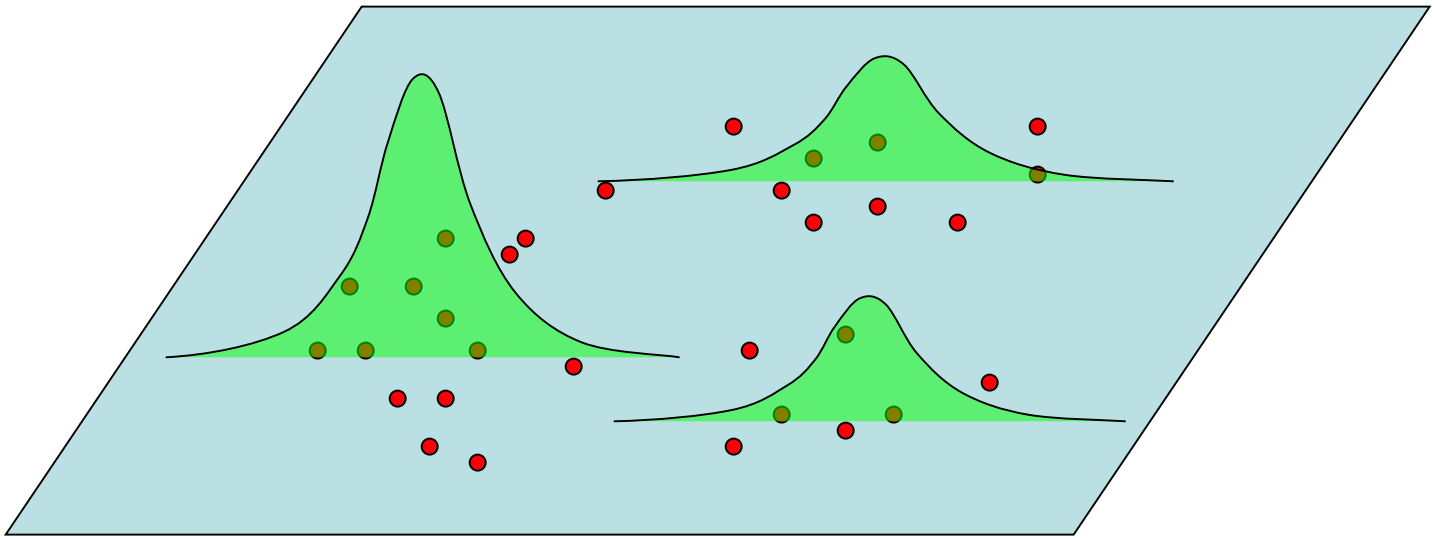
学習理論の練習 12

混合正規分布の復習

学習モデルの例として混合正規分布を
使いますが、混合〇〇分布であれば
同様の学習アルゴリズムが作れます。

例：　混合多項分布、混合ポアソン分布、
　　混合ガンマ分布、・・・。

混合正規分布



平均 b_k ,分散σ_k^2 の正規分布の重み $\{a_k\}$ の和

混合正規分布の定義式

x : M 次元ベクトル

パラメータ　$w = (a_k , b_k , \sigma_k)$

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right)$$

$$\sum_{k=1}^{K} a_k = 1$$

平均 b_k ,分散σ_k^2 の
正規分布

隠れ変数（潜在変数）の導入

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{M/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right)$$

y について和をとる

$$p(x,y|w) = \prod_{k=1}^{K} \left[a_k \frac{1}{(2\pi\sigma_k^2)^{M/2}} \exp\left(- \frac{\| x - b_k \|^2}{2\sigma_k^2} \right) \right]^{y_k}$$

$y = (y_1, y_2, .., y_k)$ はひとつだけ1で残りは0。つまり

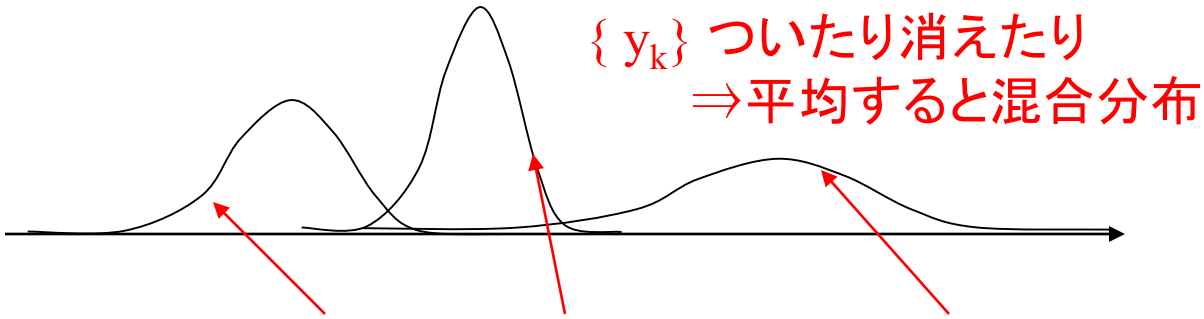
$y \in \{(1,0,..,0), (0,1,0,\ldots,0), \ldots, (0,0,\ldots,1)\} \equiv C^K$

隠れ変数を使うと確率分布が簡単になる。

$$p(x,y|w) = \prod_{k=1}^{K} \left[a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(-\frac{\| x - b_k \|^2}{2\sigma_k^2} \right) \right]^{y_k}$$

$$= \exp\left[-\sum_{k=1}^{K} y_k \{ \|x-b_k\|^2/2\sigma_k^2 - N\log\sigma_k + \log a_k \} \right]$$

(定数項は省略)



$$(y_1,y_2,..,y_K) = (1,0,..,0),\ (0,1,0,..,0),\ ..,(0,0,..,1)$$

平均化最大化繰り返し(EM) 法

EM法のアルゴリズム

EM法は最尤推定量の探索を目的とするアルゴリズムである。

$$G(w_1,w_2) = \sum_{i=1}^{n} \Sigma_y \, p(y| x_i, w_1) \log p(x_i,y \mid w_2)$$

(1) w_1 初期化
(2) $G(w_1,w_2)$を w_2 について最大化(w_1 固定）
(3) w_1 ：＝w_2 として(2)に戻る。

この繰り返しで尤度が単調非減少であることを示すことができる。

http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/joho16.pdf

混合正規分布のEM法

(1) w_1 初期化

(2) $w_2=(a_k\, b_k \sigma_k)$を次式で計算

$$
\begin{cases}
a_k = \dfrac{\Sigma\, E[y_k|\, x_i\, ,\, w_1]}{n} \\
b_k = \dfrac{\Sigma\, E[y_k|\, x_i\, ,\, w_1]\, x_i}{\Sigma\, E[y_k|\, x_i\, ,\, w_1]} \\
\sigma_k^{\,2} = \dfrac{\Sigma\, E[y_k|\, x_i\, ,\, w_1]\, \| x_i - b_k \|^2}{N\ \ \Sigma\, E[y_k|\, x_i\, ,\, w_1]}
\end{cases}
$$

(3) w_1 :＝w_2 として(2)に戻る。

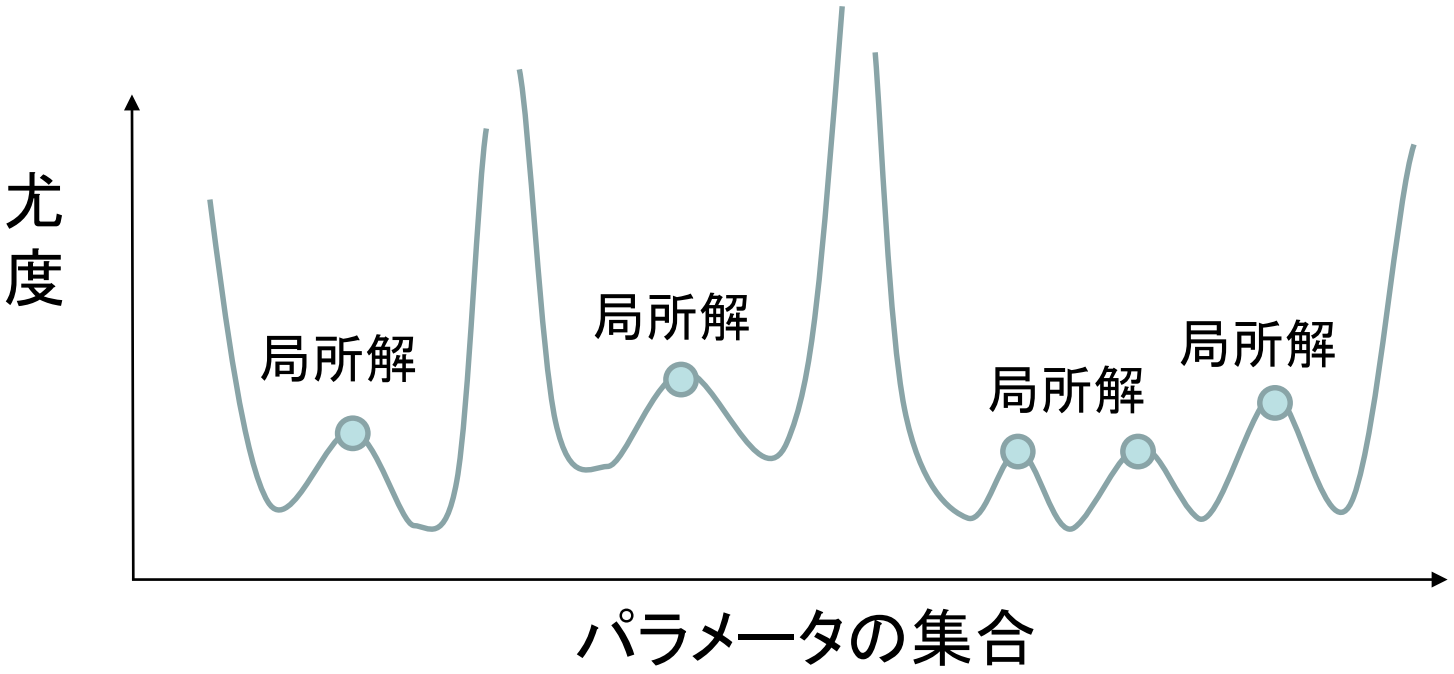
混合正規分布の対数尤度関数の形状

良い局所解を判定することができるだろうか。

(1) 実は混合正規分布では最尤推定量は発散している。($x_i=b_k$, $\sigma_k \to 0$)。

(2) 局所最適解はとてもたくさんあると思われる。

(3) 尤度が大きい局所解が汎化の意味で良いのではないので、
尤度で局所解の良さの比較はできない。AIC, TIC, BIC は使えない。
クロスバリデーション(k-fold)も使えない。
→ 最尤法やEM法は「ダマシダマシ使う」必要がある。
安心して使いたい場合は VB に変更する。EM を VB に変更して
損失になることは何もない(演算量も)。ただし数式は少し難しくなる。



実験例

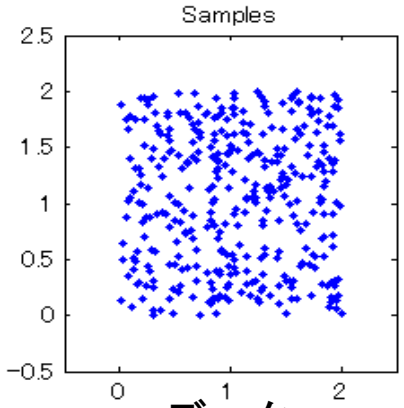
データ: $[0,2]^2$ 上の一様分布から発生(n=360)。

混合正規分布(K=9)で学習(EM法)。

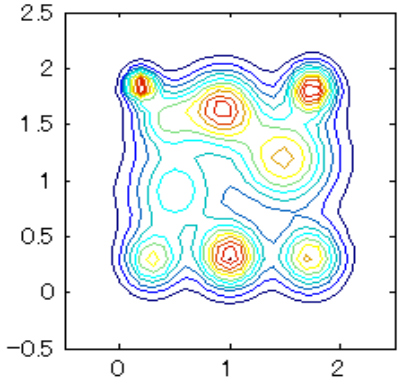
初期値

a_k=1/K,
b_k= 全体平均＋乱数
σ_k= 全体標準偏差
から初めて1000回繰り返し

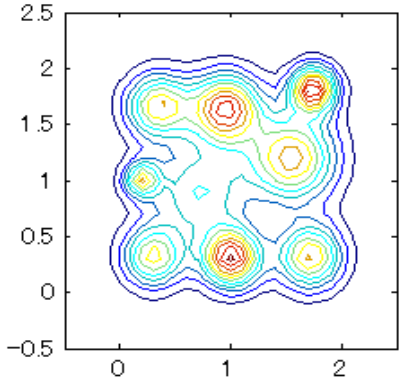
初期値を変えて11回学習した。尤度が大きい解は、σ_kが小さくなっている。



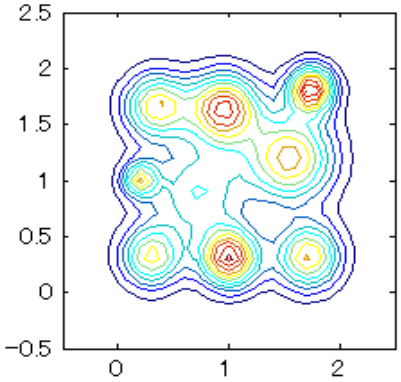
データ



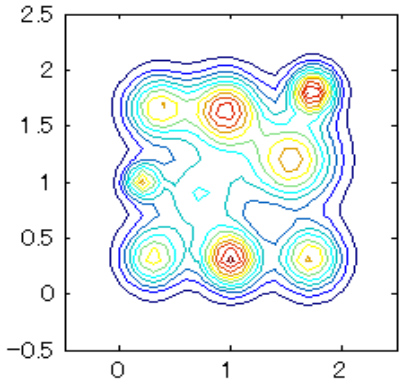
92.965150



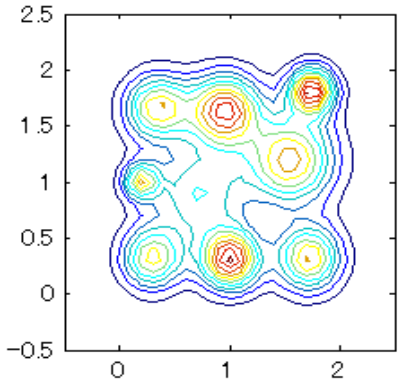
92.452349



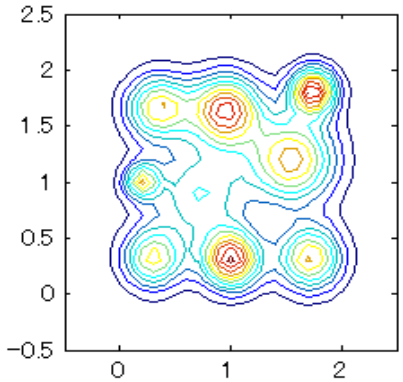
92.452349



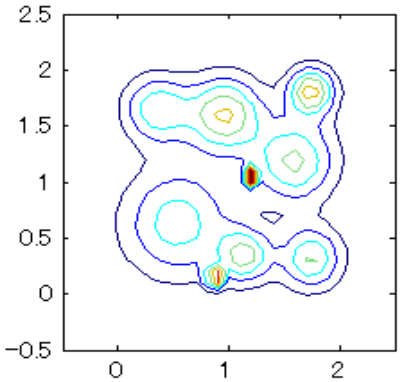
92.452349



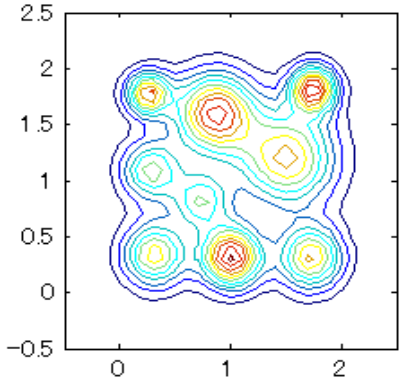
92.452349



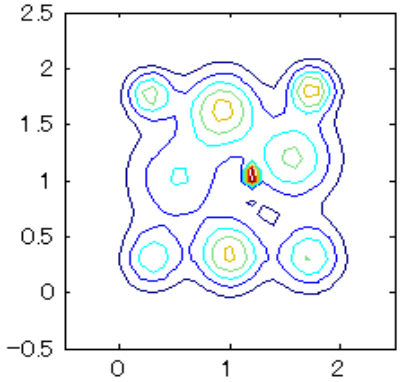
92.452349



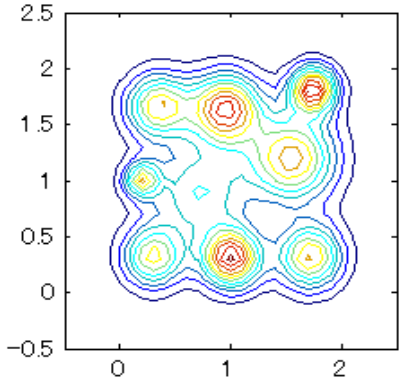
100.110198



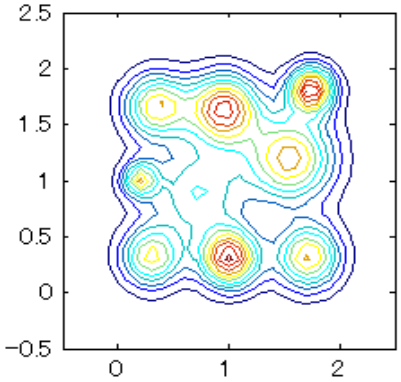
92.691436



96.076763



92.452349



92.452349

アルゴリズムの発展

(1) EM 法(1977)の問題点

EM法は最尤推定量をみつけようとする方法であるが
混合正規分布では最尤推定量は存在しないので
EM法が何をしているのかは実はよくわからない。
本気で尤度を大きくしてはだめ。いい感じでやめる。
データの揺らぎに対して弱く、汎化誤差も大きい。

(2) VB 変分ベイズ法(1999)の発見

EM法とよく似た学習アルゴリズムを与える変分ベイズ法は
ベイズ事後分布の平均場近似により導かれるが、
データの揺らぎに対して強く、汎化誤差も小さくできる。
自由エネルギーはいくら小さくしても大丈夫。
隠れ変数を持つシステムにおいて広く使われている。

変分ベイズ(VB) 法

混合正規分布の事前分布

(1) a については ハイパーパラメータ $(\alpha_1,\alpha_2,\ldots,\alpha_K)$ $(\alpha_k>0)$ により定まるディリクレ分布 (ただし $\alpha_k=\alpha>0$) を用いる。

$$\varphi(a|\alpha_k) \propto \Pi_k \left(a_k \right)^{\alpha_k-1}$$

(2) b についてはハイパーパラメータ (B_k,σ^2) により定まる正規分布(ただし$B_k=0$) を用いる。

$$\varphi(b|B_k,\sigma^2) \propto \Pi_k \exp\{ - (1/2\sigma^2) ||b_k-B_k||^2) \}$$

(注意) 変分ベイズ法では、事前分布をこのように設定すると、学習の結果得られる変分事後分布も $\varphi(a|\alpha_k)\ \varphi(b|B_k,\sigma^2)$ という形になることを示すことができる(事前分布と事後分布が同じ形式になる)。自己無矛盾条件の繰り返し代入は、(α_k,B_k) の更新則に帰着する。分散の学習も行う場合には分散の事前分布の設定も必要である。背後にある数理:ディリクレ分布と正規分布は「指数分布」であり事前分布として共役なものを選んだ。

変分ベイズ（VB）法

(1) パラメータ w の事前分布 φ(w)

(2) (w,{x_i}, {y_i}) の同時分布　　$\varphi(w) \prod_{i=1}^{n} p(x_i,y_i|w)$

(3) データ {x_i} が与えられたときの(w,{y_i})の分布

$$P(w,\{y_i\}|\{x_i\}) = (1/Z)\ \varphi(w) \prod_{i=1}^{n} p(x_i,y_i|w)$$

(4) この分布を (w,y={y_i}) が独立な分布で近似する

$H(w,y) = -\log \{ \varphi(w) \Pi\ p(x_i,y_i|w) \}$ と書くと

$P = \exp(- H(w,y)) / Z$　（Z は周辺尤度）

(5) KL情報量の最小化により最適化を行う。

平均場近似の自己無矛盾条件

確率分布 $p(w,y) = \exp(-H(w,y)) / Z$ とする。

KL情報量 $\int f(w)g(y) \log(f(w)g(y) / p(w,y)) dw\, dy$ を最小にする確率分布 $f(w)$, $g(y)$ は次の条件を満たす。

$$\begin{cases} f(w) \propto \exp(- \int H(w,y)\, g(y)\, dy) \\ g(y) \propto \exp(- \int H(w,y)\, f(w)\, dw) \end{cases}$$

この条件を満たす $(f(w),g(y))$のペアは一般的にはたくさんある。
KL情報量を最小にするものが目的のものである。

http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/var_basic.pdf

変分ベイズ（VB）法

(6) 自己無矛盾条件の繰り返し代入は、ハイパーパラメータの繰り返し代入に帰着する。この計算は数学的には難しくはないが、初めてだとメンドウくさいかもしれない。 ほとんどの本に計算が書いてあるが、次の本にも書いてある。

渡辺「ベイズ統計の理論と方法」2012,pp148-162.

(7) 初期値を変えて繰り返し D(f(w) g(y) || exp(-H(w,y))) を最小化。(= KL – log Z = F : 変分自由エネルギー）。
モデルの選択やハイパーパラメータの比較にも利用できる。
（汎化誤差の指標ではないが）。

変分ベイズ（VB）法

長所

(a)　複数の局所解のよさを比較できる。

(b)　変分自由エネルギーを規準にしてモデルや
ハイパーパラメータを決めることができる。

(c) 演算量が少ない。

短所

(d) 平均場近似理論は、初めての人には計算がメンドウ。

研究課題（カンタンではない可能性が高い）

(e) 局所解の数や性質は難しくてわかっていない。

VB法のアルゴリズム

0

α, σ： 変分ベイズ法のハイパーパラメータ
$\varepsilon > 0$ は十分小さく固定する。

1

α_k B_k t_k 初期化

Ψはディガンマ関数, N はデータの次元

2

$$L_{ik} = \Psi(\alpha_k) - N/(2t_k) - \|B_k/t_k - x_i\|^2/(2\sigma^2)$$

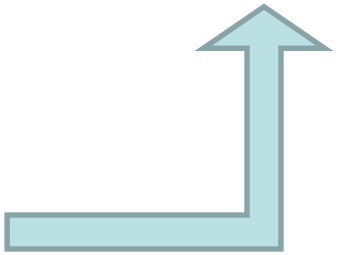
$$y_{ik} = \exp(L_{ik})/(\Sigma_k \exp(L_{ik}))$$

3

$$\alpha_k = \Sigma_i \; y_{ik} + \alpha$$

$$B_k = \Sigma_i \; y_{ik} x_i$$

$$t_k = \Sigma_i \; y_{ik} + \varepsilon$$



繰り返し

②③を繰り返すだけで学習できる

学習結果と評価

推定パラメータ:

$$a_k = \alpha_k / (n + \varepsilon K)$$

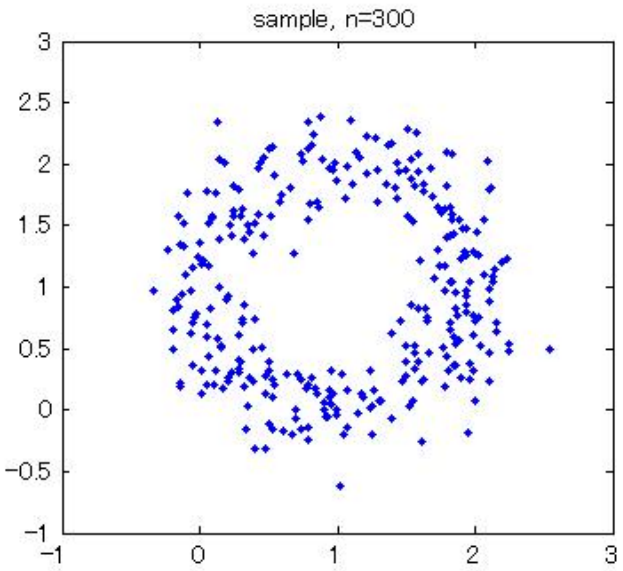
$$b_k = B_k / t_k$$

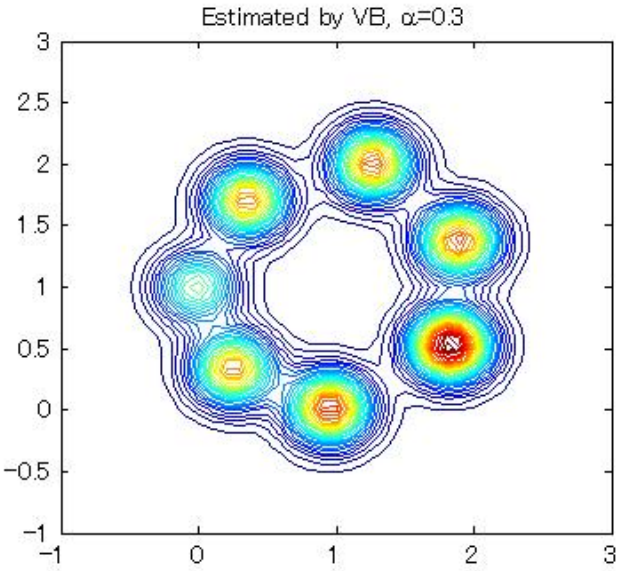
変分自由エネルギーはモデル・事前分布の適切さを示す。
最小とするハイパーパラメータが適切だと考えられる。

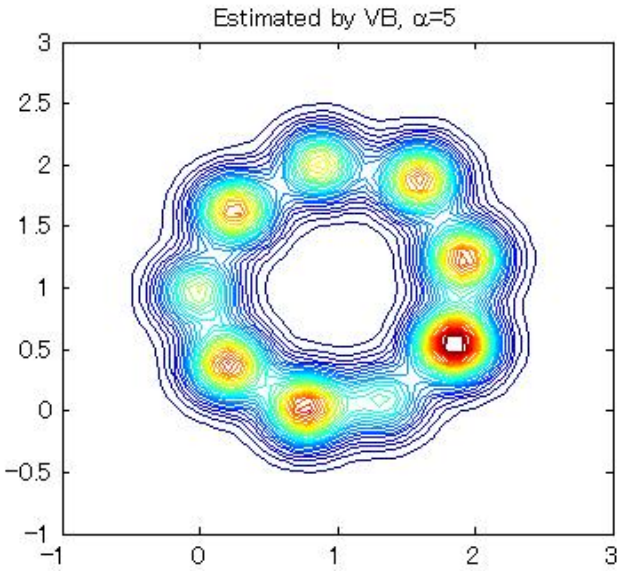
$$F = -\sum_k \log(\Gamma(\alpha_k)/\Gamma(\alpha)) + \log(\Gamma(n+K\alpha)/\Gamma(K\alpha))$$
$$- \sum_k \{ \|B_k\|^2/(2\sigma^2 t_k) - (N/2)\log t_k \}$$
$$+ \sum_i \|x_i\|^2/(2\sigma^2) + NK \log\sigma$$
$$+ \sum_i \sum_k y_{ik} \log y_{ik}$$

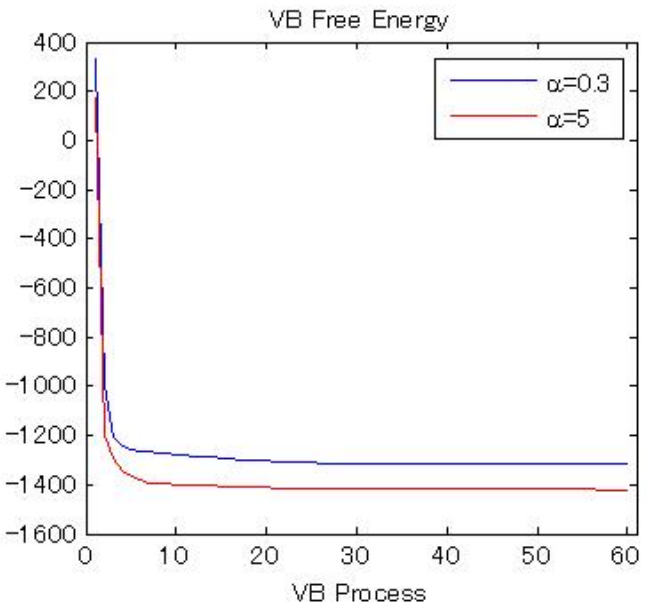
実験の例

事前分布は学習に影響する。









変分ベイズ法　あれこれ

(1)「パラメータと隠れ変数が独立な分布」から事後分布までのKL情報量を最小化する方法を平均場近似という。

(2) 事後分布の数値実現にはマルコフ連鎖モンテカルロ法が用いられるが、平均場近似は繰り返し代入法で実行できる。

(3) 平均場近似によるパラメータの分布は真の事後分布よりも縮んでいる。

(4) 平均場近似の自由エネルギーは真の自由エネルギーよりも少し大きい。ハイパーパラメータの変化による相転移が存在する（渡辺一帆）。

(5) 平均場近似の汎化誤差の挙動は未解決。縮小ランク回帰モデルではランダム行列理論を用いて数学的に解ける（中島伸一）。

(6) 自由エネルギー・汎化誤差・相転移は数値実験では解析できないので専門書でも論文でも誤った記載が多い（PRMLなども間違っている）。間違っていることを知らない研究者も多いので注意しましょう。直感や人間力だけではわかないことも多い。数理が役立つところのひとつ。

変分ベイズ法について解明されていること

(1) 渡辺一帆さんが、混合正規分布の学習に変分ベイズ法を適用する時の相転移の構造を解明されています（変分ベイズの理論研究のパイオニア）(2006)。

(2) 星野力さんが、確率文脈自由文法の自由エネルギーの漸近挙動を解明されています(2006)。

(3) 中島伸一さんが、変分ベイズ法における行列分解における相転移の構造を解明し汎化誤差の挙動を導出されています（2007)。

(4) 梶大介さんが、混合ベルヌーイ分布の学習に変分ベイズ法を適用するときの相転移の構造を解明されています(2011)。

(5) 中村文士さんが、混合正規分布の事前分布のハイパーパラメータを複数個にした場合の相転移構造を解明されています(2014)。

(6) 幸島匡宏さんが、非負値行列分解の相転移の構造を解明されています(2017)。

(注意)「物理を勉強しないと変分ベイズは研究できない」は誤り。

変分ベイズ法について研究したいかたに

渡辺一帆，Stochastic Complexities of Gaussian Mixtures
in Variational Bayesian Approximation, JMLR, Vol.7, 2006.
http://www.jmlr.org/papers/volume7/watanabe06a/watanabe06a.pdf
渡辺さんのホームページ：
http://www.lisl.cs.tut.ac.jp/wkazuho/index-j.html

中島伸一、変分ベイズ学習、機械学習プロフェッショナルシリーズ、
講談社、2016.
中島伸一、ランダム行列の数理と科学・第5章、森北出版、2014.
中島さんのホームページ：　https://sites.google.com/site/shinnkj23j/