変分ベイズ推定の漸近理論

渡辺 一帆*

平成 18 年 11 月 28 日

1 はじめに

変分ベイズ法は計算困難なベイズ事後分布を扱いやすい分布で近似する手法である。混合正規分布や隠れマルコフモデルなどの隠れ変数を持つ確率モデルに対して適用され、近似事後分布を求めるための繰り返しアルゴリズムが開発されて以来、音声認識、画像処理、遺伝子解析等の応用において、計算効率の良さや優れた汎化性能といった変分ベイズ法の有効性が検証されている [1][2]。

近年、変分ベイズ法自体の性質について理論的な側面から研究が行なわれ、近似精度やハイパーパラメータの影響が明らかにされてきている。本稿では変分ベイズ法の概要を述べ，混合正規分布の推定を例にとり、近年明らかにされた変分ベイズ法の理論的な性質について概説する.

2 ベイズ推定

まず始めに、ベイズ推定の概要を述べ、後の比較のために、その漸近理論について述べる。

n 個のデータ $X^n = \{X_1, \cdots, X_n\}$ が真の分布 $p_0(x)$ に従い独立に得られたとする。このとき、パラメータ θ を持つ学習モデル $p(x|\theta)$ のベイズ推定では、事前分布 $\varphi(\theta)$ を用意し、予測分布

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)d\theta$$

を用いて真の分布 $p_0(x)$ を推定する。ここで、

$$p(\theta|X^n) = \frac{1}{Z(X^n)} \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta) \tag{1}$$

はベイズ事後分布であり、$Z(X^n)$ は規格化定数 (周辺対数尤度) である。

また、

$$F(X^n) = -\log Z(X^n)$$

は確率的複雑さと呼ばれており、モデル選択やハイパーパラメータの最適化等において重要な量である。

確率的複雑さは以下のような漸近展開をもつことが明らかにされている [10]。

$$F(X^n) - S(X^n) = \lambda \log n - (m-1) \log\log n + R(X^n). \tag{2}$$

ここで、$S(X^n) = -\sum_{i=1}^{n} \log p_0(X_i)$ は経験エントロピー、$R(X^n)$ はある (有限な平均をもつ) 確率変数に法則収束する確率変数である。また、λ は有理数、m は自然数であり、学習モデル $p(x|\theta)$ によって決まる定数である。上式から、確率的複雑さの漸近的な振舞いは、これらの定数によって特徴づけられているのがわかる。

*東京工業大学大学院総合理工学研究科知能システム科学専攻, e-mail:kazuho23@pi.titech.ac.jp

統計的正則モデルについては $\lambda = d/2, m = 1$(d はモデルのパラメータ数) が成り立ち、これはモデル選択規準 BIC に対応している。混合正規分布、隠れマルコフモデル等のパラメータの特定不能であるモデルについても、これらの定数の値の評価が行なわれており、一般に定数 λ は $d/2$ よりも小さくなることが知られている [10][12]。

以上のように幾つかのことがベイズ推定の理論的な性質として明らかにされてきているが、一方で実際にベイズ推定を行なう際には、(1) 式のベイズ事後分布を実現することが困難であるという問題点がある。変分ベイズ法はベイズ推定を近似する手法であり、計算困難なベイズ事後分布を扱いやすい別の分布で代用することを行なう。

3 変分ベイズ推定

ここでは隠れ変数（潜在変数）y をもつ確率モデル $p(x, y|\theta)$ の変分ベイズ推定の概要を述べる。y は離散的な変数である場合を考える。データ $X^n = \{X_1, \cdots, X_n\}$ に対応する隠れ変数を $Y^n = \{Y_1, \cdots, Y_n\}$ とする。このとき n 個の隠れ変数 Y^n とパラメータ θ のベイズ事後分布 $p(Y^n, \theta|X^n) \propto \prod_{i=1}^n p(X_i, Y_i|\theta)\varphi(\theta)$ を別の分布 $q(Y^n, \theta)$ で近似することを考える。分布 q による近似の良さを測る尺度として汎関数 $\overline{F}[q]$ を以下のように定義する。確率的複雑さ $F(X^n)$ に対してイェンセンの不等式を用いて（$\sum_{Y^n}$ は Y^n の全ての出方についての和)、

$$\begin{aligned} F(X^n) &= -\log \sum_{Y^n} \int \prod_{i=1}^n p(X_i, Y_i|\theta)\varphi(\theta) d\theta \\ &\leq \sum_{Y^n} \int q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{\prod_{i=1}^n p(X_i, Y_i|\theta)\varphi(\theta)} d\theta \equiv \overline{F}[q]. \end{aligned} \tag{3}$$

(上で等号成立は $q(Y^n, \theta) = p(Y^n, \theta|X^n)$ のとき)$\overline{F}[q]$ は変分ベイズ法における近似の際の損失関数であり変分自由エネルギーとも呼ばれる。また、分布 q のエントロピーやカルバック情報量 $K(q(Y^n, \theta)||p(Y^n, \theta|X^n)) = \sum_{Y^n} \int q(Y^n, \theta) \log \frac{q(Y^n, \theta)}{p(Y^n, \theta|X^n)} d\theta$ を用いて、

$$\begin{aligned} \overline{F}[q] &= -\langle \log \prod_{i=1}^n p(X_i, Y_i|\theta)\varphi(\theta) \rangle_{q(Y^n, \theta)} + \langle \log q(Y^n, \theta) \rangle_{q(Y^n, \theta)} \\ &= F(X^n) + K(q(Y^n, \theta)||p(Y^n, \theta|X^n)) \end{aligned} \tag{4}$$

のように表すこともでき、$K(q||p)$ により近似精度を測っていることがわかる[1] 。

ここで、計算が可能な分布によりベイズ事後分布を近似するために、近似分布 q のクラスを、隠れ変数の分布 $Q(Y^n)$ とパラメータの分布 $r(\theta)$ を用いて、

$$q(Y^n, \theta) = Q(Y^n) r(\theta) \tag{5}$$

のように積の形で表されるのものに制限する。このクラスの中で $\overline{F}[q]$ を最小にする分布（変分事後分布と呼ばれる）を用いてベイズ事後分布を近似を行なう。このときの汎関数 $\overline{F}[q]$ の最小値を $\overline{F}(X^n)$ と表し、変分確率的複雑さと呼ぶことにする。すなわち、

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[Qr]. \tag{6}$$

さらに（5）式の制約の下で $\overline{F}[q]$ を最小にする分布 Q と r は、

$$Q(Y^n) = \frac{1}{C_Q} \exp\langle \log p(X^n, Y^n|\theta) \rangle_{r(\theta)} \tag{7}$$

$$r(\theta) = \frac{1}{C_r} \exp\langle \log p(X^n, Y^n|\theta) \rangle_{Q(Y^n)} \varphi(\theta) \tag{8}$$

$$\tag{9}$$

[1] $\langle \cdot \rangle_p$ により分布 p による平均を表す。

(C_Q と C_r は規格化定数) という関係式を満たすことが導出される。このことに基づき、多くのモデルに対し、分布 r と Q を繰り返し更新し変分事後分布を求める効率的なアルゴリズムが開発されている。またこのアルゴリズムは一種の自然勾配法であることが示されている [6]。

4　混合正規分布の変分ベイズ推定

ここでは混合正規分布を例にとり上述の変分ベイズ法のアルゴリズムを概観する。$x \in R^M$ として M 次元混合正規分布

$$p(x|\theta) = \sum_{k=1}^{K} a_k g(x-\mu_k) \tag{10}$$

の推定を考える。ここで、$\theta = \{\{a_k\}_{k=1}^K, \{\mu_k\}_{k=1}^K | 0 \le a_k \le 1, \sum_{k=1}^K a_k = 1, \mu_k \in R^M\}$, $g(x) = \frac{1}{\sqrt{2\pi}^M} \exp(-\frac{||x||^2}{2})$。データ x が k 番目のコンポーネントから生成したものであるとき $y^{(k)} = 1$ で、それ以外は $y^{(k)} = 0$ となる離散的な隠れ変数 $y = (y^{(1)}, \cdots, y^{(K)})^T$ を用いると、

$$p(x, y|\theta) = \prod_{k=1}^{K} \{a_k g(x-\mu_k)\}^{y^{(k)}} \tag{11}$$

と表すことができ、$\sum_y p(x, y|\theta) = p(x|\theta)$ が成り立つ。

パラメータの事前分布として、$\varphi(\theta) = \varphi(a)\varphi(\mu)$ ($a = (a_1, \cdots, a_K)^T$,$\mu = (\mu_1, \cdots, \mu_K)^T$ と表す)

$$\varphi(a) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \quad (\text{ディリクレ分布}) \tag{12}$$

$$\varphi(\mu) = \prod_{k=1}^{K} \sqrt{\frac{\xi_0}{2\pi}}^M \exp(-\frac{\xi_0 \|\mu_k - \nu_0\|^2}{2}), \quad (\text{正規分布}) \tag{13}$$

を仮定する。これらは共役事前分布であり、$\phi_0 > 0$, $\xi_0 > 0$, $\nu_0 \in R^M$ はハイパーパラメータである。変分ベイズ法においてはアルゴリズムが簡単になるため、たいてい共役事前分布を用いる。

以上の設定で変分ベイズ法のアルゴリズムを導出する。i 番目のデータに対応する隠れ変数を $Y_i = (Y_i^{(1)}, \cdots, Y_i^{(K)})^T$ として、

$$n_k = \sum_{i=1}^{n} \langle Y_i^{(k)} \rangle_{Q(Y^n)}, \quad \nu_k = \frac{1}{n_k} \sum_{i=1}^{n} \langle Y_i^{(k)} \rangle_{Q(Y^n)} X_i,$$

とおくと、(8) 式は、$r(\theta) = r(a)r(\mu)$

$$r(a) = \frac{\Gamma(n + K\phi_0)}{\prod_{k=1}^K \Gamma(\overline{a}_k(n + K\phi_0))} \prod_{k=1}^{K} a_k^{\overline{a}_k(n+K\phi_0)-1},$$

$$r(\mu) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\overline{\sigma}_k^2}^M} \exp(\frac{-\|\mu_k - \overline{\mu}_k\|^2}{2\overline{\sigma}_k^2}),$$

となる。ここで

$$\overline{a}_k = \frac{n_k + \phi_0}{n + K\phi_0}, \quad \overline{\sigma}_k^2 = \frac{1}{n_k + \xi_0}, \quad \overline{\mu}_k = \frac{n_k\nu_k + \xi_0\nu_0}{n_k + \xi_0}$$

とおいた。また、(7) 式は、ディガンマ関数 $\Psi(x) = \Gamma'(x)/\Gamma(x)$ を用いて、次のように与えられる。

$$Q(Y^n) = \frac{1}{C_Q} \prod_{i=1}^{n} \exp\Big[Y_i^{(k)}\big\{\Psi(n_k + \phi_0) - \Psi(n + K\phi_0) - \frac{\|x_i - \overline{\mu}_k\|^2}{2} - \frac{M}{2}(\log 2\pi + \frac{1}{n_k + \xi_0})\big\}\Big].$$

ここで分布 $r(\theta)$ によるパラメータの平均を $\overline{\theta} = \{\{\overline{a}_k\}, \{\overline{\mu}_k\}\}$ とし、変分ベイズパラメータと定義すると、分布 $r(\theta)$,$Q(Y^n)$ および汎関数 $\overline{F}[q]$ は $\overline{\theta}$ の関数として表される。$\overline{F}[q]$ を $\overline{\theta}$ の関数として $\overline{F}(\overline{\theta})$ と表して、

$$\overline{\theta}_{\rm VB} = \mathop{\rm argmin}_{\overline{\theta}} \overline{F}(\overline{\theta})$$

を変分ベイズ推定量と呼ぶことにする。変分ベイズ法のアルゴリズムは $\overline{\theta}$ に適当な初期値を与え、$\overline{\theta}$ を用いて分布 Q を更新し、Q により $\overline{\theta}$ を更新し、$\cdots$、を繰り返すことで $\overline{F}(\overline{\theta})$ の (局所) 最小解を探索する。

5　変分ベイズ法の性質

ここでは上記の混合正規分布の変分ベイズ推定における変分確率的複雑さの漸近的な振る舞いを示す定理を述べ、変分ベイズ法の性質を考察する。最後に他のモデルへの拡張等について、これまでに明らかにされたことを紹介する。

5.1　変分確率的複雑さ

(6) 式の変分確率的複雑さ $\overline{F}(X^n)$ について次の定理が成り立つ [7]。

定理 1 真の分布 $p_0(x)$ を K_0 個のコンポーネントを持つ混合正規分布とし、パラメータ θ_0 を持つとする。*(10)* 式のモデル（$K \geq K_0$ とする）の変分確率的複雑さ $\overline{F}(X^n)$ は以下の不等式を満たす。

$$\underline{\lambda} \log n + nH_n(\overline{\theta}_{\rm VB}) + C_1 \leq \overline{F}(X^n) - S(X^n) \leq \overline{\lambda} \log n + C_2, \tag{14}$$

ここで

$$\underline{\lambda} = \begin{cases} (K-1)\phi_0 + \frac{M}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}), \end{cases} \quad \overline{\lambda} = \begin{cases} (K-K_0)\phi_0 + \frac{MK_0+K_0-1}{2} & (\phi_0 \leq \frac{M+1}{2}), \\ \frac{MK+K-1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \tag{15}$$

であり、$S(X^n) = -\sum_{i=1}^n \log p(X_i|\theta_0)$ は経験エントロピー、$H_n(\overline{\theta}_{\rm VB}) = \frac{1}{n}\sum_{i=1}^n \log \frac{p(x_i|\theta_0)}{p(x_i|\overline{\theta}_{\rm VB})}$ は $\overline{\theta}_{\rm VB}$ と θ_0 の対数尤度比、C_1, C_2 は定数である。

5.2　近似精度

上記の変分確率的複雑さと真のベイズ推定における確率的複雑さを比較することで変分ベイズ法による近似（(5) 式の制限）による近似精度を評価することができる。(2) 式と (14) 式から、それぞれの主要項 ($\log n$ オーダーの項) の係数を比較すればよい。(2) 式の確率的複雑さの係数 λ は事前分布についての一定の条件の下で導かれており、多く場合、ここでのハイパーパラメータを $\phi_0 = 1$ とした場合に対応するため、(15) 式において $\phi_0 = 1$ とした場合と比較する。

(10) 式のモデルのパラメータ数は $MK + K - 1$ なので、BIC(統計的正則モデルの確率的複雑さ) の係数は

$$\lambda_{\rm BIC} = \frac{MK + K - 1}{2}$$

である。M が大きいときには $\phi_0 = 1$ を代入した (15) 式の $\overline{\lambda}$ は $\lambda_{\rm BIC}$ に比べてはるかに小さい。

また、上記の定理と同様の条件の下で (2) 式の λ に対する以下の評価が与えられている [11][12]。

$$\lambda \leq (K - K_0)/2 + (MK_0 + K_0 - 1)/2$$

これより、例えば $K_0 = 1$ のときは $\underline{\lambda} = \overline{\lambda}$ であり、

$$\overline{\lambda} - \lambda \geq (K - K_0)/2$$

が成り立つ。以上のように変分ベイズ法の近似精度が定量的に議論された。

5.3 ハイパーパラメータの影響

(15) 式の変分確率的複雑さの係数は事前分布 (12) のハイパーパラメータ ϕ_0 について $\phi_0 = \frac{M+1}{2}$ を境に２つの場合に分かれている。（ここでは省略したが）定理の導出の過程から、これは $\phi_0 = \frac{M+1}{2}$ を境に変分事後分布の収束先が大きく変わることと対応している。具体的には、$\phi_0 < \frac{M+1}{2}$ では冗長なコンポーネントの混合比が消えるようなパラメータのまわりに近似事後分布が集中するのに対し、$\phi_0 \geq \frac{M+1}{2}$ では全てのコンポーネントを用いて推定を行なうようになることを示唆している。

5.4 他のモデルへの拡張

上に述べた定理は M 次元の自然パラメータをもつ指数型分布族の混合モデルに対しても同様に成り立つことが示されている [8]。より多くの隠れ変数を持つモデルとして、隠れマルコフモデルやベイジアンネットワークの変分確率的複雑さの評価が与えられ、上記のような変分ベイズ法の性質が明らかにされた [3][9]。また一般のモデルについて、パラメータのベイズ事後分布を独立性のある分布で近似する場合の変分確率的複雑さの主要項の係数 $\lambda_{\rm VB}$ は

$$\lambda_{\rm VB} \leq \frac{d+r}{4}$$

を満たすことが示されている [5]。ここで d はモデルのパラメータ数、r はフィッシャー情報行列の対角成分のうち非零なものの最小数である。

変分ベイズ法を用いたときの汎化誤差（真の分布から予測分布までのカルバック情報量の平均）を評価することは今後の課題である。縮小ランク回帰モデルはパラメータの特定不能性をもつモデルであり、ベイズ事後分布の実現が同様に困難になる。このモデルについて、パラメータの独立性のある分布による変分ベイズ法においては、変分ベイズ推定量が漸近的にスタイン型の縮小推定量になることが示され、変分確率的複雑さ及び汎化誤差の振る舞いが明らかにされている [4]。

参考文献

[1] H.Attias, “Inferring parameters and structure of latent variable models by variational Bayes,” *Proc. of Uncertainty in Artificial Intelligence*, 21–30, 1999.

[2] Z.Ghahramani, M.J.Beal, “Graphical models and variational methods,” *Advanced Mean Field Methods*, MIT Press, 161–177, 2001.

[3] 星野力、渡辺一帆、渡辺澄夫、”隠れマルコフモデルの変分ベイズ学習における確率的複雑さについて,”電子情報通信学会論文誌 DII,J89-D, 6, 1279-1287, 2006.

[4] S.Nakajima, S.Watanabe, ”Variational Bayes solution of linear neural networks and its generalization performance,” *Neural Computation*, to appear.

[5] 西山悠、渡辺澄夫、”完全２部グラフ型ボルツマンマシンの平均場近似による確率的複雑さについて，” 電子情報通信学会論文誌 A, J89-A, 8, 671–678, 2006.

[6] M.Sato, ”Online model selection based on the variational Bayes,” *Neural Computation*, 13 (7), 1649-1681, 2001.

[7] K.Watanabe, S.Watanabe, ”Stochastic complexities of Gaussian mixtures in variational Bayesian approximation,” *Journal of Machine Learning Research*, 7, 625–644, 2006.

[8] K.Watanabe, S.Watanabe, ”Variational Bayesian stochastic complexity of mixture models,” *Advances in Neural Information Processing Systems*, 18, MIT Press, 1465-1472.

[9] K.Watanabe, M.Shiga, S.Watanabe, ”Upper bounds for variational stochastic complexities of Bayesian networks,” *Inter. Conf. on Intelligent Data Engineering and Automated Learning*, Burgos, Spain, 139-146, 2006.

[10] 渡辺澄夫、” 代数幾何と学習理論”、森北出版、2006.

[11] 渡辺澄夫、山崎啓介、青柳美輝、” 混合正規分布の特異点の非解析性について”、電子情報通信学会技術研究報告、IEICE Technical Report NC2004-50, 41–46, 2004.

[12] K.Yamazaki, S.Watanabe, ”Singularities in mixture models and upper bounds of stochastic complexity,” *International Journal of Neural Networks*, 16, 1029-1038, 2003.