pi^{(\gamma)}(\theta \mid X) \propto \exp\left\{-\sum_{i=1}^{n} q^{(\gamma)}(\theta, x_i)\right\} \pi^{(\gamma)}(\theta)$$

により提案した．ここで，$\pi^{(\gamma)}(\theta)$ は $f(x \mid \theta)$ における θ の事前分布ではなく，ターゲットとなるパラメータ θ^*_γ に関する事前分布であることに注意する．また，一般事後分布における learnig rate ω は 1 としている．上記の事後分布に限らず一般事後分布は正規分布モデルの場合ですら非常に複雑な形になり，共役事前分布も存在しないので，事後平均等の事後要約統計量を解析的に計算することは難しい．そのため，事後分布を MCMC 法を用いて近似することで母数の推測を行う．また，事後分布の漸近性質としては以下の定理が成り立つ．

定理 2 (Nakagawa and Hashimoto (2019))**.** いくつかの正則条件と，$\hat{\theta}^{(\gamma)}_n$ を γ-divergence 最小化推定量であるとする．このとき，事前分布 $\pi(\theta)$ が θ^*_γ において正値かつ連続ならば

$$\int \left| \pi^{(\gamma)}(t \mid X) - (2\pi)^{-p/2} \left| J^{(\gamma)}(\theta^*_\gamma) \right|^{p/2} \exp\left(-\frac{1}{2} t^\top J^{(\gamma)}(\theta^*_\gamma) t\right) \right| \mathrm{d}t \xrightarrow{p} 0 \quad (n \to \infty)$$

が成り立つ．ただし，$t = \sqrt{n}(\theta - \hat{\theta}^{(\gamma)}_n)$ であり漸近共分散は

$$J^{(\gamma)}(\theta^*_\gamma) = -\mathrm{E}_g\left[\nabla^2 q^{(\gamma)}(\theta; X_1)\right]$$

である．

この定理は，事後分布の漸近正規性に関する定理として知られている Bernstein-von Mises の定理の類似である．ここで注意すべきことは，頻度論における γ-divergence 最小化推定量の漸近正規分布の共分散行列は $J^{(\gamma)}(\theta^*_\gamma)^{-1} I^{(\gamma)}(\theta^*_\gamma) J^{(\gamma)}(\theta^*_\gamma)^{-1}$ であり，漸近事後分布の共分散行列を比較すればわかるように，モデルが誤特定されているときの事後信用区間は近似的にも頻度論における nominal coverage を達成しないことがわかる．そのため，実際に使う際には適切な方法で事後信用区間をカリブレーションする必要がある．事前分布として，客観事前分布の一つである probability matching prior を用いることにより，nominal coverage が近似的に達成されることが期待できそうであるが，モデルが誤特定されている場合にはそれでもうまくいくとは限らないことが Syring and Martin (2019) で指摘されている．probability matching prior に関しては，Datta and Mukerjee (2004) が詳しく，また，事後平均とベイズ推定量を高次のオーダーで matching するような客観事前分布もある (Ghosh and Liu (2011), Hashimoto (2019))．

一方，density power divergence もロバスト統計ではよく用いられる方法の一つであり，これに基づくベイズ推論は Ghosh and Basu (2016) で提案されている．彼らは，数値実験において正規分布の平均の推定問題のみを扱っているが，γ-divergence に基づく方法を用いると分散も未知の場合でも (潜在バイアスが小さいという意味で) 安定した推定を行うことができるということが Nakagawa and Hashimoto (2019) により確認されている．

また，推定量の感度を分析する際によく用いられる影響関数についても導出することができる．$T^{(\gamma)}_{n,\pi}$ を事後平均汎関数とすると n を固定したもとで，

$$\mathrm{IF}_n(y, T^{(\gamma)}_{n,\pi}, F) = n\mathrm{Cov}_{\pi^{(\gamma)}(\theta;F)}(\theta, k_\gamma(\theta; y, f))$$

が成り立つ．ただし，$\mathrm{Cov}_{\pi^{(\gamma)}(\theta;F)}$ は事後分布の下での共分散であり

$$\begin{aligned} k_\gamma(\theta; y, f) =& \frac{\partial}{\partial\varepsilon} Q^{(\gamma)}(\theta; G_\varepsilon, F_\theta)\Big|_{\varepsilon=0} \\ =& \left[f_\theta^\gamma(y) - \int f_\theta^\gamma(x) f(x) \mathrm{d}x \right] \left(\int f_\theta^{1+\gamma}(x) \mathrm{d}x \right)^{-\gamma/(1+\gamma)}. \end{aligned}$$

とする．ここで，$G_\varepsilon = (1-\varepsilon)F + \varepsilon\Lambda_y$ (Λ_y は y に退化した分布) であり，

$$Q^{(\gamma)}(\theta; G_\varepsilon, F_\theta) = \left(\int f_\theta^{1+\gamma} \mathrm{d}x \right)^{-\gamma/(1+\gamma)} \int f_\theta^\gamma(x) \mathrm{d}G_\varepsilon(x)$$

とする．従来手法との影響関数の比較と推定量のバイアスに関する数値実験は当日紹介する．

4. 縮小事前分布に基づくロバストなベイズ回帰モデル

ダイバージェンスに基づく回帰モデルの推定は，条件付き分布を同定することと同値であり，non-Bayesian の枠組みではスパースな場合も含めて様々な研究がなされている (例えば，Kawashima and Fujisawa (2017))．ここでは，ベイズ回帰モデルの枠組みで，一般的な損失関数に対する事後分布を構成し，縮小事前分布に基づく外れ値に対して頑健な回帰係数ベクトルの推定方法を提案する．

目的変数 y_i と共変量ベクトル $x_i = (x_{i1}, \ldots, x_{ip})^\top$ に対し，回帰モデル $f(y_i \mid x_i; \theta)$ を考える．例えば，連続変数 y_i に対して，$f(y_i \mid x_i; \theta) = \phi(y_i; x_i^\top\beta, \sigma^2)$ を考えると正規線形回帰モデルであり，二値変数 y_i に対して，$f(y_i = 1 \mid x_i; \theta) = \psi(x_i^\top\beta)^{y_i}\{1 - \psi(x_i^\top\beta)\}^{1-y_i}$ を考えるとロジスティック回帰モデルに対応する．ただし，$\phi(\cdot; \mu, \sigma^2)$ は正規分布 $\mathrm{N}(\mu, \sigma^2)$ の密度関数とし，$\psi(x) = \exp(x)/\{1 + \exp(x)\}$ とする．θ のベイズ推測に関して，次のような一般ベイズ更新を考えることができる：

$$\pi(\theta \mid y, x) \propto \exp\left\{ -\sum_{i=1}^n \ell(y_i, x_i, \theta) \right\} \pi(\beta, \sigma^2).$$

ただし，$L(y_i, x_i, \theta)$ は i 番目のデータと θ を結びつける損失関数であり，$\pi(\theta)$ は $\int L(y, x, \theta)\mathrm{d}G(x, y)$ を最小にするようなターゲット母数 θ の事前密度である．$L(y_i, x_i, \theta) = -\log f(y_i \mid x_i; \theta)$ とおくとこれは通常のベイズ更新に対応するが，仮定されたモデル $f(y_i \mid x_i; \theta)$ が外れ値の混入などにより誤特定されている場合には誤った推定結果を与えてしまう．そこで，前節と同様に単調変換された γ-cross entropy に基づく損失関数

$$L(y_i, x_i, \theta) = c_\gamma(\theta)^{-\gamma/(1+\gamma)} f(y_i \mid x_i; \theta)^\gamma$$

に基づいた一般事後分布を用いる．ただし，$c_\gamma(\theta) = n^{-1}\sum_{i=1}^n \int f(y \mid x_i; \theta)^{1+\gamma}\mathrm{d}y$ とする．

共変量ベクトル x_i の次元が大きい場合，y_i に影響を与える x_i の部分集合を選ぶことは重要である．ベイズ統計学の枠組みでは，事前分布を適切に定めることにより意味のある変数を選び出すことが可能である．ここでは，特に縮小事前分布 (shrinkage prior) と呼ばれる事前分布を用いた方法を考える．縮小事前分布としては，β に対する正規分布の尺度混合 (scale mixtures of normal) 事前分布を用いることが多い：

$$\pi(\beta) = \prod_{k=1}^{p} \int_0^\infty \phi(\beta_k; 0, u_k) g(u_k) \mathrm{d}u_k.$$

ただし，$g(\cdot)$ は mixing distribution とする．この形の事前分布は，様々なものが提案されており，対応する MCMC 法も多く開発されているが，一般事後分布の枠組みではそれらの方法をそのまま使うことができないところに難しさがある．$g(\cdot)$ を例えば，指数分布とすると Park and Casella (2008) による Bayesian lasso，半コーシー (half-Cauchy) 分布 とすると Carvalho et al. (2010) による horseshoe estimator に対応することに注意する．

当日は，縮小事前分布を用いたときのロバストベイズ回帰モデルに対するスケーラブルな MCMC アルゴリズムを提案し，数値実験と実データ解析を通して提案手法のパフォーマンスを示す．

参考文献

[1] Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society* B, **78**(5), 1103–1130.

[2] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

[3] Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, **115**, 293–346.

[4] Datta, G. S. and Mukarjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Springer, New York.

[5] Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, **99** (9), 2053–2081.

[6] Ghosh, A. and Basu, A. (2016). Robust Bayes estimation using the density power divergence. *Annals of Institute of Statistical Mathematics*, **68**, 413–437.

[7] Ghosh, M., Liu, R. (2011). Moment matching priors. *Sankhyā*, 73-A, 185–201.

[8] Hashimoto, S. (2019). Moment matching priors for non-regular models. *Journal of Statistical Planning and Inference*, **203**, 169–177.

[9] Kawashima, T. and Fujisawa, H. (2017). Robust and sparse regression via γ-divergence. *Entropy*, **19**, 608.

[10] Nakagawa, T. and Hashimoto, S. (2019). Robust Bayesian inference via γ-divergence. To appear in *Communications in Statistics–Theory and Methods*.

[11] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of American Statistical Association*, **103**(482), 681–686.

[12] Polson, N. G., Scott, J. G. and Windle, J. (2014). The Bayesian bridge. *Journal of the Royal Statistical Society: Series* B, **76**, 713-733.

[13] Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, **106**(2), 479–486.

[14] Zhang, T. (2006). From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, **34**(5), 2180–2210.