階層的なモデルにおける
学習の目的関数の大域的性質

福水健次
情報・システム研究機構　統計数理研究所
総合研究大学院大学　複合科学研究科

アウトライン

- イントロ ― 階層的なパラメータを持つモデル
  - 多層パーセプトロン、有限混合モデル
- 階層的モデルの目的関数の持つ性質
- 臨界直線の存在
- 極小点／鞍点となるための十分条件
- ガウス混合モデルの例： コンポーネント分割法
- まとめ

イントロ： 階層的なパラメータを持つモデル

■ 非線形モデルの難しさ

- 目的関数の局所解
- モデルの特異性／識別不能性

  統計的解析の困難、学習ダイナミクスの複雑さ

一般に非線形モデルの理論解析は難しい

e.g.) 多層パーセプトロンの局所解の存在

■ 階層的モデル

- 3層パーセプトロン、有限混合モデルに共通な特殊な構造を用いて、目的関数の大局的な構造を探る

階層的モデル

■ 有限混合モデル

$$f_K(x\,|\,\theta^{(K)}) = \alpha_1\, g(x\,|\,\beta_1) + \cdots + \alpha_K\, g(x\,|\,\beta_K)$$

$g(x\,|\,\beta)$: **確率密度関数、** $\sum_{j=1}^{K} \alpha_j = 1, \quad \alpha_j \geq 0.$

■ 3層ニューラルネット

$$f_K(x\,|\,\theta^{(K)}) = \alpha_1\, \varphi(x\,|\,\beta_1) + \cdots + \alpha_K\, \varphi(x\,|\,\beta_K) + d$$

$\varphi(x\,|\,\beta)$: **非線形関数 (シグモイドなど)**

■ 階層的モデルの一般系

$$\boxed{f_K(x\,|\,\theta^{(K)}) = \alpha_1\, g(x\,|\,\beta_1) + \cdots + \alpha_K\, g(x\,|\,\beta_K) + \phi(\gamma)}$$

$\theta^{(K)} = (\alpha_1, \ldots, \alpha_K; \beta_1, \ldots, \beta_K; \gamma)$ ： **パラメータ**

$\sum_{j=1}^{K} \alpha_j = 1$ **の制約は** Lagrange**乗数法で扱う (以下では略)**

階層的モデルに共通の性質

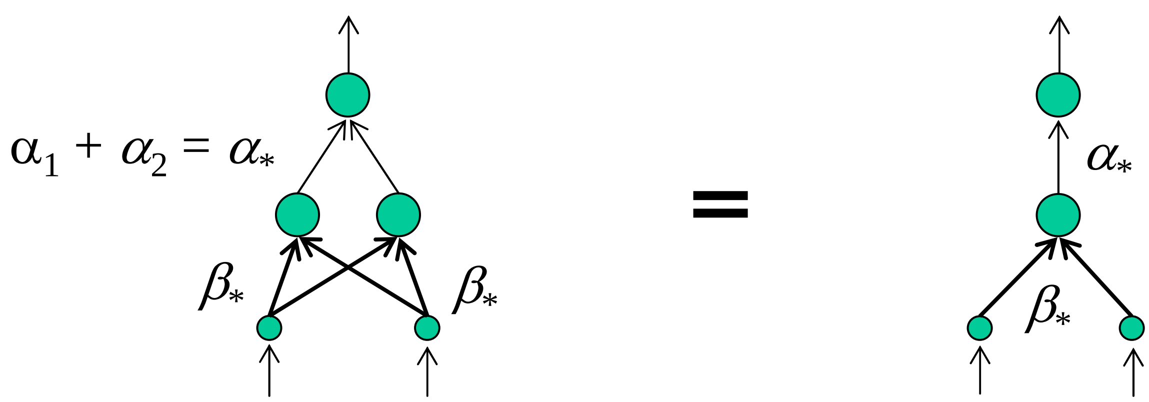
■ 同一のコンポーネント ⇒ サイズのひとつ小さいモデル

$$f_K(x\,|\,\alpha_1,\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\beta_{K-1},\beta_{K-1}})$$

$$=\alpha_1\,g(x\,|\,\beta_1)+\cdots+\alpha_{K-2}\,g(x\,|\,\beta_{K-2})+(\alpha_{K-1}+\alpha_K)g(x\,|\,\beta_{K-1})+\phi(\gamma)$$

$$=f_{K-1}(x\,|\,\alpha_1,\ldots,\alpha_{K-2},\underline{\alpha_{K-1}+\alpha_{K-2}};\beta_1,\ldots,\beta_{K-2},\beta_{K-1})$$

識別不能性の原因



階層的モデルに共通の性質2

■ 微分に関する性質

$$\frac{\partial f_K}{\partial \beta_{K-1}}(x \mid \alpha_1,\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\tilde{\beta},\tilde{\beta}}) = \boxed{\alpha_{K-1}}\underline{\frac{\partial g(x \mid \tilde{\beta})}{\partial \beta}}$$

$$\frac{\partial f_K}{\partial \beta_K}(x \mid \alpha_1,\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\tilde{\beta},\tilde{\beta}}) = \boxed{\alpha_K}\underline{\frac{\partial g(x \mid \tilde{\beta})}{\partial \beta}}$$

係数だけが異なる。

階層的モデルの目的関数

■ 目的関数の定め方

与えられたデータ $x_1,\ldots,x_n$ と損失関数 $\ell_1(t),\ldots,\ell_n(t)$ に対し

$$\boxed{\text{最小化の目的関数：}\quad L_K(\theta^{(K)}) = \sum_{i=1}^{n} \ell_i\left(f_K(x_i \mid \theta^{(K)})\right)}$$

例1) ニューラルネット： 教師データ y_i があるときは

$$\ell_i(t) = \left(y_i - t\right)^2$$

例2) 有限混合モデル： 最尤法

$$\ell_i(t) = \ell(t) = -\log(t)$$

階層的モデルの目的関数の性質

■ 目的関数の２つの性質

(P-1)
$$L_K(\alpha_1,\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\beta_{K-1},\beta_{K-1}})$$
$$=L_{K-1}(\alpha_1,\ldots,\alpha_{K-2},\underline{\alpha_{K-1}+\alpha_{K-2}};\beta_1,\ldots,\beta_{K-2},\beta_{K-1})$$

ひとつ小さいモデルとの関係式

(P-2)
$$\alpha_K\frac{\partial L_K}{\partial \beta_{K-1}}(x\,|\,\alpha_1\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\tilde{\beta},\tilde{\beta}})$$
$$-\alpha_{K-1}\frac{\partial L_K}{\partial \beta_K}(x\,|\,\alpha_1\ldots,\alpha_K;\beta_1,\ldots,\beta_{K-2},\underline{\tilde{\beta},\tilde{\beta}})=0$$

微分の線形従属性

Note: $$\frac{\partial L_K}{\partial \beta_j}(\theta^{(K)})=\sum_{i=1}^{n}\ell'_i(f_K(x_i\,|\,\theta^{(K)}))\frac{\partial f_K(x_i\,|\,\theta^{(K)})}{\partial \beta_j}$$

臨界直線の存在

■ 臨界点の埋め込み

$$\theta_*^{(K-1)} = (\alpha_{*,1}^{(K-1)}, \dots, \alpha_{*,K-1}^{(K-1)}; \beta_{*,1}^{(K-1)}, \dots, \beta_{*,K-1}^{(K-1)})$$

： $L_{K\text{-}1}(\theta^{(K\text{-}1)})$ の臨界点 (微分が0の点)

$\rho \in \mathbf{R}$ に対し、サイズ K のモデルのパラメータ θ_ρ を以下で定義

$$\alpha_{K-1}^{(K)} = \rho\alpha_{*,K-1}^{(K-1)}, \quad \alpha_K^{(K)} = (1-\rho)\alpha_{*,K-1}^{(K-1)}$$

$$\beta_{K-1}^{(K)} = \beta_K^{(K)} = \beta_{*,K-1}^{(K-1)}$$

$$\alpha_j^{(K)} = \alpha_{*,j}^{(K-1)}, \quad \beta_j^{(K)} = \beta_{*,j}^{(K-1)} \quad (1 \le j \le K-2)$$

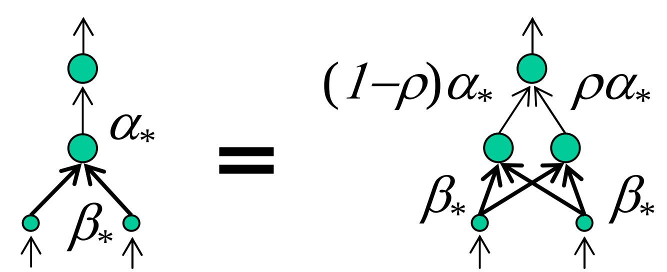
このとき、

$$f_K(x \mid \theta_\rho) = f_{K-1}(x \mid \theta_*^{(K-1)})$$

かつ

$$L_K(\theta_\rho) = L_{K-1}(\theta_*^{(K-1)})$$

(任意の ρ)



臨界直線の存在2

■ 臨界直線

定理1

$\alpha_{*,K-1}^{(K-1)} \neq 0$ のとき、任意の ρ に対し θ_ρ は $L_K(\theta^{(K\text{-}1)})$ の臨界点。

集合 $M_K(\theta^{(K-1)}{}_*) = \{\theta_\rho \mid \rho \in \mathrm{R}\}$ はパラメータ空間の中で直線 (線分) をなすので、臨界直線 (線分) と呼ぶ。

直接微分すると、すぐ証明できるのだが、、、

どういう点が特殊か？

臨界直線の存在3

■ 定理1の意味

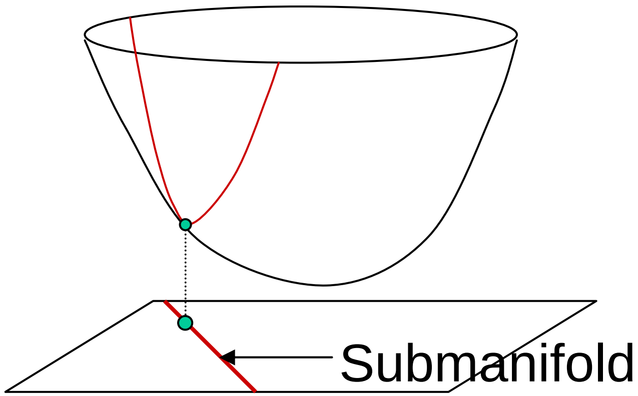
部分多様体上の臨界点は、

大きなパラメータ集合での臨界点とは

限らない。

　一般に、補空間方向の微分に

　関する情報はない。



ところが、階層モデルでは、

構造的に臨界点として埋め込まれている。

定理1の証明

性質(P-1)**を微分すると**

$$\begin{pmatrix} I_K & 0 & \cdots & & 0 \\ 0 & I_m & & O & \\ & & \ddots & & \vdots \\ & O & & \underline{I_m} & \underline{I_m} \end{pmatrix} \nabla L_K(\theta_\rho) = \begin{pmatrix} 1 & & O & 0 \\ & \ddots & & \vdots \\ O & & 1 & 0 \\ 0 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & I \end{pmatrix} \nabla L_{K-1}(\theta_*^{(K-1)}) = 0$$

一方、(P-2) **より**

$$\begin{pmatrix} 0 & \cdots & 0 & \underline{\alpha_K^{(K)} I_m} & \underline{-\alpha_{K-1}^{(K)} I_m} \end{pmatrix} \nabla L_K(\theta_\rho) = 0$$

$\alpha_{K-1}^{(K)} + \alpha_K^{(K)} = \alpha_{*,K-1}^{(K-1)} \neq 0$ **の仮定により、**

$$\frac{\partial}{\partial \alpha^{(K)}}, \frac{\partial}{\partial \beta_1^{(K)}}, \cdots, \frac{\partial}{\partial \beta_{K-2}^{(K)}}, \underline{\frac{\partial}{\partial \beta_{K-1}^{(K)}} + \frac{\partial}{\partial \beta_K^{(K)}}}$$

$$\underline{\alpha_K^{(K)} \frac{\partial}{\partial \beta_{K-1}^{(K)}} - \alpha_{K-1}^{(K)} \frac{\partial}{\partial \beta_K^{(K)}}}$$

は1階微分のすべてを張る。 $\Rightarrow$ $\nabla L_K(\theta_\rho) = 0$

定理1からわかること

■ 臨界直線は「常に」存在する

臨界直線の存在は、階層型モデルの構造のみ由来。
損失関数、コンポーネントの非線形関数、学習データには依らない性質。

■ 臨界直線は多数存在する

K-1 サイズのモデルに対し、分割するコンポーネントは任意。
Kサイズのモデルのどの2コンポーネントに分割してもよい。
$L_{K\text{-}1}$ の1個の臨界点に対し、$(K\text{-}1)\ {}_K\mathrm{C}_2$ 本の臨界直線。

■ 高次元の臨界アフィン部分集合も存在

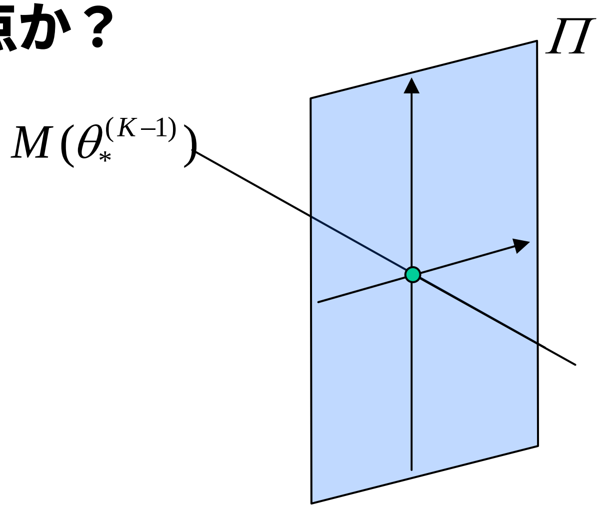
臨界直線をさらにサイズ K+1 に埋め込む ⇒ 臨界2次元平面　…

極小点/鞍点の十分条件

■ 極小点の埋め込み

$\theta^{(K\text{-}1)}_*$ ： $L_{K\text{-}1}(\theta^{(K\text{-}1)})$ の**孤立極小点**

極小点により作られた θ_ρ は、極小点か鞍点か？



直線 $M_K(\theta^{(K\text{-}1)}_*)$ 上では、

L_K **は一定値** (= $L_{K\text{-}1}(\theta^{(K\text{-}1)}_*)$)

⇒

$M_K(\theta^{(K\text{-}1)}_*)$ **上の各点で、この直線に直交する平面に制限した L_K の Hessian を考える。**

- Hessian **が正定値 → 極小点**
- Hessian **が負の固有値を持つ → 鞍点**

極小点/鞍点の十分条件2

定理2

$$R = \sum_{i=1}^{n} \ell'_i(f_{K-1}(x_i \mid \theta_*^{(K-1)})) \frac{\partial^2 g(x_i \mid \beta_{*,K-1}^{(K-1)})}{\partial\beta\partial\beta}$$ とおく。

(i) R が正負両方の固有値をもてば、直線 $M_K(\theta^{(K-1)}_*)$ 上の任意の点は鞍点。

(ii) R の固有値がすべて正(負)ならば、

$\alpha_{K-1}^{(K)}\alpha_K^{(K)}(\alpha_{K-1}^{(K)}+\alpha_K^{(K)}) > (<)\ 0$ を満たす点は極小点。

$\alpha_{K-1}^{(K)}\alpha_K^{(K)}(\alpha_{K-1}^{(K)}+\alpha_K^{(K)}) < (>)\ 0$ を満たす点は鞍点。

略証)

$\frac{\partial}{\partial\alpha^{(K)}}, \frac{\partial}{\partial\beta_1^{(K)}}, \cdots, \frac{\partial}{\partial\beta_{K-2}^{(K)}}, \frac{\partial}{\partial\beta_{K-1}^{(K)}} + \frac{\partial}{\partial\beta_K^{(K)}}$ による2階微分は $\nabla\nabla L_{K-1}(\theta_*^{(K-1)}) > 0$

$\alpha_K^{(K)}\frac{\partial}{\partial\beta_{K-1}^{(K)}} - \alpha_{K-1}^{(K)}\frac{\partial}{\partial\beta_K^{(K)}}$ による2階微分は $\alpha_{K-1}^{(K)}\alpha_K^{(K)}(\alpha_{K-1}^{(K)}+\alpha_K^{(K)})R$

3層ニューラルネットの場合

■ 鞍点と極小点の共存

行列 R が正定値、 $\alpha_{*,K-1}^{(K-1)} > 0$ と仮定する。

$$\alpha_{K-1}^{(K)} = \rho\alpha_{*,K-1}^{(K-1)}, \quad \alpha_{K}^{(K)} = (1-\rho)\alpha_{*,K-1}^{(K-1)}$$

で定まる臨界直線上の点 $\theta_\rho \in M_K(\theta^{(K\text{-}1)}{}_*)$ を考える

$\alpha_{K-1}^{(K)}\alpha_{K}^{(K)}(\alpha_{K-1}^{(K)} + \alpha_{K}^{(K)}) = (\alpha_{*,K-1}^{(K-1)})^3\rho(1-\rho)$ であるから、

(i) $0 < \rho < 1$ の線分上の点 ・・・ 極小点

(ii) $\rho \leqq 0,\ \rho \geqq 1$ の2線分上の点 ・・・ 鞍点

同一の関数値をとるパラメータ点だが、近傍での様子が異なる。

有限混合モデルの場合

系3

$$R=\sum_{i=1}^{n}\ell'_i(f_{K-1}(x_i\mid\theta_*^{(K-1)}))\frac{\partial^2 g(x_i\mid\beta_{*,K-1}^{(K-1)})}{\partial\beta\partial\beta}$$ とおく。

$\alpha_{*,K-1}^{(K-1)}>0$ のとき

(i) R が負の固有値をもてば、直線 $M_K(\theta^{(K\text{-}1)}{}_*)$ 上の任意の点は鞍点。

(ii) R の固有値がすべて正ならば、直線 $M_K(\theta^{(K\text{-}1)}{}_*)$ 上の任意の点は極小点。

＊α_j は常に非負

他のモデルの例

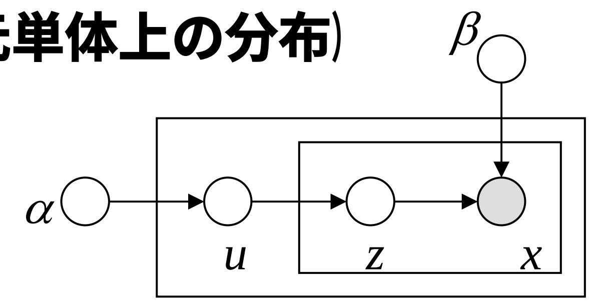
- Latent Dirichlet Allocation (Blei, Ng & Jordan 02)

$$f_K(x \mid \theta^{(K)}) = \int_{\Delta_{K-1}} D_K(u^{(K)} \mid \alpha^{(K)}) \prod_{\nu=1}^{M} \left(\sum_{j=1}^{K} u_j \, p(x_\nu \mid \beta_j) \right) du^{(K)}$$

$D_K(u^{(K)} \mid \alpha^{(K)})$ ：**ディリクレ分布** (K-1 **次元単体上の分布**)

$p(x \mid \beta)$ ：**任意の確率密度**

文書中の単語発生の確率モデル



(P-1)
$$L_K(\alpha_1, \ldots, \alpha_K; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1}, \beta_{K-1}) = L_{K-1}(\alpha_1, \ldots, \alpha_{K-2}, \alpha_{K-1} + \alpha_{K-2}; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1})$$

(P-2)'
$$\frac{\partial L_K}{\partial \beta_{K-1}}(\alpha_1 \ldots, \alpha_K; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1}, \beta_{K-1}) = \frac{\alpha_{K-1}}{\alpha_{K-1} + \alpha_K} \frac{\partial L_{K-1}}{\partial \beta_{K-1}}(\alpha_1 \ldots, \alpha_{K-1} + \alpha_K; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1})$$

$$\frac{\partial L_K}{\partial \beta_K}(\alpha_1 \ldots, \alpha_K; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1}, \beta_{K-1}) = \frac{\alpha_K}{\alpha_{K-1} + \alpha_K} \frac{\partial L_{K-1}}{\partial \beta_{K-1}}(\alpha_1 \ldots, \alpha_{K-1} + \alpha_K; \beta_1, \ldots, \beta_{K-2}, \beta_{K-1})$$

ガウス混合モデル

■ ガウス分布、PCA、FAの混合モデル

$$f_K(x;\theta^{(K)}) = \sum_{j=1}^{K} \alpha_j \, g(x \mid \beta_j),$$

$$g(x \mid \beta) = \frac{1}{\sqrt{(2\pi)^M \det V}} \exp\left\{-\tfrac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\right\} \qquad \beta = (\mu, V)$$

(MPCA や MFA の場合は, V に構造が入る)

定理4

ガウス分布、PCA、FAの混合モデルにおいては, K-1 モデルに対する尤度関数の孤立極大点は、K モデルにおける鞍点を与える。

＊分散共分散行列をパラメータに持つ必要がある。

行列 R_j の正の固有値に対応する固有方向に対して, 尤度は常に上昇

⇨ **コンポーネント分割法に応用**

コンポーネント分割法

■ 有限混合モデルに対するコンポーネント分割法

- EMアルゴリズムの初期値依存性
- 尤度無限大（分散→0）の問題 ⇒ ローカルサーチの必要性

コンポーネント分割・統合が有力な方法

■ 尤度を増大する分割法

$\theta^{(K\text{-}1)}{}_*$： $L_{K\text{-}1}(\theta^{(K\text{-}1)})$ の（孤立）極大点

$\zeta_{j*}\, g(x;\, \mu_{j*},\, V_{j*})$： $f^{(K\text{-}1)}(x;\, \theta^{(K\text{-}1)}{}_*)$ の j 番目のコンポーネント

分割法：

$$\alpha_{*,j}^{(K-1)} \to \quad \tfrac{1}{2}\alpha_{*,j}^{(K-1)},\ \tfrac{1}{2}\alpha_{*,j}^{(K-1)}$$

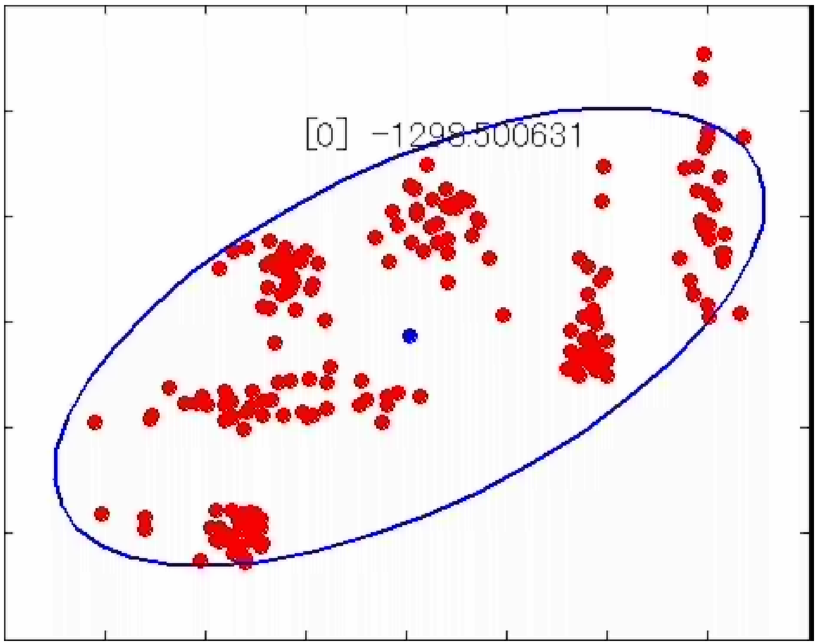
$$\mu_{*,j} \to \quad \mu_{*,j} - \varepsilon\Delta\mu_j,\ \mu_{*,j} + \varepsilon\Delta\mu_j, \qquad V_{*,j} \to \quad V_{*,j} - \varepsilon\Delta V_j,\quad V_{*,j} + \varepsilon\Delta V_j$$

$\left(\Delta\mu_j, \Delta V_j\right)$ ：R_j の最大固有値に対応する固有ベクトル

ε：小さい正の数

EM＋分割法の実験1

■ ガウス混合モデル



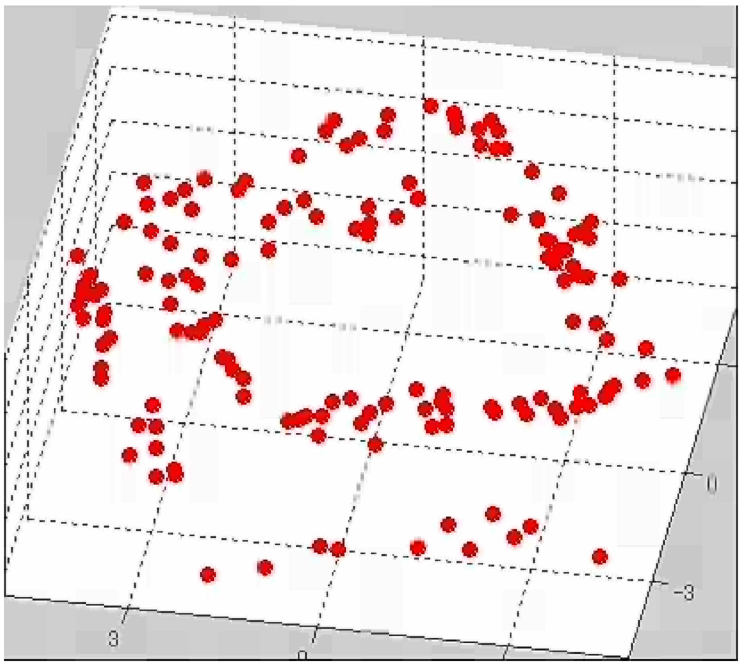
EM with random initialization
log like.：av. = – 1059.0 (30 trials)
Best (once) = –1030.9

Online-EM with random initialization
4 times failure (Singular V)
log like. : av. = – 1037.9 (26 trials)
best (7 times) = –1021.2

Online-EM + component splitting
30times – 1021.2

EM+分割法の実験2

■ PCAの混合モデル

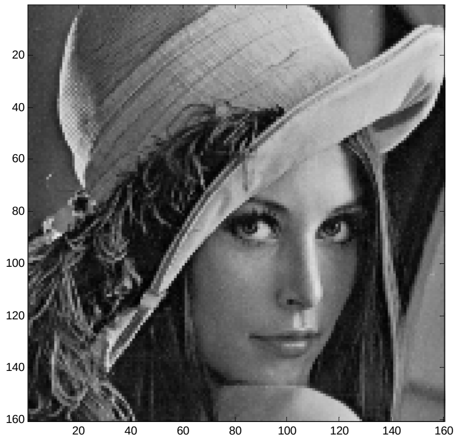


150 data, 3 dim.
Mixture of PCA
up to 8 components of rank 1

Online-EM with random initialization
log like. : av. = – 583.9 (30 trials)
best (6times) = – 534.9
worst = – 648.1

Online-EM + component splitting
log like. : av. = – 541.3 (30 trials)
best (26times) = – 534.9
worst = – 587.9

■ 実データ



“Lenna”

160 x 160 pixels
8 x 8 block = 64 dimensional vector
400 data in total

Mixture of PCA with 10 components of rank 4.

Image compression
Choose argmin$_j$ $\| X - \mu_j \|$
Reconstruction according to

$$\hat{X} = \mu_j + F_j (F_j^T F_j)^{-1} F_j^T (X - \mu_j)$$

Residual square error: $\sum_{n=1}^{400} \left\| \hat{X}_n - X_n \right\|^2$

Results (10 trials, RSE:x10^3)

EM with random initialization
Best 5.94
Worst 6.40
Av. 6.15

EM with component split
Best 5.38
Worst 6.12
Av. 5.78

まとめ

- **階層型モデルの目的関数が持つ大域的性質**
  - 有限混合モデル、多層パーセプトロンなどのモデルの目的関数は，構造から来る特別な性質を持っている。
  - 小さいモデルの臨界点が，大きいモデルの臨界直線、臨界アフィン集合として埋め込まれる。
  - 臨界直線，臨界アフィン集合は、常に多数存在している。

- **鞍点/極小点の十分条件**
  - 臨界直線上の点が，極小点/鞍点になる十分条件が容易に得られる。
  - ガウス混合モデル/MPCA/MFAでは、最尤点からコンポーネントを分割すると、尤度を上昇させる方向が常に存在する。

- **EM＋コンポーネント分割法**

参考文献

- K. Fukumizu and S. Amari. Local Minima and Plateaus in Hierarchical Structures of Multilayer Perceptrons. *Neural Networks*, 13(3) 317—327, 2000.
- K. Fukumizu, S. Akaho, and S. Amari. Critical Lines in Symmetry of Mixture Models and its Application to Component Splitting. *Advances in NIPS 15* (2003).
- **福水,栗木,竹内,赤平. 特異モデルの統計学 (統計科学のフロンティア 7) 岩波書店(2004).**