ベイズ学習法の最前線
－変分ベイズ法－

上田 修功
NTTコミュニケーション科学基礎研究所
ueda@cslab.kecl.ntt.co.jp

■最尤法からベイズ法へ■

学習法の良さは，汎化能力，すなわち，未学習データに対する予測能力で評価される．換言すれば，汎化能力の高い学習法の確立が学習研究の最重要課題となっている．従来，**最尤法**と呼ばれる学習法が広く用いられていたが，近年，**ベイズ法**が有力視されつつある．ベイズ法の原型は数十年前から存在していたのに“なぜ，今，ベイズ法か？”

ベイズ法は最尤法と異り，予測値の代わりに予測分布を求める．後述するように最尤法に対して，いくつかの利点を持つが，難解な高次元の期待値計算を伴うという実用面での問題を抱えていたため，これまでお手軽な最尤法が広く用いられていたといえる．

計算機の急速な発展に伴い，期待値計算に**マルコフ連鎖モンテカルロ（MCMC）法**と呼ばれるサンプリング手法が用いられるようになり，ベイズ法もかなり実用的な手法となった．しかしながら，MCMC 法は計算量が膨大で，最尤法に比べて効率面で圧倒的なハンディを持つ．一方，近年，**変分ベイズ法**と呼ばれる効率的な手法が提案され，ベイズ法の実用化に大きく貢献する手法として注目されつつある．本稿では，この変分ベイズ法について平易に説明する．

■最尤法■

ベイズ法を理解するためには，最尤法の基礎知識が必須である．そこで，まず最尤法についておさらいしておく．モデルの複雑さの指標mとモデルパラメータθで規定されるパラメトリックな確率分布（確率モデル）のクラスを$\mathcal{H}_m=\{p(\cdot|\theta, m)\}$とする☆1．$\mathcal{H}_m$は仮説空間と呼ばれる．統計的学習とは，観測データ$D=\{d_1, d_2,..., d_n\}$に基づいて仮説空間上で真のモデルを最良近似する仮説$p(D|\theta, m)$を“探索”することといえる．

統計的学習で通常用いられる**最尤法**ではその良さの基準として**尤度**が用いられる．尤度とは，観測データDが与えられた下で確率モデルをパラメータの関数としてみたもの，すなわち，$p(D|\theta, m)$である．ゆえに，パラメータの最尤推定値$\hat{\theta}_{ML}$は$p(D|\theta, m)$を最大ならしめるθで，最尤法による最適仮説は$\mathcal{H}_m=\{p(\cdot|\hat{\theta}_{ML})\}$で与えられる．

また，最尤法におけるモデルの複雑さの決定には，通常，AIC，MDL等の**情報量基準**が用いられる．これは直観的には同じ推定値が得られるのならできるだけ簡単なモデルを選択しようという“ケチの原理”に基づく．

しかしながら，データ数が有限かつ少数という実際の応用では最尤法は学習データにオーバーフィットするという問題がある．これは**過学習**と呼ばれており，汎化能力を減少させる元凶となっている．また，ニューラルネットワークのような非線形モデルでは一般に上記情報量基準はうまく作用しない．これは専門的になるので詳細は省略するが，要は，情報量基準の導出に必要な仮定が破綻することによる．

■ベイズ法■

事後分布という考え方

上記問題の一解決法として**ベイズ法**がある．ベイズ法では，尤度に加えてパラメータθの**事前分布**$p(\theta|m)$をも考慮する．すなわち，パラメータも確率変数として取り扱う．最尤法のように1つの仮説$p(D|\hat{\theta}, m)$を求めるのではなく，未知データd_{n+1}に対し，Dが与えられた

☆1 モデルの複雑さの指標とは，たとえば混合分布モデルでは混合数，多項式回帰モデルでは多項式の次数，また，多層ニューラルネットワークの場合は中間ユニット数に相当する．

下でのθの事後分布$p(\theta|D, m)$で仮説$p(d_{n+1}|\theta, m)$を重み付き平均した**事後の予測分布**$p(d_{n+1}|D, m)$を求めd_{n+1}がどういう値をとるかを確率的に予測する．事後の予測分布は次式で計算される．

$$p(d_{n+1}|D, m)=\int p(d_{n+1}|\theta, m)p(\theta|D, m)d\theta \qquad (1)$$

ベイズ法ではモデル指標mも確率変数として扱う．すなわち，mの事前分布$P(m)$も考慮すると，式(1)は次式のように書ける．

$$p(d_{n+1}|D)=\sum_m \int p(d_{n+1}|\theta, m)p(\theta, m|D)d\theta \qquad (2)$$

ベイズ的推論の過程（図-1参照）を形式的に整理すると以下のようになる．

1. 対象となる問題に対し，データの生成過程をある確率分布$p(\cdot|\theta, m)$でモデル化する．
2. 対象に関する事前知識を事前分布$p(\theta, m)=P(m)p(\theta|m)$として表現する．
3. 実際に実験を行いデータDを観測し，モデル分布から尤度$p(D|\theta, m)$を算出する．
4. ベイズの定理より事後分布$p(\theta, m|D)$を求める．
5. 新たなデータd_{n+1}に対する予測分布を式(2)で求める．

上記4．での事後分布はベイズ法の核となる部分である．ベイズ法と最尤法との相違点を整理すると以下の2点に集約される．

- 最尤法ではパラメータを，数学的変数と見なすのに対し，ベイズ法ではすべての未知量（パラメータ，モデル指標）を**確率変数**と見なす．
- 最尤法ではモデルに最尤推定値を代入して予測値を推定する（plug-in estimateと呼ばれる）のに対し，ベイズ法では予測値ではなく**予測分布**として推定する．

すなわち，ベイズ法では事後分布による“平均化”を行っているので最尤法に比べ過学習が緩和される．事前分布と事後分布の架け橋となるのが**ベイズの定理**である．

ベイズの定理の意味するもの

データDが与えられた下でのθの分布（**事後分布**）$p(\theta|D)$はベイズの定理より次式で得られる．

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \qquad (3)$$

式(3)の右辺の分母はパラメータによらないので，$p(\theta|D)\propto p(D|\theta)p(\theta)$，すなわち，

事後分布∝尤度×事前分布

が成り立つ．さらに$-\log p$が情報量であることおよび

$$-\log p(\theta|D)\propto -\log p(D|\theta)-\log p(\theta) \qquad (4)$$

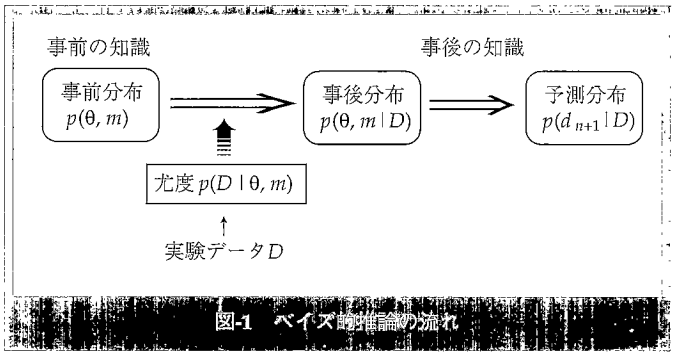


図-1 ベイズ的推論の流れ

より，

事後の情報量∝データの情報量＋事前の情報量

という直観的に自然な情報論的解釈ができる．明らかに，データ数nの増加に伴い尤度は単調に増加する．それゆえ，$n\to\infty$の時，式(4)の右辺は第2項に比べ第1項の尤度項が支配的になる．すなわち，$n\to\infty$のときはベイズ法は最尤法に縮退することが分かる．実際，データが無限個ある場合，事前情報の有無が推定に影響を及ぼさないことは直観的にも妥当である．

事前分布の設定

事前分布をどう設定するかは設計者の勘所であり，一般的に述べることは難しい．通常は，数学的な取り扱いの良さから**自然共役事前分布**（natural conjugate priors）が用いられる．自然共役事前分布とは，ある尤度に対して事前分布と事後分布が同じ分布族となる分布をいう．

最も簡単なコイン投げ（ベルヌーイ試行）の場合について説明する．表が出る確率θを持つコインをn回繰り返し投げ，その結果$D=\{d_1,\ldots, d_n\}$を得たとする．$d_i\in\{0,1\}$で，0(1)は裏（表）を表すものとする．このとき，尤度関数は$p(D|\theta)\propto\theta^{\bar{d}}(1-\theta)^{n-\bar{d}}$となる．ただし，$\bar{d}=\sum_i d_i$（表の出た回数）を表す．この問題に対する自然共役事前分布は次式のベータ分布であることが知られている．

$$p(\theta)=\mathrm{Beta}(\theta|\alpha, \beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

ここで$\Gamma(\cdot)$はガンマ関数．α, βはハイパーパラメータ（通常は定数）である．ベイズの定理より事後分布を求めると，

$$p(\theta|D)=\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\bar{d})\Gamma(\beta+n-\bar{d})}\theta^{\alpha+\bar{d}-1}(1-\theta)^{\beta+\bar{d}-1}$$
$$=\mathrm{Beta}(\theta|\alpha+\bar{d}, \beta+n-\bar{d})$$

となり，やはり事後分布もベータ分布となることが確認できる．ただし，観測データにより$\alpha\leftarrow\alpha+\bar{d}$，$\beta\leftarrow\beta+n-\bar{d}$と修正され，分布形状は観測データを反映したものに修正されている．

事前知識がほとんどない場合の事前分布は**無情報事前分布**（noninformative priors）と呼ばれる．これは，

できるだけ客観性を重視した推論を望む場合にも必要となる．具体例として，事前分布をフィッシャー情報行列の行列式の平方根に比例するように定めるジェフリーズの事前分布が著名である．

フィッシャー情報行列は学習理論業界では“お馴染み”な重要事項なので補足しておく．$D=\{d_1,...,d_n\}$の対数尤度関数を$p(D|\theta)$，ただし，パラメータはk次元ベクトル$\theta \in \mathcal{R}^k$とすると，フィッシャー情報行列Iの第(i,j)成分は次式で定義される．

$$I(\theta)_{i,j}=\mathrm{E}_D\left\{-\frac{\partial^2 \log p(D|\theta)}{\partial\theta_i \partial\theta_j}\right\} \tag{5}$$

ただし，θ_iはθの第i成分を表す．また，$E_D\{\}$はデータDの分布に関する期待値を表す．θの変化に対し対数尤度関数がどう変化するかは，$\frac{\partial}{\partial\theta_i}\log\ p(D|\theta), i=1,...,k$で評価できる．パラメータは多次元故，単一の$i$について考察するのでは不十分で，$i,j$間での変化率の相関を考慮すべく，変化率のデータ$D$に関する共分散を考える．

$$\mathrm{E}_D\left\{\frac{\partial}{\partial\theta_i}\log\ p(D|\theta)\frac{\partial}{\partial\theta_j}\log\ p(D|\theta)\right\} \tag{6}$$

式(6)は対数尤度関数がどの程度上に凸であるかの期待値を示す量と見なせる．

正則な場合，式(6)は式(5)に一致することは容易に確認できる(この証明は高校数学レベルである．各自確かめられたい)．対数尤度関数のピークが，あるθのまわりに集中していればθがその領域内にあるという意味でθに関する情報量が多いといえる．この情報論的解釈から情報行列という名前がついたものと推察される．

ベイズ法の実用上の問題

式(1)，(2)に示した事後予測分布は，高次元の多次元積分を伴うため，特殊な場合を除き解析的に求めることが困難である．これがこれまでベイズ法を実用から遠ざけていた主たる理由といえる．従来，この積分の近似法として次の2つが著名である．

- ラプラス近似法
- マルコフ連鎖モンテカルロ(MCMC)法

ラプラス近似法[1),2)]では事後分布をガウス関数近似し上記積分を解析的に求める手法である．しかしながら，この近似は$n\to\infty$を前提にしており，有限データの場合，近似の精度に問題がある．

MCMC法は，ラプラス近似に比べより正確な近似解法である．通常のモンテカルロ法との相違点は，θ空間すべてを評価するのではなく，$p(\theta|D)$を近似する有限個の$\{\theta_t\}$をマルコフ連鎖を利用したサンプリングという形式で“生成”する点にある．$\{\theta_t, t=1,...,T\}$が得られれば，たとえば式(1)は

$$p(d_{n+1}|D,m)\simeq\frac{1}{T}\sum_{t=1}^{T}p(d_{n+1}|\theta_t,m)$$

として求めることができる．サンプリングの具体的手法としてメトロポリス法，ギブスサンプリング法が著名である．これらの話題の詳細については文献2)を参照されたい．

ここで，“事後分布は式(3)のベイズの定理で求まっているのではないか？”と疑問を抱くかもしれないが，そう単純ではない．つまり，式(3)の分母はそれ自身が積分を伴うため計算困難である．すなわち，ベイズの定理では，$p(\theta|D)\propto p(D|\theta)\,p(\theta)$により規格化されていない分布(improperと呼ばれる)が得られるに過ぎないことに注意．

これらMCMC法はサンプリングに膨大な時間を要し，また，収束判定も一般には容易ではないという問題がある．近年，ラプラス近似よりも近似精度が高く，MCMC法に比べ遥かに効率的なベイズ学習法の第三のアプローチとして変分ベイズ法(Variational Bayesian method)が注目されている．この原型はWaterhouseら[3)]により提案されたが，その後，モデル指標の導入によるベイズ的モデル選択法[4)]，また，局所解の回避とモデル選択を同時実現するベイズ的モデル探索法[5)]等へと発展している．次章ではこの変分ベイズ法について説明する．

■変分ベイズ法■

基本原理

変分ベイズ法では，すべての未知量Z,θ,mを周辺化した次式のアンサンブル対数尤度(log-evidenceとも呼ばれる)を出発点とする．

$$\mathcal{L}(D)=\log\ p(D)$$
$$=\log\sum_m\sum_z\int p(D,Z,\theta,m)d\theta \tag{7}$$

ここにZは潜在変数の集合を表す．潜在変数とは，本来観測できない変数の総称をいう．たとえば混合数mの混合分布の場合，データ$\{\boldsymbol{x}_n, n=1,...,N\}$がどの混合要素から生起したかは観測できない．したがってこの場合，$Z=\{z_1,...,z_N\}$．ただし，z_nは$\boldsymbol{x}_n$に対応する潜在変数で$z_n\in\{1,...,m\}$．

すべての確率変数の結合分布$p(D,Z,\theta,m)$は

$$p(D,Z,\theta,m)=p(D,Z|m)p(\theta|m)P(m) \tag{8}$$

として算出される．式(8)の右辺第1項はモデル指標が与えられた下での完全データ尤度に，第2項はパラメータθの事前分布，そして，第3項はモデル指標の事前分布に，各々対応している．

変分ベイズ法の次のステップは，新たな分布Qの導入と，対数関数に対するJensenの不等式の適用によるア

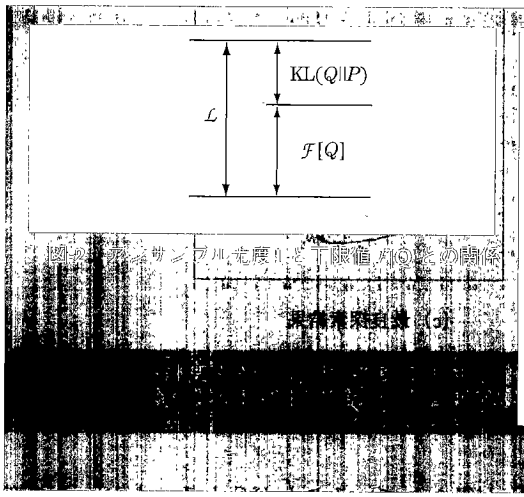


図-2 アンサンブル尤度と下限値 $\mathcal{F}[Q]$ との関係

表-1 各学習手法の比較

	確率的手法	確定的手法	
	MCMC法 (Hastings, '70)	ラプラス近似法 (MacKay, '92)	変分ベイズ法 (Waterhouse et al., '95)
特徴	$E\{f(x)\}\cong\sum_t f(x\|\Theta^{(t)})/T$ としてモンテカルロ近似	事後分布を正規分布で近似	新たな事後分布Qを導入し，期待値の下限値をQに関して最大化する
長所	Tが無限大で厳密	期待値が解析的に求まる学習が確定的	Qの分布形を仮定しない学習が確定的
短所	学習時間が膨大	近似精度が悪い	導出がやや困難
	統計の分野で考案	ニューラルネットの分野で考案	

ンサンブル尤度の下限値を求めることである．

$$\mathcal{L}(D)=\log\sum_m\sum_z\int Q(Z,\theta,m)\frac{p(D,Z,\theta,m)}{Q(Z,\theta,m)}d\theta$$
$$\geq\sum_m\sum_z\int Q(Z,\theta,m)\log\frac{p(D,Z,\theta,m)}{Q(Z,\theta,m)}d\theta$$
$$\equiv\mathcal{F}[Q] \qquad (9)$$

$\mathcal{F}[Q]$はQを変関数とする汎関数で，アンサンブル対数尤度の下限値となっている．そして$\mathcal{L}$と$\mathcal{F}$の間には次式に示す重要な関係式が成り立つ．

$$\mathcal{L}(D)=\mathcal{F}[Q]+\mathrm{KL}(Q(Z,\theta,m)\,\|\,p(Z,\theta|D)) \qquad (10)$$

式（10）の右辺第2項は，分布Qと真の事後分布$p(Z,\theta|D)$との**カルバックダイバージェンス**（KL情報量）を表す．KL情報量とは，分布間の距離で，$\mathrm{KL}(p(x)\|q(x))=\mathrm{E}_x\{\log\frac{p(x)}{q(x)}\}$ で定義される非負量で，$p\equiv q$のとき，零となる．

式（10）で$\mathcal{L}$がDのみに依存する定数であることに注意すると，図-2に示すように，下限値を最大化すべく，$\mathcal{F}[Q]$をQに関して最大化することは，Qと真の事後分布$p(\cdot|D)$とのKL情報量を最小化することと等価である．換言すれば，$\mathcal{F}$を最大化する分布Qは真の事後分布の最良の近似となっている．Qは変分事後分布（variational posterior）と呼ばれる．Qは事後分布ゆえ，本来は$Q(\cdot|D)$と書くべきであるが表記を簡単にするため，Dを省略している．

Qとして，通常，各未知変量ごとに分解した

$$Q(Z,\theta,m)=Q(Z|m)\,Q(\theta|m)\,Q(m) \qquad (11)$$

の形を仮定するが，各分布のクラスは任意とする．式（11）の制約された形で真の事後分布を推定するため一般には真の分布に一致しないが，全パラメータの同時事後分布を単一の正規分布で近似するラプラス近似法に比べればはるかに近似精度が高いといえる．なお，式（11）のような分解は統計物理の**平均場近似**の考え方に通じる．

$\mathcal{F}[Q]$の最大化は，実際には以下に示すように，$\mathcal{F}$を$Q(Z|m)$と$Q(\theta|m)$の交互の最大化を行い，収束後，$Q(m)$に関する最大化を行うという2ステップにより実現される．

[VB学習アルゴリズム]

1. 各$m\in\mathcal{M}$に対し，$Q(Z|m)^{(0)}, Q(\theta|m)^{(0)}$を設定し，$t\leftarrow 0$とし，以下を収束するまで実行．

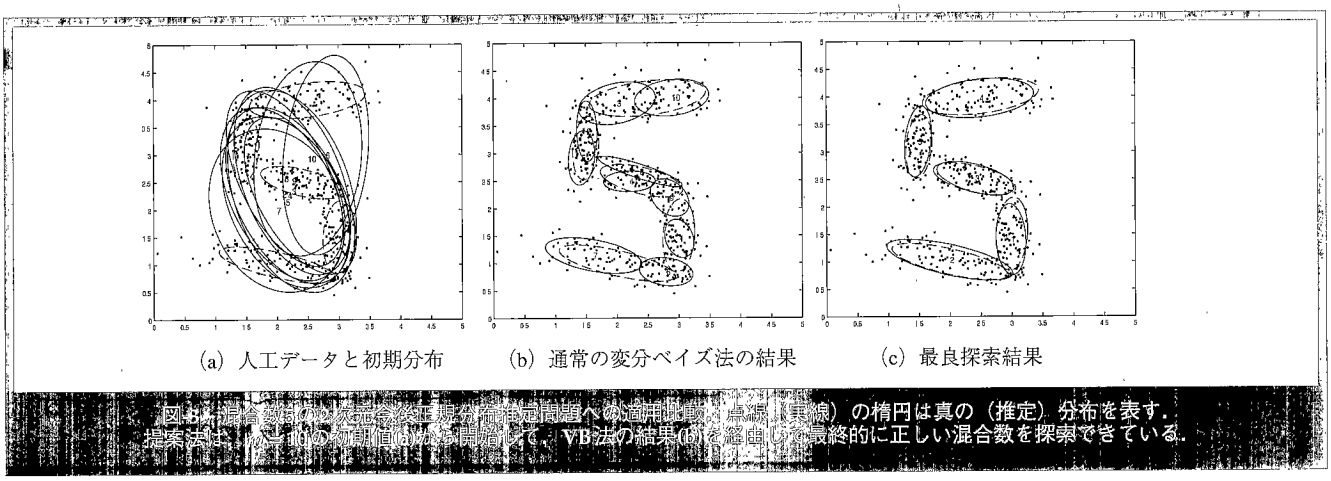
$$Q(Z|m)^{(t+1)}=\arg\max_{Q(Z|m)}\mathcal{F}[Q(Z|m),Q(\theta|m)^{(t)}]$$
$$Q(\theta|m)^{(t+1)}=\arg\max_{Q(\theta|m)}\mathcal{F}[Q(Z|m)^{(t+1)},Q(\theta|m)]$$
$$t\leftarrow t+1$$

2. 各$m\in\mathcal{M}$に対し，$\mathcal{F}$を$Q(Z)$に関し最大化する．ただし，$\mathcal{M}$は候補モデル指標集合を表す．

このように変分ベイズ法では，ベイズの定理から事後分布を求めるのではなく，$\mathcal{F}$のQに関する最大化により間接的に真の事後分布を最良近似する事後分布Qを求めるという方策をとる．その最大の利点はMCMC法のような確率的（サンプリング）手法とは異なり，上に示したように確定的なアルゴリズムが導出される点にある．これにより，MCMC法やラプラス近似に比べはるかに効率よく事後分布が推定できる．以上，ベイズ法の代表手法を表-1に整理しておく☆2．

実際の応用問題に変分ベイズ法を適用する際，$\mathcal{F}$をQで偏微分して零とおき，$Q(Z|m), Q(\theta|m)$の関数形を具体的に導出することになる．関数形が定まれば，上記アルゴリズムを実行すればよい．実はこの関数形の導出の手計算が結構面倒なのである．換言すれば，それなりの腕力がないと導出が困難といえる．その意味で一般向きではなくややマニア向けの手法かもしれない．これが表-1に記した変分ベイズ法の短所といえる．一方，MCMC法では事前分布さえ設定すればよいのでユーザの立場からいえば“お手軽”ではある．

☆2 表-1中で，ラプラス近似法がニューラルネットの分野により考案されたとしているが，正確にはラプラス近似は統計の分野で既に存在した手法である．しかしながら，多層ニューラルネットワークを対象に，より具体的な形でMacKayにより定式化されたという点を考慮して表のように分類した．



(a) 人工データと初期分布　(b) 通常の変分ベイズ法の結果　(c) 最良探索結果

図-3 混合数の同時決定を伴う正規分布推定問題への適用比較．点線（実線）の楕円は真の（推定）分布を表す．提案法は，初期の冗長値(m=10)から開始して，VB法の結果(b)を経由して最終的に正しい混合数を探索できている．

最良モデル探索への拡張

変分ベイズ法は，前述したように，汎化能力の観点で最尤法より優れているが，確定的なアルゴリズムゆえ，最尤法と同様，局所最適性の問題に悩まされる．

最尤法に関しては，混合モデルを対象に，EMアルゴリズムに要素モデルの併合分割操作を導入したSMEMアルゴリズムという局所最適性の問題解決法が提案されている[6]．併合操作とは，2つの要素モデルを新たな1つの要素モデルとして統合（併合）する操作で，分割操作とは，1つの要素モデルを新たな2つの要素モデルに分割する操作である．この考え方は変分ベイズ法にも自然に導入できる．

最尤法ではモデル指標mの推定はできないので，mを固定すべく，要素モデルの併合と分割を同時に行っていた．これに対し変分ベイズ法ではパラメータとモデルの複雑さの最適化が同時実行できる．

そこで，同時併合分割操作だけでなく，“併合操作のみ”，あるいは，“分割操作のみ”，も試みる．“併合（分割）操作のみ”は要素モデル数が1つ減る（増える）のでmを1だけ減少（増加）させることになる．したがって，これら3種類の操作を繰り返し実行することにより，低品質の局所解から脱出しながら，最良なモデル指標を同時に探索できる．すなわち，局所最適性の問題と最適なモデルの複雑さの決定の問題が同時解決可能となる[5]．ただし，この探索はあくまで“欲張り法”（greedy search）に基づいているので大域的な最適性は理論保証されないことに注意．

最良モデル探索の例

図-3 (a) の点線の楕円群に示すように，5個の混合正規分布から人工的にデータを発生し，混合数m=10としてモデル探索無しの通常の変分ベイズ法[4]とモデル探索有りの変分ベイズ法[5]を実行した．m=10と固定すると図-3 (b) の結果が得られた．この結果からさらにモデル探索を行うと，この場合，数回の要素モデルの統合操作を経て最終的に図-3 (c) に示すように最適な混合数 (m=5) が得られていることが分かる．なお，この変分ベイズ法に基づく最適モデル探索法は，混合分布推定のみならず，混合回帰モデルでの有効性も確認している．興味のある読者は文献5) を参照されたい．

■展　望■

本稿では，ベイズ法の基本的な考え方について説明し，次いで，実践的ベイズ法として近年注目されつつある変分ベイズ法についてその基本原理を概説した．変分ベイズ法は手法としての完成度は高いものの，導出が面倒ということもあり，サポートベクトルマシンのように爆発的に研究者の人気を得るには至っていない．実際，統計の分野でも，変分ベイズ法の知名度は低い．統計の分野では古くから慣れ親しんだMCMC法に根強い人気があり，他手法にはあまり興味がないようにも見える．海外の知人からの情報も含め，現在のところ変分ベイズ法の研究者は世界中でまだ十数人程度と推察される．それゆえ，実用面での有効性の評価は現在のところ十分とは言い難い．特に，MCMC法との近似精度のきちんとした比較評価は今後重要であろう．また，理論的にどの程度の近似精度が保証されるのかについても検討すべきである．

以上から分かるように，変分ベイズ法の研究はまだまだこれからといえる．興味のある読者はぜひこの分野に参入されたい．当事者としては，変分ベイズ法がサポートベクトルマシンのような賑わいを見せる日がくることを密かに期待している．

参考文献

1) MacKay, D.: Bayesian Interpolation, Neural Computation, Vol.4, pp.405-447 (1992).
2) Geleman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B.: Bayesian Data Analysis, Chapman & Hall (1995).
3) Waterhouse, S.R., MacKay, D. and Robinson, A.J.: Bayesian Methods for Mixture of Experts, Advances in Neural Information Processing Systems (NIPS8) (1995).
4) Attias, H.: Learning Parameters and Structure of Latent Variable Models by Variational Bayes, in proc. Uncertainty in Articial Intelligence (1999).
5) Ueda, N. and Ghahramani, Z.: Optimal Model Search Based on Varatinal Bayesian Approach, to appear in IEEE Neural Networks for Signal Processing (NNSP2000) (2000).
6) 上田修功, 中野良平: 混合モデルのための併合分割操作付きEMアルゴリズム, 信学論, Vol.J82-D-II, No.5, pp.930-940 (1999).

(平成12年12月5日受付)