日本応用数理学会論文誌. *vol.23, no.3, pp.453–483, 2013.*

変分ベイズ学習理論の最新動向

Recent Advances in Variational Bayesian Learning Theory

中島 伸一
ニコン

Shinichi Nakajima
Nikon Corporation

杉山 将
東京工業大学

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi

概 要

変分ベイズ学習は，行列分解モデル，混合分布モデルや隠れマルコフモデルなど，ベイズ学習の計算が困難なモデルにおける有力な近似学習手法として知られており，その良い性能が様々なアプリケーションにおいて実験的に示されてきた．実験的成功に伴って理論解析も活発に行われ，解のスパース性を誘起する相転移現象などの興味深い性質が解明されている．本論文では，変分ベイズ学習理論の最新動向を紹介する．

The variational Bayesian (VB) learning is known to be a promising approximation method to Bayesian learning for many practical models, such as matrix factorization models, mixture models, and hidden Markov models, where Bayesian learning is computationally hard. The VB learning has been empirically demonstrated to perform excellently in many applications, which stimulated theoretical analysis. Interesting properties, including phase transition phenomena that induce sparsity, have been revealed. In this paper, we review recent advances in VB learning theory.

1 はじめに

パラメトリックモデルによる統計的学習においては，未知パラメータを持つ確率モデルが観測データを説明するために利用される．未知パラメータ上の事前分布が与えられたとき，ベイズ学習によって確率の基本法則に基づいた未知パラメータ推定法が得られる．しかし，ベイズ学習は尤度関数の積分演算を含むため，行列分解や混合分布等の実用的なモデルに対しては計算が困難な場合が多い．

変分ベイズ法は，そのようなモデルに対するベイズ学習の効率的な近似法として提案された [3, 8, 9, 54]．多くのアプリケーションにおいてその良い性能が実験的に示され，それにともなって変分ベイズ法の解の性質に関する理論解析も進んでいる．確率的主成分分析 [19, 53] や縮小ランク回帰モデル [5, 42] を含む行列分解モデルに対しては，自由エネルギーの最小値を与える変分ベイズ大域解が解析的に導出されている [39]．解析解の振る舞いから，モデル起因正則化と呼ばれる意図しない正則化現象，スパース性を誘起するメカニズムやベイズ学習との違いなど，多くの興味深い性質が解明された [36]．また，変分ベイズ行列分解による主成分次元選択性能についても解析され [40]，正しい次元数を推定するための十分条件が得られている．

一方，大サンプル極限での変分ベイズ法の振る舞いを解析する漸近学習理論は，比較的多くのモデルに対して適用された．混合分布 [57, 58]，隠れマルコフモデル [18] やベイジアンネットワーク [56] に対しては，自由エネルギーの挙動の解析を通して相転移現象が発見され，事前分布が解にもたらす影響などが解明された．また，縮小ランク回帰モデルに対しては汎化誤差の漸近形が導出され，ベイズ学習で成立していた汎化誤差と自由エネルギーの漸近形に関する単純な関係が，近似的にも成立しないことなどが明らかになった [41]．

本論文では，このように近年発展が著しい変分ベイズ学習の漸近および非漸近学習理論を紹介する．2 節で変分ベイズ法の枠組みについて述べたあと，3 節および 4 節で非漸近理論および漸近理論をそれぞれ紹介する．

2 変分ベイズ学習

本節では，変分ベイズ法の枠組みを示す．

2.1 ベイズ法

観測値 $\boldsymbol{v} \in \mathbb{R}^d$ が，パラメータ $\boldsymbol{\theta} \in \mathbb{R}^K$ を持つ確率モデル $p(\boldsymbol{v}|\boldsymbol{\theta})$ に従うと仮定する．n 個の i.i.d. サンプル $\mathcal{V}^n = (\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(n)})$ が学習データとして与えられたとき，ベイズ事後分布は

$$p(\boldsymbol{\theta}|\mathcal{V}^n) = \frac{p(\mathcal{V}^n|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{V}^n)} \tag{1}$$

で与えられる．ここで，

$$p(\mathcal{V}^n|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{v}^{(i)}|\boldsymbol{\theta})$$
$$p(\mathcal{V}^n) = \int p(\mathcal{V}^n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \langle p(\mathcal{V}^n|\boldsymbol{\theta})\rangle_{p(\boldsymbol{\theta})} \tag{2}$$

であり，$p(\boldsymbol{\theta})$ はパラメータの事前分布，$\langle\cdot\rangle_p$ は確率分布 p に関する期待値を表す．式 (2) は周辺尤度，確率的複雑さ，エビデンスなどと呼ばれる量であり，学習データに対するモデルと事前分布の組の尤度と解釈できる．

式 (1) の分子はモデル分布（尤度関数）と事前分布の積であり，分母はパラメータ $\boldsymbol{\theta}$ に依存しない．従って，ベイズ事後分布の（比例定数を除いた）形状は，尤度と事前分布との積で表されることがわかる．一方，規格化因子である周辺尤度 $p(\mathcal{V}^n)$ を計算するためには分子 $p(\mathcal{V}^n|\boldsymbol{\theta})p(\boldsymbol{\theta})$ を積分する必要がある．この積分は限られた場合にしか解析的に計算することができず，またパラメータ次元 K が大きくなってくると数値的に近似計算することすら困難である．そのため，例えば行列分解，混合分布，隠れマルコフモデルやベイジアンネットなどの確率モデルを実用的なサイズで用いようとすると，学習が著しく困難になる．

2.2 変分ベイズ法

このような場合の効率的な近似法として，変分ベイズ法が提案された [3, 8, 9, 54]．

$r(\boldsymbol{\theta})$（ または r と略す）を試行分布とする．そして，次式で定義される r の汎関数を自由エネルギーと呼ぶ：

$$F(r) = \left\langle \log \frac{r(\boldsymbol{\theta})}{p(\mathcal{V}^n|\boldsymbol{\theta})p(\boldsymbol{\theta})} \right\rangle_{r(\boldsymbol{\theta})} = \left\langle \log \frac{r(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{V}^n)} \right\rangle_{r(\boldsymbol{\theta})} - \log p(\mathcal{V}^n) \tag{3}$$

式 (3) の最後の式の第 1 項は試行分布とベイズ事後分布とのカルバック擬距離 [30] であり，第 2 項は r に依存しない定数である．したがって，自由エネルギー (3) を最小化することは，カルバック擬距離の意味でベイズ事後分布に最も近い分布を見つけることに相当する．変分ベイズ法では，r に何らかの制約を課して自由エネルギーを最小化することにより，積分演算可能な分布を得る．

積分可能な分布クラスを制約として直接指定することもできるが，パラメータ間の独立性を課すだけで積分可能な分布が得られることも多い．パラメータ $\boldsymbol{\theta}$ の成分を S 個のグループ $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$ に分割することを考える．そして，すべての $s = 1, \ldots, S$ に対して，$\{\boldsymbol{\theta}_{s'}\}_{s'\neq s}$ を固定したときの尤度関数と事前分布との積

$$p(\mathcal{V}^n|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\mathcal{V}^n|\boldsymbol{\theta}_s; \{\boldsymbol{\theta}_{s'}\}_{s'\neq s})p(\boldsymbol{\theta}_s)$$

が $\boldsymbol{\theta}_s$ に関して積分可能であると仮定する．このとき事前分布 $p(\boldsymbol{\theta}_s)$ として，モデル分布を $\boldsymbol{\theta}_s$ の関数として見たときの尤度 $p(\mathcal{V}^n|\boldsymbol{\theta}_s;\{\boldsymbol{\theta}_{s'}\}_{s'\neq s})$ に対する共役事前分布を選ぶとよい．すると，事後分布にパラメータグループ間の独立性制約

$$r(\boldsymbol{\theta}) = \prod_{s=1}^{S} r_s(\boldsymbol{\theta}_s)$$

を課すことにより，変分ベイズ解

$$\widehat{r} = \mathop{\rm argmin}_{r} F(r) \qquad \text{s.t.} \qquad r(\boldsymbol{\theta}) = \prod_{s=1}^{S} r_s(\boldsymbol{\theta}_s) \tag{4}$$

は積分演算が容易な分布となる．モデル分布と事前分布との共役性により，各パラメータグループの変分ベイズ事後分布 $r_s(\boldsymbol{\theta}_s)$ は事前分布 $p(\boldsymbol{\theta}_s)$ と同じ形になる．ただし，変分ベイズ事後分布は自由エネルギーの最小化問題を通して学習データに依存することに注意する．このことを陽に表す場合には，$\widehat{r} = \widehat{r}(\boldsymbol{\theta}|\mathcal{V}^n)$ と表記する．

行列分解，混合分布，隠れマルコフモデルやベイジアンネットなどの多くのモデルは，正規分布や多項分布を組み合わせた形をしている．そのようなモデルに対しては，上記の積分可能性を満たす S 個のグループへの分割を容易に見つけることができる．

最小化問題 (4) の停留条件は変分法を用いて導出することができ，これによって繰り返しアルゴリズムが得られる．3 節および 4 節で，具体的な確率モデルに対する変分ベイズアルゴリズムを紹介する．

2.3 経験変分ベイズ法

事前分布は，その共役性の仮定により関数形が規定される．そこで，パラメータ $\boldsymbol{\eta}$ のみで定められる事前分布 $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\eta})$ を考えることにする．$\boldsymbol{\eta}$ のような 1 階層上のパラメータはハイパーパラメータと呼ばれる．多くのアプリケーションでは，スケールを含めた適切な事前分布を予め仮定することは難しい．そのような場合，パラメータ $\boldsymbol{\theta}$ と同時にハイパーパラメータ $\boldsymbol{\eta}$ も観測データから推定すれば良い．この方法は経験ベイズ法 [13] と呼ばれる．

変分ベイズ法においては，自由エネルギー (3) を事後分布とハイパーパラメータの両方に関して同時最小化することによって経験ベイズ法を実現することができる．すなわち

$$(\widehat{r}, \widehat{\boldsymbol{\eta}}) = \mathop{\rm argmin}_{r,\boldsymbol{\eta}} F(r, \boldsymbol{\eta}) \qquad \text{s.t.} \qquad r(\boldsymbol{\theta}) = \prod_{s=1}^{S} r_s(\boldsymbol{\theta}_s), \boldsymbol{\eta} \in \mathcal{S} \tag{5}$$

ここで，$\mathcal{S}$ はハイパーパラメータ $\boldsymbol{\eta}$ の定義域である．式 (5) で示される方法は経験変分ベイズ法と呼ばれる．

3 変分ベイズ学習の非漸近論

本節では，変分ベイズ法の厳密な振る舞いが解明されている行列分解モデルについて述べる．3.1 節および 3.2 節で行列分解モデルとその変分ベイズ法を導入したのち，3.3 節で理論解析結果を紹介する．

3.1 行列分解モデル

行列分解モデルでは通常，$n = 1$ 個の行列サンプル $V \in \mathbb{R}^{L\times M}$ が観測値として与えられる．観測行列 V は，低ランクの信号行列 $U \in \mathbb{R}^{L\times M}$ とノイズ行列 $\mathcal{E} \in \mathbb{R}^{L\times M}$ との和

$$V = U + \mathcal{E}$$

で表されると仮定する．行列 U を低ランクに制限するためには，積の形

$$U = BA^{\top}$$

に分解すると都合が良い．ここで，$A \in \mathbb{R}^{M\times H}$，$B \in \mathbb{R}^{L\times H}$ であり，$\top$ は行列あるいはベクトルの転置を表す．このように表現すると，行列 U のランクは高々 $H \leq \min(L, M)$ に制限される．

$\mathcal{E}$ の各成分が独立にガウス分布に従うと仮定すると，V の分布は以下のように表すことができる：

$$p(V|A, B) \propto \exp\left(-\frac{1}{2\sigma^2}\|V - BA^{\top}\|_{\mathrm{Fro}}^2\right) \tag{6}$$

ここで，σ^2 はノイズ分散であり，$\|\cdot\|_{\mathrm{Fro}}$ は行列のフロベニウスノルムを表す．任意の正則行列 $T \in \mathbb{R}^{H\times H}$ に対して

$$BA^{\top} = BT^{-1}TA^{\top} \tag{7}$$

が成立するため，このモデルは変数変換 $(A, B) \to (AT^{\top}, BT^{-1})$ に関して不変であることに注意する．

本節では，一般性を失うこと無く $L \leq M$ を仮定する．$L > M$ である場合には，$V^{\top}$ を V と取り直せば良い．行列の列ベクトルを太小文字，行ベクトルをチルダ付きの太小文字で表すことにする．すなわち

$$A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_H) = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M)^{\top} \in \mathbb{R}^{M\times H}$$
$$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_H) = \left(\widetilde{\boldsymbol{b}}_1, \ldots, \widetilde{\boldsymbol{b}}_L\right)^{\top} \in \mathbb{R}^{L\times H}$$

行列分解モデルは，行列 V の全要素が観測される場合（全観測）と，一部の要素が未観測である場合（部分観測）とに分けられる．全観測行列分解は，確率的主成分分析 [53]

および縮小ランク回帰モデル [42] を特別な場合として含み，主に多変量解析における次元削減法として用いられる．一方，部分観測行列分解は，推定された低ランク行列による欠損値予測が主目的である場合が多く，映画や書籍等の推薦システムなどに応用される [28, 16]．本節で紹介する理論解析は全観測の場合を対象としており，部分観測問題には直接適用できないことに注意する．

3.2 変分ベイズアルゴリズム

行列分解モデル (6) は指数関数の中にパラメータに関する 4 次の項を含むため積分が難しく，ベイズ学習が困難である．しかし明らかに，B を定数と考えれば A についてガウス分布であり，A を定数と考えれば B についてガウス分布である．従って，2 節で述べた手順に従って変分ベイズ法を導出できる．

まず，A および B それぞれに関する共役事前分布であるガウス事前分布を採用する：

$$p(A) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(AC_A^{-1}A^\top\right)\right) \tag{8}$$

$$p(B) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(BC_B^{-1}B^\top\right)\right) \tag{9}$$

ここで，C_A および C_B は事前分布の共分散に対応するハイパーパラメータであり，$\mathrm{tr}(\cdot)$ は行列のトレースを表す．さらに，A と B との独立性制約

$$r(A,B) = r_{\mathrm{A}}(A) r_{\mathrm{B}}(B) \tag{10}$$

を事後分布に課すと，尤度と事前分布との共役性によって事後分布がガウス分布となる [7, 32]．

変分法を用いて自由エネルギー最小化問題 (4) を解析すると，事後分布が

$$r(A,B) = \prod_{m=1}^{M} \mathcal{N}_H(\widetilde{\boldsymbol{a}}_m; \widetilde{\widehat{\boldsymbol{a}}}_m, \Sigma_A) \prod_{l=1}^{L} \mathcal{N}_H(\widetilde{\boldsymbol{b}}_l; \widetilde{\widehat{\boldsymbol{b}}}_l, \Sigma_B) \tag{11}$$

の形で表され，また，ガウス分布の平均と共分散行列は以下の連立方程式を満たすことがわかる [7, 32]：

$$\widehat{A} = \left(\widetilde{\widehat{\boldsymbol{a}}}_1, \ldots, \widetilde{\widehat{\boldsymbol{a}}}_M\right)^\top = V^\top \widehat{B} \frac{\Sigma_A}{\sigma^2} \tag{12}$$

$$\widehat{B} = \left(\widetilde{\widehat{\boldsymbol{b}}}_1, \ldots, \widetilde{\widehat{\boldsymbol{b}}}_L\right)^\top = V \widehat{A} \frac{\Sigma_B}{\sigma^2} \tag{13}$$

$$\Sigma_A = \sigma^2 \left(\widehat{B}^\top \widehat{B} + L\Sigma_B + \sigma^2 C_A^{-1}\right)^{-1} \tag{14}$$

$$\Sigma_B = \sigma^2 \left(\widehat{A}^\top \widehat{A} + M\Sigma_A + \sigma^2 C_B^{-1}\right)^{-1} \tag{15}$$

ここで，$\mathcal{N}_d(\cdot;\boldsymbol{\mu},\Sigma)$ は平均が $\boldsymbol{\mu}$，共分散行列が Σ の d 次元ガウス分布である．

式 (11) は一般のガウス分布ではなく，A（および B）の行ベクトル $\{\widetilde{\boldsymbol{a}}_m\}$ （$\{\widetilde{\boldsymbol{b}}_l\}$）が互いに独立であり，共通の共分散行列 Σ_A （Σ_B）を持つような特別なガウス分布であることに注意する [7].

式 (12) は，変数 $(\widehat{B},\Sigma_A,\Sigma_B)$ を固定したときに，$\widehat{A}$ について自由エネルギーを最小化する解となっている．式 (13)–(15) も同様に，右辺に現れる変数を固定したときの左辺に関する自由エネルギー最小解となっている．式 (12)–(15) を繰り返すことによって，最小化問題 (4) の局所解が得られることが知られている．このように，変数をひとつずつ最適化するアルゴリズムは ICM(iterated conditional modes) アルゴリズム [6, 8, 9] と呼ばれる．

以下の議論では，事前分布の共分散行列 C_A および C_B は正定値対角であると仮定する．すなわち

$$C_A = \mathrm{diag}(c_{a_1}^2,\ldots,c_{a_H}^2)$$
$$C_B = \mathrm{diag}(c_{b_1}^2,\ldots,c_{b_H}^2)$$

また，積 C_AC_B の対角成分が非増加順に並んでいることも仮定する．すなわち，すべてのペア $h<h'$ に対して

$$c_{a_h}c_{b_h} \geq c_{a_{h'}}c_{b_{h'}}$$

任意の C_A および C_B に対してこのような並び替えが可能であるので，この仮定は一般性に影響しない．

経験変分ベイズ法では，ハイパーパラメータ (C_A,C_B) も観測値から推定するために，C_A および C_B に関しても最小化問題 (5) を解く．ハイパーパラメータに関する停留条件は，次式で与えられる：

$$c_{a_h}^2 = \|\widehat{\boldsymbol{a}}_h\|^2/M + (\Sigma_A)_{hh} \tag{16}$$
$$c_{b_h}^2 = \|\widehat{\boldsymbol{b}}_h\|^2/L + (\Sigma_B)_{hh} \tag{17}$$

実際の応用問題ではノイズ分散 σ^2 も未知である場合が多いが，自由エネルギー最小化原理を用いれば σ^2 も観測値から推定することができる．σ^2 に関する停留条件は，次式で与えられる：

$$\sigma^2 = \frac{\|V\|_{\mathrm{Fro}}^2 - \mathrm{tr}(2V^\top\widehat{B}\widehat{A}^\top) + \mathrm{tr}\left((\widehat{A}^\top\widehat{A}+M\Sigma_A)(\widehat{B}^\top\widehat{B}+L\Sigma_B)\right)}{LM} \tag{18}$$

ハイパーパラメータやノイズ分散が未知である場合，式 (12)–(18) を繰り返すことによってすべての未知変数を推定することができる．

3.3 理論解析結果

全観測行列分解モデルに対しては変分ベイズ法の多くの性質が明らかにされており，特に自由エネルギー最小化問題 (4) の大域解析解が得られることが知られている [39].

3.3.1 変分ベイズ行列分解の大域解析解

成分がすべて正である d 次元ベクトルの集合を $\mathbb{R}^d_{++}$ で，$d \times d$ 正定値対称行列の集合を $\mathbb{S}^d_{++}$ でそれぞれ表す．行列分解の変分ベイズ解は，以下の最適化問題を解くことによって得られる：

$$\begin{aligned} &\text{Given} \quad (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}^2_{++} \ (\forall h = 1, \ldots, H), \ \sigma^2 \in \mathbb{R}_{++} \\ &\quad \min \ \ F(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B) \quad \text{s.t.} \quad \widehat{A} \in \mathbb{R}^{M \times H}, \ \widehat{B} \in \mathbb{R}^{L \times H}, \ \Sigma_A \in \mathbb{S}^H_{++}, \ \Sigma_B \in \mathbb{S}^H_{++} \end{aligned} \tag{19}$$

ただし，$F(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B) = F(r)$ は自由エネルギーであり，

$$\begin{aligned} F(r) &= \langle \log r_{\mathrm{A}}(A) + \log r_{\mathrm{B}}(B) - \log p(V|A,B)p(A)p(B) \rangle_{r_{\mathrm{A}}(A) r_{\mathrm{B}}(B)} \\ &= \frac{\|V\|^2_{\mathrm{Fro}}}{2\sigma^2} + \frac{LM}{2} \log \sigma^2 + \frac{M}{2} \log \frac{|C_A|}{|\Sigma_A|} + \frac{L}{2} \log \frac{|C_B|}{|\Sigma_B|} \\ &\quad + \frac{1}{2} \mathrm{tr} \left\{ C_A^{-1} \left(\widehat{A}^\top \widehat{A} + M\Sigma_A \right) + C_B^{-1} \left(\widehat{B}^\top \widehat{B} + L\Sigma_B \right) \right. \\ &\qquad \left. + \sigma^{-2} \left(-2\widehat{A}^\top V^\top \widehat{B} + \left(\widehat{A}^\top \widehat{A} + M\Sigma_A \right) \left(\widehat{B}^\top \widehat{B} + L\Sigma_B \right) \right) \right\} + \mathrm{const.} \end{aligned} \tag{20}$$

で与えられる．ここで，$|\cdot|$ は行列式を表す．式 (12)–(15) は自由エネルギー (20) の変数 $(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B)$ に関する停留（必要十分）条件になっていることに注意する．

最適化問題 (19) は非凸最適化問題であり，一般の凸解法では効率的に解くことはできない．しかし以下で示すように，$O(MH)$ 個の変数を含む最小化問題 (19) は，$O(1)$ 個の変数の最小化問題に分解できる．これにより，(19) は非凸最適化問題であるにもかかわらず，大域解析解を得ることができる．

定理 1 (Nakajima et al. (2013) [39])**.** 共分散行列 (Σ_A, Σ_B) が対角である解を対角解と呼ぶ．最小化問題 (19) のすべての解は対角解であるか，あるいは冗長性 (7) を通して対角解と等価な解である． ♢

この定理は，大域最適解が停留点であることを示した後，自由エネルギー (20) の最適解まわりの摂動を調べることによって証明できる．

(Σ_A, Σ_B) が対角であるならば，変分ベイズ事後分布 (11) は (A, B) のすべての要素が独立なガウス分布となる．実はこの解は，単純変分ベイズ法 [22] の解と一致することが知られている．単純変分ベイズ法とは，A および B の各列ベクトルの独立性

$$r^{\mathrm{VB}}(A, B) = \prod_{h=1}^{H} r_{a_h}^{\mathrm{VB}}(\boldsymbol{a}_h) \prod_{h=1}^{H} r_{b_h}^{\mathrm{VB}}(\boldsymbol{b}_h) \tag{21}$$

を課して自由エネルギーを最小化する方法である．単純変分ベイズ法では事後分布の共分散行列の非対角成分を考慮する必要がないため，メモリ量および計算量を大幅に節約できる．列ベクトルごとの独立性制約 (21) は行列間独立性制約 (10) よりも強い制約であるが，

定理 1 はこの強い制約が変分ベイズ解には影響を与えないことを示している．ただし，この定理は全観測行列分解に対して導かれたものであり，部分観測行列分解に対しては一般には成立しない．

制約 (21) のもとで，変分ベイズ解 $\widehat{U}^{\rm VB} = \widehat{B}\widehat{A}^{\top}$ が縮小特異値分解となることを示すことができる．定理 1 により，これが制約 (10) においても成り立つことがわかる．

補題 1 (Nakajima and Sugiyama (2011) [36])**.** 観測行列 V の h 番目に大きい特異値およびその右左特異ベクトルを $(\gamma_h, \boldsymbol{\omega}_{a_h}, \boldsymbol{\omega}_{b_h})$ で表す．すなわち，

$$V = \sum_{h=1}^{H} \gamma_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top}$$

変分ベイズ解は，$4 \times H$ 個のスカラー変数 $\{a_h, b_h, \sigma_{a_h}^2, \sigma_{b_h}^2\}_{h=1}^{H}$ を用いて以下の形で表現することができる．

$$\begin{aligned}
\boldsymbol{a}_h &= a_h \boldsymbol{\omega}_{a_h} \\
\boldsymbol{b}_h &= b_h \boldsymbol{\omega}_{b_h} \\
\Sigma_A &= diag(\sigma_{a_1}^2, \ldots, \sigma_{a_H}^2) \\
\Sigma_B &= diag(\sigma_{b_1}^2, \ldots, \sigma_{b_H}^2)
\end{aligned}$$

$\diamondsuit$

補題 1 の表現を自由エネルギー (20) および停留条件 (12)–(15) に代入すると，次の補題が得られる：

補題 2. 変分ベイズ解は，以下の *4* 変数最小化問題を $h = 1, \ldots, H$ に対してそれぞれ解くことによって得られる．

$$\begin{aligned}
&Given\ (c_{a_h}^2, c_{b_h}^2) \in \mathbb{R}_{++}^2, \sigma^2 \in \mathbb{R}_{++} \\
&\quad \min \quad F_h(a_h, b_h, \sigma_{a_h}^2, \sigma_{b_h}^2) \quad s.t. \quad (a_h, b_h) \in \mathbb{R}^2, (\sigma_{a_h}^2, \sigma_{b_h}^2) \in \mathbb{R}_{++}^2
\end{aligned} \tag{22}$$

ここで

$$\begin{aligned}
F_h(a_h, b_h, \sigma_{a_h}^2, \sigma_{b_h}^2) = &-M \log \sigma_{a_h}^2 + \frac{a_h^2 + M\sigma_{a_h}^2}{c_{a_h}^2} - L \log \sigma_{b_h}^2 + \frac{b_h^2 + L\sigma_{b_h}^2}{c_{b_h}^2} \\
&- \frac{2}{\sigma^2} \gamma_h a_h b_h + \frac{1}{\sigma^2} \left(a_h^2 + M\sigma_{a_h}^2\right) \left(b_h^2 + L\sigma_{b_h}^2\right)
\end{aligned} \tag{23}$$

であり，その停留条件は次式で与えられる．

$$a_h = \frac{1}{\sigma^2}\sigma_{a_h}^2 \gamma_h b_h \tag{24}$$

$$b_h = \frac{1}{\sigma^2}\sigma_{b_h}^2 \gamma_h a_h \tag{25}$$

$$\sigma_{a_h}^2 = \sigma^2 \left(b_h^2 + L\sigma_{b_h}^2 + \frac{\sigma^2}{c_{a_h}^2} \right)^{-1} \tag{26}$$

$$\sigma_{b_h}^2 = \sigma^2 \left(a_h^2 + M\sigma_{a_h}^2 + \frac{\sigma^2}{c_{b_h}^2} \right)^{-1} \tag{27}$$

$\diamondsuit$

こうして補題 2 によって，$O(ML)$ 個の変数に関する自由エネルギー最小化問題 (19) を，H 個の 4 変数問題 (22) に分解することができた．

連立方程式 (24)–(27) は解析的に解くことができるため，結果として変分ベイズ解が解析的に得られる．

定理 2 (Nakajima et al. (2013) [39]). *$\widehat{\gamma}_h$ に関する 4 次方程式*

$$\widehat{\gamma}_h^4 + \xi_3 \widehat{\gamma}_h^3 + \xi_2 \widehat{\gamma}_h^2 + \xi_1 \widehat{\gamma}_h + \xi_0 = 0 \tag{28}$$

の 2 番目に大きい正の実解を $\widehat{\gamma}_h^{\mathrm{second}}$ とする．ただし，係数は

$$\xi_3 = \frac{(L-M)^2 \gamma_h}{LM}$$

$$\xi_2 = -\left(\xi_3 \gamma_h + \frac{(L^2+M^2)\eta_h^2}{LM} + \frac{2\sigma^4}{c_{a_h}^2 c_{b_h}^2} \right)$$

$$\xi_1 = \xi_3 \sqrt{\xi_0}$$

$$\xi_0 = \left(\eta_h^2 - \frac{\sigma^4}{c_{a_h}^2 c_{b_h}^2} \right)^2$$

$$\eta_h^2 = \left(1 - \frac{\sigma^2 L}{\gamma_h^2} \right) \left(1 - \frac{\sigma^2 M}{\gamma_h^2} \right) \gamma_h^2$$

で与えられる．このとき，変分ベイズ行列分解の大域解は次式で与えられる：

$$\widehat{U}^{\mathrm{VB}} \equiv \langle BA^\top \rangle_{r(A,B)} = \widehat{B}\widehat{A}^\top = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^\top$$

$$\widehat{\gamma}_h^{\mathrm{VB}} = \begin{cases} \widehat{\gamma}_h^{\mathrm{second}} & \mathit{if}\ \gamma_h > \widetilde{\gamma}_h \\ 0 & \mathit{otherwise} \end{cases}$$

$$\widetilde{\gamma}_h = \sqrt{\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2} + \sqrt{\left(\frac{(L+M)\sigma^2}{2} + \frac{\sigma^4}{2c_{a_h}^2 c_{b_h}^2} \right)^2 - LM\sigma^4}}$$

♢

4 次方程式の解はフェラーリ法 [17] などを用いて解析的に求めることができるため，定理 2 によって行列分解の変分ベイズ解を解析的に求めることができる．ただし実際に変分ベイズ行列分解を実装する際には，例えば MATLAB® の ‘roots’ コマンドなどを用いて数値的に解いても問題はない．なお，事後分布の分散 $(\sigma^2_{a_h}, \sigma^2_{b_h})$ も解析的に得られるため，変分ベイズ事後分布を明示的に描画することも可能である [39].

3.3.2 経験変分ベイズ行列分解の大域解析解

経験変分ベイズ解は，以下の最適化問題を解くことによって得られる：

$$\begin{aligned} \text{Given} \quad & \sigma^2 \in \mathbb{R}_{++} \\ \min \quad & F(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B, \{c^2_{a_h}, c^2_{b_h}; h = 1, \ldots, H\}) \\ \text{s.t.} \quad & \widehat{A} \in \mathbb{R}^{M \times H},\ \widehat{B} \in \mathbb{R}^{L \times H},\ \Sigma_A \in \mathbb{S}^H_{++},\ \Sigma_B \in \mathbb{S}^H_{++}, \\ & (c^2_{a_h}, c^2_{b_h}) \in \mathbb{R}^2_{++}\ (\forall h = 1, \ldots, H) \end{aligned}$$

ただし，$F(\widehat{A}, \widehat{B}, \Sigma_A, \Sigma_B, \{c^2_{a_h}, c^2_{b_h}; h = 1, \ldots, H\})$ は式 (20) で与えられる自由エネルギーである．式 (20) は (A, B) 間の相対スケール変換に関して不変であるため，ハイパーパラメータの比 c_{a_h}/c_{b_h} は不定である [36]．そこで，一般性を失うことなく $c_{a_h}/c_{b_h} = 1$ と仮定することにする．

3.3.1 節と同様の分解法を適用すれば，変分ベイズ法の停留条件 (24)–(27) に (16) および (17) を加えたものを解いて得られる停留点上で自由エネルギーの値を評価することによって，経験変分ベイズ解を得ることができる．

定理 3 (Nakajima et al. (2013) [39])**.** 行列分解モデルの経験変分ベイズ解は以下で与えられる：

$$\begin{aligned} \widehat{U}^{\mathrm{EVB}} &= \sum_{h=1}^{H} \widehat{\gamma}^{\mathrm{EVB}}_h \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}^\top_{a_h} \\ \widehat{\gamma}^{\mathrm{EVB}}_h &= \begin{cases} \breve{\gamma}^{\mathrm{VB}}_h & \mathit{if}\ \gamma_h > \underline{\gamma}_h\ \mathit{and}\ \Delta_h \leq 0 \\ 0 & \mathit{otherwise} \end{cases} \\ \underline{\gamma}_h &= (\sqrt{L} + \sqrt{M})\sigma \\ \breve{c}^2_{a_h} \breve{c}^2_{b_h} &= \frac{1}{2LM} \left(\gamma_h^2 - (L+M)\sigma^2 + \sqrt{(\gamma_h^2 - (L+M)\sigma^2)^2 - 4LM\sigma^4} \right) \\ \Delta_h &= M \log \left(\frac{\gamma_h}{M\sigma^2} \breve{\gamma}^{\mathrm{VB}}_h + 1 \right) + L \log \left(\frac{\gamma_h}{L\sigma^2} \breve{\gamma}^{\mathrm{VB}}_h + 1 \right) + \frac{1}{\sigma^2} \left(-2\gamma_h \breve{\gamma}^{\mathrm{VB}}_h + LM \breve{c}^2_{a_h} \breve{c}^2_{b_h} \right) \end{aligned}$$

ただし，$\breve{\gamma}^{\mathrm{VB}}_h$ は $c_{a_h} c_{b_h} = \breve{c}_{a_h} \breve{c}_{b_h}$ が与えられたときの変分ベイズ解である． ♢

定理 2 及び定理 3 を用いると，(12)–(15) あるいは (12)–(17) を繰り返し解く ICM アルゴリズムよりも高速かつ確実に変分ベイズ解が得られるため，これらの定理は実用上非常に有用である．ノイズ分散 σ^2 が未知の場合には，これらの定理を用いて σ^2 以外のパラメータの解析解を得ながら，σ^2 に関する 1 次元の最適化を行えば良い [39].

3.3.3 モデル起因正則化

定理 2 および定理 3 では，複雑な 4 次方程式を解くことによって変分ベイズ解を得るため，必ずしも直感的な解釈がしやすいとはいえない．そこで以下では，解が非常に簡単な形で表現できる 2 つの場合を考え，変分ベイズ法の振る舞いについてより詳細に議論することにする．

事前分布の分散を無限に大きくとる（$c_{a_h}c_{b_h} \to \infty$）と，事前分布は殆ど平坦になる．このとき，4 次方程式 (28) は以下のように表現することができる．

$$\lim_{c_{a_h}c_{b_h}\to\infty} f(\widehat{\gamma}_h) = \left(\widehat{\gamma}_h + \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\gamma_h\right)\left(\widehat{\gamma}_h + \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h\right) \cdot \left(\widehat{\gamma}_h - \left(1-\frac{\sigma^2 M}{\gamma_h^2}\right)\gamma_h\right)\left(\widehat{\gamma}_h - \frac{M}{L}\left(1-\frac{\sigma^2 L}{\gamma_h^2}\right)\gamma_h\right) = 0$$

定理 2 によれば，$\gamma_h > \lim_{c_{a_h}c_{b_h}\to\infty} \widetilde{\gamma}_h = \sqrt{M\sigma^2}$ が成立するとき，4 次方程式の 2 番目に大きい解が変分ベイズ解となる．このことから次の系が得られる．

系 1 (Nakajima et al. (2013) [39])**.** 平坦事前分布 *(* $c_{a_h}c_{b_h} \to \infty$*)* に対する変分ベイズ解は

$$\lim_{c_{a_h}c_{b_h}\to\infty} \widehat{\gamma}_h^{\mathrm{VB}} = \widehat{\gamma}_h^{\mathrm{PJS}} = \max\left\{0, \left(1 - \frac{M\sigma^2}{\gamma_h^2}\right)\gamma_h\right\} \tag{29}$$

で与えられる． ♢

系 1 より，変分ベイズ解の各特異値は *positive-part James-Stein* (PJS) 推定量 [24, 29] の形で縮小されることがわかる．平坦事前分布を用いているにもかかわらずこのような強い正則化がかかることは一見すると直感に反するかも知れないが，フィッシャー計量の体積要素（すわなちジェフリーズ事前分布 [25]）が顕著に不均一であることを考えれば，自然な結果である．この正則化は事前分布ではなく確率モデルの構造に起因するため，モデル起因正則化と呼ばれる [36].

$L = M$ の場合にも，4 次方程式 (28) を因数分解することによって解が単純な形で得られる．$\gamma_h > \sqrt{M\sigma^2}$ の場合，

$$f^{\mathrm{square}}(\widehat{\gamma}_h) = \left(\widehat{\gamma}_h + \widehat{\gamma}_h^{\mathrm{PJS}} + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\left(\widehat{\gamma}_h + \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right) \cdot \left(\widehat{\gamma}_h - \widehat{\gamma}_h^{\mathrm{PJS}} + \frac{\sigma^2}{c_{a_h}c_{b_h}}\right)\left(\widehat{\gamma}_h - \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h}c_{b_h}}\right) = 0$$

の 2 番目に大きい解が大域解であることを利用すれば，以下の系が得られる．

系 2 (Nakajima et al. (2013) [39]). $L = M$ のとき，変分ベイズ大域解は

$$\widehat{\gamma}_h^{\mathrm{VB-square}} = \max\left\{0, \widehat{\gamma}_h^{\mathrm{PJS}} - \frac{\sigma^2}{c_{a_h} c_{b_h}}\right\} \tag{30}$$

で与えられる． ♢

式 (30) から，正方行列の場合にはモデル起因正則化（$\widehat{\gamma}_h^{\mathrm{PJS}}$）と事前分布に起因する正則化（$-\sigma^2/(c_{a_h} c_{b_h})$）とが分離できることがわかる．

実は，行列分解モデル (6)，(8) および (9) に対する MAP 推定量は以下で与えられることが知られている [36]：

$$\widehat{\gamma}_h^{\mathrm{MAP}} = \max\left\{0, \gamma_h - \frac{\sigma^2}{c_{a_h} c_{b_h}}\right\} \tag{31}$$

$\gamma_h^{\mathrm{PJS}} < \gamma_h$ であるので，変分ベイズ解 (30) は MAP 解 (31) によって上からバウンドされることがわかる．なお，MAP 解 (31) はトレースノルム正則化によるノンベイズなスパース推定

$$\min_U \|V - U\|_{\mathrm{Fro}}^2 + \lambda \|U\|_{\mathrm{tr}}$$

において，$\lambda = \frac{2\sigma^2}{c_{a_h} c_{b_h}}$ としたときの解に一致することも知られている [50, 10]．

変分ベイズ法を必要とするモデルの殆どは，確率分布とパラメータとが 1 対 1 対応しない特異モデルに属し（4.1.3 節参照），そこでは一般に，フィッシャー計量の不均一性によって起こるモデル起因正則化が顕著に現れる．モデル起因正則化は，ユーザーの意図と無関係に起こるという意味でモデリングによるアーティファクトであると捉えることもできるが，これをジェフリーズ事前分布を用いて抑制することは，以下に述べる二つの理由によって推奨されない．第 1 に，特異モデルのジェフリーズ事前分布の多くが，無限遠で発散するような規格化不可能（improper）な分布であり，近似的にもベイズ学習を行うことは困難である．第 2 に，モデル起因正則化は適切な正則化やモデル選択に貢献する場合が多い．たとえば平坦事前分布における行列分解の変分ベイズ解は PJS 推定量に一致する（系 1）が，この推定量は次元とノイズとのバランスをとる優れた推定量であることが知られている [29]．また，3.3.5 節で議論するように，変分ベイズ主成分分析のモデル起因正則化による次元選択が，ある条件下で非常に良い性能を発揮することが理論的にも証明されている．

3.3.4 変分ベイズ法の相転移現象

系 1 に見られるように，変分ベイズ法は小さい特異値成分を無視してスパースな解を出力する．これはモデルの「枝狩り」機能として作用し，変分ベイズ法の便利な特徴のひと

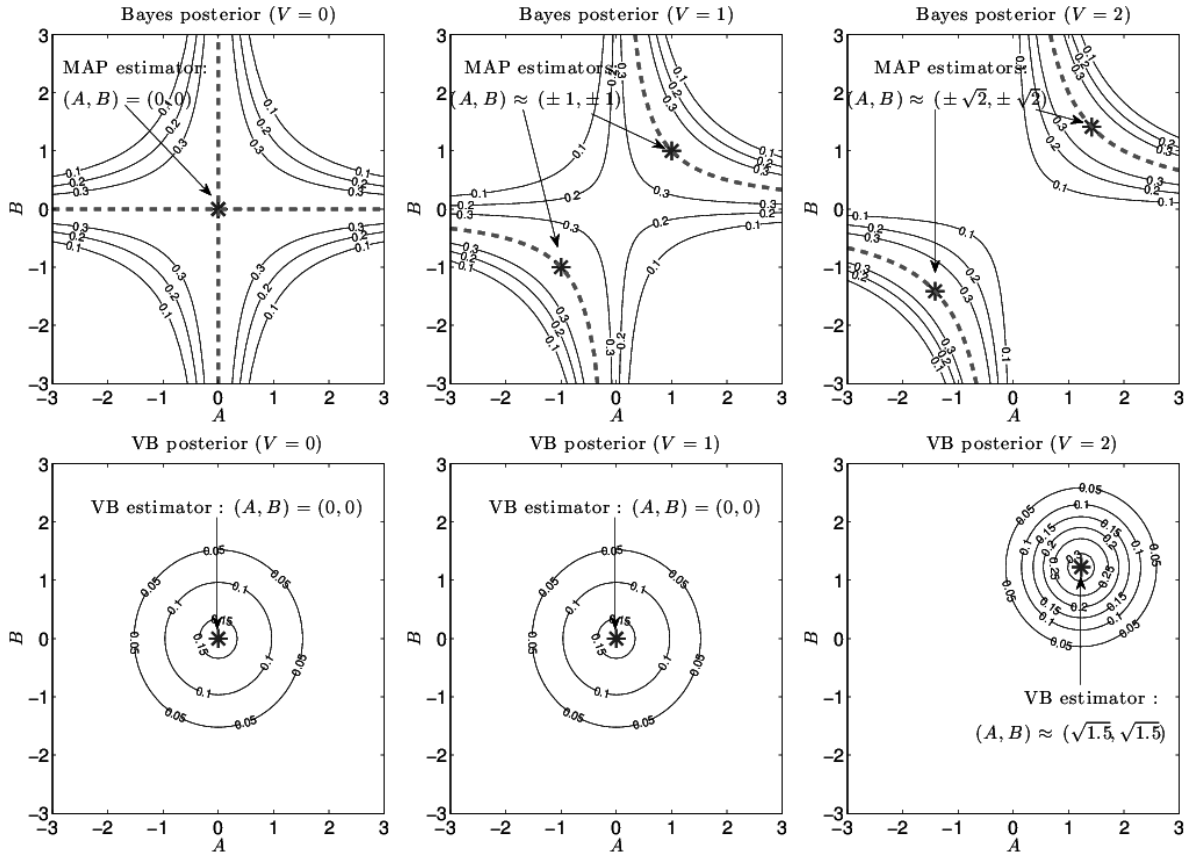

図 1: Bayes posteriors (top row) and the VB posteriors (bottom row) of a *scalar factorization* model (i.e., a MF model for $L = M = H = 1$) with $\sigma^2 = 1$ and $c_a = c_b = 100$ (almost flat priors), when the observed values are $V = 0$ (left), $V = 1$ (middle), and $V = 2$ (right), respectively. In the top row, the asterisks indicate the MAP estimators, and the dashed lines the ML estimators (the modes of the contour). In the bottom row, the asterisks indicate the VB estimators.

つとみなされている．ところが，実は厳密なベイズ学習にはこの枝狩り機能がないことが知られている [36].

枝狩りは，自由エネルギー最小化問題の相転移現象に起因する．混合分布モデルにおいては，対称性が自発的に破れたときにモデルが枝狩りされることが報告されている [33]. 一方，行列分解モデルにおいては対称性が破れない場合に枝狩りが起こる．Fig. 1 に，$L = M = H = 1$ の場合のベイズ事後分布（上段）と変分ベイズ事後分布（下段）を示す．ここでは単位ノイズ分散 $\sigma^2 = 1$ およびほぼ平坦な事前分布 $c_a = c_b = 100$ を仮定している．

Fig. 1 からわかるように，ベイズ事後分布は $V = 0$ の場合を除いて 2 つのピークを持ち，ピーク間距離は観測の絶対値 $|V|$ に従って増大する．一方，変分ベイズ事後分布はベイズ事後分布を (A, B) 間の独立性を保持しながら近似するため，$|V|$ が十分大きくなるま

で（$V \leq 1$）は原点から離れられない．$|V|$ が十分大きくなると（$V = 2$），対称性が自発的に破れて 2 つのピークのうちのいずれかを近似するように原点から移動する．ここで，$(A, B) \approx (-\sqrt{1.5}, -\sqrt{1.5})$ も等価な解であることに注意する．

式 (29) より，対称性の破れは $V > \widetilde{\gamma}_h \sim \sqrt{M\sigma^2} = 1$ で起こることがわかる．この量は，特異値に混入するノイズの（全特異成分にわたる）期待値である．この効果により，変分ベイズ法ではノイズが支配的な成分が枝狩りされる．

3.3.5 変分ベイズ主成分分析の次元推定性能

古典的な多変量解析法である主成分分析 [19] を確率的に解釈 [53] すると，行列分解モデルが得られる．具体的にはまず，観測値 $\boldsymbol{v} \in \mathbb{R}^L$ が本質的には隠れ変数 $\widetilde{\boldsymbol{a}} \in \mathbb{R}^H$ にのみ以下の形で依存すると仮定する：

$$\boldsymbol{v} = B\widetilde{\boldsymbol{a}} + \boldsymbol{\varepsilon} \tag{32}$$

ここで，$B \in \mathbb{R}^{L \times H}$ は入出力間の線形関係を記述するローディング行列である．ノイズはガウス分布 $\boldsymbol{\varepsilon} \sim \mathcal{N}_L(\boldsymbol{0}, \sigma^2 I_L)$ に従うと仮定する．

M 個のサンプル $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M)$ が与えられる場合を考え，これらが $\widetilde{\boldsymbol{a}} \sim \mathcal{N}_H(\boldsymbol{0}, I_H)$ に従う隠れ変数 $A^\top = (\widetilde{\boldsymbol{a}}_1, \ldots, \widetilde{\boldsymbol{a}}_M)$ に式 (32) の形で依存すると仮定する．これは行列分解モデル (6), (8) および (9) において，$C_A = I_H$ と設定したものと一致する．

確率的主成分分析に変分ベイズ法を適用すると，いくつかの推定特異値が自動的に 0 となり，主成分の次元数の選択を行えることが知られている [7]．この効果の有用性は実験的に示されているが [37]，厳密なベイズ学習では起こらない変分ベイズ法固有の効果であるため，その正当性には議論の余地があった．この疑問に応えるべく，ノイズ分散 σ^2 を含むすべての未知数をデータから推定したときの次元数選択性能が理論的に調べられた [40]．そこでは，σ^2 の推定値の上界および下界を求めたうえでランダム行列理論 [34, 4, 20, 35] を適用することによって，変分ベイズ法がある条件のもとで高い確率で正しい次元数を選択できることが証明された．

3.3.6 他のモデルへの拡張

全観測変分ベイズ行列分解モデルの大域解析解導出には，変分ベイズ事後共分散が対角であること（定理 1）および変分ベイズ推定量が縮小特異値分解になっていること（補題 1）を用いて，同時に考えなければならない未知変数の数を $O(1)$ 個にまで減らせた（補題 2）ことが本質的である．

残念ながら，このような性質が成立するモデルは全観測行列分解の他には見つかっておらず（部分観測行列分解では，定理 1，補題 1 および補題 2 のいずれも成立しない），大域解析解導出の見通しは立っていない．しかし，定理 3 をサブルーチンとして利用するこ

とにより，標準的な手法によって導出される ICM アルゴリズムよりも効率的に良い局所解を出力するアルゴリズムが，いくつかのモデルにおいて提案されている．

主成分分析に外れ値項を追加したロバスト主成分分析においては，部分問題に対して定理 1 を繰り返し適用する期待値逐次更新法（mean update）と呼ばれるアルゴリズムが提案されている [38]．また，部分観測行列分解においても定理 3 がサブルーチンとして利用され，さらに非ガウスノイズへの拡張も行われた [48]．今後の更なる発展が期待される．

4 変分ベイズ学習の漸近理論

本節では，変分ベイズ学習の漸近論を紹介する．4.1 節で解析対象である汎化誤差および自由エネルギーの漸近形を示し，4.2 節で最新の解析結果を紹介する．

4.1 漸近学習理論の基礎

2 節の冒頭では，ベイズ学習と変分ベイズ学習を導出するために $\boldsymbol{v}$ が $p(\boldsymbol{v}|\boldsymbol{\theta})$ に従うと仮定した．しかし，実際に統計的学習を行う場合，仮定するモデルが正しいかどうかはわからない場合が殆どである．そのような一般的な状況で客観的にモデルと学習方法の良さを評価するために，統計的学習理論では $\boldsymbol{v}$ が「本当に」従う真の分布 $q(\boldsymbol{v})$ を仮定する．ただし，この $q(\boldsymbol{v})$ は統計的学習のユーザーには未知である．統計的学習の目的は，学習データ $\mathcal{V}^n$ から真の分布 $q(\boldsymbol{v})$ を推定することであり，統計的学習理論の目的は，$q(\boldsymbol{v})$ がどのような分布の場合に学習がうまくいくかを解明することである．

4.1.1 汎化誤差および自由エネルギーの漸近形

パラメータの事後分布 $\widehat{r}(\boldsymbol{\theta}|\mathcal{V}^n)$ が得られたとき，$q(\boldsymbol{v})$ は予測分布

$$p(\boldsymbol{v}|\mathcal{V}^n) = \langle p(\boldsymbol{v}|\boldsymbol{\theta}) \rangle_{\widehat{r}(\boldsymbol{\theta}|\mathcal{V}^n)} \tag{33}$$

によって推定される．事後分布は，ベイズ学習の場合には

$$\widehat{r}^{\,\mathrm{Bayes}}(\boldsymbol{\theta}|\mathcal{V}^n) = p(\boldsymbol{\theta}|\mathcal{V}^n)$$

であり，変分ベイズ学習の場合は最小化問題 (4) の解，事後確率最大化法の場合はデルタ関数となる：

$$\widehat{r}^{\,\mathrm{MAP}}(\boldsymbol{\theta}|\mathcal{V}^n) = \delta(\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}})$$

通常，独立なサンプルの数 n が多ければ多いほど $q(\boldsymbol{v})$ に関する多くの情報が観測されるため，予測分布 (33) は真の分布 $q(\boldsymbol{v})$ に近づく．この近さをカルバック擬距離 [30] で測っ

た量

$$G(\mathcal{V}^n) = D(q(\boldsymbol{v}) \| p(\boldsymbol{v}|\mathcal{V}^n)) = \left\langle \log \frac{q(\boldsymbol{v})}{p(\boldsymbol{v}|\mathcal{V}^n)} \right\rangle_{q(\boldsymbol{v})} \tag{34}$$

を汎化誤差と呼ぶ.

汎化誤差 (34) は 1 回の学習における評価値であり，学習データの実現値 $\mathcal{V}^n$ に依存する．統計的学習理論では，学習モデルと学習方法の一般的な性能を調べるために，真の分布 $q(\mathcal{V}^n) = \prod_{i=1}^n q(\boldsymbol{v}^{(i)})$ に従う学習データに関する期待値

$$G(n) = \langle G(\mathcal{V}^n) \rangle_{q(\mathcal{V}^n)} \tag{35}$$

の振る舞いを解析する．この量は，サンプル数 n，仮定するモデル（モデル分布と事前分布の組）および学習方法に依存する量である．

モデルが真の分布を含む場合，すなわち $q(\boldsymbol{\theta}) = p(\boldsymbol{v}|\boldsymbol{\theta}^*)$ を満たす $\boldsymbol{\theta}^*$ が存在する場合を考える．このとき，適切な学習方法を用いる限り，汎化誤差は $n \to \infty$ の漸近極限で以下のオーダーで 0 に収束する：

$$G(n) = \lambda n^{-1} + o(n^{-1})$$

主要項の係数 λ は汎化係数と呼ばれる．λ が小さいほど優秀な学習方法と言えるので，これを理論的に求めることによって学習方法の良さを評価できる．

自由エネルギー (3) の解析も重要である．ベイズ学習の場合，自由エネルギーは周辺対数尤度（の符号反転）に一致し，その挙動は汎化誤差の挙動と強く関連している．また，自由エネルギーは変分ベイズ法が最小化する目的関数であり，その解析を通して変分ベイズ解の振る舞いに関する知見を得ることができる．

自由エネルギー $F(\widehat{r})$ から真のエントロピー $-\log q(\mathcal{V}^n)$ を引いたものを，規格化自由エネルギーと呼ぶ．規格化自由エネルギーの学習サンプルの出方に関する期待値

$$F(n) = \langle F(\widehat{r}) + \log q(\mathcal{V}^n) \rangle_{q(\mathcal{V}^n)} \tag{36}$$

は，サンプル数 n を増やしたとき以下のように漸近展開することができる [60]：

$$F(n) = \lambda' \log n + o(\log n)$$

λ' は自由エネルギー係数と呼ばれる．

4.1.2 正則モデルの学習理論

ここでは，真のパラメータ $\boldsymbol{\theta}^*$ が $\boldsymbol{\theta}$ の定義域の内点に存在し，また，$\boldsymbol{\theta}^*$ のまわりでモデル分布 $p(\boldsymbol{v}|\boldsymbol{\theta})$ とパラメータ $\boldsymbol{\theta}$ とが 1 対 1 対応する場合を考えることにする．また，$\boldsymbol{\theta}$ から $p(\boldsymbol{v}|\boldsymbol{\theta})$ への対応が，$\boldsymbol{\theta}^*$ のまわりでなめらかであることも仮定する．これらの仮定のもと，

汎化誤差 (35) および規格化自由エネルギー (36) をテイラー展開することにより，これらの量の n が大きい場合の漸近的な振る舞いを解析することができる [11, 45, 46]．具体的には，汎化係数は最尤法，MAP 法およびベイズ学習に共通して，

$$2\lambda_{\rm Regular} = K \tag{37}$$

で与えられることがわかっている．ここで，K は $\boldsymbol{\theta}$ の次元数を表す．式 (37) は，汎化誤差の漸近的な主要項がパラメータの次元数のみに依存することを示唆しており，赤池情報量規準 (AIC; Akaike's information criterion)[1] やその拡張の理論的根拠となっている [27, 49, 52, 26, 51]．

自由エネルギー係数については

$$2\lambda'_{\rm Regular} = K \tag{38}$$

が成り立つことが知られており，これをもとにしたモデル選択規準がベイズ情報量規準 (BIC; Bayesian information criterion) [47] である．ベイズ情報量規準は，情報理論の文脈で提案された記述長最小化 (MDL; minimum description length) 規準 [43, 23] と等価である．

4.1.3 特異モデルのベイズ学習理論

確率分布とパラメータとが 1 対 1 対応しないモデルは，特異モデルと呼ばれる [15]．多くの特異モデルでは，真の分布 $q(\boldsymbol{v})$ を表現するために最低限必要な数以上の自由度をモデル分布 $p(\boldsymbol{v}|\boldsymbol{\theta})$ が持つとき，$p(\boldsymbol{v}|\boldsymbol{\theta}^*) = q(\boldsymbol{v})$ を満たす $\boldsymbol{\theta}^*$ が 1 点に定まらない．そのような場合，$p(\boldsymbol{v}|\boldsymbol{\theta}^*) = q(\boldsymbol{v})$ を満たす $\boldsymbol{\theta}^*$ の集合上でフィッシャー計量が特異となり，$n \to \infty$ における漸近挙動を調べるために汎化誤差や自由エネルギーを $\boldsymbol{\theta}^*$ のまわりでテイラー展開することができない．

特異モデルに対しては，一般に式 (37) や式 (38) のような単純な関係が成り立たないことが知られている．近年，代数幾何学を駆使した特異学習理論 [61, 60] によって自由エネルギー係数 $\lambda'^{\rm Bayes}$ を計算する方法が確立し，縮小ランク回帰モデル，混合分布モデル，隠れマルコフモデルなど，多くの特異モデルに対して $\lambda'^{\rm Bayes}$ の上界および下界が求められた [59, 62, 64, 63, 44, 2]．

任意のモデルに対して，ベイズ学習の汎化誤差と規格化自由エネルギーとの間に以下の関係が成り立つことが知られている [31]：

$$G(n) = F(n+1) - F(n)$$

このことから，自由エネルギー係数と汎化係数が一致する事がわかる：

$$\lambda^{\rm Bayes} = \lambda'^{\rm Bayes} \tag{39}$$

そのため，特異学習理論によって自由エネルギー係数を求めることにより，特異モデルに対するベイズ学習の汎化性能が評価できる．

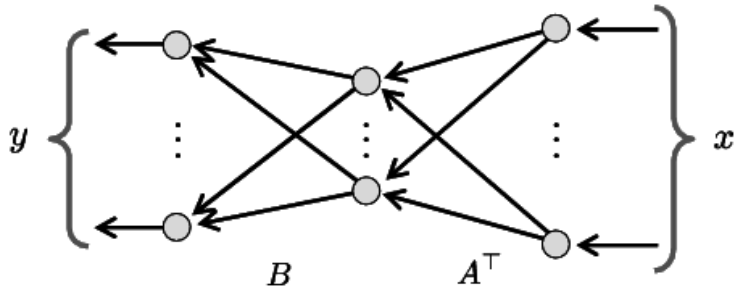

図 2: Linear neural network.

4.2 変分ベイズ学習の漸近解析

本節では，変分ベイズ学習の汎化係数および自由エネルギー係数の漸近解析の結果を紹介する．ベイズ学習では，自由エネルギー係数の振る舞いを解析することによって，式 (39) を通して汎化係数の振る舞いを明らかにすることができた．しかし変分ベイズ法では，ベイズ学習の汎化係数と自由エネルギー係数との関係 (39) が成立しないため，多くのモデルに対して自由エネルギー係数の上界および下界が明らかにされているにもかかわらず，汎化係数の振る舞いに関しては未だ解明されていない点が多い．以下では，汎化係数が解明されている縮小ランク回帰モデルに対する研究結果を紹介するとともに，混合ガウス分布に対する理論解析の結果を紹介する．

4.2.1 縮小ランク回帰モデル

縮小ランク回帰モデル [5, 42] は，多次元入力ベクトル $\boldsymbol{x} \in \mathbb{R}^M$ と多次元出力ベクトル $\boldsymbol{y} \in \mathbb{R}^L$ との間の線形関係を学習するために用いられる：

$$\boldsymbol{y} = BA^\top \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{40}$$

パラメータは行列 $A \in \mathbb{R}^{M\times H}$ および $B \in \mathbb{R}^{L\times H}$ であり，ノイズは独立なガウス分布に従うと仮定する．

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_L(\mathbf{0}, \sigma'^2 I_L)$$

ここで，I_d は d 次元の単位行列を表す．このモデルのパラメータの自由度はみかけ上 $(L+M)H$ であるが，式 (7) と同様の冗長性により，実質的な自由度は

$$K = (M+L)H - H^2 \tag{41}$$

である．縮小ランク回帰モデルは，Fig. 2 のような線形神経回路網として表すこともできる．

ここで，n 組の学習データ $\mathcal{V}^n = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\}$ が与えられる場合を考える．この学習データに対する尤度は，次式で与えられる．

$$p(\mathcal{V}^n|A,B) \propto \exp\left(-\frac{1}{2\sigma'^2}\sum_{i=1}^{n}\|\boldsymbol{y}^{(i)} - BA\boldsymbol{x}^{(i)\top}\|^2\right) \tag{42}$$

学習データ $\mathcal{V}^n$ は事前に中心化されており，また，入力は白色化されていると仮定する [21]：

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)} = \boldsymbol{0} \qquad \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}^{(i)} = \boldsymbol{0} \qquad \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top} = I_M$$

このとき入出力間の共分散行列

$$V = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top} \tag{43}$$

を用いると，確率モデル (42) はパラメータ A, B の関数として

$$p(\mathcal{V}^n|A,B) \propto \exp\left(-\frac{n}{2\sigma'^2}\|V - BA^\top\|_{\mathrm{Fro}}^2\right) \tag{44}$$

と表すことができる．確率モデル (44) は，3.1 節で紹介した行列分解モデルにおいてノイズ分散を

$$\sigma^2 = \sigma'^2/n \tag{45}$$

とおいたものと等価であることに注意する．従って，正定値対角共分散 (C_A, C_B) を持つガウス事前分布

$$p(A) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(AC_A^{-1}A^\top\right)\right)$$
$$p(B) \propto \exp\left(-\frac{1}{2}\mathrm{tr}\left(BC_B^{-1}B^\top\right)\right)$$

を仮定し，パラメータ A, B 間の独立性制約

$$r(A,B) = r_{\mathrm{A}}(A)r_{\mathrm{B}}(B)$$

を課すことにより，変分ベイズアルゴリズム (12)–(15) および自由エネルギー (20) が得られる．ただし，これらは式 (45) を通してサンプル数 n に依存することに注意する．以下，簡潔な記述のために $L \leq M$ を仮定する．$L > M$ の場合は，3 節の場合と同様の操作により，$L \leq M$ の場合の結果から容易に推察することができる．

ハイパーパラメータ (C_A, C_B) およびノイズ分散 σ'^2 が既知である場合，以下の定理が成り立つ．

定理 4 (Nakajima and Watanabe (2007) [41]). 大サンプル極限 $n \to \infty$ において，縮小ランク回帰モデルの変分ベイズ解は

$$\widehat{U}^{\mathrm{VB}} = \widehat{B}\widehat{A}^{\top} = \sum_{h=1}^{H} \widehat{\gamma}_h^{\mathrm{VB}} \boldsymbol{\omega}_{b_h} \boldsymbol{\omega}_{a_h}^{\top} + O_p(n^{-1})$$

と表すことができる．ただし

$$\widehat{\gamma}_h^{\mathrm{VB}} = \widehat{\gamma}_h^{\mathrm{PJS}} = \max\left\{0, \left(1 - \frac{M\sigma'^2}{n\gamma_h^2}\right)\gamma_h\right\}$$

♢

サンプル数が増えるにつれて事前分布の影響は相対的に小さくなることから，縮小ランク回帰モデルの漸近解は，行列分解モデルの均一事前分布解（3.3.3 節の系 1）に収束することがわかる．

予測分布に関しては，以下の補題が成り立つ．

補題 3 (Nakajima and Watanabe (2007) [41]). 縮小ランク回帰モデルの予測分布は，大サンプル極限において以下のように表現できる：

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{V}^n) = \mathcal{N}_L(\boldsymbol{y}; \widehat{\Sigma}\widehat{U}^{\mathrm{VB}}\boldsymbol{x}, \widehat{\Sigma}) + O_p(n^{-3/2})$$

ここで，予測共分散行列 $\widehat{\Sigma}$ は

$$\widehat{\Sigma} = I_L + O_p(n^{-1})$$

で与えられる． ♢

次に，汎化誤差および自由エネルギーを評価する．学習対象の真の分布は，パラメータ (A^*, B^*) を持つ縮小ランク回帰モデルであり，$B^*A^{*\top}$ のランクは H^* であると仮定する．$H < H^*$ の場合はモデルの表現力が足りないため，汎化誤差は 0 に収束せず，汎化係数は意味を成さない．よって $H \geq H^*$ の場合のみを考えることにする．

3.3.1 節で述べたように，行列分解モデルの変分ベイズ事後分布は完全に記述することができる [39]．解析解を式 (20) に代入することによって以下の定理が得られる．

定理 5 (Nakajima and Watanabe (2007) [41]). 縮小ランク回帰モデルの規格化自由エネルギー (36) は，大サンプル極限において以下のように表すことができる：

$$F = \lambda' \log n + O(1)$$

ただし，

$$2\lambda' = H^*(L + M) + (H - H^*)L$$

♢

汎化係数の解析は，真の行列 B^*A^* を表現するために必要な成分 $\{(\boldsymbol{a}_h, \boldsymbol{b}_h); h = 1, \ldots, H^*\}$ からの寄与と，冗長な成分 $\{(\boldsymbol{a}_h, \boldsymbol{b}_h); h = H^*+1, \ldots, H\}$ からの寄与を別々に考えることによって行われる．必要な成分は正則モデルの場合と同様に振るまうため，それらの汎化係数への寄与は自由度 (41) のみから決まる．一方，冗長な成分は Wishart 分布に従うため，その固有値分布が汎化係数に影響する．行列サイズ (L, M) が十分大きいことを仮定すると，Wishart 行列の固有値分布を記述する Marčenko-Pastur 則 [34, 55] を用いることができ，汎化係数を計算することができる．

定理 6 (Nakajima and Watanabe (2007) [41])**.** (L, M, H, H^*) が互いの比を保って無限大に近づく大スケール極限において，縮小ランク回帰モデルの変分ベイズ汎化係数は次式で与えられる：

$$2\tilde{\lambda} = (H^*(M+L) - H^{*2}) + \frac{(M-H^*)(L-H^*)}{2\pi\alpha}\left\{J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1)\right\}$$

ここで

$$\begin{aligned}
\alpha &= (L-H^*)/(M-H^*) \\
\beta &= (H-H^*)/(L-H^*) \\
\kappa &= M/(M-H^*) \\
J(s;1) &= 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s) \\
J(s;0) &= -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1} s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s + 2\alpha}{2\alpha s + \sqrt{\alpha}(1+\alpha)} \\
J(s;-1) &= \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1}s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0<\alpha<1) \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1) \end{cases}
\end{aligned}$$

であり，s_t は $J(s;0)$ の逆関数 $J^{-1}(\cdot;0)$ を用いて

$$s_t = \max\left((\kappa - (1+\alpha))/2\sqrt{\alpha}, J^{-1}(2\pi\alpha\beta;0)\right)$$

で与えられる． ♢

定理 5 で得られた変分ベイズ自由エネルギー係数の振る舞いを，代数幾何学的手法によって計算されたベイズ自由エネルギー係数 [2] とともに Fig. 3 に示す．この図より，自由エネルギー係数の振る舞いが，変分ベイズ法とベイズ学習とで非常に似ていることがわかる．一方，定理 6 によって得られた変分ベイズ汎化係数を，最尤法の汎化係数 [14] およびベイズ学習の汎化係数とともに Fig. 4 に示す．ここで，関係 (39) よりベイズ学習の汎化係数は Fig. 3 のベイズ自由エネルギー係数と一致することに注意する．変分ベイズ汎化誤差はベイズ汎化誤差と同程度に小さいが，H に対する依存性はむしろ最尤法に似ていることがわかる．自由エネルギー係数の挙動の類似性（Fig. 3）と汎化係数の挙動の不一致性（Fig. 4）より，ベイズ学習において成立する単純な関係 (39) が，変分ベイズ法においては近似的にも成立しないことが示唆される．

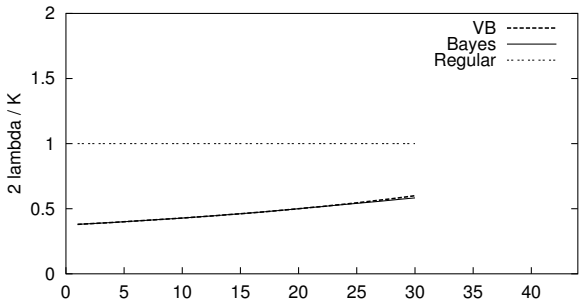

図 3: The variational Bayes (VB) and the Bayes free energy coefficients in the reduced rank regression model with $L = 30$, $M = 50$, $H = 1, \ldots, 30$, and $H^* = 0$. The VB and the Bayes free energies almost coincide with each other. 'Regular' indicates the free energy coefficient in a regular statistical model with the same degrees of freedom K, which is given by Eq.(41).

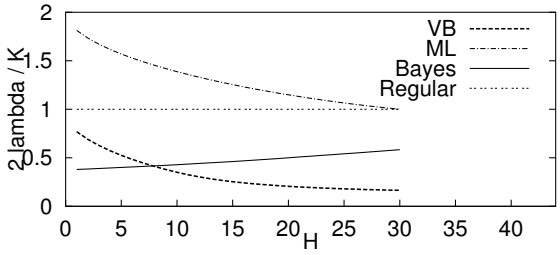

図 4: The variational Bayes (VB), the maximum likelihood (ML), and the Bayes generalization coefficients when $L = 30$, $M = 50$, $H = 1, \ldots, 30$, and $H^* = 0$.

4.2.2 混合ガウス分布

観測値 $\boldsymbol{x} \in \mathbb{R}^M$ が，混合数 H の混合ガウス分布

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{h=1}^{H} a_h \mathcal{N}_M(\boldsymbol{x}; \boldsymbol{\mu}_h, \Sigma_h) \tag{46}$$

に従うと仮定する．簡単のために各ガウス成分の共分散はすべて $\Sigma_h = I_M$ であると仮定すると，モデルのパラメータは

$$\boldsymbol{\theta} = \left\{ (a_h, \boldsymbol{\mu}_h); a_h \in \mathbb{R}, \boldsymbol{\mu}_h \in \mathbb{R}^M, h = 1, \ldots, H, a_h \geq 0, \sum_{h=1}^{H} a_h = 1 \right\}$$

である．また，n 個の学習サンプル $\mathcal{X}^n = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)})$ に対する尤度は，

$$p(\mathcal{X}^n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\sum_{h=1}^{H} a_h \mathcal{N}_M(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_h, I_M) \right) \tag{47}$$

で与えられる．式 (47) は混合分布の積であり，一般に $O(H^n)$ 個の項に展開されるためそのまま計算することは困難である．しかし，もし各サンプル $\boldsymbol{x}^{(i)}$ が H 個のガウス成分のうちのどれから生成されたかがわかっているとすると，異なる成分間の相関が無くなる．そこで，$\{(i,h); 1 \leq i \leq n, 1 \leq h \leq H\}$ に対して，以下の隠れ変数を定義することにする：

$$y_h^{(i)} = \begin{cases} 1 & i \text{ 番目のサンプルが } h \text{ 番目のガウス成分から生成された} \\ 0 & \text{それ以外} \end{cases} \tag{48}$$

n 個すべてのサンプルに対して $\mathcal{Y}^n = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)})$ が与えられていると仮定すると，尤度関数を以下のように単一の項で表すことができる：

$$p(\mathcal{X}^n, \mathcal{Y}^n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{h=1}^{H} \left(a_h \mathcal{N}_M(x^{(i)}; \boldsymbol{\mu}_h, I_M) \right)^{y_h^{(i)}} \tag{49}$$

観測値と隠れ変数の組 $(\boldsymbol{x}, \boldsymbol{y})$ は完全データと呼ばれ，式 (49) は完全データに対する尤度関数と呼ばれる．実際には隠れ変数 $\mathcal{Y}^n$ は未知であるため，隠れ変数を導入したことよって推定しなければならない変数は増えてしまうが，式 (47) が含む和 $\sum_{h=1}^{H}$ を積 $\prod_{h=1}^{H}$ に変えるメリットは非常に大きい．なぜならば，単項の尤度関数 (49) に対しては，パラメータの部分集合に対する共役事前分布が存在するからである．これは，混合分布モデルに対して最尤推定を行う EM アルゴリズムを導出する際に用いられたトリックである [12].

混合ガウスモデルに対して変分ベイズ法を導出するには，完全尤度 (49) を，$\{\boldsymbol{\mu}_h\}$ を固定して $\{a_h\}$ の分布関数とみたときの共役事前分布

$$p(\{a_h\}) = \mathcal{D}(\{a_h\}; \{\alpha_0, \ldots, \alpha_0\}) \tag{50}$$

と，$\{a_h\}$ を固定して $\{\boldsymbol{\mu}_h\}$ の分布関数とみたときの共役事前分布

$$p(\boldsymbol{\mu}_h) = \mathcal{N}(\boldsymbol{\mu}_h; \boldsymbol{\beta}_0, c^2) \tag{51}$$

を用いて，以下の独立性制約を仮定する [3]：

$$r(\mathcal{Y}^n, \boldsymbol{\theta}) = r(\mathcal{Y}^n) r(\boldsymbol{\theta})$$

ここで，

$$\mathcal{D}(\{a_h\}; \{\alpha_h\}) = \frac{\Gamma(\sum_{h=1}^{H} \alpha_h)}{\prod_{h=1}^{H} \Gamma(\alpha_h)} \prod_{h=1}^{H} a_h^{\alpha_h - 1}$$

は Dirichlet 分布であり，$\Gamma(\cdot)$ はガンマ関数である．$\{a_h\}$ と $\{\boldsymbol{\mu}_h\}$ の間の独立性は完全尤度と事前分布 (49)–(51) によって自動的に満たされるため，明示的に制約を加える必要はない．

変分法によって自由エネルギー

$$F(r) = \sum_{\mathcal{Y}^n} \int r(\mathcal{Y}^n, \boldsymbol{\theta}) \log \frac{r(\mathcal{Y}^n, \boldsymbol{\theta})}{p(\mathcal{Y}^n, \boldsymbol{\theta}|\mathcal{X}^n)} d\boldsymbol{\theta}$$

を最小化すると，変分ベイズ事後分布が以下の形に分解できることがわかる：

$$r(\mathcal{Y}^n, \boldsymbol{\theta}) = \left(\prod_{i=1}^{n} r(\boldsymbol{y}^{(i)}) \right) \left(\prod_{h=1}^{H} r(a_h) r(\boldsymbol{\mu}_h) \right)$$

ただし

$$\begin{aligned} r(\{a_h\}) &= \mathcal{D}(\{a_h\}; \{\widehat{\alpha}_h\}) \\ r(\boldsymbol{\mu}_h) &= \mathcal{N}_M(\boldsymbol{\mu}_h; \widehat{\boldsymbol{\mu}}_h, \sigma_h^2) \\ r(\boldsymbol{y}^{(i)}) &\propto \exp\left(\sum_{h=1}^{H} y_h^{(i)} \left\{ \Psi(\widehat{\alpha}_h) - \Psi\left(\sum_{h'=1}^{H} \widehat{\alpha}_{h'} \right) - \frac{1}{2} \left(\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_h\|^2 + M\widehat{\sigma}_h^2 \right) \right\} \right) \end{aligned}$$

であり，分布を記述するパラメータは以下のように与えられる：

$$\widehat{\alpha}_h = \overline{N}_h + \alpha_0 \tag{52}$$

$$\widehat{\boldsymbol{\mu}}_h = \overline{\sigma}_h^2 \left(\overline{N}_h \overline{\boldsymbol{x}}_h + \boldsymbol{\beta}_0 c^{-2} \right) \tag{53}$$

$$\widehat{\sigma}_h^2 = \frac{1}{\overline{N}_h + c^{-2}} \tag{54}$$

$$\overline{N}_h = \sum_{i=1}^{n} \overline{y}_h^{(i)} \tag{55}$$

$$\overline{\boldsymbol{x}}_h = \frac{1}{\overline{N}_h} \sum_{i=1}^{n} \overline{y}_h^{(i)} \boldsymbol{x}^{(i)} \tag{56}$$

$$\overline{y}_h^{(i)} = r(y_h^{(i)} = 1) \propto \exp\left(\Psi(\widehat{\alpha}_h) - \frac{1}{2} \left(\|\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_h\|^2 + M\widehat{\sigma}_h^2 \right) \right) \tag{57}$$

ここで，

$$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

はディガンマ関数である．式 (52)–(57) を繰り返し適用することによって自由エネルギーの局所最小解が得られる [3].

混合分布モデルにおける変分ベイズ法の性質を調べるために，真の分布が

$$p(\boldsymbol{x}|\boldsymbol{\theta}^*) = \sum_{h=1}^{H^*} a_h^* \mathcal{N}_M(\boldsymbol{x}; \boldsymbol{\mu}_h^*, I_M)$$

であると仮定する．このとき，自由エネルギー係数に関して以下の定理が成り立つ．

定理 7 (Watanabe and Watanabe (2006) [57]). 分散既知の混合ガウス分布の自由エネルギー係数は以下のようにバウンドされる：

$$(H-1)\alpha_0 + \frac{M}{2} \leq \lambda' \leq (H-H^*)\alpha_0 + \frac{MH^*+H^*-1}{2} \qquad for\ \alpha_0 \leq \frac{M+1}{2} \tag{58}$$

$$\lambda' = \frac{MH+H-1}{2} \qquad for\ \alpha_0 > \frac{M+1}{2} \tag{59}$$

♢

定理 7 より，混合係数 $\{a_h\}$ 上の事前分布 (50) が持つハイパーパラメータ α_0 が，変分ベイズ解の振る舞いに強く影響することがわかる．ここで，$\alpha_0 \leq (M+1)/2$ のときに

$$(H-H^*)\alpha_0 + \frac{MH^*+H^*-1}{2} \leq \frac{MH+H-1}{2}$$

が成り立つことに注意すると，$\alpha_0 \leq (M+1)/2$ に対する自由エネルギー係数 (58) が，$\alpha_0 > (M+1)/2$ に対する自由エネルギー係数 (59) 以下であることがわかる．

実は (58) の上界を導出する過程で，$\alpha_0 \leq (M+1)/2$ のときは冗長なガウス成分の混合係数がすべて 0 になることが発見された（すなわち $\widehat{\alpha}_h = 0$ for $h = H^*+1, \ldots, H$）[57]. 一方，$\alpha_0 > (M+1)/2$ の場合には冗長な成分の混合係数は消えずに，複数のガウス成分の平均値パラメータ $\widehat{\boldsymbol{\mu}}_h$ が重複することによって H^* 個のガウス成分が表現される．この相転移現象はシミュレーションによっても確認されている．この結果より，変分ベイズ法の枝狩り効果によって混合成分の数を決定する際には，事前分布を $\alpha_0 \leq (M+1)/2$ となるように設定することが重要であることがわかる．

4.2.3 他のモデルへの拡張

混合ガウス分布において自由エネルギー係数（定理 7）を導出した方法は，混合指数分布族 [58]，隠れマルコフモデル [18] およびベイジアンネットワーク [56] など，比較的広い範囲のモデルに適用され，自由エネルギー係数のバウンドが導出された．また，混合ガウス分布の場合と同様に，自由エネルギー係数の解析を通して Dirichlet 事前分布のハイパーパラメータ設定に関する指針も与えられている．

一方，縮小ランク回帰モデルの汎化係数（定理 6）は大域解析解（定理 2）を利用して得られた結果であるため，この方法を他のモデルに適用することは現在のところ困難である．また，厳密なベイズ学習の場合と異なり，自由エネルギー係数の解明は汎化係数の解明に直接つながらない．一般のモデルに適用可能な新たな汎化性能解析法の開発が待たれる．

5 まとめ

本論文では，近年発展してきた変分ベイズ法の学習理論研究を紹介した．全観測行列分解モデルに対しては，変分ベイズ法のモデル起因正則化やスパース性誘起メカニズムが解明され，ベイズ法との違いなどの多くの興味深い性質が明らかになっている．また，自由エネルギーの漸近解析は，混合分布モデル，隠れマルコフモデルやベイジアンネットワークモデルにおける事前分布設定に指針を与えている．

一方，縮小ランク回帰モデルを除く多くの特異モデルに対しては，変分ベイズ法の汎化性能は未解明である．本論文で紹介した解析手法のさらなる発展により，より多くのモデルに対して変分ベイズ法の性質を明らかにすることが今後の課題である．

謝辞

本研究は科研費新学術領域研究「予測と意思決定」（23120004）の助成を受けたものである．

参考文献

[1] H. Akaike. A new look at statistical model. *IEEE Trans. on Automatic Control*, Vol. 19, No. 6, pp. 716–723, 1974.

[2] M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, Vol. 18, No. 7, pp. 924–933, 2005.

[3] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 21–30, San Francisco, CA, 1999. Morgan Kaufmann.

[4] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, Vol. 97, No. 6, pp. 1382–1408, 2006.

[5] P. F. Baldi and K. Hornik. Learning in linear neural networks: A survey. *IEEE Transactions on Neural Networks*, Vol. 6, No. 4, pp. 837–858, 1995.

[6] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, Vol. 48, pp. 259–302, 1986.

[7] C. M. Bishop. Variational principal components. In *Proc. of International Conference on Artificial Neural Networks*, Vol. 1, pp. 514–509, 1999.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

[9] C.M. ビショップ（著）, 元田浩・栗田多喜雄・樋口智之・松本裕治・村田昇（監訳）. パターン認識と機械学習. 丸善出版, 2007.

[10] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, Vol. 20, No. 4, pp. 1956–1982, 2010.

[11] H. Cramer. *Mathematical Methods of Statistics*. University Press, Princeton, 1949.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Vol. 39-B, pp. 1–38, 1977.

[13] B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, Vol. 68, pp. 117–130, 1973.

[14] K. Fukumizu. Generalization error of linear neural networks in unidentifiable cases. In *Proc. of International Conference on Algorithmic Learning Theory*, pp. 51–62. Springer, 1999.

[15] 福水健次, 栗木哲, 竹内啓, 赤平昌文. 特異モデルの統計学. 岩波書店, 東京, 2004.

[16] S. Funk. Try this at home. http://sifter.org/˜simon/journal/20061211.html, 2006.

[17] M. Hazewinkel, editor. *Encyclopaedia of Mathematics*. Springer, 2002.

[18] 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, Vol. J89-D, No. 6, pp. 1279–1287, 2006.

[19] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, Vol. 24, pp. 417–441, 1933.

[20] D. C. Hoyle. Automatic PCA dimension selection for high dimensional data and small sample sizes. *Journal of Machine Learning Research*, Vol. 9, pp. 2733–2759, 2008.

[21] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, 2001.

[22] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research*, Vol. 11, pp. 1957–2000, 2010.

[23] 伊藤秀一. モデル選択（第 2 部：情報圧縮と確率的複雑さ—MDL 原理）. 岩波書店, 東京, 2004.

[24] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 361–379, Berkeley, CA., USA, 1961. University of California Press.

[25] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, Vol. 186, pp. 453–461, 1946.

[26] S. Konishi and G. Kitagawa. Generalized information criteria in model selection. *Biometrika*, Vol. 83, pp. 875–890, 1996.

[27] 小西貞則, 北川源四郎. 情報量規準. 朝倉書店, 東京, 2004.

[28] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, Vol. 40, No. 3, pp. 77–87, 1997.

[29] 久保川達也. モデル選択（第 3 部：スタインのパラドクスと縮小推定の世界）. 岩波書店, 東京, 2004.

[30] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, Vol. 22, pp. 79–86, 1951.

[31] E. Levin, N. Tishby, and S. A. Solla. A statistical approaches to learning and generalization in layered neural networks. In *Proc. of IEEE*, Vol. 78, pp. 1568–1674, 1990.

[32] Y. J. Lim and T. W. Teh. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, 2007.

[33] D. J. C. Mackay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Available from `http://www.inference.phy.cam.ac.uk/mackay/minima.pdf`. 2001.

[34] V. A. Marcenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, Vol. 1, No. 4, pp. 457–483, 1967.

[35] 永尾太郎. ランダム行列の基礎. 東京大学出版会, 東京, 2005.

[36] S. Nakajima and M. Sugiyama. Theoretical analysis of Bayesian matrix factorization. *Journal of Machine Learning Research*, Vol. 12, pp. 2579–2644, 2011.

[37] S. Nakajima, M. Sugiyama, and S. D. Babacan. On Bayesian PCA: Automatic dimensionality selection and analytic solution. In *Proceedings of 28th International Conference on Machine Learning (ICML2011)*, pp. 497–504, Bellevue, WA, USA, Jun. 28–Jul.2 2011.

[38] S. Nakajima, M. Sugiyama, and S. D. Babacan. Variational Bayesian sparse additive matrix factorization. *Machine Learning (Special Issue of Selected Papers of ACML 2012)*, Vol. 92, pp. 319–1347, 2013.

[39] S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational Bayesian matrix factorization. *Journal of Machine Learning Research*, Vol. 14, pp. 1–37, 2013.

[40] S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Perfect dimensionality recovery by variational Bayesian PCA. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 980–988, 2012.

[41] S. Nakajima and S. Watanabe. Variational Bayes solution of linear neural networks and its generalization performance. *Neural Computation*, Vol. 19, No. 4, pp. 1112–1153, 2007.

[42] G. R. Reinsel and R. P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York, 1998.

[43] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, Vol. 14, No. 3, pp. 1080–1100, 1986.

[44] D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, Vol. 6, pp. 1–35, 2005.

[45] 坂本慶之, 石黒真木夫, 北川源四郎. 情報量統計学. 共立出版, 東京, 1983.

[46] T. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. D.Reidel Publishing Company, 1986.

[47] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, 1978.

[48] M. Seeger and G. Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, La Palma, Spain, 2012.

[49] 下平英寿. モデル選択（第 1 部：情報量規準によるモデル選択とその信頼性評価）. 岩波書店, 東京, 2004.

[50] N. Srebro, J. Rennie, and T. Jaakkola. Maximum margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

[51] M. Sugiyama and H. Ogawa. Subspace information criterion for model selection. *Neural Computation*, Vol. 13, No. 8, pp. 1863–1889, 2001.

[52] 竹内啓. 情報統計量の分布とモデルの適切さの基準. 数理科学, Vol. 153, pp. 12–18, 1976.

[53] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, Vol. 61, pp. 611–622, 1999.

[54] 上田修功. ベイズ学習. 電子情報通信学会誌, Vol. 85, No. 4,6,7,8, April–August 2002.

[55] K. W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability*, Vol. 6, pp. 1–18, 1978.

[56] K. Watanabe, M. Shiga, and S. Watanabe. Upper bound for variational free energy of Bayesian networks. *Machine Learning*, Vol. 75, No. 2, pp. 199–215, 2009.

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

[58] K. Watanabe and S. Watanabe. Stochastic complexities of general mixture models in variational Bayesian learning. *Neural Networks*, Vol. 20, No. 2, pp. 210–219, 2007.

[59] S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, Vol. 13, No. 4, pp. 899–933, 2001.

[60] 渡辺澄夫. 代数幾何学と学習理論. 森北出版, 東京, 2006.

[61] S. Watanabe. *Algebraic Geometry and Statistical Learning*. Cambridge University Press, Cambridge, UK, 2009.

[62] K. Yamazaki and S. Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, Vol. 16, No. 7, pp. 1029–1038, 2003.

[63] K. Yamazaki and S. Watanabe. Stochastic complexity of Bayesian networks. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pp. 592–599, Acapulco, Mexico, 2003.

[64] K. Yamazaki and S. Watanabe. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, Vol. 69, pp. 62–84, 2005.