GMMとEMアルゴリズム

本資料では，識別モデルの代表的な方法であるGMMを紹介するとともに，そのパラメタ推定に必要となる，EMアルゴリズムについて解説する。

生成モデルによるパターン認識とは
MAP復号と最尤復号
混合正規分布 (GMM)
最尤推定とEMアルゴリズム

§ 識別モデル と 生成モデル

パターン認識には，識別モデルという考え方と，生成モデルという考え方がある。

識別モデルに基づくパターン認識：

- 特徴空間に対し，各クラスの境界を求めておき，クラス未知のデータの特徴ベクトルが，特徴空間上のどこに位置するか（どのクラスのクラス境界内に位置するか）を調べることでデータのクラスを定める方法。

生成モデルに基づくパターン認識

- クラス毎に，特徴ベクトルの出力確率を与える，確率モデルを用意し，クラス未知のデータの特徴ベクトルが，それぞれの確率モデルから出力する確率を調べることでクラスを定める方法。

（クラス境界は，各クラスの確率モデルの与える出力確率を比較することで得られる。識別モデルによる方法より，１ステップ手続きが多い。）

§ 識別モデル と 生成モデル

◘ 識別モデルに基づくパターン認識：

- 特徴空間に対し，各クラスの境界を求めておき，クラス未知のデータの特徴ベクトルが，特徴空間上のどこに位置するか（どのクラスのクラス境界内に位置するか）を調べることでデータのクラスを定める方法。

識別モデルに基づくパターン認識では，認識対象データの特徴ベクトルの位置をもとに，そのクラスを直接出力する。これまでにこの授業で扱ってきたニューラルネットやSVMは識別モデルにあたる。

◘ 生成モデ

- クラス毎に，特徴ベクトルの出力確率を与える，確率モデルを用意し，クラス未知のデータの特徴ベクトルが，それぞれの確率モデルから出力する確率を調べることでクラスを定める方法。

（クラス境界は，各クラスの確率モデルの与える出力確率を比較することで得られる。識別モデルによる方法より，１ステップ手続きが多い。）

§ 識別モデル と 生成モデル

識別モデルに基づくパターン認識：

- 特徴空間に対し，各クラスの境界を求めておき，クラス未知のデータの特徴ベクトルが，特徴空間上のどこに位置するか（どのクラスのクラス境界内に位置するか）を調べることでデータのクラスを定める方法。

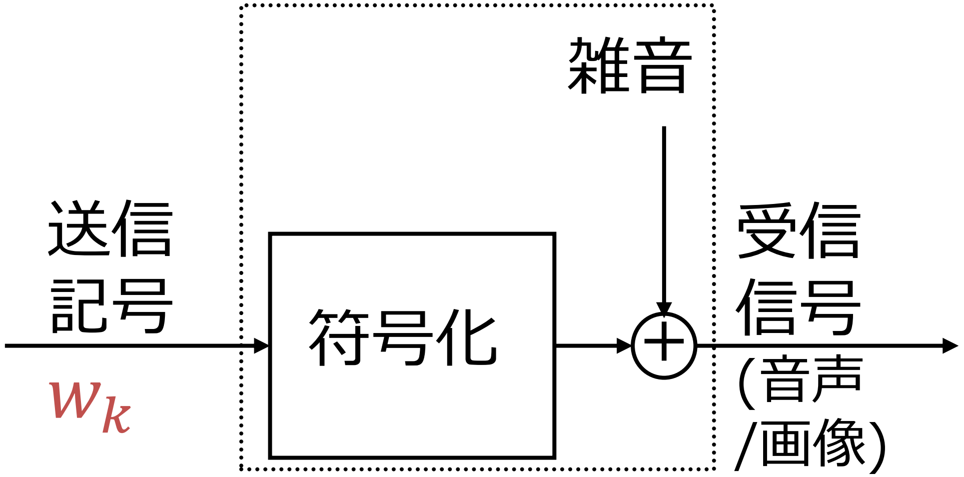
生成モデルに基づくパターン認識

- クラス毎に，特徴ベクトルの出力確率を与える，確率モデルを用意し，クラス未知のデータの特徴ベクトルが，それぞれの確率モデルから出力する確率を調べることでクラスを定める方法。

（クラス境界は，各クラスの確率モデルの与える出力確率を比較することで得られる。識別モデルによる方法より，１ステップ手続きが多い。）

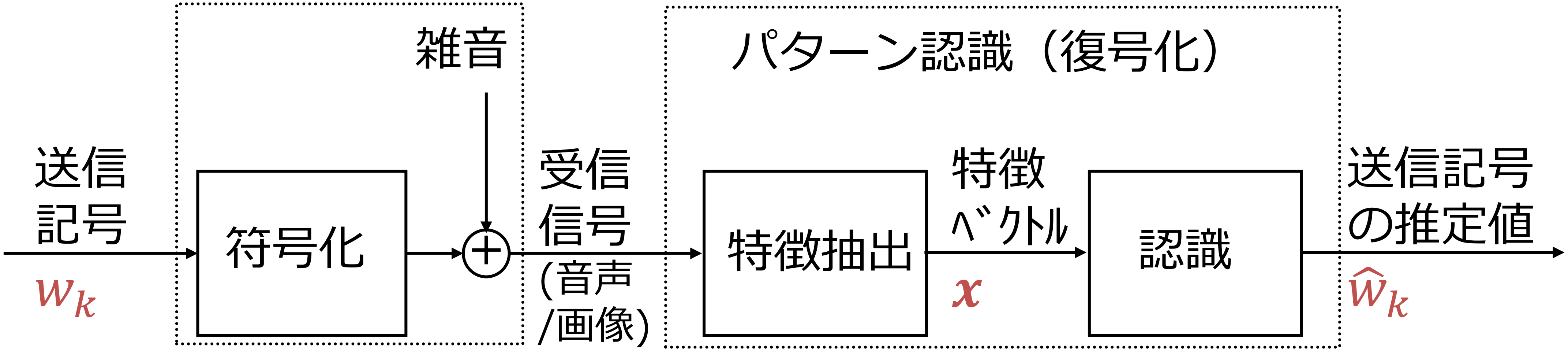
生成モデルでは，確率モデルを用意して，クラスごとに特徴ベクトルが出力する確率をモデル化しておく。認識対象データが与えられれば，各モデルがそのデータを出力する確率を求め，最大確率を与えたモデルに対応するクラスを認識結果とする。ここで紹介するGMMはその代表的方法である。

生成モデルによるパターン認識のブロック図



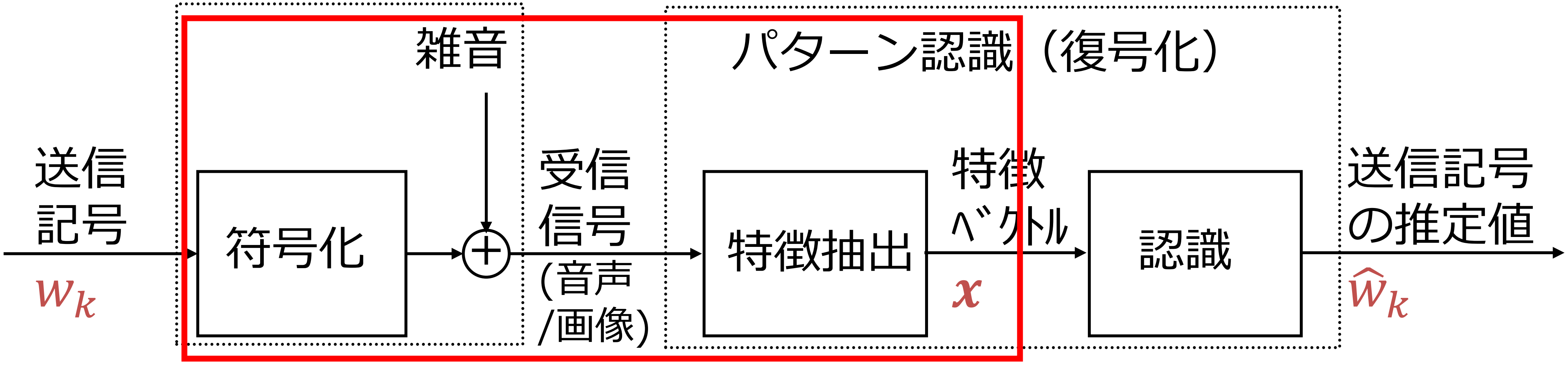
パターン認識の問題では，まず，情報の送り手が，送りたいクラスの情報を，画像や音声の媒体で送信する。
例えば，「早稲田」という情報を送るために，「わせだ」と声に出したり，「早稲田」と文字に書いたりする。

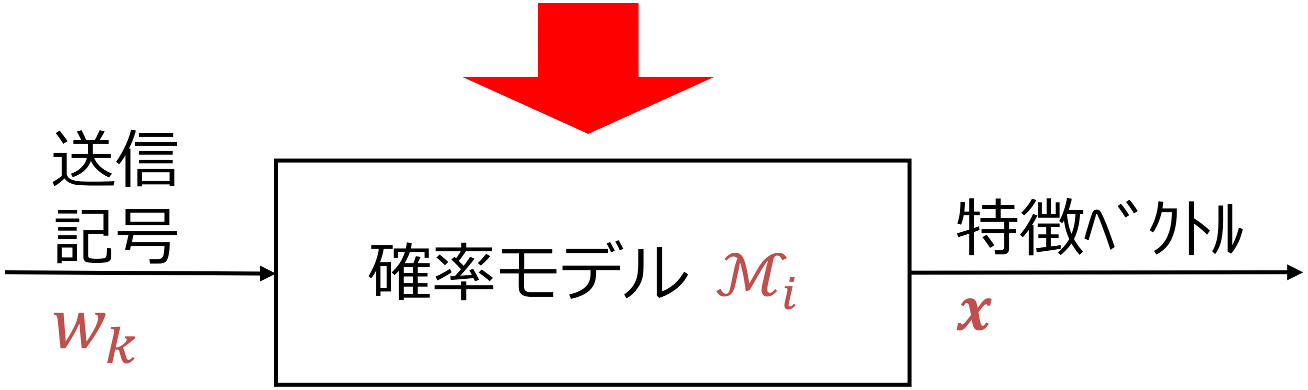
生成モデルによるパターン認識のブロック図



情報の受け手は，音声や画像を受けとって，これを特徴ベクトル化を経て，識別器にかけて，送り手が送ったクラス情報を推定する。

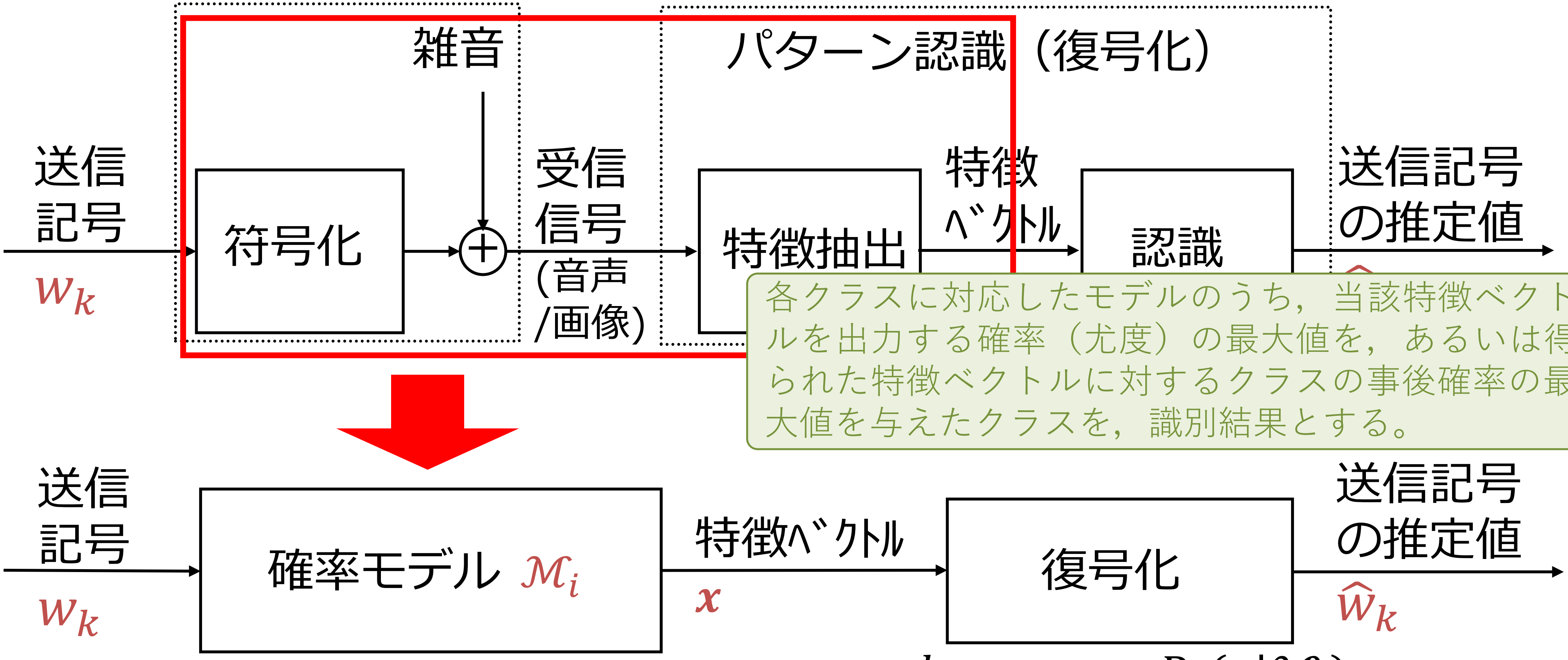
生成モデルによるパターン認識のブロック図





生成モデルに基づくパターン認識では，これらのプロセスのうち，クラス情報を送信したとき，特徴ベクトルが生成する確率をモデル化する。

生成モデルによるパターン認識のブロック図



各クラスに対応したモデルのうち，当該特徴ベクトルを出力する確率（尤度）の最大値を，あるいは得られた特徴ベクトルに対するクラスの事後確率の最大値を与えたクラスを，識別結果とする。

$$k = \underset{i}{\operatorname{argmax}} \Pr(\boldsymbol{x}|\mathcal{M}_i)$$

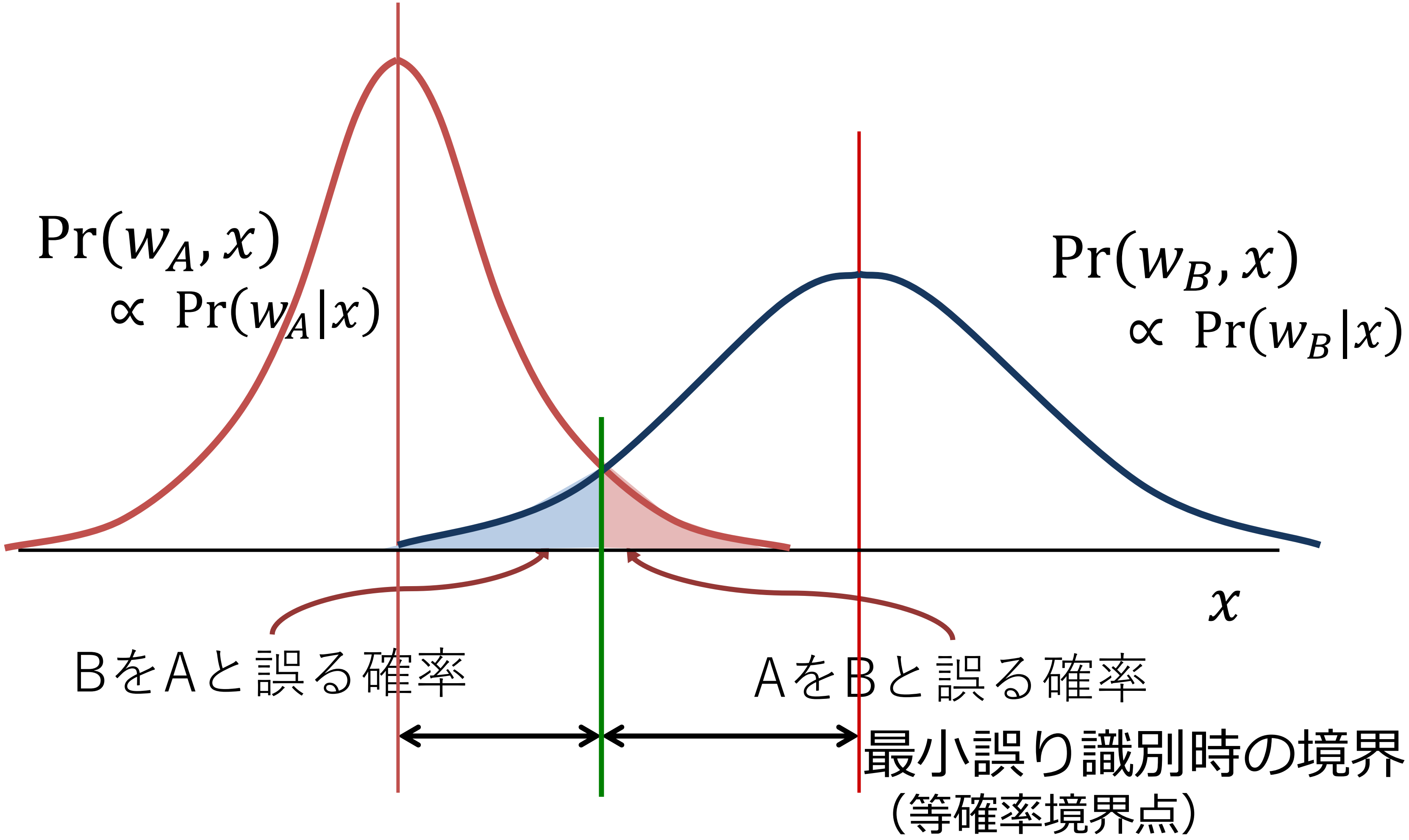
$$k = \underset{i}{\operatorname{argmax}} \Pr(\mathcal{M}_i|\boldsymbol{x})$$

§ 誤り率最小化と事後確率

誤り率最小化は,事後確率最大のクラスを選ぶことで実現できる

事後確率最大のクラスを選ぶことで,誤り率は最小化できる。

右図の赤い線はクラスAのヒストグラム，青い線はクラスBのそれとする。
事後確率最大の基準でクラスを決めることは，右の図で，上にあるクラスを選ぶことに相当する。

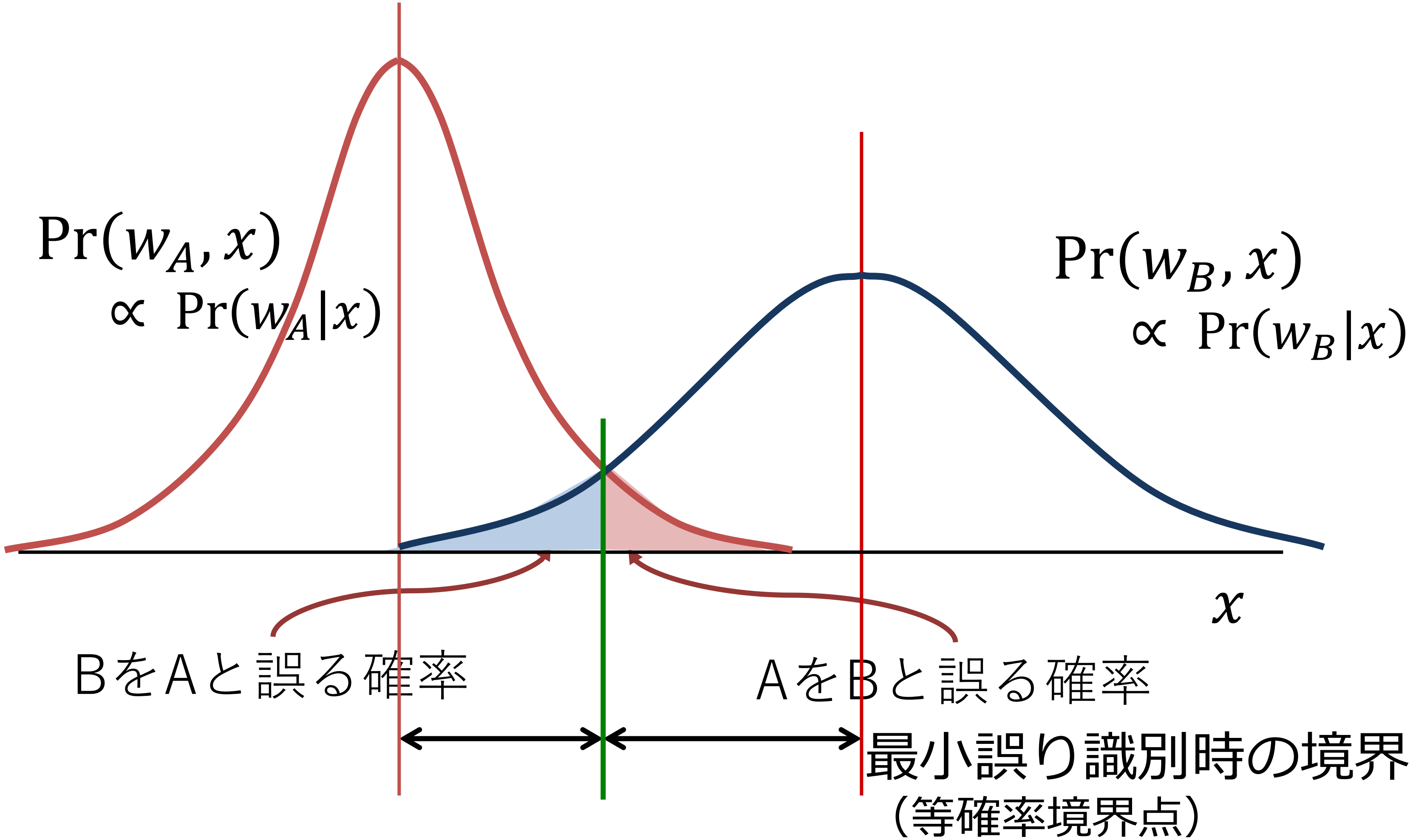


§ 誤り率最小化と事後確率

誤り率最小化は,事後確率最大のクラスを選ぶことで実現できる

このとき，青の領域がクラスBをAと誤る量，赤の領域がクラスAをBと誤る量に相当する。

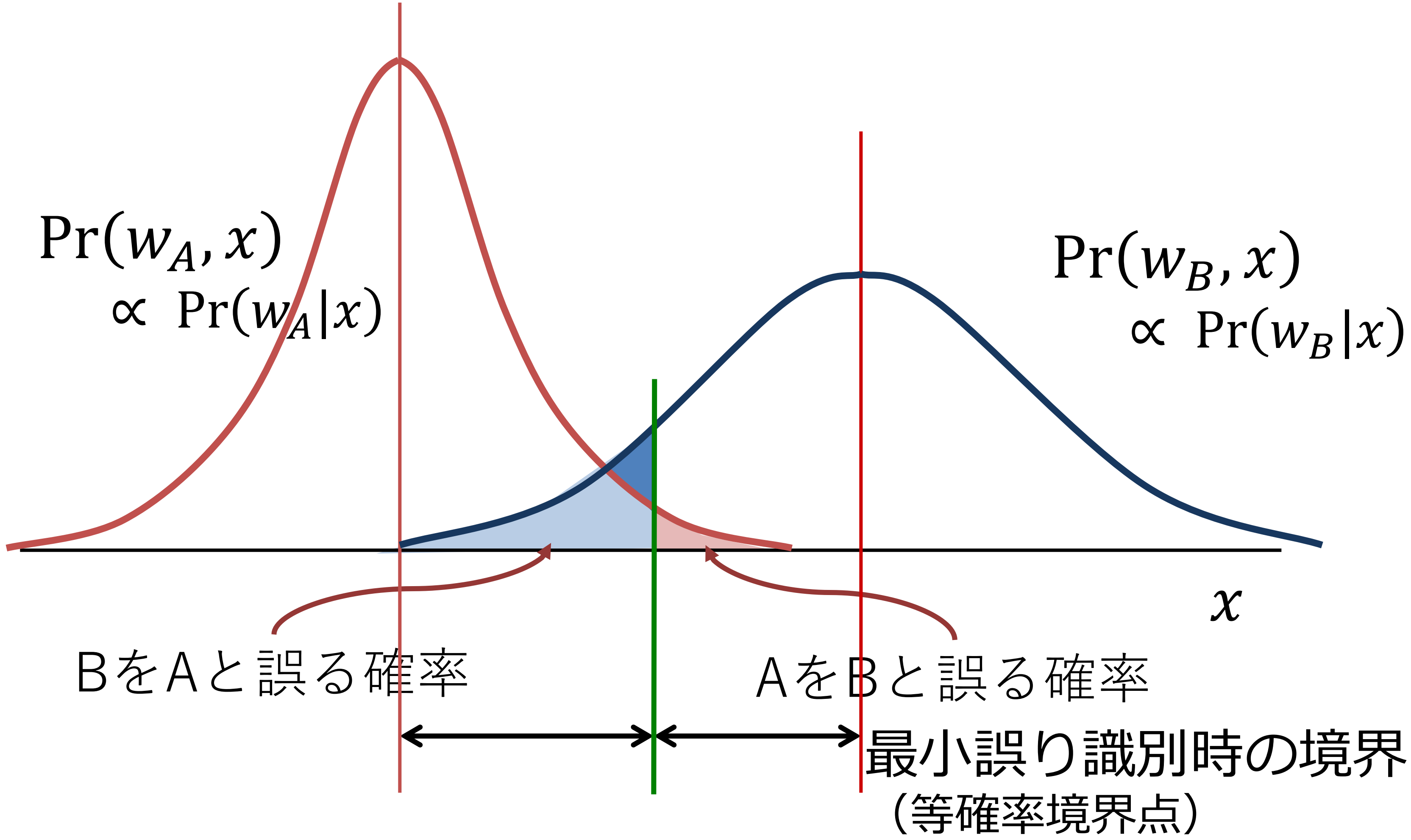
最大事後確率基準で誤り最小となることは，境界を少しずらした図を作ることで，容易に理解できる。



誤り率最小化と事後確率

誤り率最小化は,事後確率最大のクラスを選ぶことで実現できる

境界をずらすと き,誤りは右図で濃い青で示した領域の分だけ増えている。



事後確率最大化基準による復号

- 最大事後確率復号(MAP:Maximum a posteriori probability decoding):

$$k = \operatorname*{argmax}_i \Pr(w_i|\boldsymbol{x}) \qquad (1)$$

とするとき、 w_kが送信されたと判断する復号法。

事後確率が最大のクラスを識別結果とする基準による復号を，最大事後確率復号あるいはMAP復号とよぶ。

事後確率最大化基準による復号

- 最大事後確率復号(MAP:Maximum a posteriori probability decoding):

$$k = \underset{i}{\operatorname{argmax}} \Pr(w_i|\boldsymbol{x}) \qquad (1)$$

とするとき、 w_kが送信され

生成モデルの立場では，(1)をそのままモデル化するのではなく，ベイズの定理を使って(2)を得，さらに$\Pr(x)$はクラスw_iの最大化に無関係であることから，(3) の形にしてパターン認識を行う。

- Bayesの定理によって、

$$\Pr(w_i|\boldsymbol{x}) = \frac{\Pr(w_i, \boldsymbol{x})}{\Pr(\boldsymbol{x})} = \frac{\Pr(\boldsymbol{x}|w_i)\ \Pr(w_i)}{\Pr(\boldsymbol{x})} \qquad (2)$$

右辺分母の$\Pr(x)$はクラスw_iの最大化に無関係。よって、

$$\underset{i}{\operatorname{argmax}} \Pr(w_i|\boldsymbol{x}) = \underset{i}{\operatorname{argmax}} \Pr(\boldsymbol{x}|w_i)\ \Pr(w_i) \qquad (3)$$

と変形して，認識を行う。

最尤基準による復号

- 最尤復号：

$$k = \underset{i}{\operatorname{argmax}} \ \Pr(\boldsymbol{x}|w_i) \qquad (4)$$

とするとき、 w_kが送信されたと判断する復号法。

尤度（モデルがデータを生成する確率）により識別を行う方法を，最尤復号と呼ぶ。

最尤基準による復号

- 最尤復号：

$$k = \underset{i}{\operatorname{argmax}}\ \Pr(\boldsymbol{x}|w_i) \quad (4)$$

とするとき、 w_kが送信されたと判断する復号法。

- $\Pr(w_i)$がクラスによって変わらないとき、最尤複合は，最大事後確率複合と等価

$$\underset{i}{\operatorname{argmax}}\ \Pr(w_i|\boldsymbol{x}) = \underset{i}{\operatorname{argmax}}\ \Pr(\boldsymbol{x}|w_i)\ \Pr(w_i) \quad (5)$$

$$\underset{i}{\operatorname{argmax}}\ \Pr(w_i|\boldsymbol{x}) = \underset{i}{\operatorname{argmax}}\ \Pr(\boldsymbol{x}|w_i) \quad (6)$$

全クラスで事前確率が変わらないとすれば，最尤復号は最大事後確率復号に一致する。

生成モデルによるパターン認識における学習の問題

$\Pr(\boldsymbol{x}|w_i)$ が求まれば，パターン認識ができる。

．．．ところで．．．

いかに $\Pr(\boldsymbol{x}|w_i)$ を求めることができるか

以上のように，尤度が求まればパターン認識ができる。よって，生成モデルに基づく方法では，尤度をいかに得るかが問題となる。

生成モデルによるパターン認識における学習の問題

$\Pr(\boldsymbol{x}|w_i)$ が求まれば，パターン認識ができる。

．．．ところで．．．

いかに $\Pr(\boldsymbol{x}|w_i)$ を求めることができるか

．．．常套手段．．．

w_iから$\boldsymbol{x}$への写像を確率モデル$\mathcal{M}_i = \{a_1^i, a_2^i, \cdots, a_P^i\}$($a_p^i$はモデルパラメータ)で表現し，この確率モデルのパラメータを大量の学習データを用いて推定する。

$\Pr(\boldsymbol{x}|w_i)$を$\{a_1^i, a_2^i, \cdots, a_P^i\}$をパラメータとする確率モデル$\mathcal{M}_i$で表現し，これらのパラメータを大量の学習データを用いて推定することが常套手段。

§ 最尤推定（ML）

- クラスw_iのデータ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$を生成して，収集する。

学習データからパラメタを推定する典型的な方法として，最尤推定があげられる。

まず，それぞれのクラスのデータを多数生成して（「早稲田」の音声認識用のモデルを作るのであれば，「早稲田」と多数発声して），そのデータを収集する。

§ 最尤推定（ML)

- クラスw_iのデータ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$を生成して，収集する。

記号 w_i → 符号化＋観測雑音 → $\boldsymbol{x}_1$ $\boldsymbol{x}_2$

こうして得た学習データを，そのモデルが出力する確率（尤度）を最も高くするようモデルのパラメタを調整する。

- 各々のデータ$\boldsymbol{x}_1$が，カテゴリw_iの確率モデルから生起する確率が高くなるよう，モデルのパラメタ$\{a_1^i, a_2^i, \cdots, a_P^i\}$を調整する。

記号 w_i → 確率モデル $\mathcal{M}_i = \{a_1^i, a_2^i, \cdots, a_P^i\}$ →

$\boldsymbol{x}_1$ ： $\Pr(\boldsymbol{x}_1 | \{a_1^i, a_2^i, \cdots, a_P^i\})$

$\boldsymbol{x}_2$ ： $\Pr(\boldsymbol{x}_2 | \{a_1^i, a_2^i, \cdots, a_P^i\})$

⋮

$\boldsymbol{x}_N$ ： $\Pr(\boldsymbol{x}_N | \{a_1^i, a_2^i, \cdots, a_P^i\})$

$$\{\hat{a}_p^i\} = \underset{\{a_p^i\}}{\operatorname{argmax}} \prod_{j=1}^{N} \Pr(\boldsymbol{x}_j | \{a_p^i\})$$

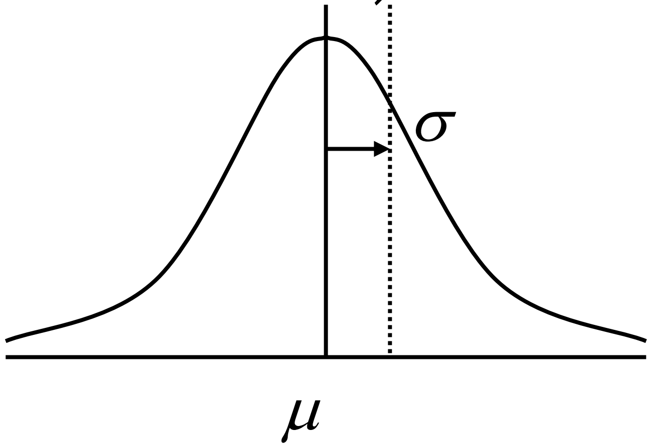
正規分布のパラメータ推定

正規分布を例にとって，最尤推定の考え方を見てみよう。
ここで，モデルのパラメタは，平均値μと分散σ^2　である。

確率分布が正規分布

$$\mathcal{M} = \{\mu, \sigma^2\}$$

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



で与えられるとき

正規分布のパラメータ推定

モデルから，学習データ$x_j, j=1,2,\cdots,N$が出力する確率は(2),その対数である対数尤度は(3)式で表される。

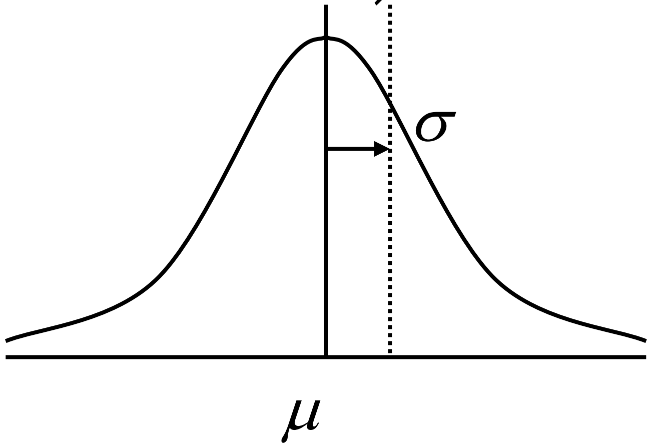
確率分布が正規分布

$$\mathcal{M} = \{\mu, \sigma^2\}$$

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

で与えられるとき

$$P = \prod_{j=1}^{N}\Pr\left(x_j|\mu,\sigma^2\right) \quad (2)$$



$$\log P = \sum_{j=1}^{N}\log\Pr\left(x_j|\mu,\sigma^2\right)$$

$$= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \sum_{j=1}^{N}\frac{1}{2}\frac{\left(x_j-\mu\right)^2}{\sigma^2} \quad (3)$$

ここで， $\log P$を各パラメタで偏微分して0とおく

$$\frac{\partial \log P}{\partial \mu} = -\sum_{j=1}^{N} \frac{x_j - \mu}{\sigma^2} = 0 \tag{4}$$

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{5}$$

$$\frac{\partial \log P}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} - \sum_{j=1}^{N} \frac{1}{2}\left(x_j - \mu\right)^2 \frac{-1}{(\sigma^2)^2} = 0 \tag{6}$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} \left(x_j - \mu\right)^2 \tag{7}$$

対数尤度を各パラメタで偏微分してゼロとおけば，平均値は(5)、分散は(7)として求まる。
これは，よく知られた平均値，分散の定義に一致している。

§ 混合正規分布 GMM (Gaussian Mixture Model)

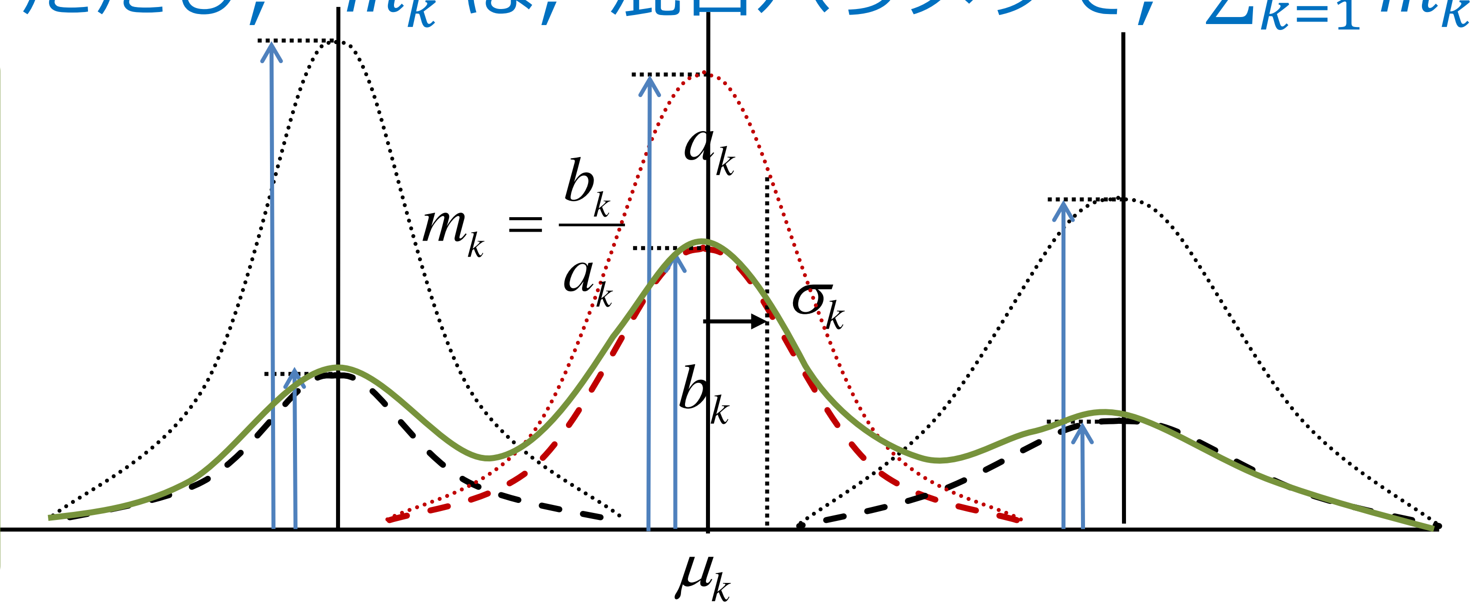
- GMM: 複数の正規分布の合成で作られる分布

$$\Pr(x|\mathcal{M}) = \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2) \quad (1)$$

$$\mathcal{M} = \{m_1, \cdots, m_K, \mu_1, \cdots, \mu_K, {\sigma_1}^2, \cdots, {\sigma_K}^2\}$$

ただし， m_k は，混合パラメタで，$\sum_{k=1}^{K} m_k = 1$ (2)

正規分布は，単純な分布であって，実際のデータの分布が正規分布で表されることはむしろ稀である。よって，もっと複雑な分布を表現できるモデルが必要となる。



§ 混合正規分布 GMM (Gaussian Mixture Model)

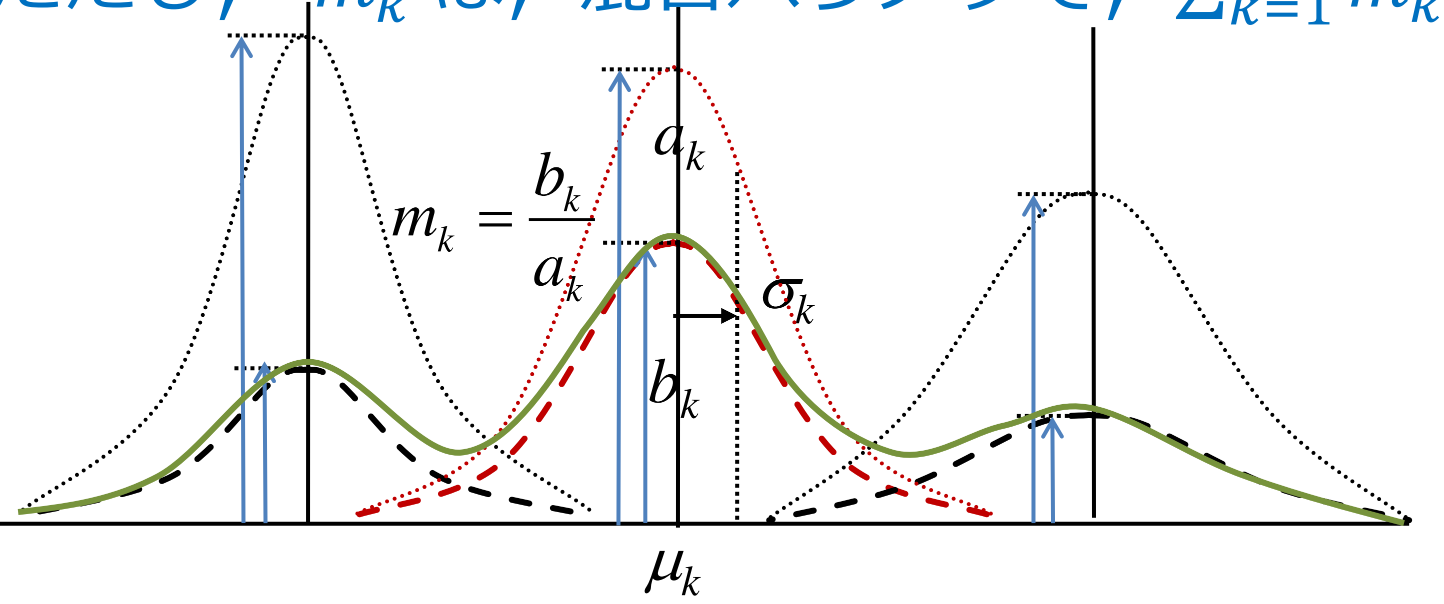
◘ GMM: 複数の正規分布の合成で作られる分布

$$\Pr(x|\mathcal{M}) = \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2) \quad (1)$$

$$\mathcal{M} = \{m_1, \cdots, m_K, \mu_1, \cdots, \mu_K, {\sigma_1}^2, \cdots, {\sigma_K}^2\}$$

ただし， m_k は，混合パラメタで， $\sum_{k=1}^{K} m_k = 1$ (2)

混合正規分布は，そのようなニーズに答えるもので，複数の正規分布を重み付きで加算した形で全体の分布を表現する。



$\mu_k, {\sigma_k}^2$ は，k番目の正規分布のパラメタ，m_k がその重みを表す。

潜在変数を用いたGMMの表現

ここで，GMMの振る舞いを，潜在変数を導入して解釈することを考える。

Y：データがどの分布から出力するかを表す確率変数

$$\Pr(Y = k|\mathcal{M}) = m_k \qquad (3)$$

：x がk 番目の分布から出力する確率 は m_k

X：データを表す確率変数

$$\Pr(X = x|Y = k, \mathcal{M}) = \mathcal{N}(x; \mu_k, {\sigma_k}^2) \qquad (4)$$

：x がk 番目の分布から出力するとき，

xは，$\mu_k, {\sigma_k}^2$ で決まる正規分布に従う。

このとき，

$$\Pr(x|\mathcal{M}) = \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2) \qquad (1)\text{再掲}$$

$$= \sum_{k=1}^{K} \Pr(Y = k|\mathcal{M}) \Pr(X = x|Y = k, \mathcal{M}) \qquad (5)$$

潜在変数を用いたGMMの表現

Y：データがどの分布から出力するかを表す確率変数

$$\Pr(Y = k|\mathcal{M}) = m_k \qquad (3)$$

：x がk 番目の分布から出力する確率 は m_k

X：データを表す確率変数

$$\Pr(X = x|Y = k, \mathcal{M}) = \mathcal{N}(x; \mu_k, {\sigma_k}^2) \qquad (4)$$

：x がk 番目の分布から出力するとき，

xは，$\mu_k, {\sigma_k}^2$ で決まる正規分布に従う。

このとき，

$$\Pr(x|\mathcal{M}) = \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2)$$

$$= \sum_{k=1}^{K} \Pr(Y = k|\mathcal{M}) \Pr($$

混合正規分布とは，xが，K個の正規分布を重みを持って合成された複雑な分布関数から出力するものであるが，これを，「x は，N個ある正規分布のうち，どれかひとつの分布に従って出力する。そして，どの正規分布に従うかは確率変数Yの値によって確率的に決まる。」と考える。

潜在変数を用いたGMMの表現

Y：データがどの分布から出力するかを表す

$$\Pr(Y = k|\mathcal{M}) = m_k$$

：x がk 番目の分布か

ここで，$Y = k$となる(xがk 番目の正規分布から出力する）確率 が m_kであれば，両者は同じ確率分布を意味することになる。ここで導入した確率変数 Y を潜在変数と呼ぶ。
（(1）は，xは複雑なひとつの分布に従って出力するとする考え方，(5)は，xは確率変数Yの値に従って選ばれる正規分布から出力するとする考え方）

X：データを表す確率変数

$$\Pr(X = x|Y = k, \mathcal{M}) = \mathcal{N}(x; \mu_k, {\sigma_k}^2) \quad (4)$$

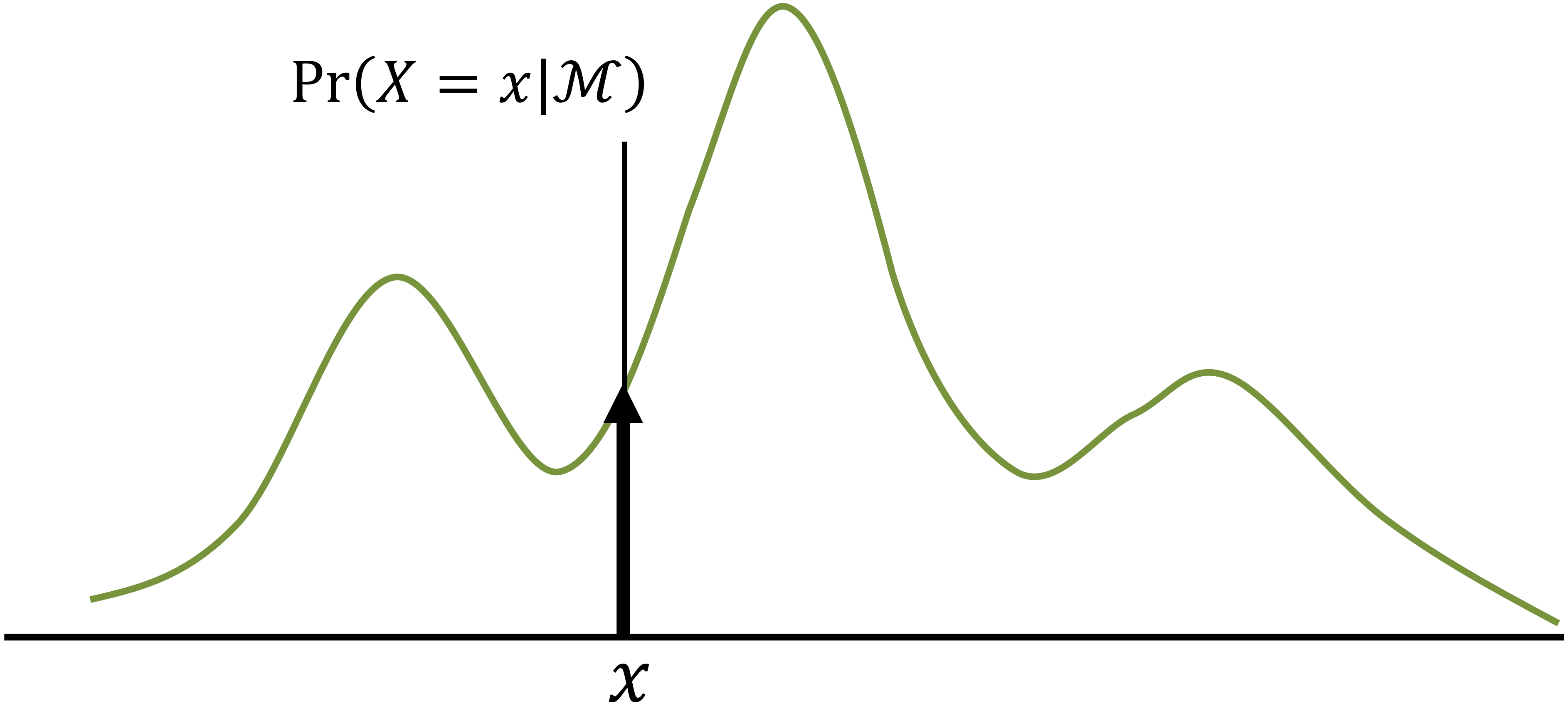
：x がk 番目の分布から出力するとき，
xは，$\mu_k, {\sigma_k}^2$ で決まる正規分布に従う。

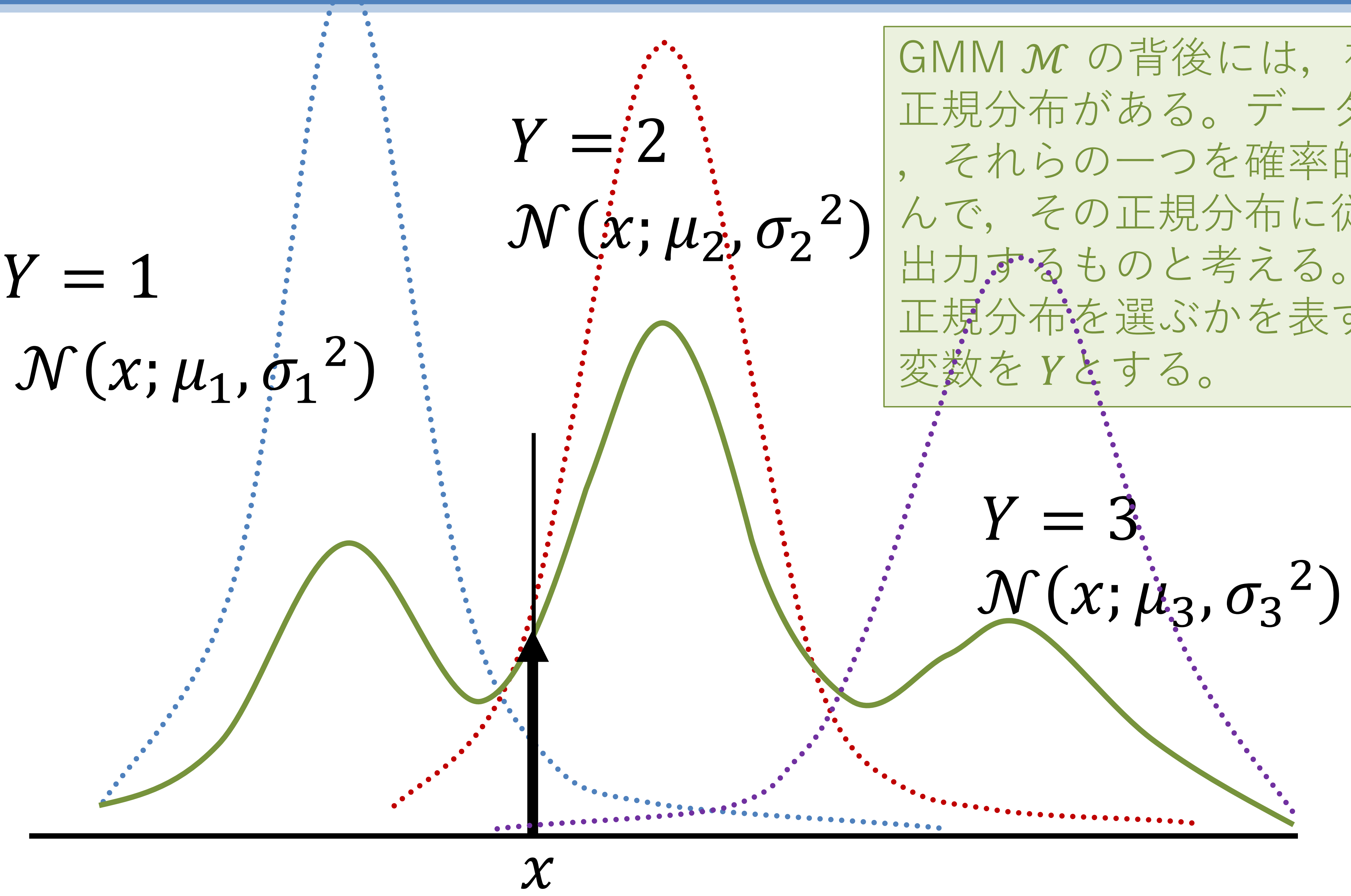
このとき，

$$\Pr(x|\mathcal{M}) = \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2) \quad (1)\text{再掲}$$

$$= \sum_{k=1}^{K} \Pr(Y = k|\mathcal{M}) \Pr(X = x|Y = k, \mathcal{M}) \quad (5)$$

GMM $\mathcal{M}$ から
xが出力する確率を考える





GMM $\mathcal{M}$ の背後には，複数の正規分布がある。データ x は，それらの一つを確率的に選んで，その正規分布に従って出力するものと考える。どの正規分布を選ぶかを表す確率変数を Y とする。

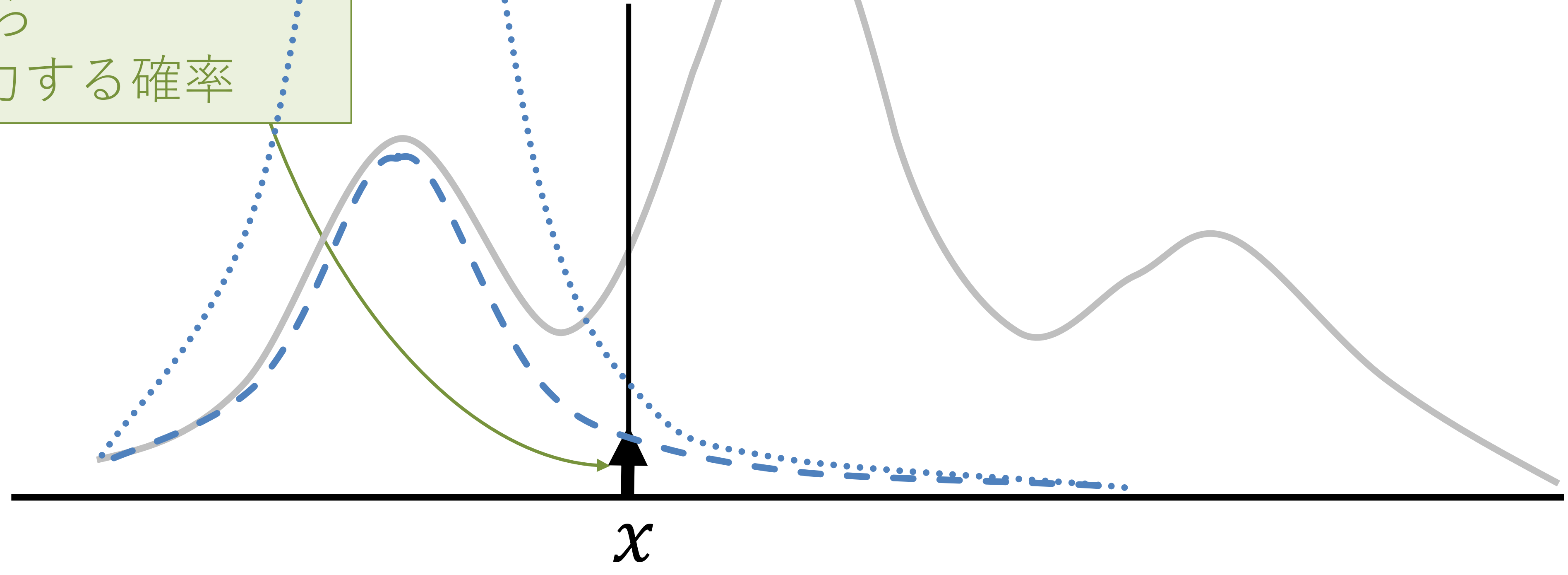
1番目の分布を選ぶ確率

1番目の分布を選んだとき x が出力する確率

$$\Pr(X = x, {\color{blue}Y = 1} | \mathcal{M}) = \Pr({\color{blue}Y = 1 |} \mathcal{M}) \Pr(X = x | {\color{blue}Y = 1,} \mathcal{M})$$

$$= m_1 \mathcal{N}(x; \mu_1, {\sigma_1}^2)$$

1番目の分布を選んでそこから x が出力する確率

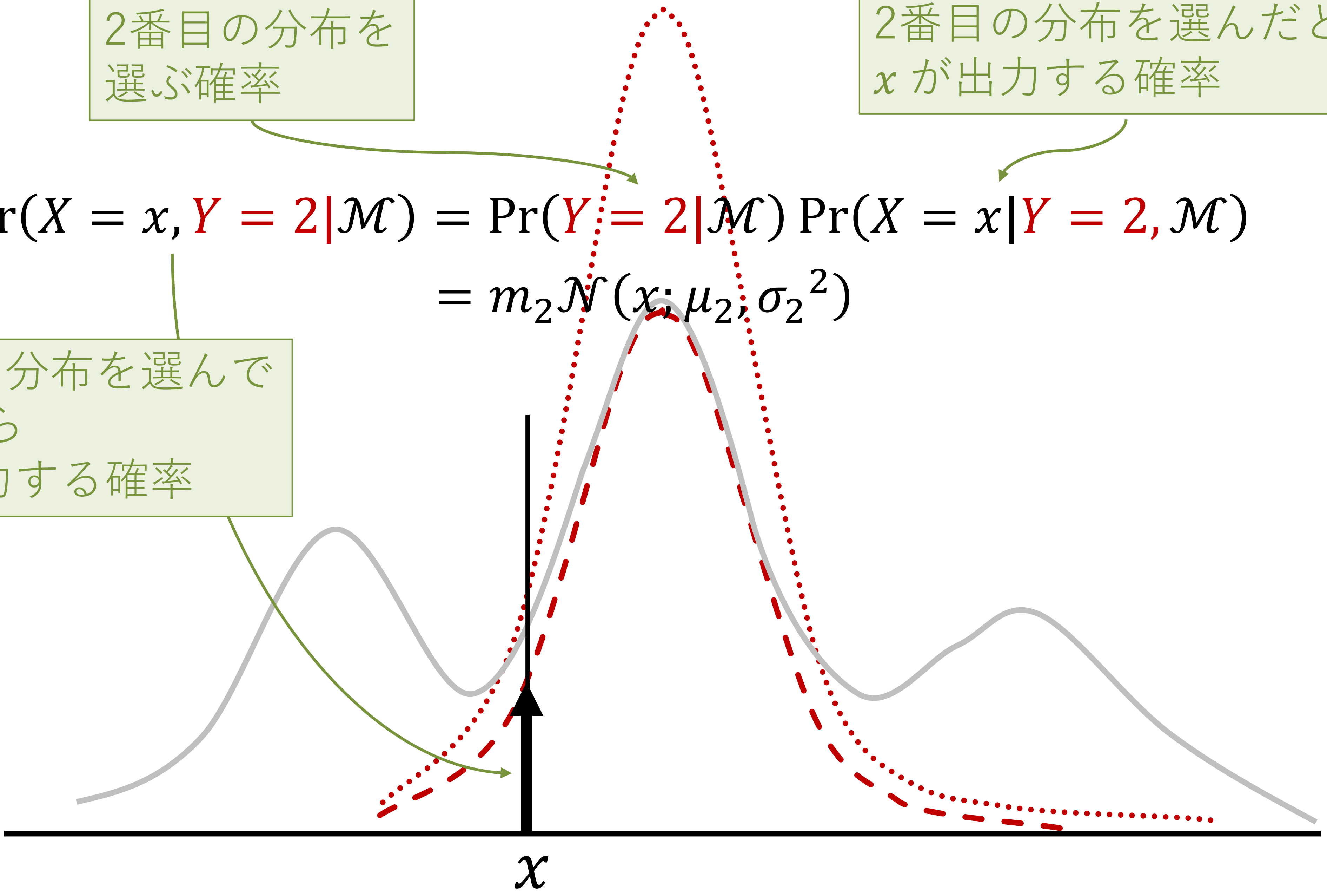


2番目の分布を選ぶ確率

2番目の分布を選んだとき
x が出力する確率

$$\Pr(X = x, Y = 2 | \mathcal{M}) = \Pr(Y = 2 | \mathcal{M}) \Pr(X = x | Y = 2, \mathcal{M})$$
$$= m_2 \mathcal{N}(x; \mu_2, {\sigma_2}^2)$$

2番目の分布を選んで
そこから
x が出力する確率

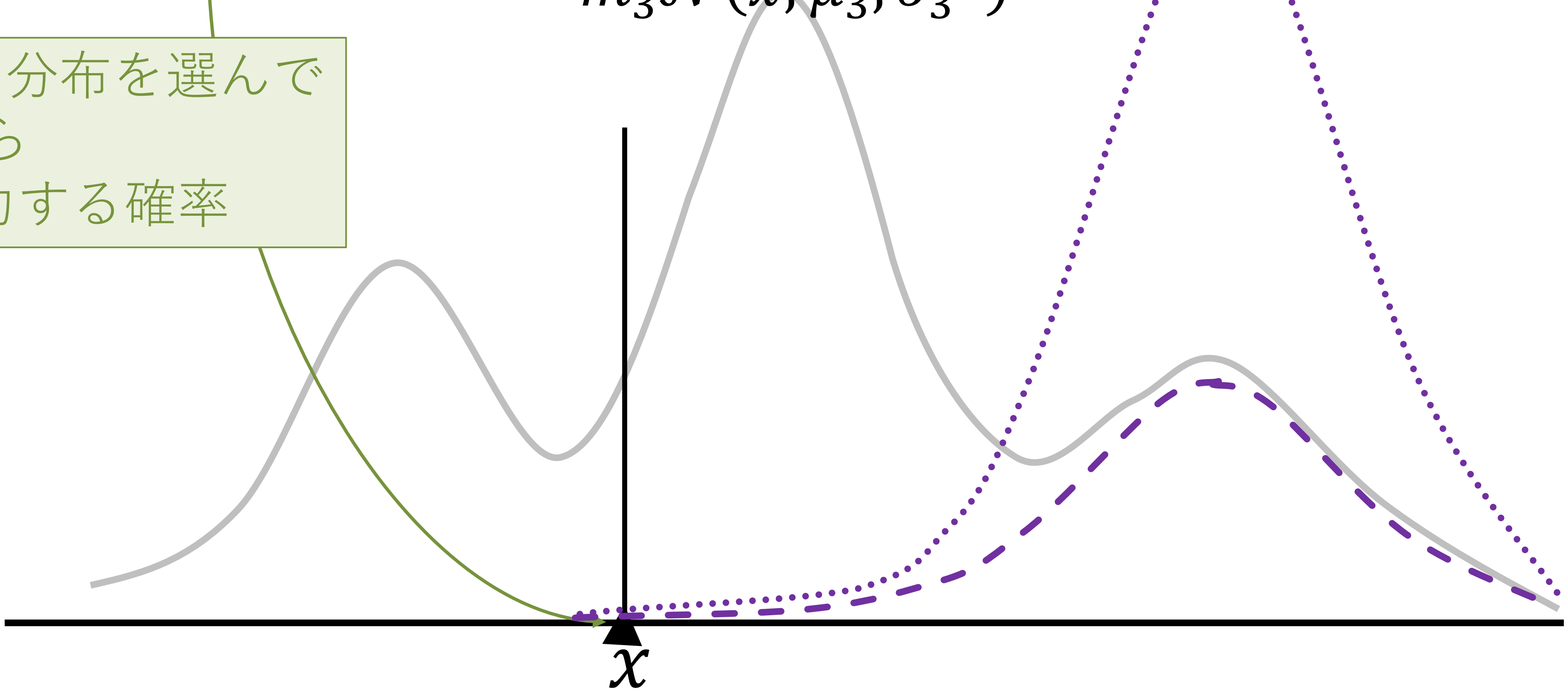


3番目の分布が
選ばれる確率

3番目の分布を選んだとき
x が出力する確率

$$\Pr(X = x, Y = 3 | \mathcal{M}) = \Pr(Y = 3 | \mathcal{M}) \Pr(X = x | Y = 3, \mathcal{M})$$
$$= m_3 \mathcal{N}(x; \mu_3, {\sigma_3}^2)$$

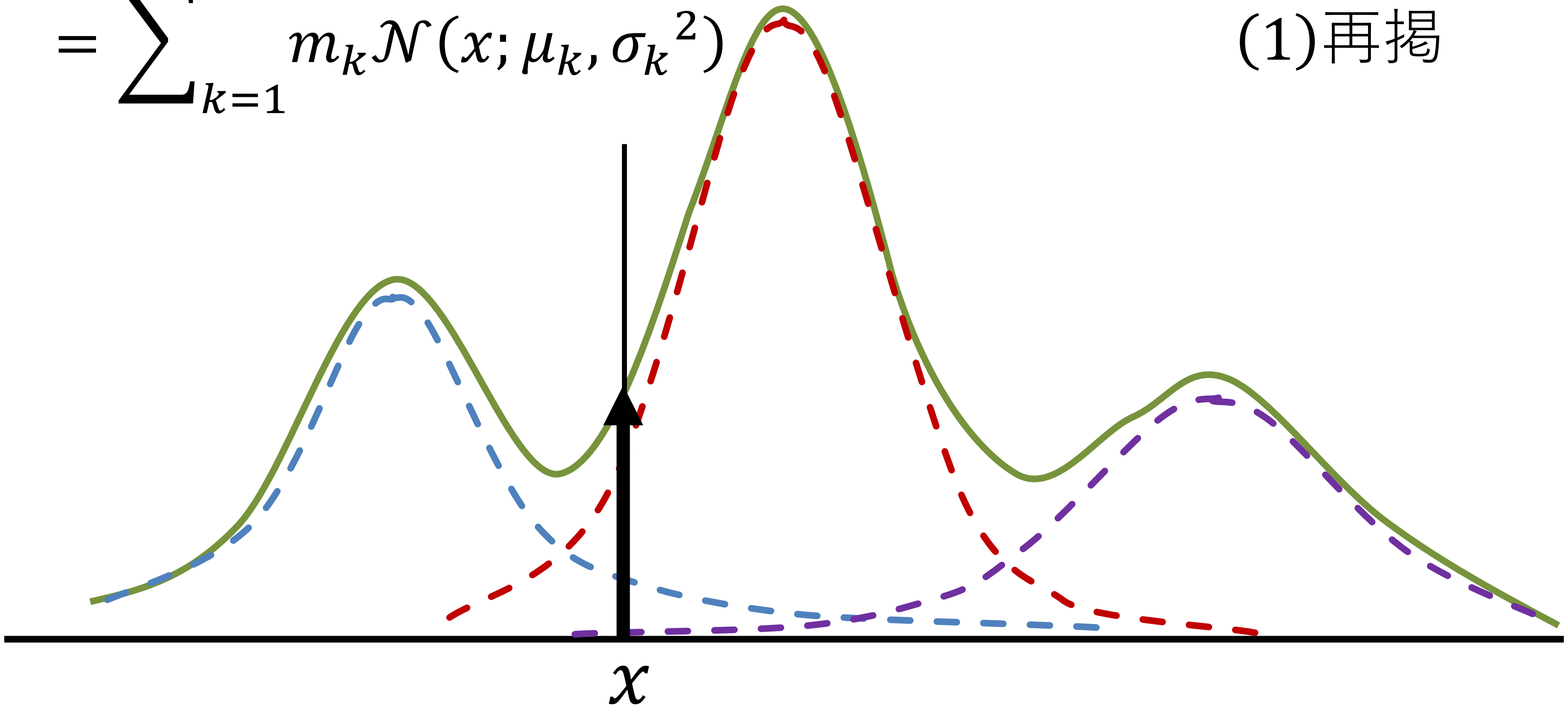
3番目の分布を選んで
そこから
x が出力する確率



$$\Pr(X = x|\mathcal{M}) = \sum_{k=1}^{K} \Pr(X = x, Y = k|\mathcal{M})$$

$$= \sum_{k=1}^{K} \Pr(Y = k|\mathcal{M}) \Pr(X = x|Y = k, \mathcal{M}) \quad \text{(5)再掲}$$

$$= \sum_{k=1}^{K} m_k \mathcal{N}(x; \mu_k, {\sigma_k}^2) \quad \text{(1)再掲}$$



§ GMMのパラメタ推定

このとき，

$$\log P = \sum_{j=1}^{N} \log \sum_{k=1}^{K} m_k \mathcal{N}\left(x_j; \mu_k, {\sigma_k}^2\right)$$

とおいて， $\log P$ を各パラメタで偏微分して0とおくと，

$$0 = \sum_{n=1}^{N} \frac{m_k \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} m_j \mathcal{N}\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

もはや閉形式では解が求まらない → EMアルゴリズム の登場

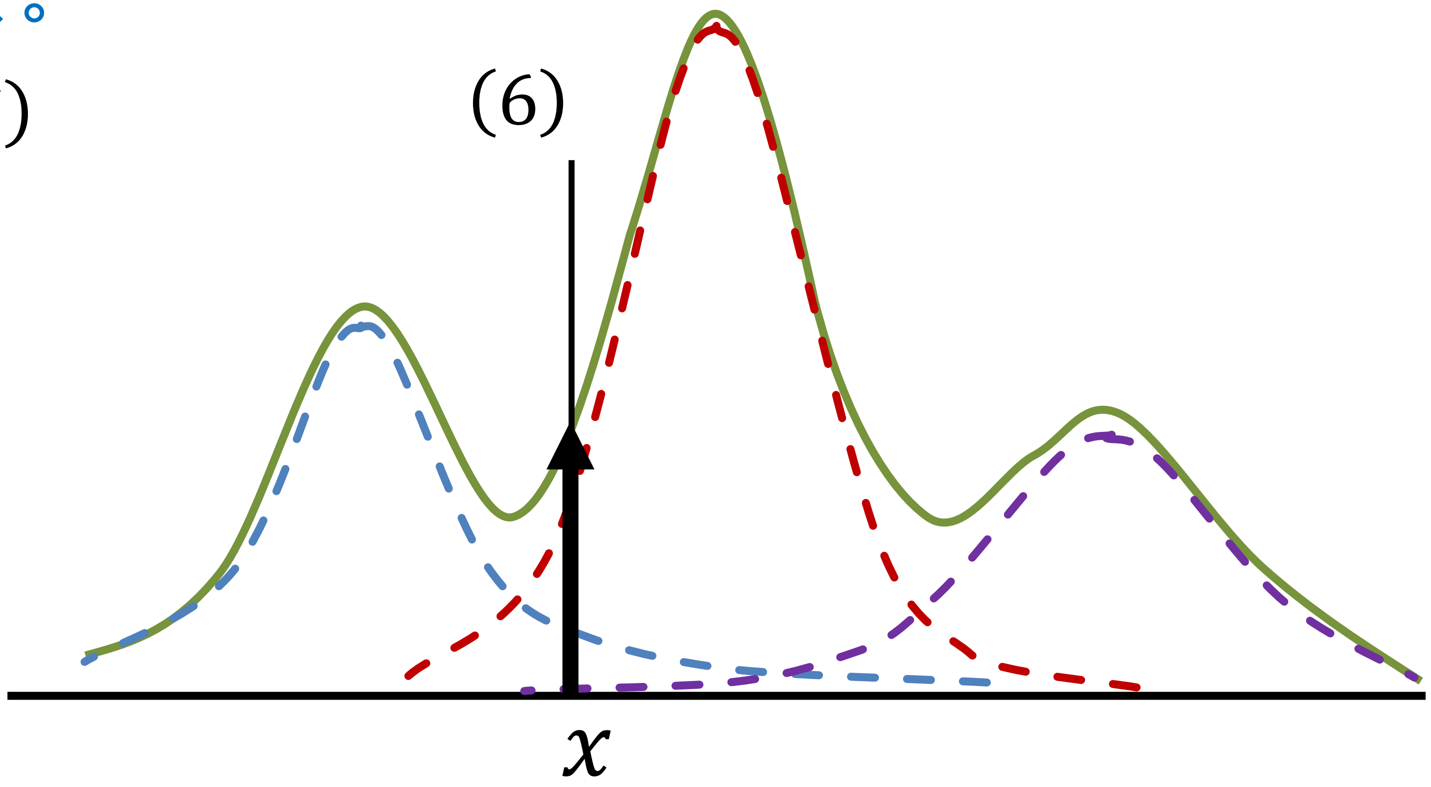
GMMのパラメタを学習データから学習することを考える。単純に対数尤度をパラメタで偏微分してゼロとおいても解は求まらない。よって，反復計算で求めることが必要となる。

EMアルゴリズム

モデル $\mathcal{M}$ から x が出力する確率を決めるのに，観測できないデータ（潜在変数） y が関与する枠組みにおいて，観測できるデータ x のみを用いて，尤度 $\Pr(x|\mathcal{M})$ を最大化するパラメタ $\mathcal{M}$ を求める問題を解く。

$$\max_{\mathcal{M}} \Pr(x|\mathcal{M}) \tag{6}$$

EMアルゴリズムは，尤度計算に潜在変数が介在する枠組みにおいて，最尤推定を行うためのアルゴリズムである。



EMアルゴリズム

モデル $\mathcal{M}$ から x が出力する確率を決めるのに，観測できないデータ（潜在変数） y が関与する枠組みにおいて，観測できるデータ x のみを用いて，尤度 $\Pr(x|\mathcal{M})$ を最大化するパラメタ $\mathcal{M}$ を求める問題を解く。

$$\max_{\mathcal{M}} \Pr(x|\mathcal{M}) \tag{6}$$

観測不能なyが$\Pr(x|\mathcal{M})$に関与するから，上式は簡単には求まらない。しかし， y の出力を仮定すると， (x,y) の同時確率は計算できるため，下式は比較的容易に実行できる。

$$\max_{\mathcal{M}} \Pr(x,y|\mathcal{M}) \tag{7}$$

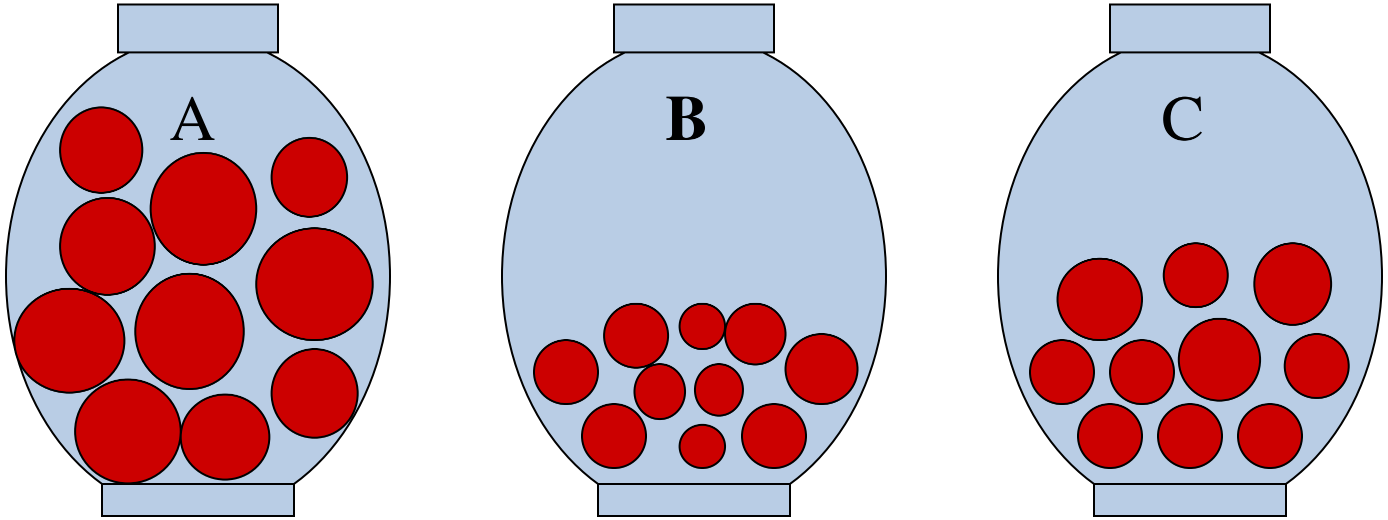
そこで，(6)の問題を(7)で置き換えて解く。

（ここで，xを不完全データ，(x,y)のデータ対を完全データと呼ぶ。）

不完全データを持つ典型的な問題例

- 報告者は，A,B,Cの壺のどれかを選び，
- その壺から一つボールを選び，
- そのボールの重さのみを観察者に伝える。

報告者と観察者の二人がいて，
観察者は報告者が報告するデータから，壺の中のボールの分布を推定するものとする。
観察者は左のようなルールに従って，報告をするものとする。

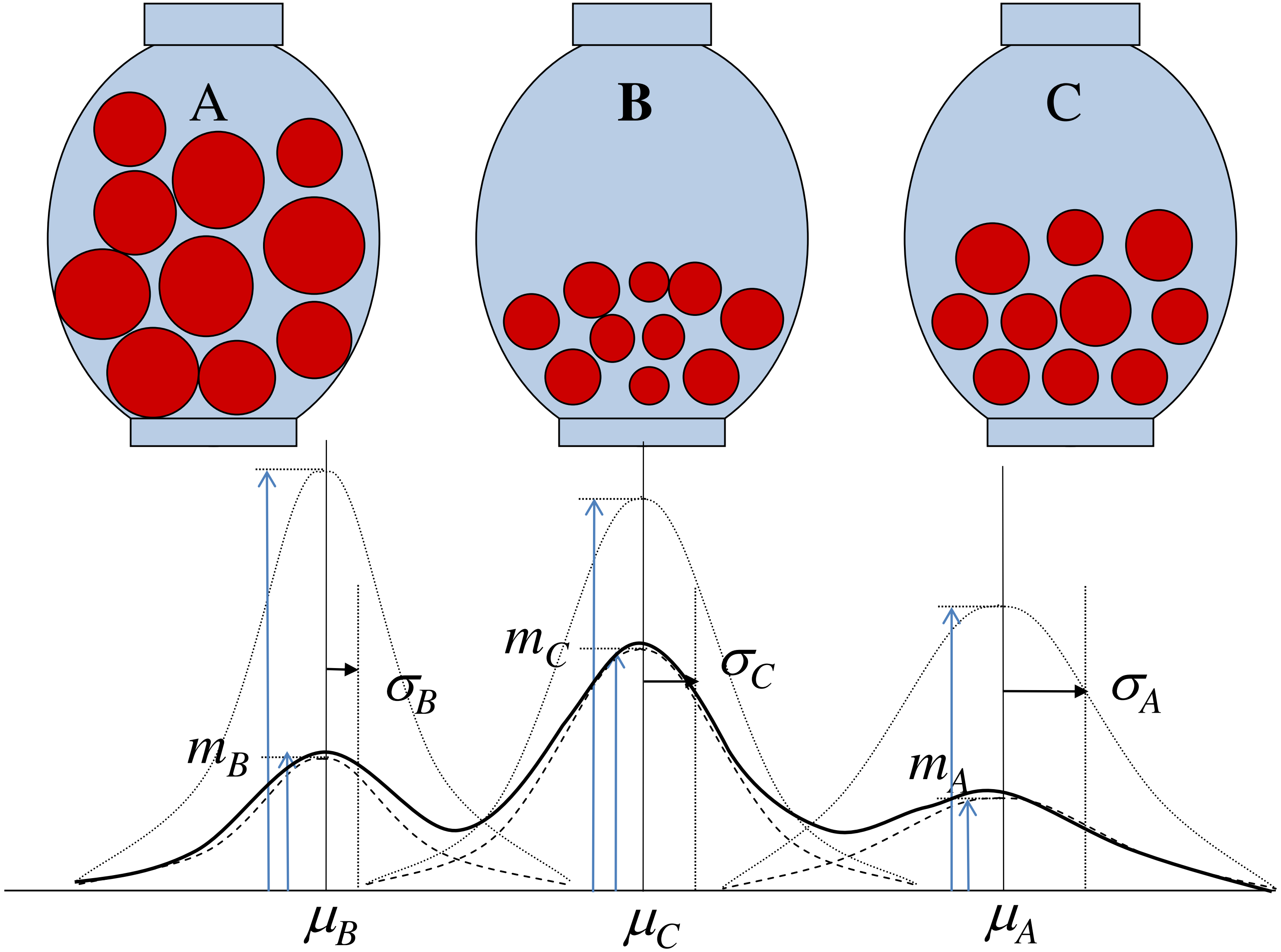


観測できるデータ：　ボールの重み
観測できないデータ（潜在変数）：　ボールがどの壷から出たか

（ボールの重み）　・・・　不完全データ
（ボールの重み，そのボールがどの壷からでたか）・・・完全データ

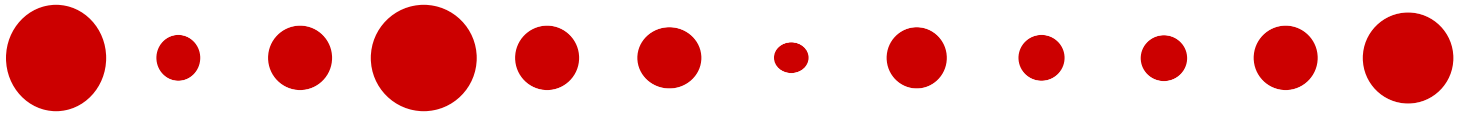
各壷の中のボールの重さが正規分布に従うとすると,
全体の分布は，混合正規分布。

報告されるボールの重さの分布は混合正規分布そのものとなる。



不完全データを持つ典型的な問題例

- 報告（つぼ+ボール）：

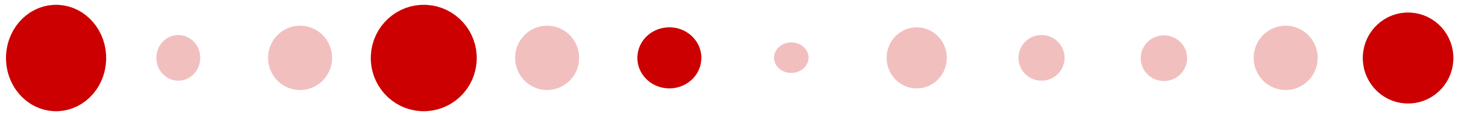


重さ:	8	3	6	9	7	6	2	6	4	4	5	7
壷:	A	B	C	A	C	A	B	B	C	B	C	A

ボールの重さと壷の組が与えられる場合（完全データ）を考える

不完全データを持つ典型的な問題例

- 報告（つぼ+ボール）：



重さ:	8	3	6	9	7	6	2	6	4	4	5	7
壷:	A	B	C	A	C	A	B	B	C	B	C	A

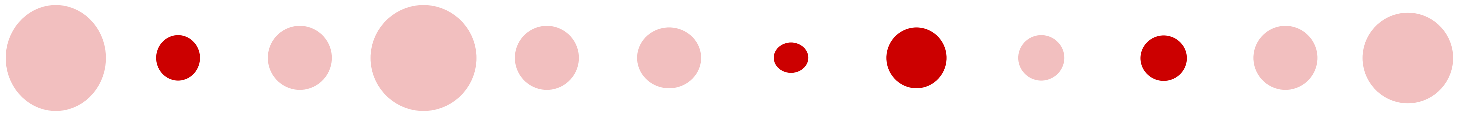
A　　4/12　　μ : $(8+9+6+7)/4 = 7.5$

σ^2: $(8^2+9^2+6^2+7^2)/4 - 7.5^2 = 1.5$

Aから出たボールのデータだけ集めれば，Aの確率分布のパラメタが分かる

不完全データを持つ典型的な問題例

□ 報告（つぼ+ボール）：



重さ:	8	3	6	9	7	6	2	6	4	4	5	7
壷:	A	B	C	A	C	A	B	B	C	B	C	A

A 4/12 μ: $(8+9+6+7)/4 = 7.5$

σ^2: $(8^2+9^2+6^2+7^2)/4 - 7.5^2 = 1.5$

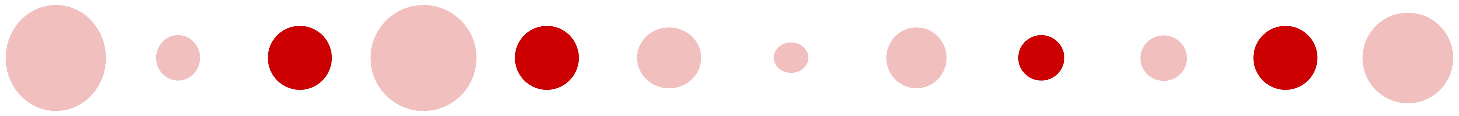
B 4/12 μ: $(3+2+6+4)/4 = 3.5$

σ^2: $(3^2+2^2+6^2+4^2)/4 - 3.5^2 = 1.25$

Bから出たボールのデータだけ集めれば，Bの確率分布のパラメタが分かる

不完全データを持つ典型的な問題例

報告（つぼ+ボール）：



重さ:	8	3	6	9	7	6	2	6	4	4	5	7
壷:	A	B	C	A	C	A	B	B	C	B	C	A

A　4/12　μ: $(8+9+6+7)/4 = 7.5$

σ^2: $(8^2+9^2+6^2+7^2)/4 - 7.5^2 = 1.5$

B　4/12　μ: $(3+2+6+4)/4 = 3.5$

σ^2: $(3^2+2^2+6^2+4^2)/4 - 3.5^2 = 1.25$

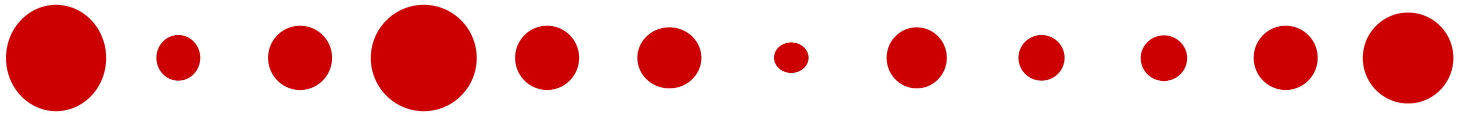
C　4/12　μ: $(6+7+4+5)/4 = 5.5$

σ^2: $(6^2+7^2+4^2+5^2)/4 - 5.5^2 = 1.25$

Cから出たボールのデータだけ集めれば，Cの確率分布のパラメタが分かる

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ: 8 3 6 9 7 6 2 6 4 4 5 7

各ボールがどの壷から出たかは分からない。

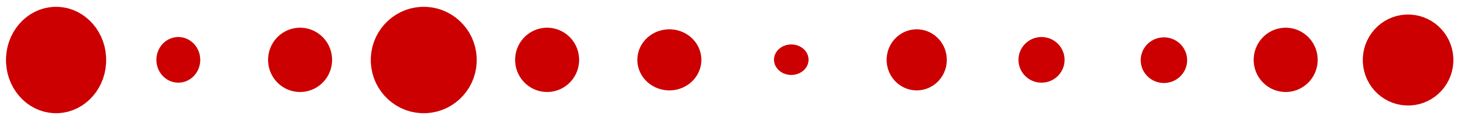
そこで，どの壷から出たかに関する推定を行う（事後確率を求める）。

事後確率は，各壷のボールの重みの分布を表す確率モデルの推定値をもとに求めることにする。

ボールの重さだけが与えられる場合（不完全データ）を考える

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:	8	3	6	9	7	6	2	6	4	4	5	7

例：

$$P(A)=1/3,\ M_A=\{\mu_A,\sigma_A^2\}=\{7,1.0\}$$

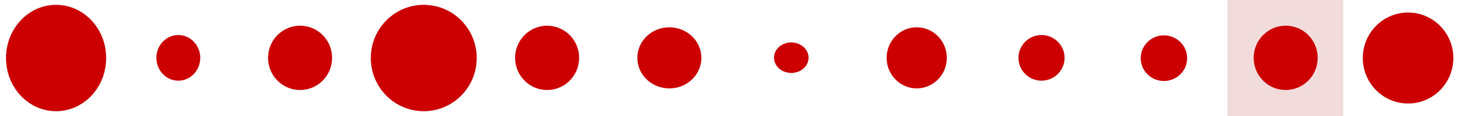
$$P(B)=1/3,\ M_B=\{\mu_B,\sigma_B^2\}=\{3,1.0\}$$

$$P(C)=1/3,\ M_C=\{\mu_C,\sigma_C^2\}=\{5,1.0\}$$

確率モデルのパラメタが与えられたとする

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ: 8 3 6 9 7 6 2 6 4 4 5 7

例：

P(A)=1/3, $M_A=\{\mu_A,\sigma_A^2\}=\{7,1.0\}$

P(B)=1/3, $M_B=\{\mu_B,\sigma_B^2\}=\{3,1.0\}$

P(C)=1/3, $M_C=\{\mu_C,\sigma_C^2\}=\{5,1.0\}$

$$P(A|5)=\frac{P(5|A)P(A)}{P(5)}$$

$$=\frac{P(5|A)P(A)}{\sum_{X=A,B,C} P(5|X)P(X)}$$

$P(5 \mid A)=0.054$

$P(5 \mid B)=0.054$

$P(5 \mid C)=0.399$

$P(A \mid 5)=0.107$

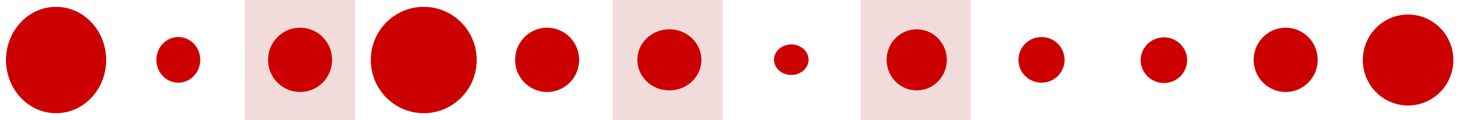
$P(B \mid 5)=0.107$

$P(C \mid 5)=0.787$

例えば，重み５のボールに対する潜在変数（どの壷から出たか）の期待値は，上のようになる。

不完全データを持つ典型的な問題例

□ 報告（ボールのみ）：



重さ: 8 3 6 9 7 6 2 6 4 4 5 7

例：

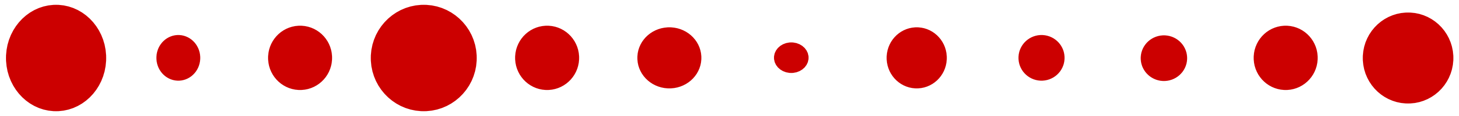
$$P(A)=1/3,\ M_A=\{\mu_A,\sigma_A^2\}=\{7,1.0\}$$
$$P(B)=1/3,\ M_B=\{\mu_B,\sigma_B^2\}=\{3,1.0\}$$
$$P(C)=1/3,\ M_C=\{\mu_C,\sigma_C^2\}=\{5,1.0\}$$

$$P(6\,|\,A)=0.242 \qquad P(A\,|\,6)=0.495$$
$$P(6\,|\,B)=0.004 \quad\Rightarrow\quad P(B\,|\,6)=0.009$$
$$P(6\,|\,C)=0.242 \qquad P(C\,|\,6)=0.495$$

重み６のボールの潜在変数の期待値は，上のようになる。

不完全データを持つ典型的な問題例

- 報告（ボールのみ）：

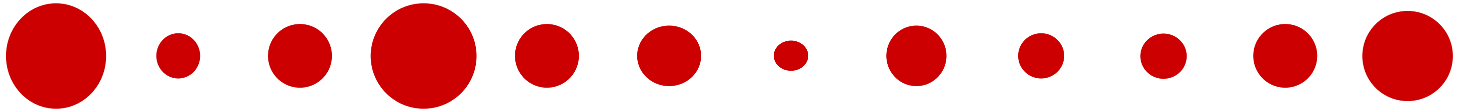


重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.98	.00	.50	1.0	.88	.50	.00	.50	.00	.01	.11	.88
	B	.00	.88	.00	.00	.00	.00	.98	.00	.50	.50	.11	.00
	C	.02	.12	.50	.00	.12	.50	.02	.50	.50	.50	.78	.12

それぞれのボールの潜在変数の期待値は，上のようになる。

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.98	.00	.50	1.0	.88	.50	.00	.50	.00	.00	.11	.88
	B	.00	.88	.00	.00	.00	.00	.98	.00	.50	.50	.11	.00
	C	.02	.12	.50	.00	.12	.50	.02	.50	.50	.50	.78	.12

A：確率的度数の合計：　5.352

$P(A) = 5.352/12 = 0.446$

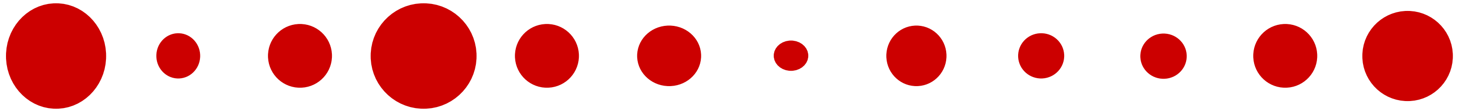
$\mu_A = (.98 \times 8 + .00 \times 3 + .50 \times 6 + \cdots) / 5.352 = 7.228$

$\sigma_A^2 = (.98 \times (8-7.23)^2 + .00 \times (3-7.23)^2 + \cdots) / 5.352 = 1.266$

Aの期待値をAの確率的な度数としてみると（重さ８のボールはAから0.98回出たと考える），Aの分布のパラメタを更新することができる

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.98	.00	.50	1.0	.88	.50	.00	.50	.00	.00	.11	.88
	B	.00	.88	.00	.00	.00	.00	.98	.00	.50	.50	.11	.00
	C	.02	.12	.50	.00	.12	.50	.02	.50	.50	.50	.78	.12

B：確率的度数の合計：　2.988

$P(B) = 2.988/12 = 0.249$

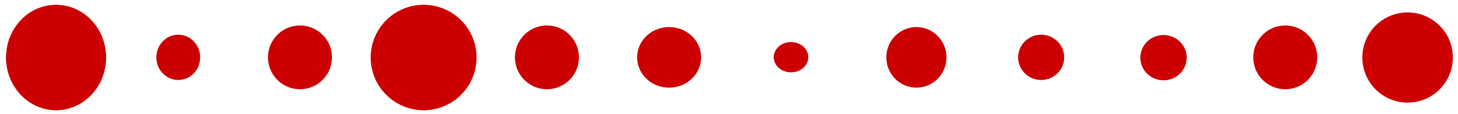
$\mu_B = (.00 \times 8 + .88 \times 3 + .00 \times 6 + \cdots) / 2.988 = 3.102$

$\sigma_B^2 = (.00 \times (8-3.08)^2 + .88 \times (3-3.08)^2 + \cdots) / 2.988 = 0.878$

同様にBの期待値から，Bの分布パラメタを更新できる

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.98	.00	.50	1.0	.88	.50	.00	.50	.00	.00	.11	.88
	B	.00	.88	.00	.00	.00	.00	.98	.00	.50	.50	.11	.00
	C	.02	.12	.50	.00	.12	.50	.02	.50	.50	.50	.78	.12

C: 確率的度数の合計： 3.660

$P(C) = 3.660/12 = 0.305$

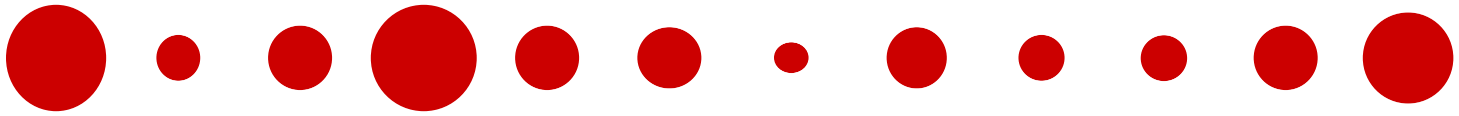
$\mu_C = (.02 \times 8 + .12 \times 3 + .50 \times 6 + \cdots) / 3.660 = 5.203$

$\sigma_C^2 = (.02 \times (8\text{-}5.02)^2 + .12 \times (3\text{-}5.02)^2 + \cdots) / 3.660 = 1.125$

同様にCの期待値から，Cの分布パラメタを更新できる

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:	8	3	6	9	7	6	2	6	4	4	5	7

モデルパラメタの新たな推定値

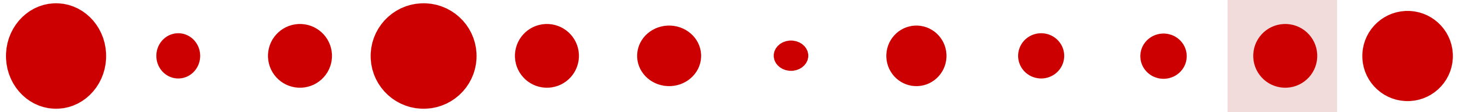
$$P(A)=0.44,\ \ M_A=\{\mu_A,\sigma_A{}^2\}=\{7.23,1.27\}$$

$$P(B)=0.25,\ \ M_B=\{\mu_B,\sigma_B{}^2\}=\{3.10,0.89\}$$

$$P(C)=0.31,\ \ M_C=\{\mu_C,\sigma_C{}^2\}=\{5.20,1.13\}$$

不完全データを持つ典型的な問題例

- 報告（ボールのみ）：



重さ: 8 3 6 9 7 6 2 6 4 4 5 7

$$P(A|5)=\frac{P(5|A)P(A)}{P(5)}=\frac{P(5|A)P(A)}{\sum_{X=A,B,C}P(5|X)P(X)}$$

モデルパラメタの新たな推定値

$P(A)=0.44,\ M_A=\{\mu_A,\sigma_A^2\}=\{7.23,1.27\}$

$P(B)=0.25,\ M_B=\{\mu_B,\sigma_B^2\}=\{3.10,0.89\}$

$P(C)=0.31,\ M_C=\{\mu_C,\sigma_C^2\}=\{5.20,1.13\}$

$P(5\,|\,M_A)=0.050$ → $P(A\,|\,5,M)=0.150$

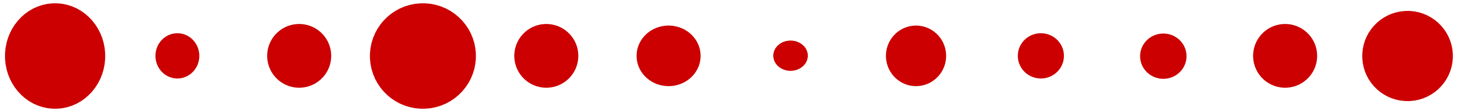
$P(5\,|\,M_B)=0.055$ → $P(B\,|\,5,M)=0.092$

$P(5\,|\,M_C)=0.369$ → $P(C\,|\,5,M)=0.758$

分布パラメタが更新されると，潜在変数の期待値も更新できる

不完全データを持つ典型的な問題例

□ 報告（ボールのみ）：

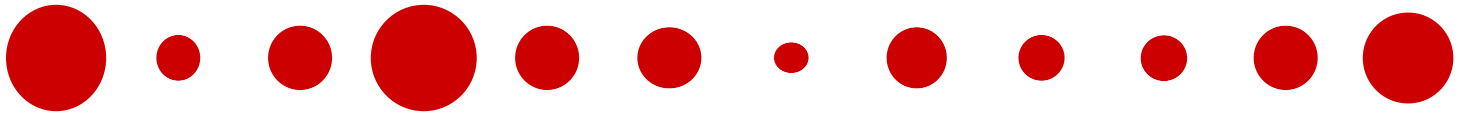


重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.97	.00	.50	1.0	.85	.50	.00	.50	.02	.02	.15	.85
	B	.00	.89	.00	.00	.00	.00	.98	.00	.52	.52	.09	.00
	C	.03	.11	.50	.00	.15	.50	.02	.50	.46	.46	.76	.15

それぞれのボールの潜在変数の期待値は，上のように変更される。

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ: 8 3 6 9 7 6 2 6 4 4 5 7

モデルパラメタの新たな推定値

$$P(A)=0.45,\ M_A=\{\mu_A,\sigma_A^2\}=\{7.20,1.35\}$$

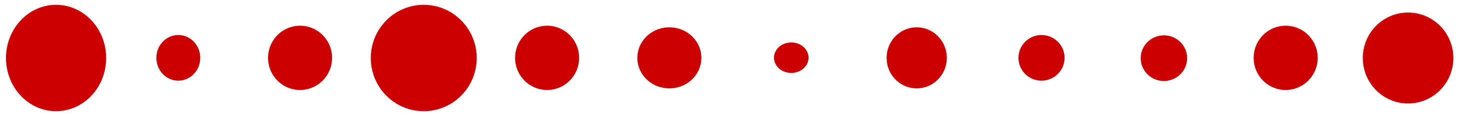
$$P(B)=0.25,\ M_B=\{\mu_B,\sigma_B^2\}=\{3.10,0.83\}$$

$$P(C)=0.30,\ M_C=\{\mu_C,\sigma_C^2\}=\{5,27,1.19\}$$

潜在変数の期待値が更新できると，分布パラメタも再度更新できる

不完全データを持つ典型的な問題例

□ 報告（ボールのみ）：

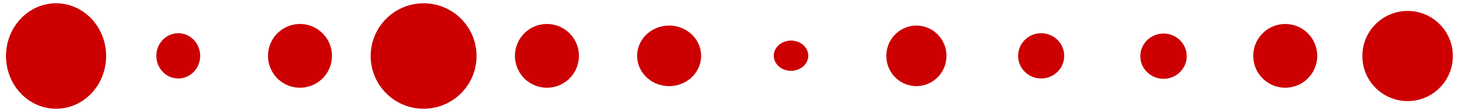


重さ:		8	3	6	9	7	6	2	6	4	4	5	7
確率的度数	A	.96	.00	.50	.99	.83	.50	.00	.50	.03	.03	.18	.83
	B	.00	.89	.00	.00	.00	.00	.98	.00	.53	.53	.08	.00
	C	.04	.11	.50	.01	.17	.50	.02	.50	.44	.44	.74	.17

分布パラメタが更新されると，潜在変数の期待値も再度更新できる

不完全データを持つ典型的な問題例

報告（ボールのみ）：



重さ:　8　3　6　9　7　6　2　6　4　4　5　7

モデルパラメタの新たな推定値

$$P(A)=0.44,\ M_A=\{\mu_A,\sigma_A^2\}=\{7.17,1.40\}$$

$$P(B)=0.25,\ M_B=\{\mu_B,\sigma_B^2\}=\{3.09,0.81\}$$

$$P(C)=0.30,\ M_C=\{\mu_C,\sigma_C^2\}=\{5,31,1.23\}$$

分布パラメタ　と　潜在変数の事後確率　の推定を交互に行うことで，
分布パラメタを更新し続けることができる

→　このとき，尤度は単調増加することが証明できる

以上が，壺とボールを用いた例題における，EMアルゴリズムの具体的手続きにあたる。

EMアルゴリズムの狙い

観測できないデータ y_i が，モデル $\mathcal{M}$ から観測可能なデータ x_n を生成する過程に関わっているような確率モデルを考える.

このとき，以下の尤度関数を最大化するパラメータ$\mathcal{M}$を求めたい.

$$\underbrace{\log p(\mathbf{x}|\mathcal{M})}_{\text{不完全データの対数尤度}} = \sum_n \sum_j \underbrace{\log p\left(x_n, y_j|\mathcal{M}\right)}_{\text{完全データの対数尤度}}$$

- $\log p\left(x_n, y_j|\mathcal{M}\right)$ は計算できる.
- ただし，(x_n, y_i) は与えられず，与えられるのは x_n のみであり，y_i は事後確率 $p\left(y_j|x_n, \mathcal{M}\right)$ によってのみ与えられる.
- $p\left(y_j|x_n, \mathcal{M}\right)$ のもとでの $p(x_n, y_i|\mathcal{M})$ の期待値を考える.

$$\mathcal{Q}\left(\mathcal{M}^{\mathrm{old}}, \mathcal{M}\right) = \sum_n \sum_j p\left(y_j|x_n, \mathcal{M}^{\mathrm{old}}\right) \log p\left(x_n, y_j|\mathcal{M}\right)$$

$$\mathcal{L}(\mathbf{x},\mathcal{M}) = \sum_n \log p(x_n|\mathcal{M}) = \sum_n \sum_j \log p\left(x_n, y_j|\mathcal{M}\right)$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log p(x_n|\mathcal{M})$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log \frac{p\left(y_j \middle| x_n, \mathcal{M}\right) p(x_n|\mathcal{M})}{p\left(y_j \middle| x_n, \mathcal{M}\right)}$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log \frac{p\left(x_n, y_j|\mathcal{M}\right)}{p\left(y_j \middle| x_n, \mathcal{M}\right)}$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log p\left(x_n, y_j|\mathcal{M}\right)$$

$$- \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log p\left(y_j \middle| x_n, \mathcal{M}\right)$$

$$\mathcal{L}(\mathbf{x}, \mathcal{M}) = \mathcal{Q}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}) - \mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M})$$

$$\mathcal{Q}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}) = \sum_n \sum_j p(y_j | x_n, \mathcal{M}^{\mathrm{old}}) \log p(x_n, y_j | \mathcal{M})$$

$$\mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}) = \sum_n \sum_j p(y_j | x_n, \mathcal{M}^{\mathrm{old}}) \log p(y_j | x_n, \mathcal{M})$$

$\mathcal{Q}(\mathcal{M}^{\mathrm{old}}, \mathcal{M})$ は観測不可能な変数 y の事後分布下での完全データの対数尤度の期待値を表す.

$\mathcal{H}$関数はモデルパラメータの修正に応じて単調に減少する.

$$\mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}) \leq \mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}^{\mathrm{old}})$$

等号は $\mathcal{M}^{\mathrm{old}} = \mathcal{M}$ のとき成立する.

Proof:

$$\mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}) - \mathcal{H}(\mathcal{M}^{\mathrm{old}}, \mathcal{M}^{\mathrm{old}})$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \left\{\log p(y_j | x_n, \mathcal{M}) - p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right)\right\}$$

$$= \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \log \frac{p(y_j | x_n, \mathcal{M})}{p(y_j | x_n, \mathcal{M}^{\mathrm{old}})}$$

$$\leq \sum_n \sum_j p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \left(\frac{p(y_j | x_n, \mathcal{M})}{p(y_j | x_n, \mathcal{M}^{\mathrm{old}})} - 1\right) = \sum_n 1 - 1 = 0 \qquad (\because \ \log z \leq z - 1)$$

$$\mathcal{L}(\mathbf{x},\mathcal{M}) = \mathcal{Q}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right) - \mathcal{H}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$$

$\mathcal{H}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$ 関数は $\mathcal{M}$ の修正に対して単調減少する（前スライド参照）

→ $\mathcal{Q}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$ が増加するように $\mathcal{M}$ を修正すれば，$\mathcal{L}(\mathbf{x},\mathcal{M})$ は必ず増加する.

→ $\mathcal{Q}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$ を最大にする $\mathcal{M}$ を求めては $\mathcal{M}^{\mathrm{old}} = \mathcal{M}$ とし，$\mathcal{Q}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$ を最大にする $\mathcal{M}$ を求める，という処理を繰り返せば，$\mathcal{M}$ は $\mathcal{L}(\mathbf{x},\mathcal{M})$ の極大値を与える $\mathcal{M}$ に収束する.

不完全データ対数尤度を最大化する問題が，完全データ対数尤度の期待値（$\mathcal{Q}$関数）の最大化を繰り返す問題に置換えられた.

$$\mathcal{M} = \operatorname*{argmax}_{\mathcal{M}} \mathcal{Q}\left(\mathcal{M}^{\mathrm{old}},\mathcal{M}\right)$$

$$\mathcal{M}^{old} \leftarrow \mathcal{M}$$

EMアルゴリズムの概要

Input　学習データ $\mathrm{x}_1, \mathrm{x}_2, \cdots, x_N$，初期モデル $\mathcal{M}^0$

Initialization　適当な初期モデル $\mathcal{M}^0$ を選び，$\mathcal{M}^{\mathrm{old}}$ とする.

Repeat　収束するまでE-step（期待値計算）とM-step（最大化）を繰り返す.

E-step (Expectation step): $Q\left(M^{\mathrm{old}}, \mathcal{M}\right)$ を計算する.

$$Q\left(M^{\mathrm{old}}, \mathcal{M}\right) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{K} p\left(y_j \middle| x_n, \mathcal{M}^{\mathrm{old}}\right) \cdot \log p\left(x_n, y_j \middle| \mathcal{M}\right)$$

M-step (Maximization step): $Q\left(M^{\mathrm{old}}, \mathcal{M}\right)$ を最大にする $\mathcal{M}$ を推定する.

$$\mathcal{M} = \operatorname*{argmax}_{\mathcal{M}} Q\left(M^{\mathrm{old}}, \mathcal{M}\right)$$

$$\mathcal{M}^{old} \leftarrow \mathcal{M}$$

Output　収束したパラメータ $\mathcal{M}$

EMアルゴリズムによる最尤推定の例

混合正規分布のパラメタ推定問題をEMアルゴリズムで解く．

Y は，$m_1, m_2, \cdots, m_K$ の確率で $1, 2, \cdots, K$ の値をとる確率変数.

$$\Pr(Y = k | \mathcal{M}) = m_k$$

Xは，$Y = k$ のとき，正規分布 $\mathcal{N}\left(X; \mu_k, \sigma_k^2\right)$ に従う確率変数.

$$\Pr(X = x_i | Y = k, \mathcal{M}) = \mathcal{N}\left(x_i; \mu_k, \sigma_k^2\right)$$

このとき，混合正規分布は，観測可能な確率変数 X と観測不可能な確率変数 Y を用いて，次のように表される．

$$\Pr(X = x_i | M) = \sum_k m_k \mathcal{N}(x_i; \mu_k, \sigma_k{}^2)$$

$$= \sum_k \Pr(Y = k | \mathcal{M}) \Pr(X = x_i | Y = k, \mathcal{M})$$

Q関数は，

$$Q(\mathcal{M}, \mathcal{M}') = E[\log \Pr(\boldsymbol{x}, y | \mathcal{M}') | \boldsymbol{x}, \mathcal{M}]$$

$$= \sum_i \sum_k \Pr(Y = k | X = x_i, \mathcal{M}) \log \Pr(X = x_i, Y = y | \mathcal{M}')$$

$$Q(\mathcal{M},\mathcal{M}') = \sum_i \sum_k \Pr(Y=k|x_i,\mathcal{M}) \log \Pr(x_i, Y=k|\mathcal{M}')$$

$$= \sum_i \sum_k a(k;x_i,\mathcal{M}) \log m'_k \mathcal{N}\left(x_i;\mu'_k,{\sigma'_k}^2\right)$$

$$= \sum_i \sum_k a(k;x_i,\mathcal{M}) \left\{\log m'_k - \frac{\log 2\pi}{2} - \frac{\log {\sigma'_k}^2}{2} - \frac{(x_i-\mu'_k)^2}{2{\sigma'_k}^2}\right\}$$

x_iの出現におけるk番目の分布の寄与度。
x_i が k番目の分布から出現した確率。

ここで,

$$a(k;x_i,\mathcal{M}) = \Pr(Y=k|X=x_i,\mathcal{M}) = \frac{\Pr(X=x_i,Y=k|\mathcal{M})}{\Pr(X=x_i|\mathcal{M})}$$

$$= \frac{m'_k \mathcal{N}\left(x_i;\mu'_k,{\sigma'_k}^2\right)}{\sum_l m'_k \mathcal{N}\left(x_i;\mu'_l,{\sigma'_l}^2\right)}$$

新モデルのパラメタ　μ'_k, σ'_k, m'_k は，以下のように求まる

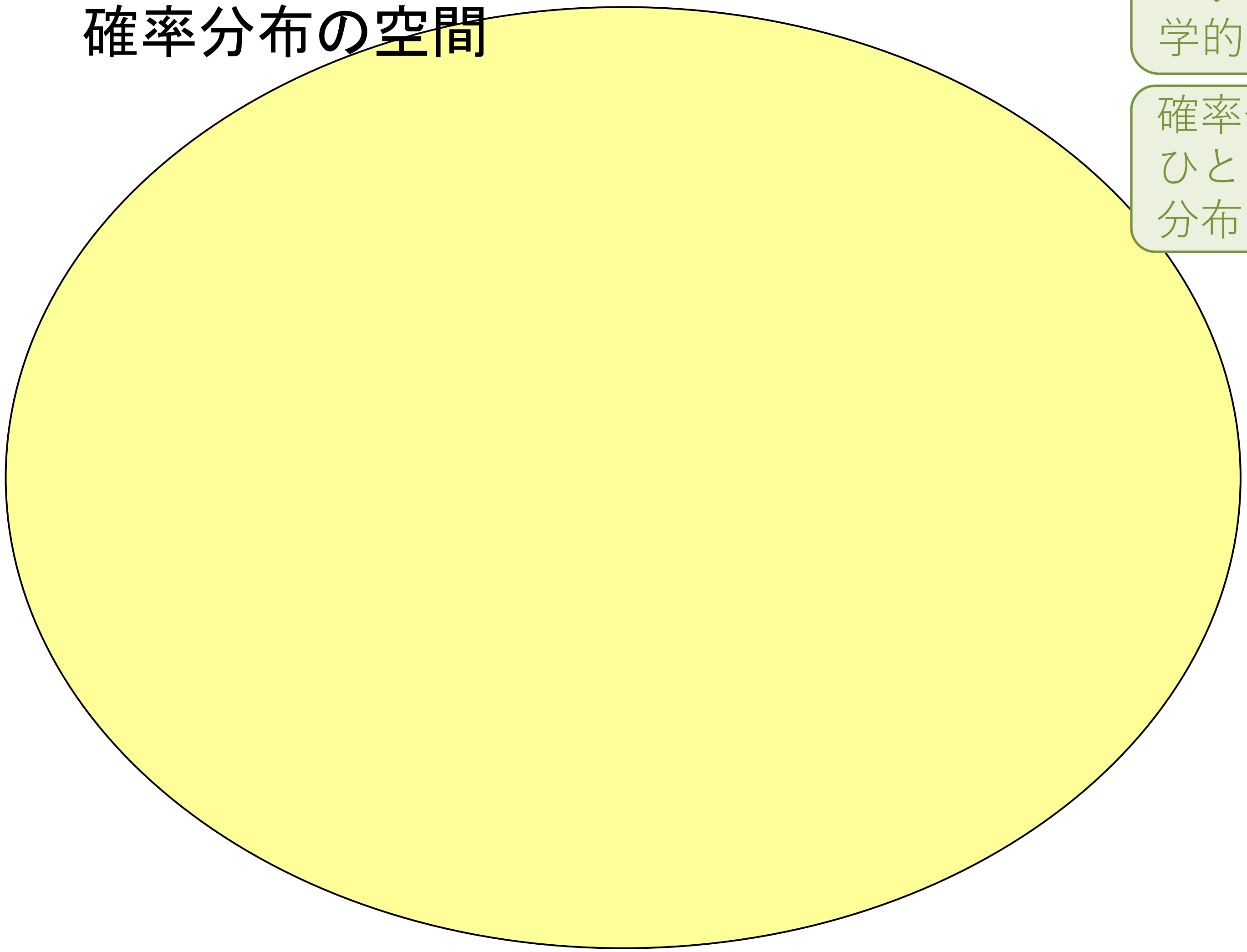
$$\mu'_k = \frac{\sum_i a(k; x_i, \mathcal{M}) x_i}{\sum_i a(k; x_i, \mathcal{M})}$$

$$\sigma'_k = \frac{\sum_i a(k; x_i, \mathcal{M})(x_i - \mu'_k)^2}{\sum_i a(k; x_i, \mathcal{M})}$$

$$m'_k = \frac{\sum_i a(k; x_i, \mathcal{M})}{\sum_l \sum_i a(l; x_i, \mathcal{M})}$$

上は，壺とボールの問題と全く同じ処理に相当する。

推定問題の情報幾何学的描像



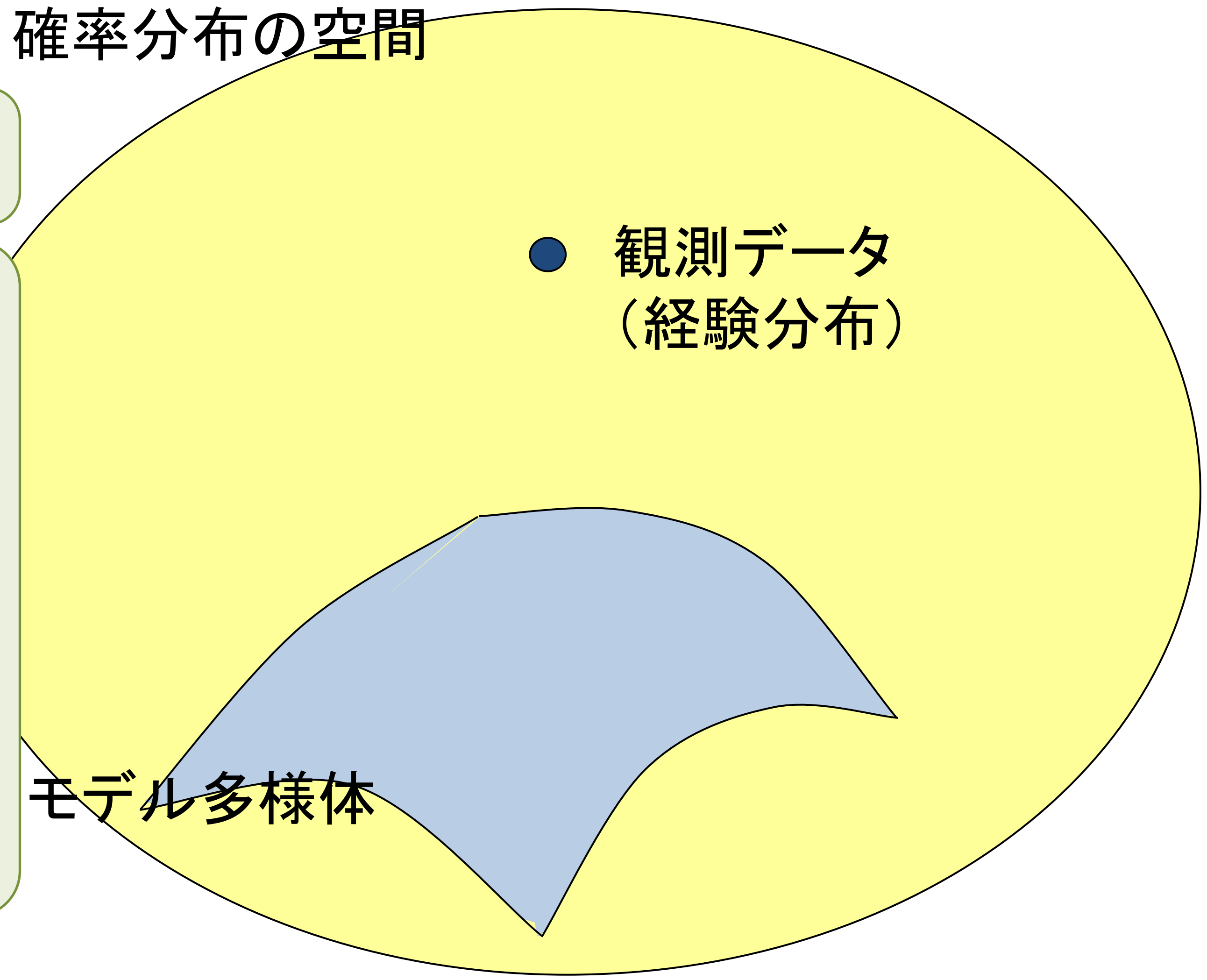
パラメタ推定の問題を幾何学的に解釈してみよう。

確率分布の空間を考える。ひとつひとつの点が，確率分布を与える。

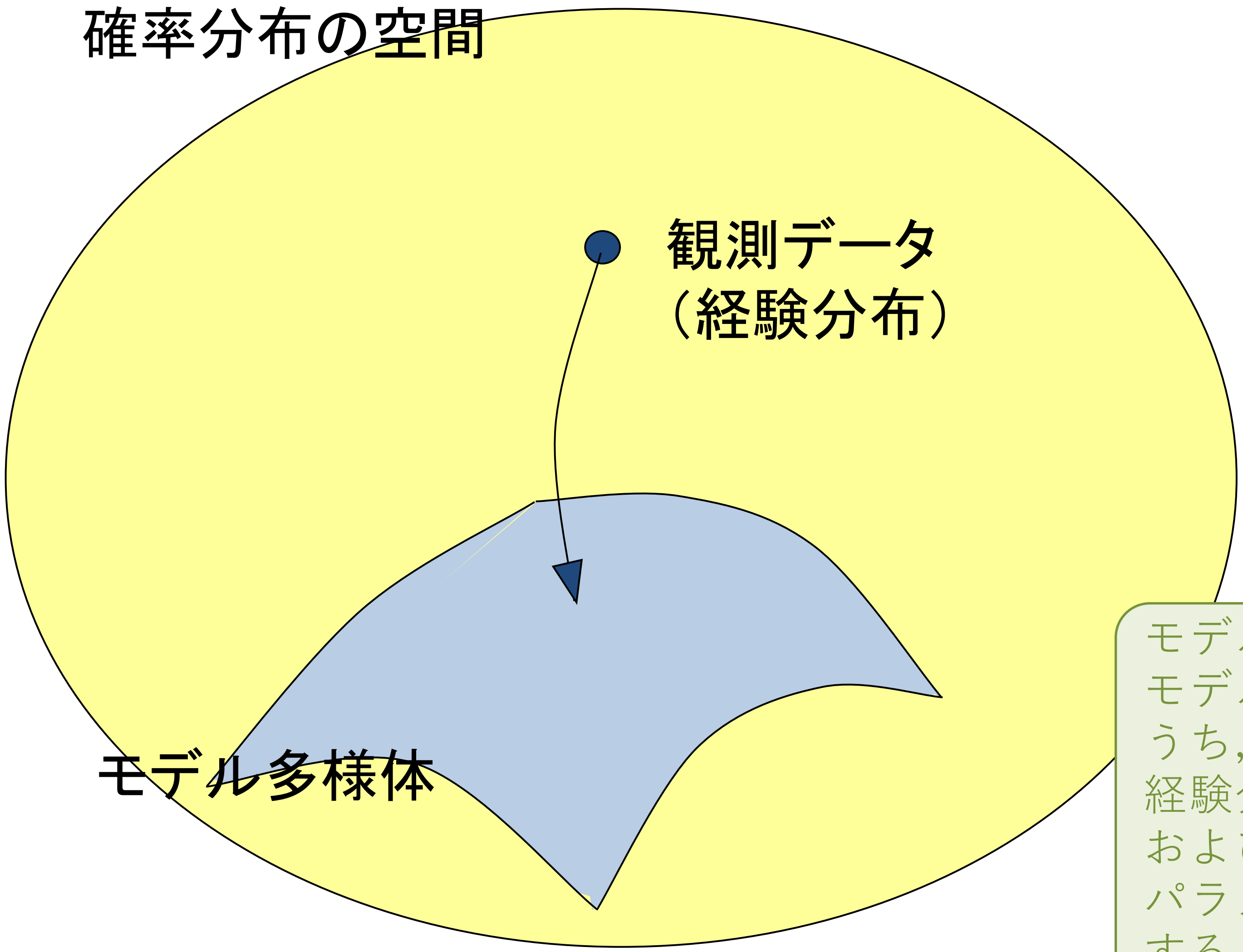
推定問題の情報幾何学的描像

データを観測すれば，それは一つの分布を与える。

また，モデルの形を与えれば，それは確率分布の空間の部分空間(多様体)を与える。すなわち，分布の形を決めれば，パラメタに応じて様々な分布を表現することができるが，モデルですべての分布を表現できるわけではないので，表現できるものは，分布空間の部分空間となる。

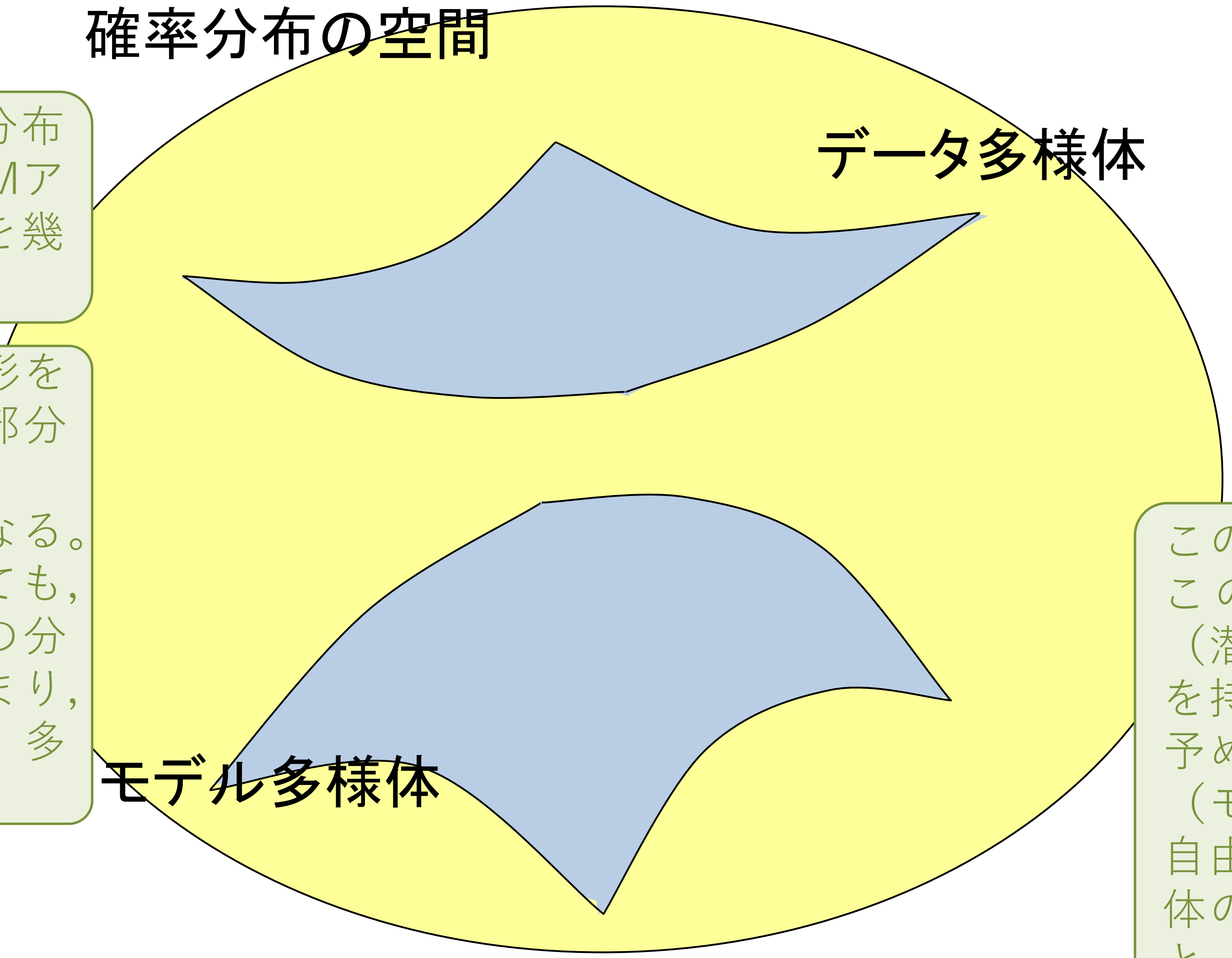


推定問題の情報幾何学的描像



モデルのパラメタ推定とは，モデルの表現しうる分布のうち，観測データが与える経験分布に一番近いもの，およびその分布を表現するパラメタを探す処理に相当する。

ＥＭアルゴリズムの情報幾何学的描像



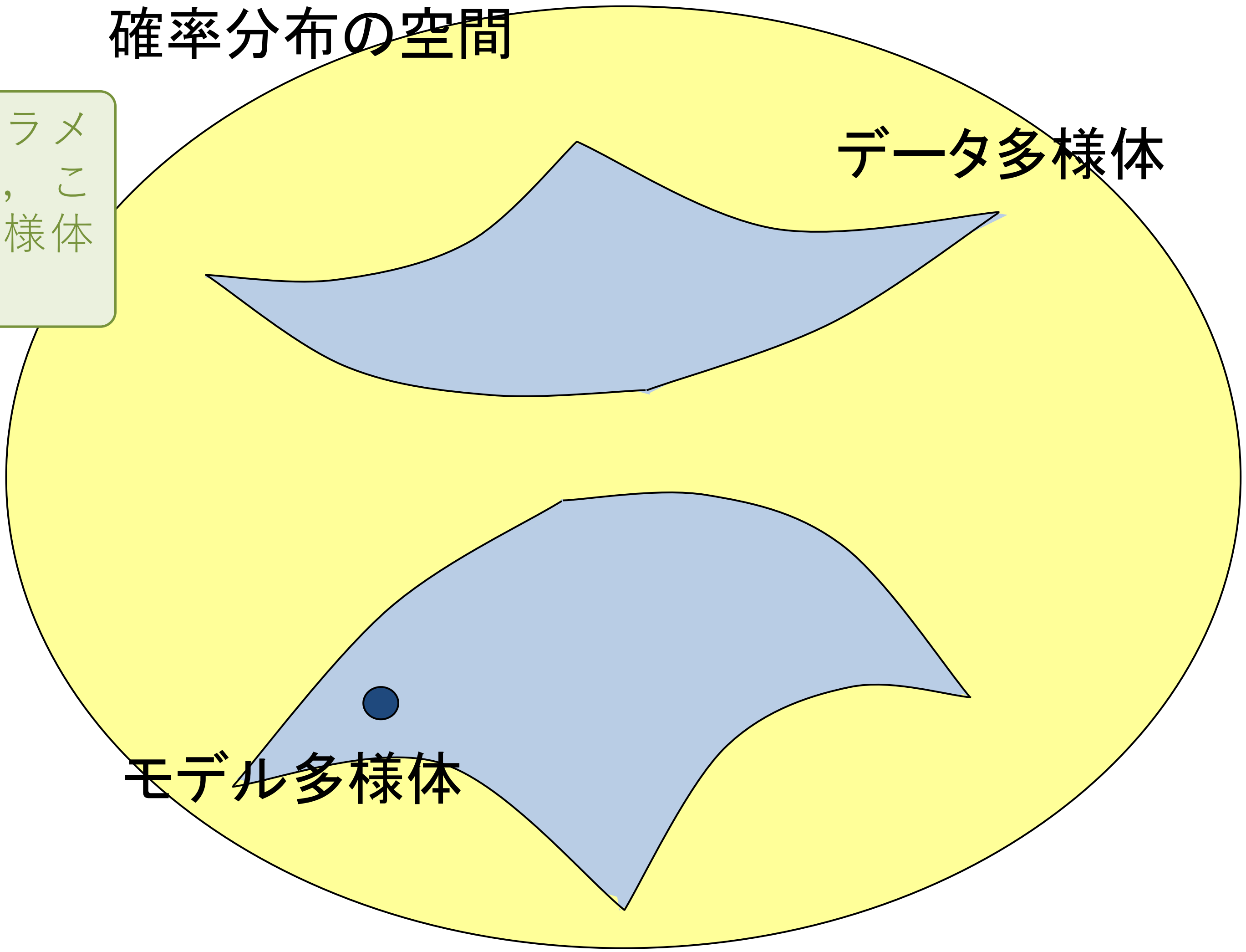
同様に潜在変数を含む分布のパラメタ推定問題をEMアルゴリズムで解くことを幾何学的に解釈する。

先と同様に，モデルの形を決めれば，分布空間の部分空間を与える。
経験分布は，先とは異なる。同じ観測データであっても，それを与える潜在変数の分だけ多様性がある。つまり，データも（点ではなく）多様体となる。

このとき，推定の問題は，この観測データが与える（潜在変数の分だけ自由度を持つ）データ多様体と，予め定めたモデルが与える（モデルパラメタの分だけ自由度を持つ）モデル多様体の最も近い分布を探すこととなる。

ＥＭアルゴリズムの情報幾何学的描像

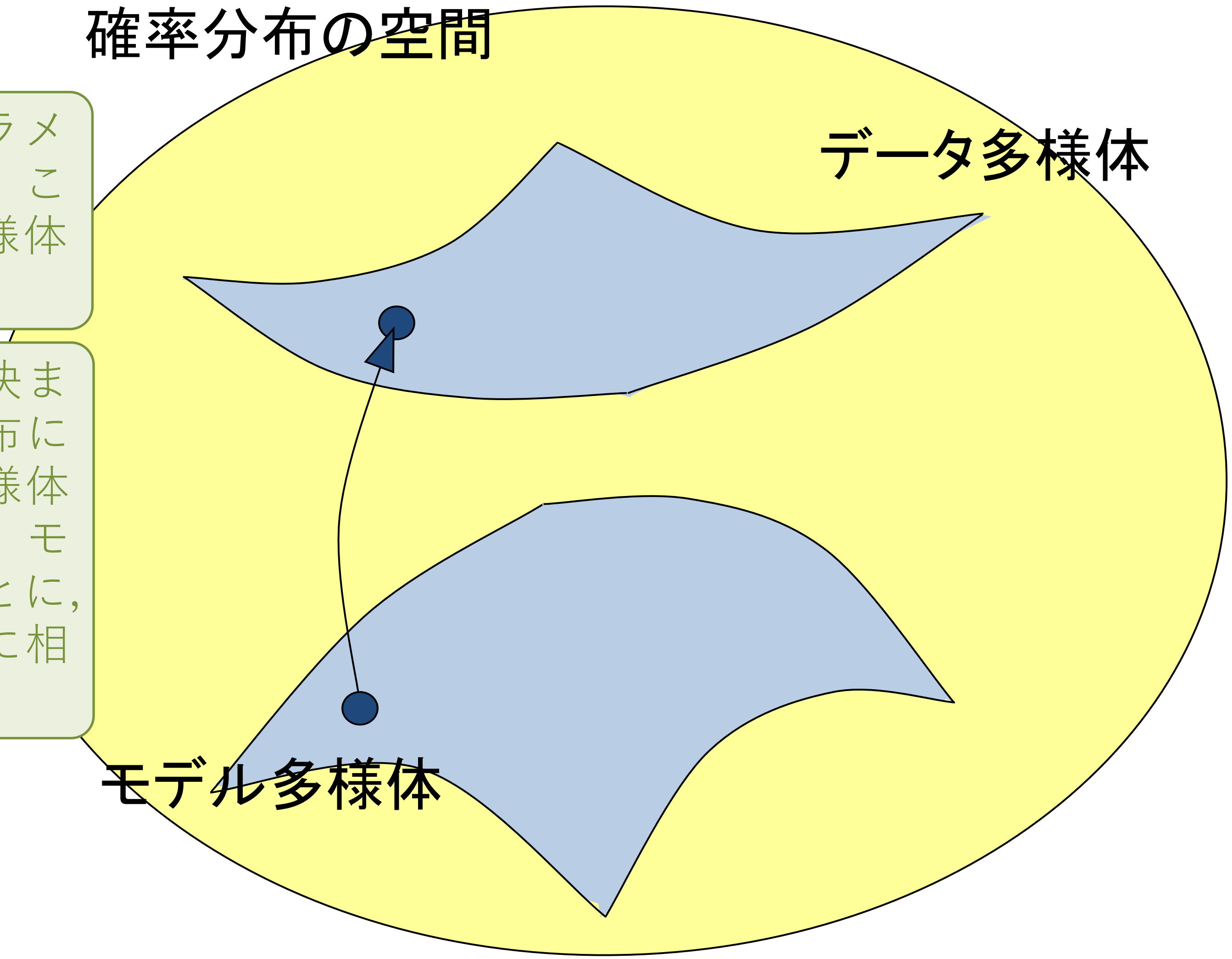
まず，適当なモデルパラメタを初期値として与え，これによって，モデル多様体上の一点を決める。



ＥＭアルゴリズムの情報幾何学的描像

まず，適当なモデルパラメタを初期値として与え，これによって，モデル多様体上の一点を決める。

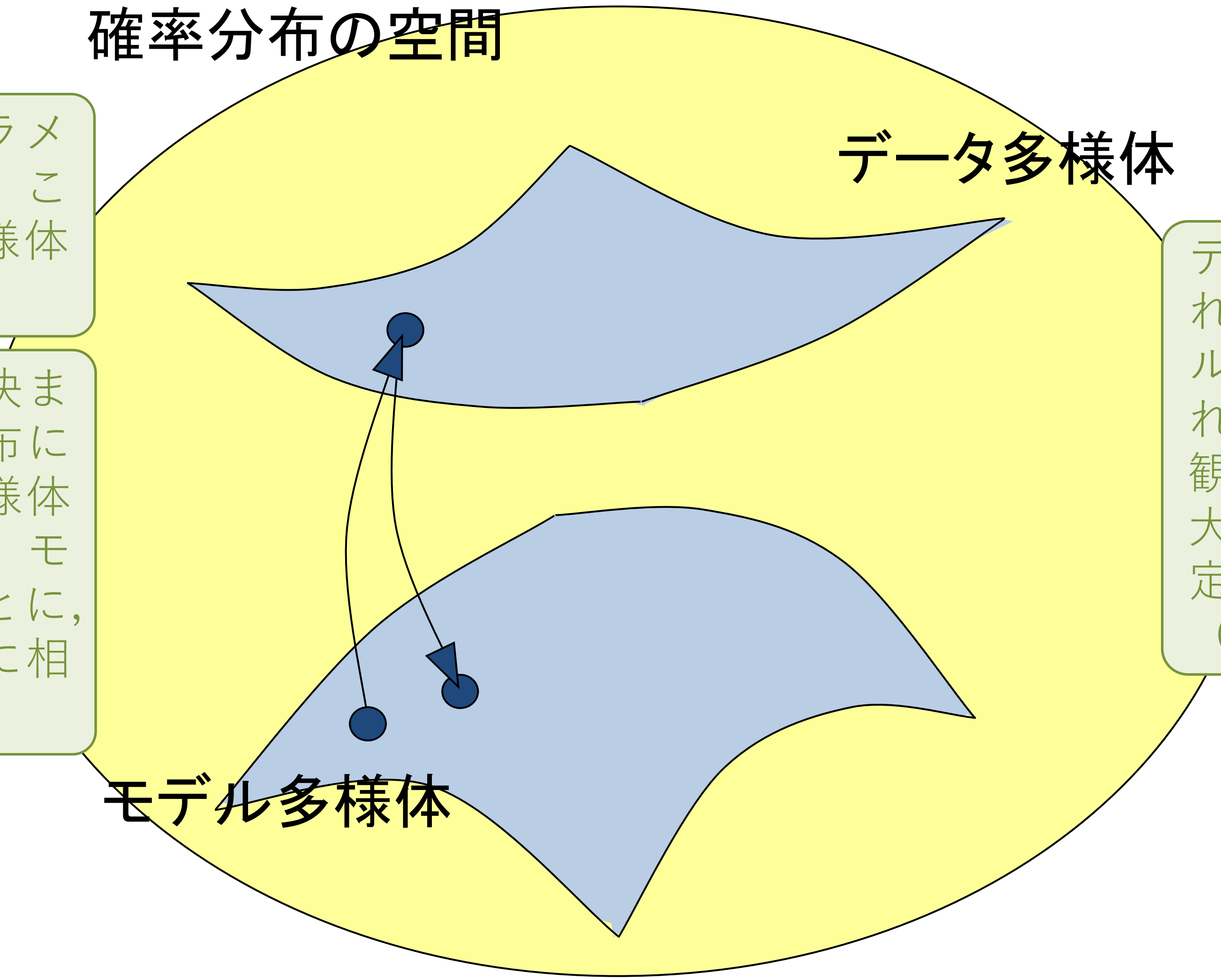
モデル多様体上の点が決まれば，それが与える分布にもっとも近いデータ多様体上の点を探す。これは，モデルパラメタの値をもとに，潜在変数を定めることに相当する。（E-step）



ＥＭアルゴリズムの情報幾何学的描像

まず，適当なモデルパラメタを初期値として与え，これによって，モデル多様体上の一点を決める。

モデル多様体上の点が決まれば，それが与える分布にもっとも近いデータ多様体上の点を探す。これは，モデルパラメタの値をもとに，潜在変数を定めることに相当する。（E-step）

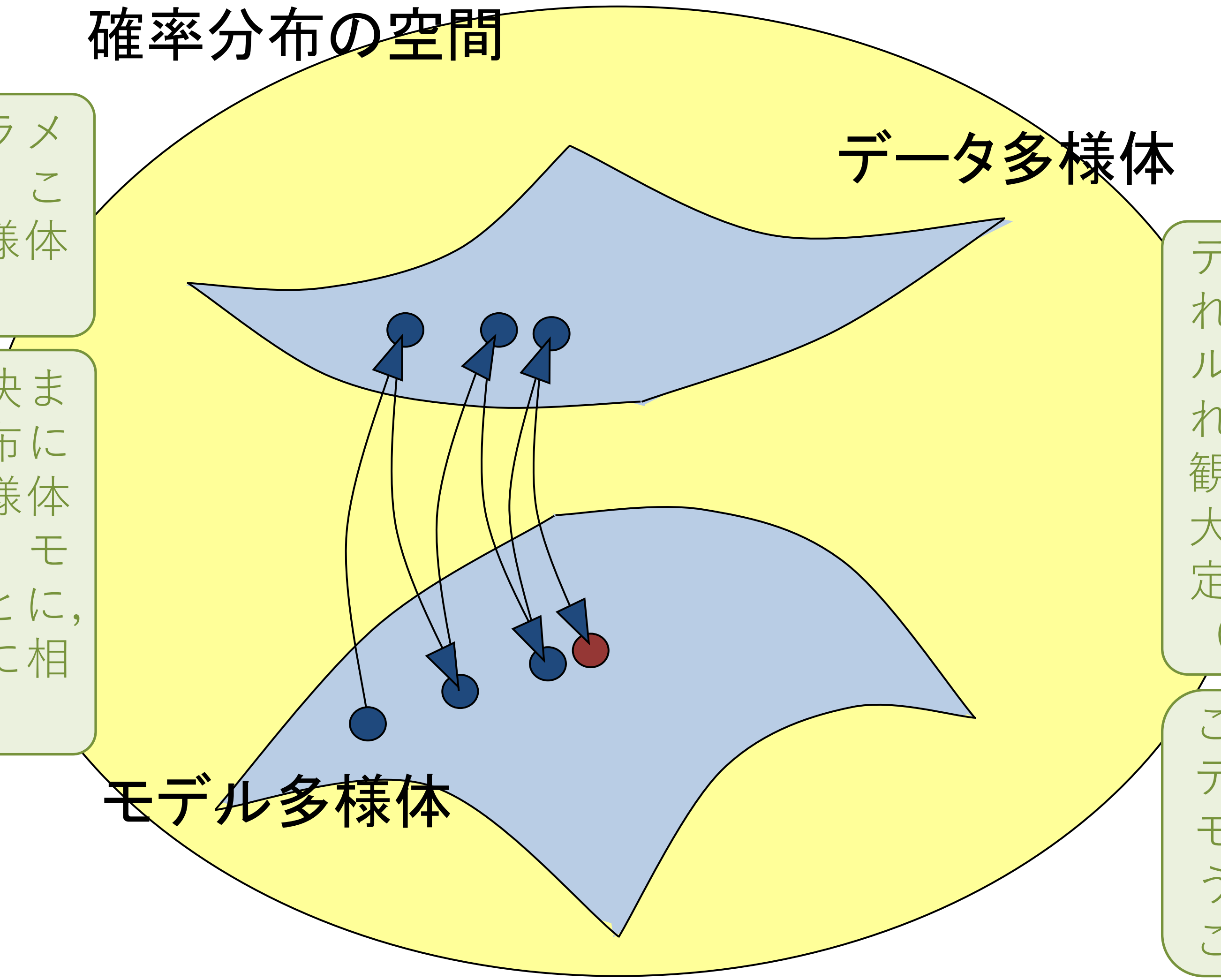


データ多様体上の点が決まれば，これに最も近いモデル多様体上の点を探す。これは完全データ(潜在変数と観測データの対)の尤度を最大化するモデルパラメタを定めることに相当する。（M-step)。

ＥＭアルゴリズムの情報幾何学的描像

まず，適当なモデルパラメタを初期値として与え，これによって，モデル多様体上の一点を決める。

モデル多様体上の点が決まれば，それが与える分布にもっとも近いデータ多様体上の点を探す。これは，モデルパラメタの値をもとに，潜在変数を定めることに相当する。（E-step）



データ多様体上の点が決まれば，これに最も近いモデル多様体上の点を探す。これは完全データ(潜在変数と観測データの対)の尤度を最大化するモデルパラメタを定めることに相当する。（M-step)。

これらの処理を繰り返せば，データ多様体が表す分布と，モデル多様体が表す分布のうち，一番近いものを探すことができる。

まとめ

- 識別対象のデータがモデルから生起する確率をもとに識別を行う方法を「生成モデルに基づくパターン認識」という。
- 生成モデルに基づくパターン認識では，複雑な分布を表現する枠組みが必要になる。混合正規分布（GMM）は，そのような用途に用いられる代表的な確率モデルである。
- GMMのパラメタ推定には，EMアルゴリズムが用いられる。
- EMアルゴリズムは，モデルパラメタと，潜在変数の事後確率を交互に推定する処理を繰り返すことで，モデルパラメタの最尤推定を行う。