

混合ガウスモデル

Gaussian Mixture Model

GMM

明治大学 理工学部 応用化学科
データ化学工学研究室 金子 弘昌

GMM とは？

- ✓ クラスタリング手法の一つ
- ✓ 与えられたデータセットを、複数の正規分布の重ね合わせで表現する
- ✓ 確率密度関数が得られる (確率分布として表現できる)
- ✓ サンプルごとに、各クラスターに所属する確率が得られる
- ✓ クラスター数を自動的に決められる

どんなときに GMM を使うか？

✓理想

- データセットが、複数の正規分布の重ね合わせで表現できることが分かっているとき

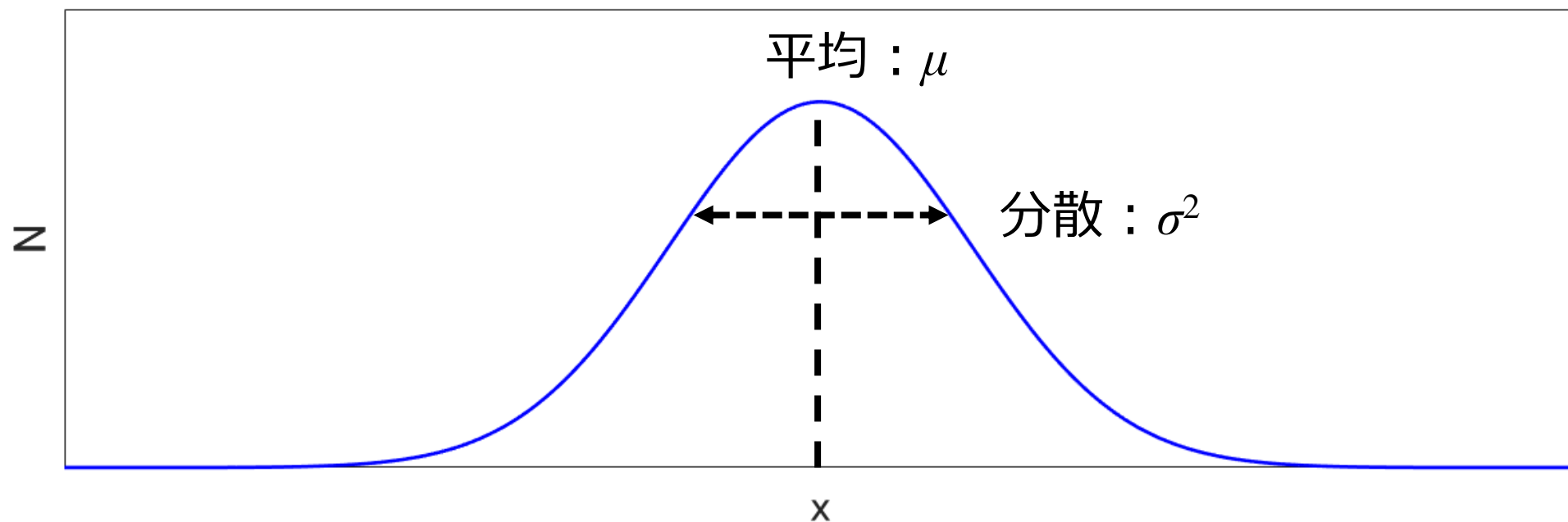
✓現実

- クラスターを自動的に決めながらクラスタリングしたいとき
- データセットの確率密度関数が欲しいとき
 - 確率密度関数の応用例)
 - 確率密度関数に基づいたサンプリング
 - 説明変数 X の事前分布として利用

正規分布 (ガウス分布, Gaussian distribution)

3

✓ データが、平均値付近に一番固まっていて、ばらつきのある確率分布



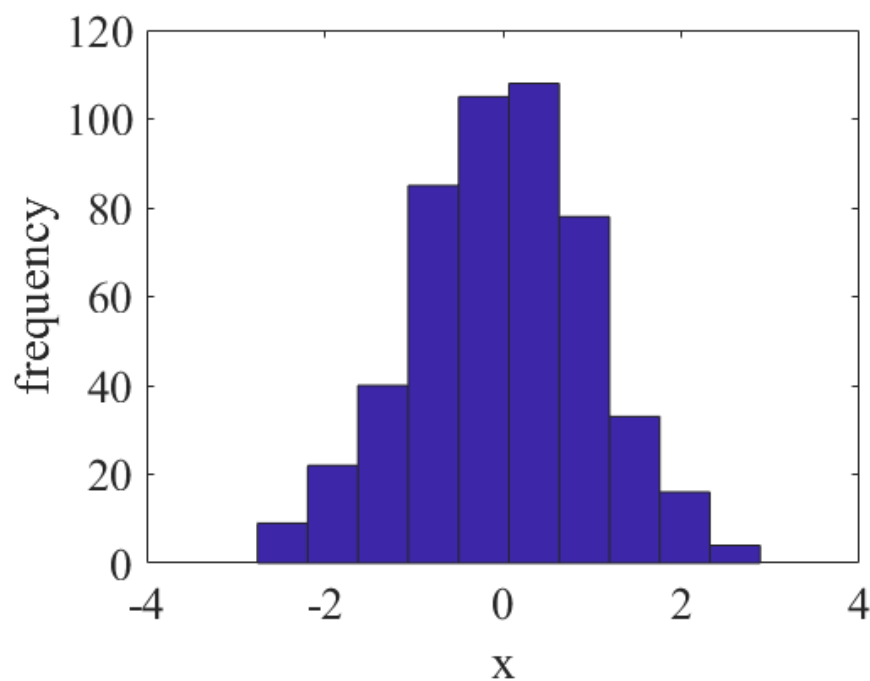
$$N(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

正規分布の例

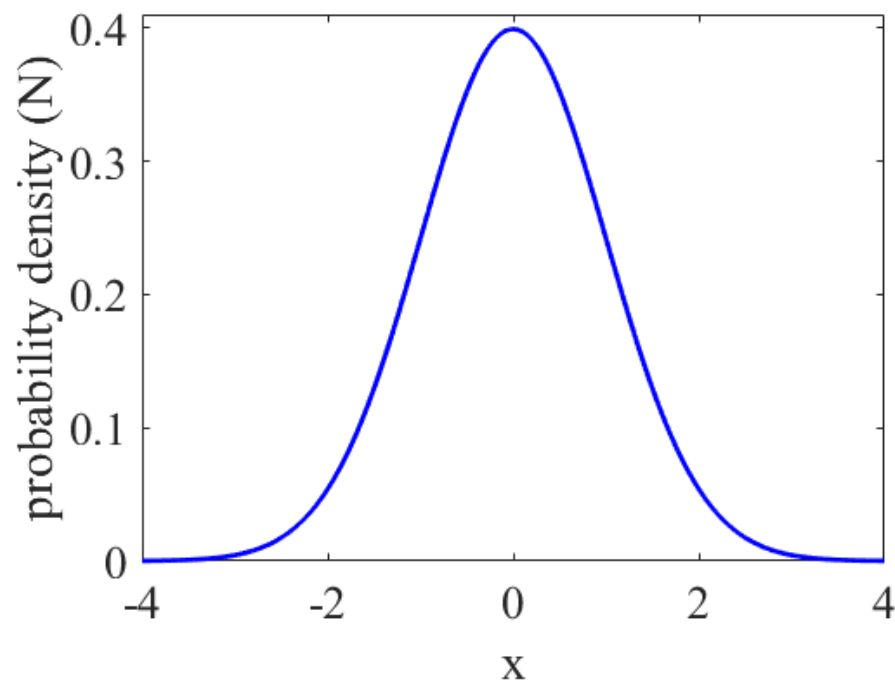
$$\checkmark \mu = 0$$

$$\checkmark \sigma = 1$$

$$N(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$



ヒストグラム



確率密度関数

多変量正規分布

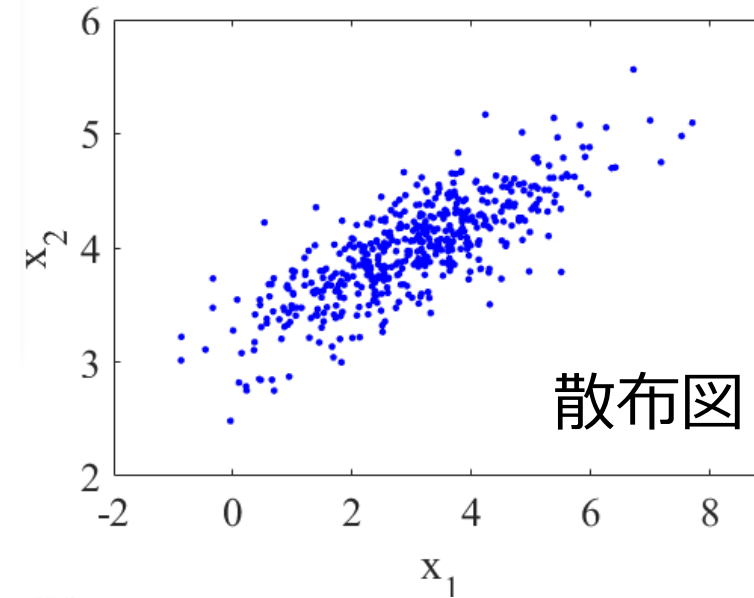
- ✓ 正規分布を複数の変数 (x_1, x_2, x_3, \dots) がある場合に拡張したもの
- ✓ 各変数の平均・分散だけでなく、変数間の共分散も必要
 - x_1 と x_2 の共分散が 2 とか
- ✓ 変数の数を m とすると、

$$N(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{m}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

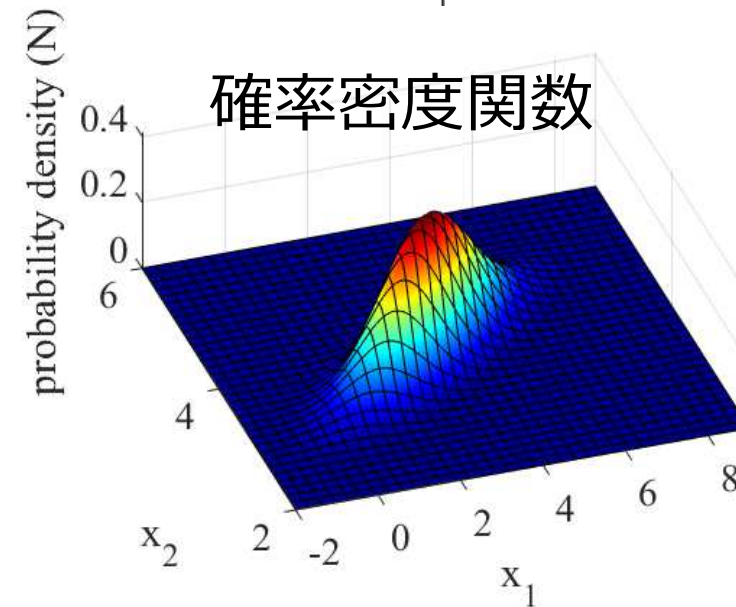
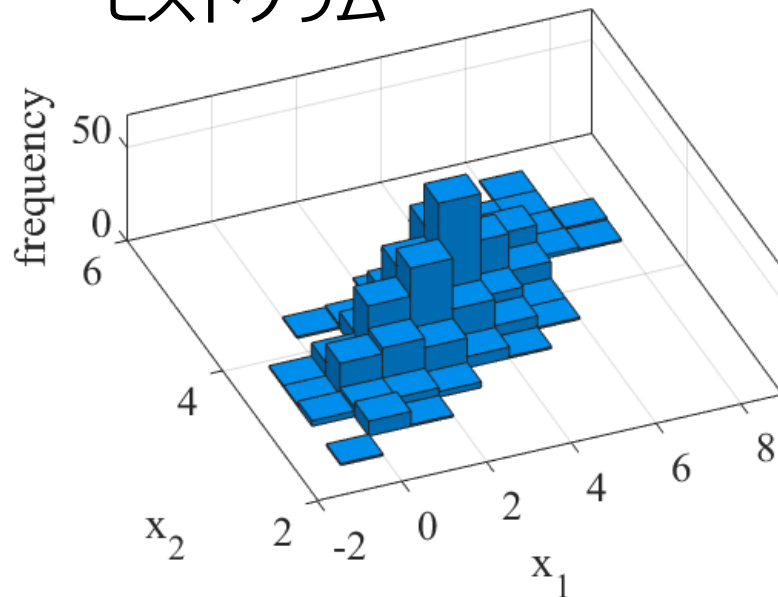
- $\mathbf{x} : [x_1, x_2, x_3, \dots, x_m]$
- $\boldsymbol{\mu} : 1 \times m$ の平均ベクトル
- $\boldsymbol{\Sigma} : m \times m$ の分散共分散行列

多変量正規分布の例 2変数

- ✓ x_1 の平均 3, 分散 2
- ✓ x_2 の平均 4, 分散 0.2
- ✓ x_1 と x_2 の共分散 0.5

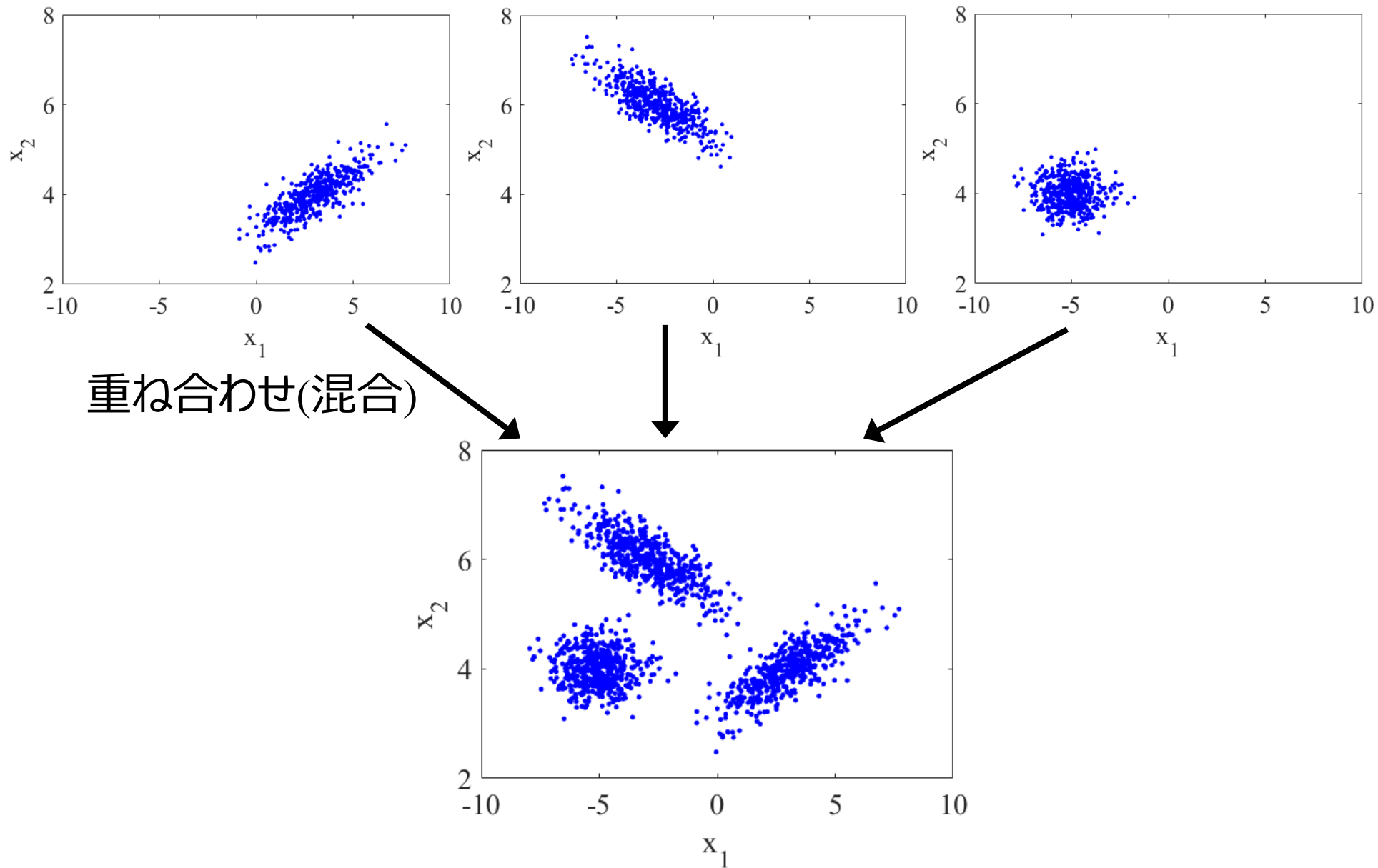


ヒストグラム

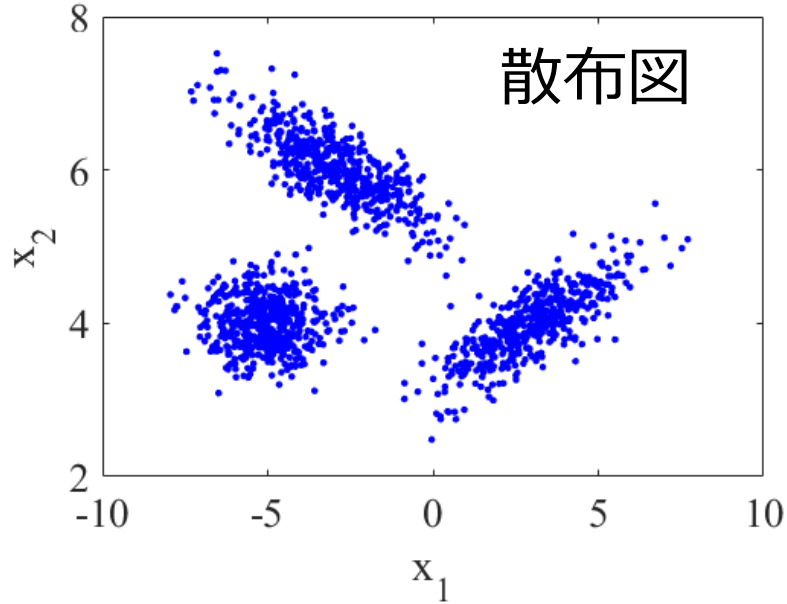


(多変量)正規分布の重ね合わせとは？ 例

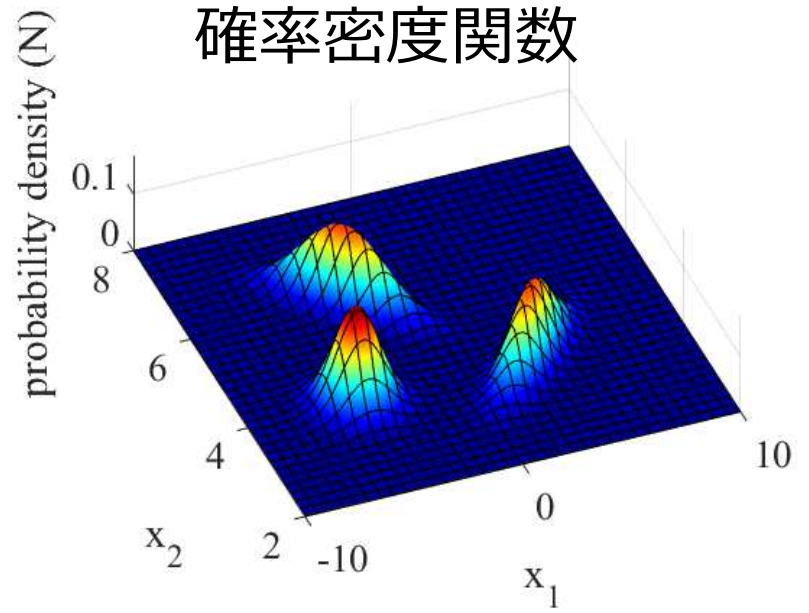
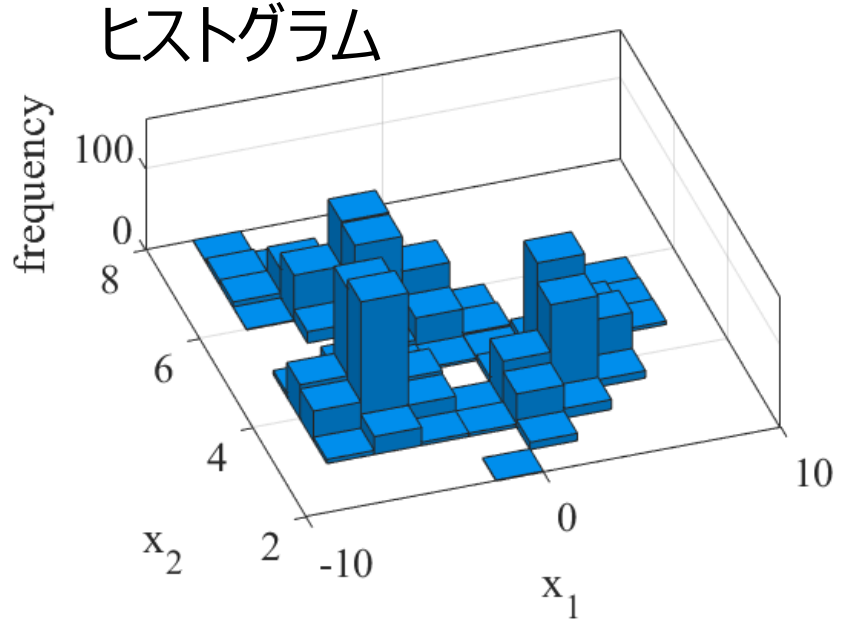
7



混合正規分布 (混合ガウス分布)



混合正規分布
(混合ガウス分布,
mixtures of Gaussians)



混合正規分布 (混合ガウス分布) 式

✓ 変数の数を m , 正規分布の数を n とすると、

$$p(\mathbf{x}) = \sum_{k=1}^n \pi_k N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{m}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}$$

- $\mathbf{x} : [x_1, x_2, x_3, \dots, x_m]$
- $\boldsymbol{\mu}_k : k$ 番目の正規分布における $1 \times m$ の平均ベクトル
- $\boldsymbol{\Sigma}_k : k$ 番目の正規分布における $m \times m$ の分散共分散行列
- $\pi_k : \text{混合係数 (各正規分布の重み)}$

$$\sum_{k=1}^n \pi_k = 1$$

データセットが与えられたとき、最尤推定法で

- μ_k : k 番目の正規分布における $1 \times m$ の平均ベクトル
- Σ_k : k 番目の正規分布における $m \times m$ の分散共分散行列
- π_k : 混合係数 (各正規分布の重み)

を求めよう！

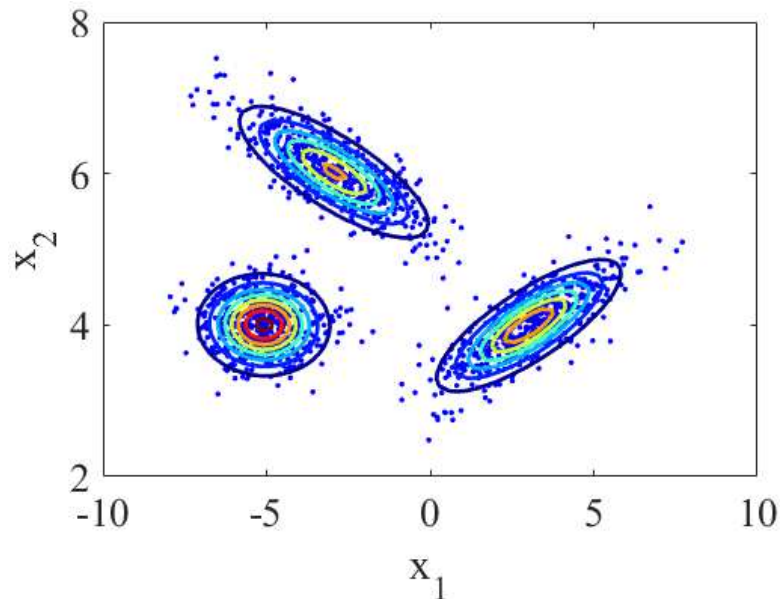
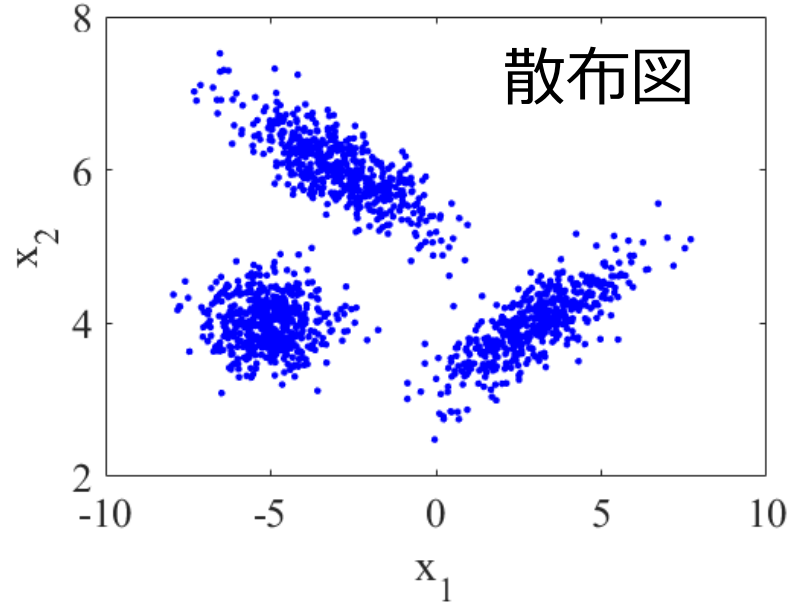
最尤推定法については、

<http://datachemeng.com/maximumlikelihoodestimation/> にあります

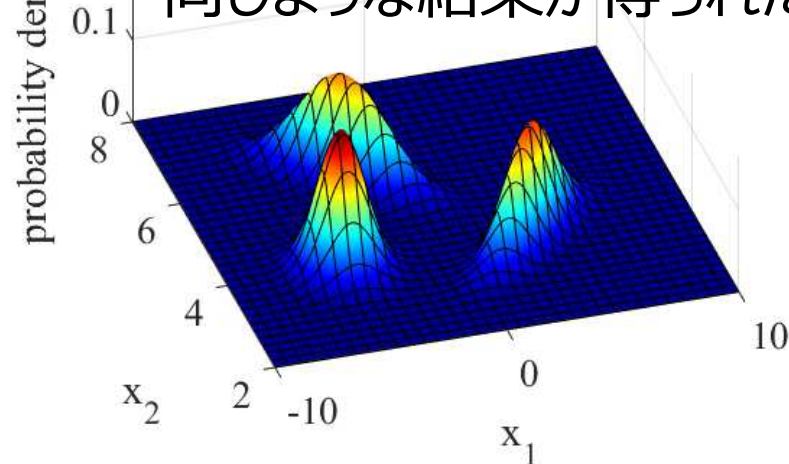
具体的な求め方については、p. 18 以降の [補足] にあります

実際に GMM をやってみる

右のデータセットを用いて
 $n = 3$ としてGMMを行うと、



p. 8 にある実際の確率密度関数と
同じような結果が得られた！



各サンプルがどのクラスターになるか考える 1/3 ¹²

- ✓ GMM では、各サンプルの割り当てられた正規分布が、そのサンプルのクラスター
 - n 個の正規分布があるとき、クラスター数も n 個ある

✓ クラスター変数 z を用いる

✓ ある k 番目の z_k だけ値が 1 で、他は 0

✓ $z_k = 1$ のとき、 k 番目のクラスターに属するという事

✓ サンプルに関する情報がないとき、 $z_k = 1$ となる確率は π_k (混合係数)

$$p(z_k = 1) = \pi_k$$

各サンプルがどのクラスターになるか考える 2/3 ¹³

知りたいのは、あるサンプル \mathbf{x} が与えられたときに、 $z_k = 1$ となる確率

$$p(z_k = 1 | \mathbf{x})$$

ベイズの定理より、

$$\begin{aligned} p(z_k = 1 | \mathbf{x}) &= \frac{p(z_k = 1) p(\mathbf{x} | z_k = 1)}{\sum_{i=1}^n p(z_i = 1) p(\mathbf{x} | z_i = 1)} \\ &= \frac{\pi_k p(\mathbf{x} | z_k = 1)}{\sum_{i=1}^n \pi_i p(\mathbf{x} | z_i = 1)} \end{aligned}$$

各サンプルがどのクラスターになるか考える 3/3 ¹⁴

$p(\mathbf{x} | z_k = 1)$ とは、 $z_k = 1$ 、つまり k 番目の正規分布、における

\mathbf{x} の確率 $N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

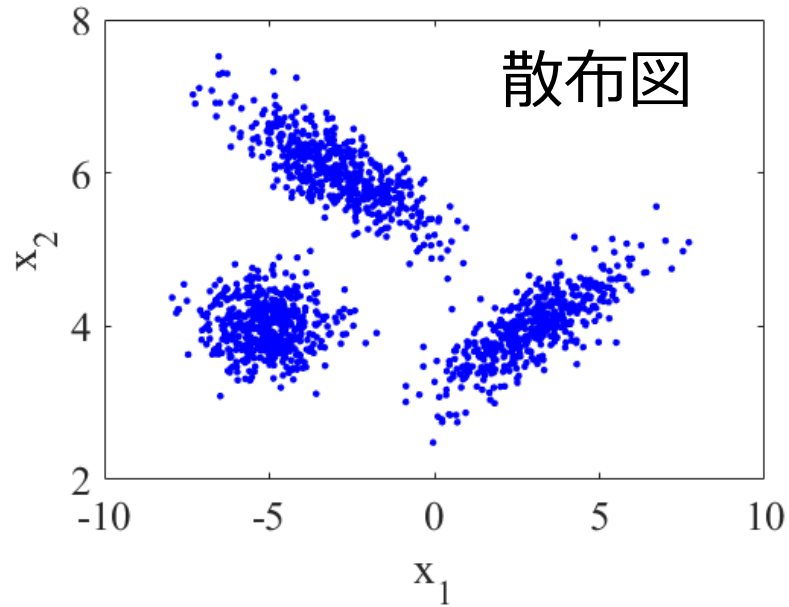
よって、

$$p(z_k = 1 | \mathbf{x}) = \frac{\pi_k N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

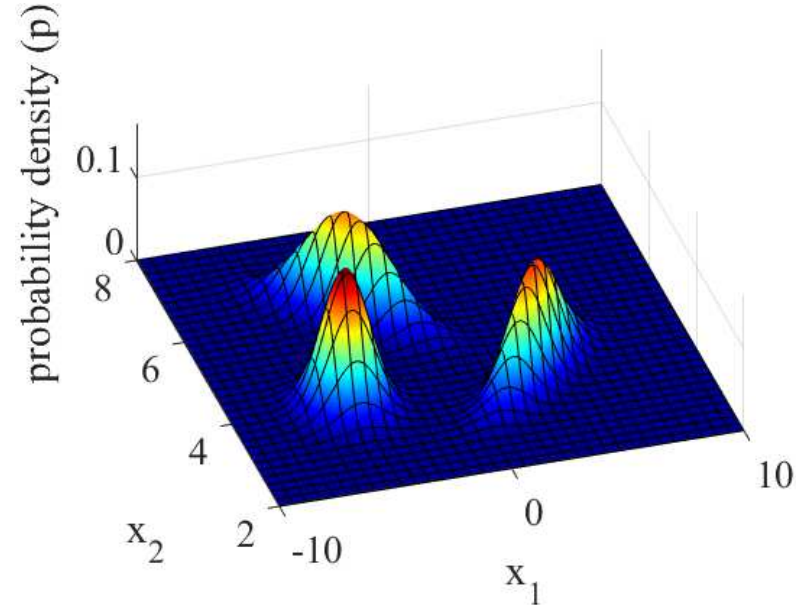
k について、1 から n まで計算し、最も大きい $p(z_k = 1 | \mathbf{x})$ をもつ

クラスターを、 \mathbf{x} が属するクラスターとする

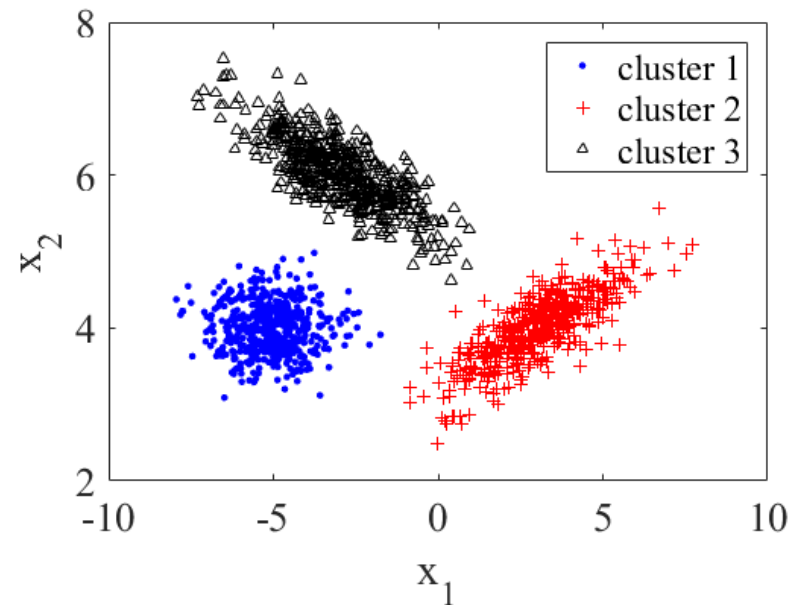
実際にクラスターを割り振る



GMM



各サンプルにクラスターを割り振ると、



クラスター数をどう決めるか？

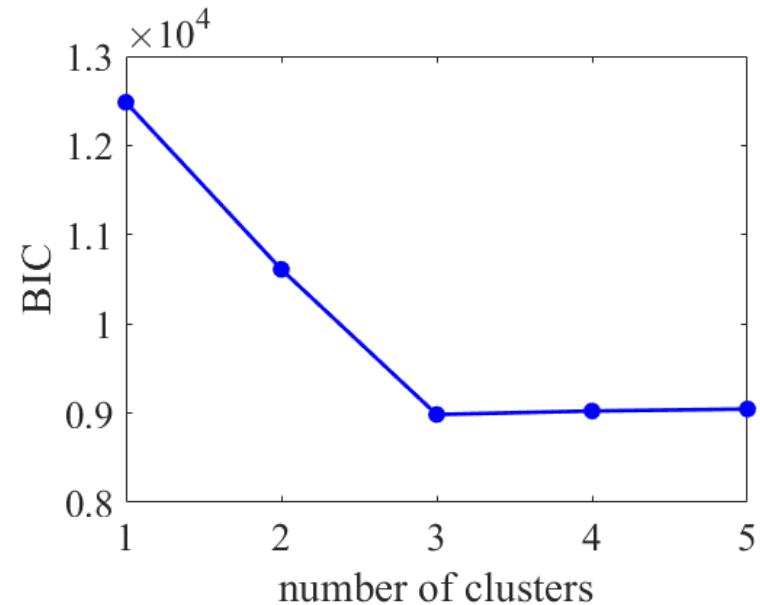
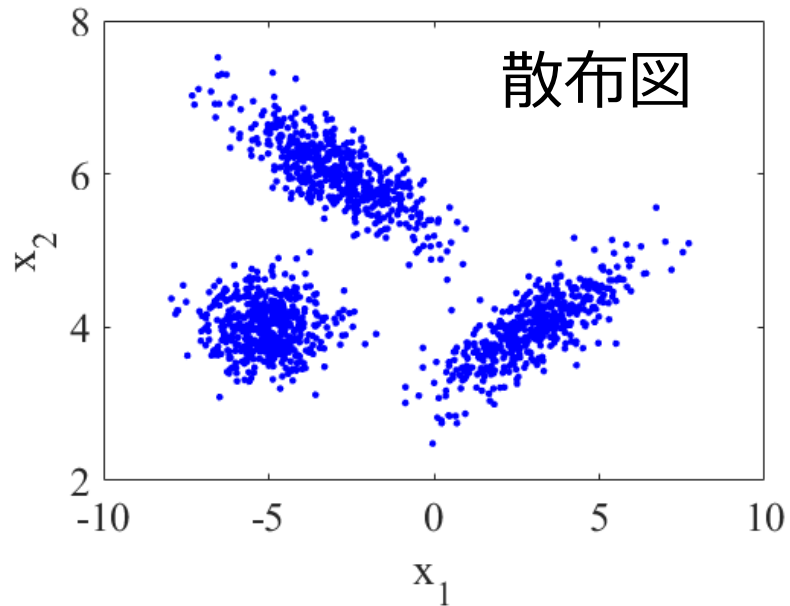
- ✓ クラスター数を 1, 2, 3, ... と振って GMM を行い、それぞれ
ベイズ情報量規準 (Bayesian Information Criterion, BIC) を
計算する

$$BIC = -2 \log L + M \log N$$

- L : 尤度 (<http://datachemeng.com/maximumlikelihoodestimation/>)
 - M : 推定するパラメータの数
 - 今回は詳細を記載しないが、分散共分散行列 Σ_k に制限を
与えることで、 M が変化する (制限しないときは考えなくてよい)
 - N : サンプル数
-
- ✓ BIC の値が最小となるクラスター数とする
 - ✓ データセットを確率密度関数として表せるため、最適クラスター数の
推定ができる

ベイズ情報量規準 (BIC) を計算してみた

17



少し見えにくいですが、クラスター数が3でBICの値が最小になっており、適切なクラスター数を推定できた

[補足] EM アルゴリズム 対数尤度関数

GMM のパラメータ推定には、EM (Expectation-Maximization) アルゴリズムが用いられることが多い

対数尤度関数 (<http://datachemeng.com/maximumlikelihoodestimation>)

$$\begin{aligned}\log L(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log \left\{ \prod_{j=1}^N \sum_{k=1}^n \pi_k N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \\ &= \sum_{j=1}^N \log \left\{ \sum_{k=1}^n \pi_k N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}\end{aligned}$$

[補足] EM アルゴリズム 最大 → 極大

対数尤度関数が、 μ_k, Σ_k, π_k それぞれで最大になるために満たされるべき条件を探す

最大 → 極大

対数尤度関数を μ_k, Σ_k, π_k それぞれで微分して 0 とする

ただし、 π_k は制約条件 $\sum_{k=1}^n \pi_k = 1$ があるため、

Lagrange の未定乗数法を用いる

[補足] EM アルゴリズム μ で微分

対数尤度関数を μ_k で微分して 0 とすると、

$$\sum_{j=1}^N \frac{\pi_k N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_k) = 0$$

上の式中の $\frac{\pi_k N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$ は、p. 14 における、

\mathbf{x}_j が与えられたときの正規分布 k の事後確率に等しい

これを、負担率 $\gamma(z_{j,k})$ をする

$$\gamma(z_{j,k}) = \frac{\pi_k N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \quad \text{とすると、}$$

$$\sum_{j=1}^N \gamma(z_{j,k}) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_k) = 0 \quad \boldsymbol{\Sigma}_k^{-1} \text{ を左からかけると、}$$

$$\sum_{j=1}^N \gamma(z_{j,k}) (\mathbf{x}_j - \boldsymbol{\mu}_k) = 0$$

[補足] EM アルゴリズム μ の計算

よって、

$$\boldsymbol{\mu}_k = \frac{\sum_{j=1}^N \gamma(z_{j,k}) \mathbf{x}_j}{\sum_{j=1}^N \gamma(z_{j,k})} = \frac{\sum_{j=1}^N \gamma(z_{j,k}) \mathbf{x}_j}{N_k}$$

ここで、 $N_k = \sum_{j=1}^N \gamma(z_{j,k})$ は、 k 番目のクラスターに

割り当てられたサンプル数

[補足] EM アルゴリズム Σ の計算

対数尤度関数を Σ_k で微分して 0 とする

整理すると、

$$\Sigma_k = \frac{1}{N_k} \sum_{j=1}^N \gamma(z_{j,k}) (\mathbf{x}_j - \boldsymbol{\mu}_k) (\mathbf{x}_j - \boldsymbol{\mu}_k)^T$$

[補足] EM アルゴリズム π の計算

π_k について、Lagrange の未定乗数法より、

$$G = \log L(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^n \pi_k - 1 \right) \quad \text{を最大化する}$$

G を π_k で微分して 0 とすると、

$$\sum_{j=1}^N \frac{N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} + \lambda = 0$$

[補足] EM アルゴリズム Ⅱ

$$\sum_{j=1}^N \frac{N(\mathbf{x}_j | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^n \pi_i N(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} + \lambda = 0$$

両辺に π_k をかけて k について和を取ると、 $\sum_{k=1}^n \pi_k = 1$ より、

$$\lambda = -N$$

これを使い、一番上の式 π_k をかけて変形すると、

$$\pi_k = \frac{N_k}{N}$$

[補足] EM アルゴリズム まとめ

- ① $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ を初期化する
- ② E ステップ: 負担率 $\gamma(z_{j,k})$ を計算する
- ③ M ステップ: 負担率 $\gamma(z_{j,k})$ を用いて、 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ を再計算する

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{\sum_{j=1}^N \gamma(z_{j,k}) \mathbf{x}_j}{N_k} \quad \pi_k^{\text{new}} = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{j=1}^N \gamma(z_{j,k}) (\mathbf{x}_j - \boldsymbol{\mu}_k^{\text{new}}) (\mathbf{x}_j - \boldsymbol{\mu}_k^{\text{new}})^T$$

- ④ ②③ を繰り返す

[補足] 変分ベイズ法 μ, Σ の事前分布

✓ GMM のパラメータを変分ベイズ法で推定
→ Variational Bayesian GMM

✓ μ, Σ の事前分布として Gaussian–Wishart 分布を導入

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{k=1}^n N(\boldsymbol{\mu}_k \mid \mathbf{m}_0, \alpha_0 \boldsymbol{\Sigma}_k) W(\boldsymbol{\Sigma}_k \mid \mathbf{W}_0, \beta_0)$$

W : Wishart 分布

\mathbf{m}_0 : 正規分布の平均 (X の平均値)

α_0 : 正規分布の分散共分散行列のパラメータ ($\alpha_0 = 1$)

\mathbf{W}_0, β_0 : Wishart 分布のパラメータ (\mathbf{W}_0 は X の分散共分散行列、
 β_0 は X の変数の数)

[補足] 変分ベイズ法 π の事前分布

✓ π の事前分布としてディリクレ分布やディリクレ過程を導入

- ディリクレ分布

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\gamma}_0) \quad \begin{array}{l} \text{Dir} : \text{ディリクレ分布} \\ \boldsymbol{\gamma}_0 : \text{ディリクレ分布のハイパーパラメータ} \end{array}$$

- ディリクレ過程

- 無限次元のディリクレ分布と考えることができ、stick-breaking 過程や中華料理店過程で推定

$$\pi_k = \nu_k \prod_{i=1}^k (1 - \nu_i) \quad \begin{array}{l} \text{Beta} : \text{ベータ分布} \\ \boldsymbol{\gamma}_0 : \text{ベータ分布のハイパーパラメータ} \end{array}$$

$$\nu_k \sim \text{Beta}(1, \boldsymbol{\gamma}_0)$$

- 詳細は D.M. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures. Bayesian Anal. 1 (2006) 121–143.

参考文献

- ✓ C.M. ビショップ, パターン認識と機械学習 下, 丸善出版 (2012)
- ✓ D.M. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures. *Bayesian Anal.* 1 (2006) 121–143.