一般ベイズ更新に基づく統計的推論の最近の発展について

広島大学大学院理学研究科　橋本真太郎
Shintaro Hashimoto
Department of Mathematics, Hiroshima University

1　はじめに

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おける事前分布や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，その枠組みのレビューを行い，問題点や応用について考える．

2　General Bayesian updating

まず，ベイズ統計学について簡単に振り返ることにする．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

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

となる．この場合，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．ベイズ法はこの事後分布のみに基づいて統計的推論を行うという意味で一貫性のある方法であり，階層ベイズモデルのように事前分布を変えることにより柔軟なモデリングが可能となる (しかし，事前分布の選択は場合によっては重要となる)．事後分布の正規化定数である $\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta$ はエビデンスともよばれ，モデル選択等で重要な役割を果たすが，一般に高次元の積分計算を行うことになり，従来はそれが困難であることが多く，ベイズ法を用いることには限界があった．しかし，1990 年代のモンテカルロ革命以降，それらの効率的な計算が可能になってきており，近年統計学におけるベイズ法は大いに発展してきている．

さて，データ生成過程 G 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，また尤度関数に基づく上記の事後分布は何を更新したものなのか，

という疑問が生じる．例えば，モデルが誤特定されているときの最尤法では

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

をターゲットとしこれを推定することを考え，最尤推定量の漸近正規分布の共分散行列の形がサンドイッチ型になることはよく知られている (もし，モデルが正しく特定されている場合には θ^* は 真のパラメータ θ_0 に等しくなる)．Bissiri et al. (2016) では，この $-\log f(x \mid \theta)$ の代わりに一般の損失関数 $\ell(\theta, x)$ に対して，$\int \ell(\theta, x)\mathrm{d}G(x)$ を最小にするパラメータに関する合理的かつ妥当なベイズ更新を考え，これを一般ベイズ更新 (general Bayesian updating) とよび，対応する事後分布を一般事後分布 (general posterior) とよんだ．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推論に関する理論は実際にはもう少し前からあるが (例えば，Chernozhukov and Hong (2003) や Zhang (2006))，Bissiri et al. (2016) の論文を起点に様々な方面への応用が考えられており，関連する方法論も多く提案されてきている．

3 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，$\theta = \theta(P)$ $(\theta \in \Theta \subseteq \mathbb{R}^d)$ 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数 $\ell_\theta(x) := \ell(\theta, x)$ を準備する．また，推定のターゲットとなる母数を $\theta^* = \arg\min_{\theta\in\Theta} \mathrm{E}_P \ell_\theta(X)$ とする．ここで，$\mathrm{E}_P \ell_\theta(X)$ を θ の関数としてみたものをリスク関数とよび $R(\theta)$ とかき，$R(\theta)$ をデータを用いて経験推定した経験リスク関数を

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(X_i)$$

により定義する．θ の事前分布を Π とするとき，一般事後分布は次で定義される:

$$\Pi_{n,\omega}(\mathrm{d}\theta) \propto \exp\{-\omega n R_n(\theta)\}\Pi(\mathrm{d}\theta)$$

ここで，$\omega > 0$ は事前に定める scale parameter であり，learning rate とよばれ，事後分布の広がりをコントロールする役割を果たす．また，この一般事後分布は Gibbs posterior ともよばれる．損失関数は，分析の目的に応じて決めればよく，例えば $\ell_\theta(x) = |x - \theta|$ と選べば $R(\theta) = \mathrm{E}_P|X - \theta|$ の最小値としてのメディアンの推定を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，一般事後分布は M-推定とベイズ法の長所を組み合わせたものである．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経験リ

スク関数によりリンクしており，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をとることにより点推定だけではなく，分布の推定を与えるため例えば信用領域 (credible region) の形で母数の不確かさの定量化 (uncertainly quantification) を行うことができる．

しかし，モデルが誤特定されているもとでの事後信用領域の構成に関しては注意が必要である．ここで一つ簡単な例を考えよう．いま，$X_1, \ldots, X_n$ を真のデータ生成分布 $\mathrm{N}(\theta, \psi^2)$ からの無作為標本とする．モデルとして $\mathrm{N}(\theta, \sigma^2)$ を考え，$\sigma^2 \neq \psi^2$ は既知とし θ の推定問題を考える．つまり，これはモデルが誤特定されている状況である．このとき，θ に対する flat な事前分布と learnig rate ω を与えたもとで一般事後分布 $\Pi_{n,\omega}$ を構成し，事後信用区間を求めると，$\bar{X} \pm z_\alpha(\omega^{-1}\sigma^2 n^{-1})^{1/2}$ となる．ここで，z_α は $\mathrm{N}(0,1)$ の上側 α 分位点であり，$\bar{X} = n^{-1}\sum_{i=1}^n X_i$ とする．この場合，$\omega = \sigma^2/\psi^2$ とおけば頻度論における被覆確率を (理論的には) 達成することになる．

さて，もう少し一般的な状況で考えることにしよう． 事後分布 $\Pi_{n,\omega}$ は，n が十分大きいとき平均 $\hat{\theta} = \arg\min R_n(\theta)$，共分散行列 $\omega^{-1}(nV_{\hat{\theta}})^{-1}$ をもつ d 次元正規分布で近似できる事が証明されている (Chernozhukov and Hong, 2003)．ただし，V_θ は $R(\theta)$ の 2 次導関数の行列とする．一方で，M-推定量の漸近正規性により $\hat{\theta}$ は n が十分大きいとき，平均 θ^*，共分散行列 $n^{-1}V_{\hat{\theta}}^{-1}E_P[\dot{\ell}_{\hat{\theta}}\dot{\ell}_{\hat{\theta}}^\top]V_{\hat{\theta}}^{-1}$ をもつ d 次元正規分布に従うことが知られている (例えば，van der Vaart (1998))．ただし，$\dot{\ell}_\theta$ は ℓ_θ の θ に関する導関数を表す．両者の漸近共分散行列を見ればわかるように，事後信用領域は頻度論における被覆確率を近似的にも達成しないことがわかる (このことについては，Kleijn and van der Vaart (2012) でも言及されている)．なお，モデルが正しく特定されている場合は適当な正則条件のもとで Bernstein-von Mises の定理により近似的に両者は一致することを示すことができる．そこで，Syring and Martin (2019) は，この現象に着目して近似的な事後信用領域が noninal coverage probability を近似的に達成するように learning rate ω をカリブレーションする方法を提案している．また，Bissiri et al. (2016) や Holmes and Walker (2017)，Lyddon et al. (2019) では ω の選択に関する別の方法が提案されている．

4　ロバスト統計への応用

一般事後分布の考え方に基づき，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パラメータ推定を行うことができる．古典的なロバスト統計の枠組みでは，損失関数として Huber 損失等を用いた一般事後分布を考えることは可能である．さて，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，ここでは統計モデル $\{f(x \mid \theta) \mid \theta \in \Theta\}$ を一つ想定し，損失関数として $\ell_\theta(x) = \ell(x, f(x \mid \theta))$ を考えると，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を最小化するものであると考えることもできる．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，先の general Bayesian updating の枠組みに乗ることに注意する．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いる．特に，Fujisawa and Eguchi (2008) により提案された γ-divergence は理論と実装の両面で優れた点をもつダイバージェンスである．

ベイズ統計学の枠組みでの研究としては，例えば，Hooker and Vidyashankar (2014), Ghosh and Basu (2016), Jewson et al. (2018), Nakagawa and Hashimoto (2019) などで行われている．これらの文献では，learning rate ω は 1 としているが本来ならば適切な選択方法により選ばれるべきでありそこは今後の課題である．また，各ダイバージェンス自体に含まれる tuning parameter の選択方法は現在特に決め手となる選択方法はないが一般事後分布の枠組みで選択方法が提案できるか否かも興味深い話題である．

ダイバージェンスに基づく事後分布は一般に複雑な形であり共役事前分布も存在しないため，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，マルコフ連鎖モンテカルロ (MCMC) 法により事後サンプルを得るのであるが，ターゲットとなる母数の次元が大きい場合は一般に定常分布への収束が遅くなるため，効率的な MCMC 法の提案が必要である．一般事後分布に対する，次元に関してスケーラブルな MCMC 法がここ数年でいくつか提案されているがまだそこまで実用化はされてきてはおらず今後の進展が期待されている (Lyddon et al., 2018)．一方で，MCMC 法と競合する方法として，変分ベイズ法がある (例えば，Blei et al. (2017))．この方法は事後分布の近似を高速に行うことが可能であるが，いわゆる mean-field family の仮定に問題があったり，近似精度は適切にデザインされた MCMC 法に劣ることがあるため，どちらを用いるべきかは重要かつ難しい問題である．なお，変分ベイズ法により近似された事後信用領域の適切なカリブレーションは現時点では未解決問題となっている．

また，事前情報がない場合には客観事前分布が用いられることが多いが，scoring rule に基づく事後分布の文脈では Giummolè et al. (2019) により，参照事前分布 (reference prior) が導出されており，よく知られている Jeffreys の事前分布に類似したものが得られている．客観事前分布の一つとして知られている，probability matching

prior (例えば，Datta and Mukerjee (2004)) に基づく事後信用領域を用いることで近似的には nominal coverage probability を達成でき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線形回帰モデルの例において指摘されている (Syring and Martin, 2019).

5　まとめと今後の課題

本論では，非常に駆け足ではあったが，モデルが誤特定されているときに有効な一般事後分布に基づくベイズ推論の枠組みについて説明した．特に，ロバスト統計への応用に焦点を当てたが，もちろん他にも多くの研究がある．例えば， Lyddon et al. (2019) や Syring and Martin (2019) では，サポートベクターマシンや分位点回帰モデルに対する適用も行なっている．この分野の入門としては Bissiri et al. (2016) が詳しく，機械学習の分野においても多数文献が出ている．また，Miller and Dunson (2018) では，別の形の一般事後分布に基づくロバスト推定を提案しており，興味深い．

今後の課題としては，特にロバスト統計への応用では高次元のベイズ回帰モデルへの拡張や，外れ値検出の問題に一般事後分布を用いることなどが考えられ，それらは現在進行中である．また，一般事後分布を他の統計的問題に適用すること，さらに一般事後分布に基づくベイズ予測分布の解析について考えることなども課題として挙げられる．理論のみではなく，使える方法論としての発展が今後さらに期待される．

謝辞

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．また，RIMS 研究集会「高次元量子雑音の統計モデリング」でお世話になった大阪大学の田中冬彦先生，広島大学の伊森晋平先生に感謝の意を表する．

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**, 1103–1130.

[2] Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of American Statistical Association*, **112**, 859–877.

[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**, 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] Giummolè, F., Mameli, V., Ruli, E. and Ventura, L. (2019). Objective Bayesian inference with proper scoring rules. *Test*, in press.

[8] Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, **104**, 497–503.

[9] Hooker, G. and Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, **23**, 556–584.

[10] Jewson, J., Smith, J. Q. and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, **20**, 442.

[11] Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, **6**, 354–381.

[12] Lyddon, S. P., Holmes, C. C. and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, **106**, 465–478.

[13] Lyddon, S., Walker, S. and Holmes, C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *32nd Conference on Neural Information Processing Systems* (NIPS 2018).

[14] Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of American Statistical Association*, in press.

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

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

[17] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

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