

第 17 回先端的データベースとWeb技術動向講演会  
(ACM SIGMOD 日本支部 第 54 回支部大会)  
[2013年10月5日]

# クラスタリング技術の基礎と応用

日本電信電話(株)  
NTT サービスエボリューション研究所  
機械学習・データ科学センタ  
澤田 宏

# 機械学習技術

---

- 教師あり学習

- » データと教師ラベルのペアを使って学習する

- 顔画像認識, 音声認識, 機械翻訳

- » 教師ラベル付加コストへの対処

- 半教師あり学習, クラウドソーシング

- 教師なし学習

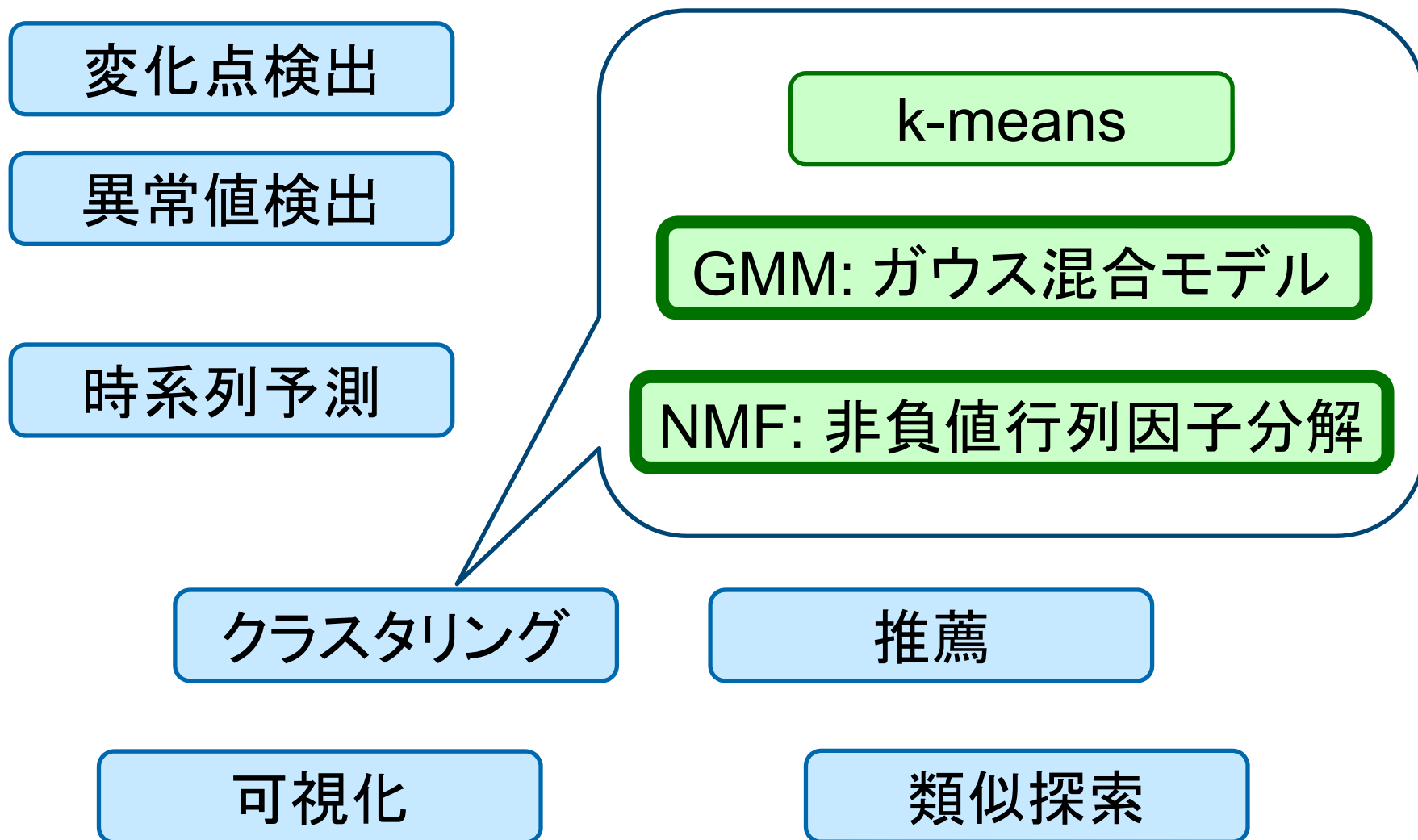
- » データのみ, 教師ラベル無し

- » ビッグデータ時代に合っている？

- データ量が膨大過ぎて, 人手介入できない

- データが多様過ぎて, 何を学習したいか分からない

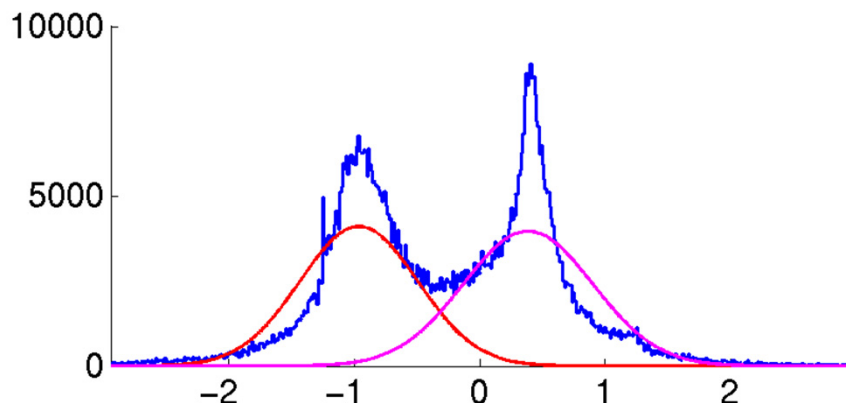
# 教師なし機械学習技術



# 2種類のクラスタリング手法

GMM:

Gaussian Mixture Model



代表的なクラスタリング手法  
であるk-meansの一般化

NMF:

Non-negative Matrix Factorization

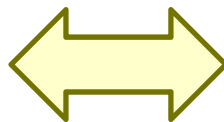
$$\mathbf{X} = \mathbf{T} \mathbf{V}$$

特異値分解SVD等の  
行列分解に非負制約を導入



クラスタリング的解釈が可能に

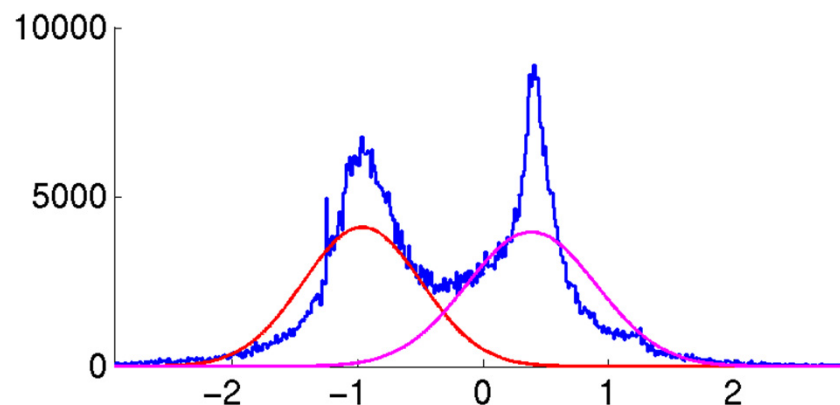
- 機械学習の教科書
- 確率の基礎



- 教科書はまだあまりない
- 実用性の高いツール

# ガウス混合モデル

## GMM: Gaussian Mixture Model

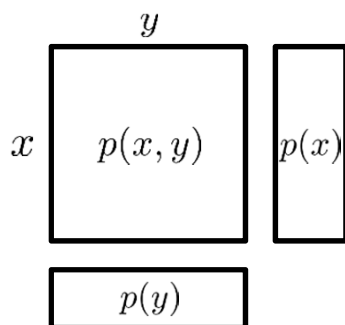


# アウトライン(GMM)

## 確率の基礎

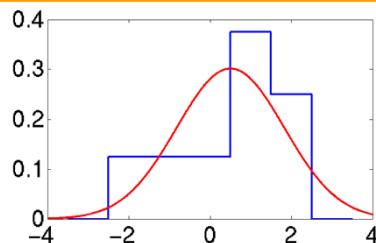
①

同時確率  
周辺確率  
条件付き確率



②

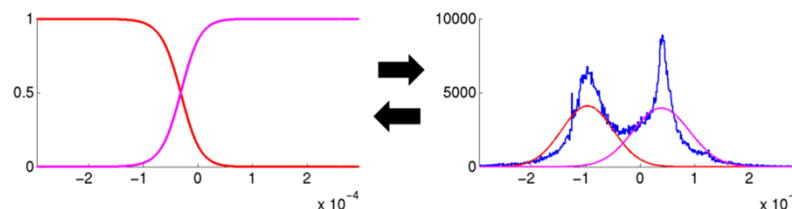
ガウス分布  
最尤推定



## 音源分離手法として

③

ガウス混合モデル(GMM)  
EMアルゴリズム



# 非常に単純化された例

各ユーザ(群)が各商品(群)を何個買ったかというデータ

|        |                                                                                   | 商品(群)                                                                             |                                                                                   |                                                                                   | 各ユーザの<br>購買数 |
|--------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------|
|        |                                                                                   |  |  |  |              |
| ユーザ(群) |  | 2                                                                                 | 0                                                                                 | 0                                                                                 |              |
|        |  | 0                                                                                 | 3                                                                                 | 1                                                                                 |              |
|        |  | 1                                                                                 | 1                                                                                 | 2                                                                                 |              |

|   |   |   |
|---|---|---|
| 3 | 4 | 3 |
|---|---|---|

各商品の販売数

全部で10(万)個売れました

# 確率として考えてみましょう

各ユーザが各商品を何個買ったかというデータ

| $p(x, y)$  |       | 商品 $y$ |       |       | $p(x)$ |         |
|------------|-------|--------|-------|-------|--------|---------|
|            |       | $b_1$  | $b_2$ | $b_3$ |        |         |
| ユーザ<br>$x$ | $a_1$ | 0.2    | 0     | 0     | 0.2    | 各ユーザの購買 |
|            | $a_2$ | 0      | 0.3   | 0.1   | 0.4    |         |
|            | $a_3$ | 0.1    | 0.1   | 0.2   | 0.4    |         |

| $p(y)$ | 0.3 | 0.4 | 0.3 |
|--------|-----|-----|-----|
|--------|-----|-----|-----|

各商品の販売

誰かが何かを1個買った場合  
どれに当るかの確率

# それぞれの確率の意味

$p(x, y)$  同時確率

» 誰か  $x$  が何か  $y$  を買った  
確率  $p(x = a_3, y = b_3) = 0.2$

$p(x)$  周辺確率

» (何が売れたかは気にせずに)  
誰か  $x$  が買った確率  
 $p(x = a_3) = 0.4$

$p(y)$  周辺確率

» (誰が買ったかは気にせずに)  
何か  $y$  が売れた確率  
 $p(y = b_3) = 0.3$

|       | $b_1$ | $b_2$ | $b_3$ |     |
|-------|-------|-------|-------|-----|
| $a_1$ | 0.2   | 0     | 0     | 0.2 |
| $a_2$ | 0     | 0.3   | 0.1   | 0.4 |
| $a_3$ | 0.1   | 0.1   | 0.2   | 0.4 |
|       | 0.3   | 0.4   | 0.3   |     |

# 同時確率と周辺確率の関係

ユーザ  $a_3$  が買った確率

$$p(x = a_3) = p(x = a_3, y = b_1) + p(x = a_3, y = b_2) + p(x = a_3, y = b_3)$$

$$= \sum_{y=b_1}^{b_3} p(x = a_3, y)$$

一般的には **和の法則**

$$p(x) = \sum_{y=b_1}^{b_3} p(x, y)$$

|       | $b_1$ | $b_2$ | $b_3$ |     |
|-------|-------|-------|-------|-----|
| $a_1$ | 0.2   | 0     | 0     | 0.2 |
| $a_2$ | 0     | 0.3   | 0.1   | 0.4 |
| $a_3$ | 0.1   | 0.1   | 0.2   | 0.4 |
|       | 0.3   | 0.4   | 0.3   |     |

さらに  $\sum_{x=a_1}^{a_3} p(x) = 1$  (確率の要件)

# 3番目のユーザが買った場合に着目

何が売れたかを確率として考える

|     |                                                                                   | 商品                                                                                |                                                                                   |                                                                                   |
|-----|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
|     |                                                                                   |  |  |  |
| ユーザ |  | 2                                                                                 | 0                                                                                 | 0                                                                                 |
|     |  | 0                                                                                 | 3                                                                                 | 1                                                                                 |
|     |  | 1                                                                                 | 1                                                                                 | 2                                                                                 |

このユーザは  
全部で4個買いました

|            |       | 商品 $y$ |       |       |
|------------|-------|--------|-------|-------|
|            |       | $b_1$  | $b_2$ | $b_3$ |
| ユーザ<br>$x$ | $a_1$ |        |       |       |
|            | $a_2$ |        |       |       |
|            | $a_3$ | 0.25   | 0.25  | 0.5   |

$$p(y|x = a_3)$$

条件付き確率

# それぞれの人が買った場合に着目

何が売れたかを確率として考える

|     |                                                                                   | 商品                                                                                |                                                                                   |                                                                                   |    |
|-----|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|----|
|     |                                                                                   |  |  |  |    |
| ユーザ |  | 2                                                                                 | 0                                                                                 | 0                                                                                 | ÷2 |
|     |  | 0                                                                                 | 3                                                                                 | 1                                                                                 | ÷4 |
|     |  | 1                                                                                 | 1                                                                                 | 2                                                                                 | ÷4 |

それぞれのユーザが  
買った数で割る

|            |       | 商品 $y$ |       |       |
|------------|-------|--------|-------|-------|
|            |       | $b_1$  | $b_2$ | $b_3$ |
| ユーザ<br>$x$ | $a_1$ | 1      | 0     | 0     |
|            | $a_2$ | 0      | 0.75  | 0.25  |
|            | $a_3$ | 0.25   | 0.25  | 0.5   |

$p(y|x)$

条件付き確率

商品の売れ行きはユーザ(の嗜好)による

# 同時確率の分解

同時確率

周辺確率

条件付き確率

$p(x, y)$

|     |     |     |
|-----|-----|-----|
| 0.2 | 0   | 0   |
| 0   | 0.3 | 0.1 |
| 0.1 | 0.1 | 0.2 |

$p(x)$

|     |
|-----|
| 0.2 |
| 0.4 |
| 0.4 |

$p(y|x)$

|      |      |      |
|------|------|------|
| 1    | 0    | 0    |
| 0    | 0.75 | 0.25 |
| 0.25 | 0.25 | 0.5  |

=

×

$$p(x = a_2, y = b_2) = p(x = a_2) p(y = b_2 | x = a_2)$$

一般的には

$$p(x, y) = p(x)p(y|x)$$

積の法則

# ①のまとめ

---

同時確率

$$p(x, y)$$

周辺確率

$$p(x) \quad p(y)$$

条件付き確率

$$p(y|x) \quad p(x|y)$$

和の法則

$$p(x) = \sum_{y=b_1}^{b_3} p(x, y)$$

積の法則

$$p(x, y) = p(x)p(y|x)$$

ベイズの定理

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

説明はしませんでした  
が、③で  
用います

# ガウス分布(正規分布)

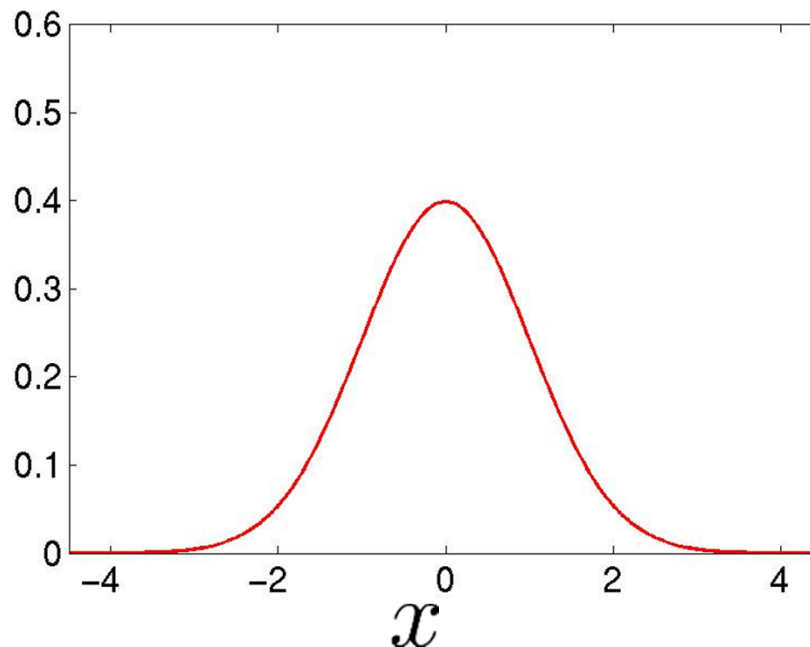
平均  $\mu$  と分散  $\sigma^2$  が決まれば一意に定まる

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

平均  $\mu = 0$

分散  $\sigma^2 = 1$

のガウス分布の  
確率密度

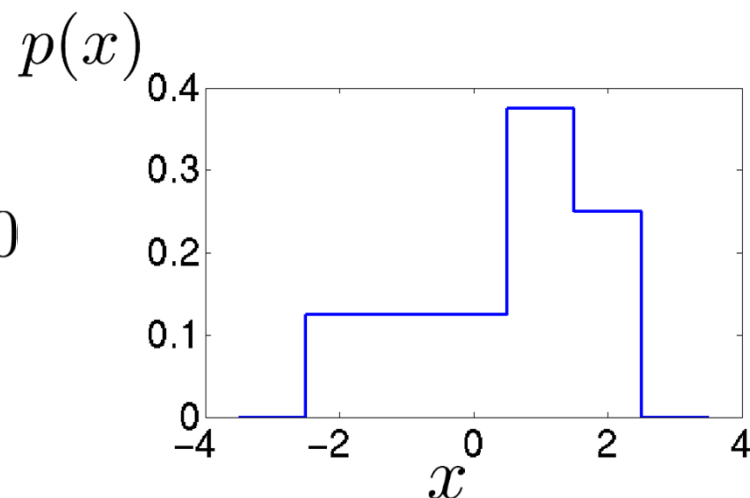


$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

# データをガウス分布でモデル化

8個の数値からなるデータ

$$x_1, \dots, x_8 = 1, 2, -1, 1, -2, 2, 1, 0$$

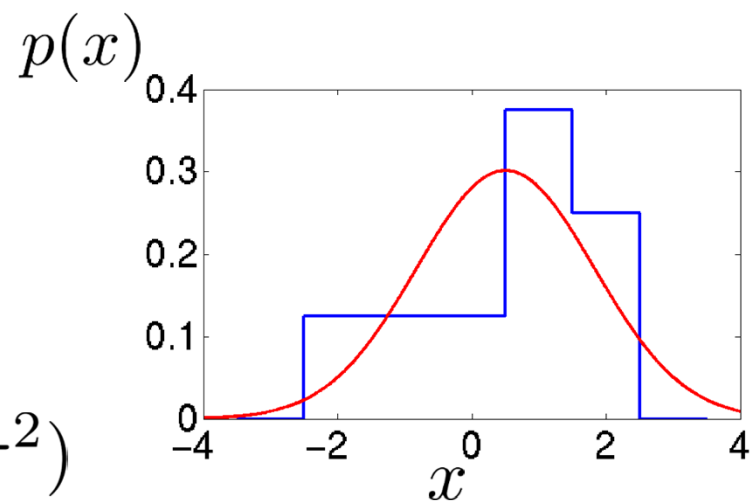


平均

$$\mu = \frac{1}{8} \sum_{n=1}^8 x_n = 0.5$$

分散

$$\sigma^2 = \frac{1}{8} \sum_{n=1}^8 (x_n - \mu)^2 = 1.75$$



赤色のガウス分布  $\mathcal{N}(x|\mu, \sigma^2)$

# 平均と分散は最尤推定による

## 最尤推定

$x_1, \dots, x_8 = 1, 2, -1, 1, -2, 2, 1, 0$  の

尤度  $\prod_{n=1}^8 p(x_n | \mu, \sigma^2) = \prod_{n=1}^8 \mathcal{N}(x_n | \mu, \sigma^2)$

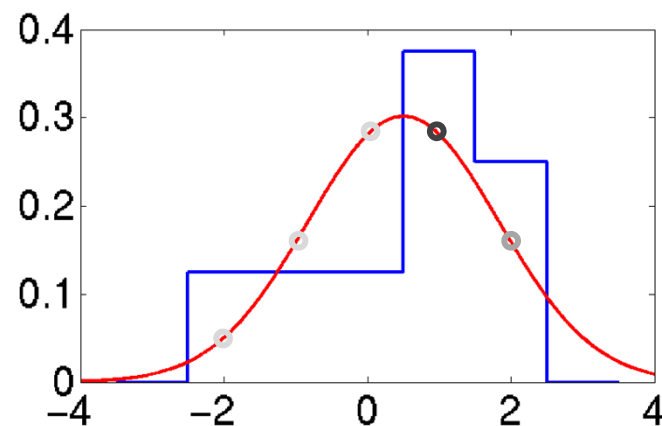
を最大化するパラメータ  $\mu, \sigma^2$  として定義される

↓ 一般には計算の簡便さから

## 対数尤度

$$\begin{aligned} & \sum_{n=1}^8 \log p(x_n | \mu, \sigma^2) \\ &= \sum_{n=1}^8 \log \mathcal{N}(x_n | \mu, \sigma^2) \\ &= - \sum_{n=1}^8 \left[ \frac{1}{2} \log 2\pi\sigma^2 + \frac{(x_n - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

を最大化する

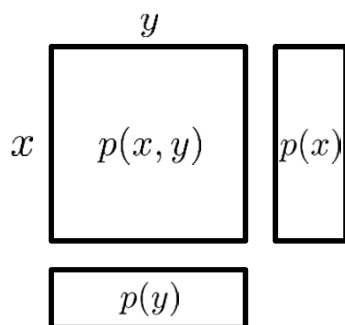


# アウトライン(GMM)

## 確率の基礎

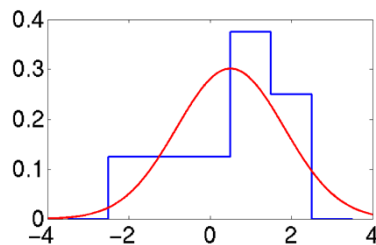
①

同時確率  
周辺確率  
条件付き確率



②

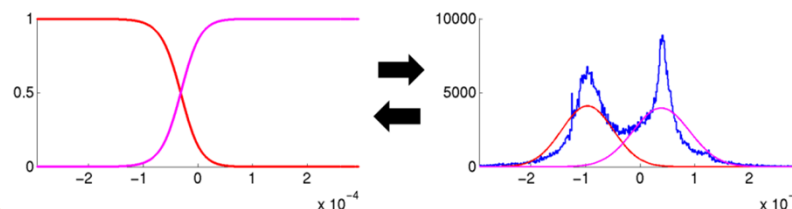
ガウス分布  
最尤推定



## 音源分離手法として

③

ガウス混合モデル(GMM)  
EMアルゴリズム



### ③の流れ: 音源分離手法

---

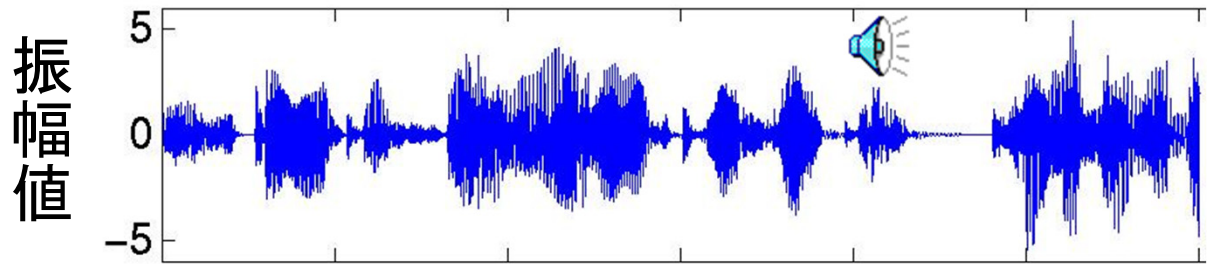
1. 音の時間周波数表現を説明する
2. マイクロホンを2個用いることで, 音の到達時間差  
→ 到来方向が推定できることを説明する
3. 時間差をGMMでモデル化する
4. GMMをEMアルゴリズムで解く
5. 混ざった二つの音の音源分離を行う

# 音の時間周波数表現

波形

$t$  時間

$x(t)$  振幅値



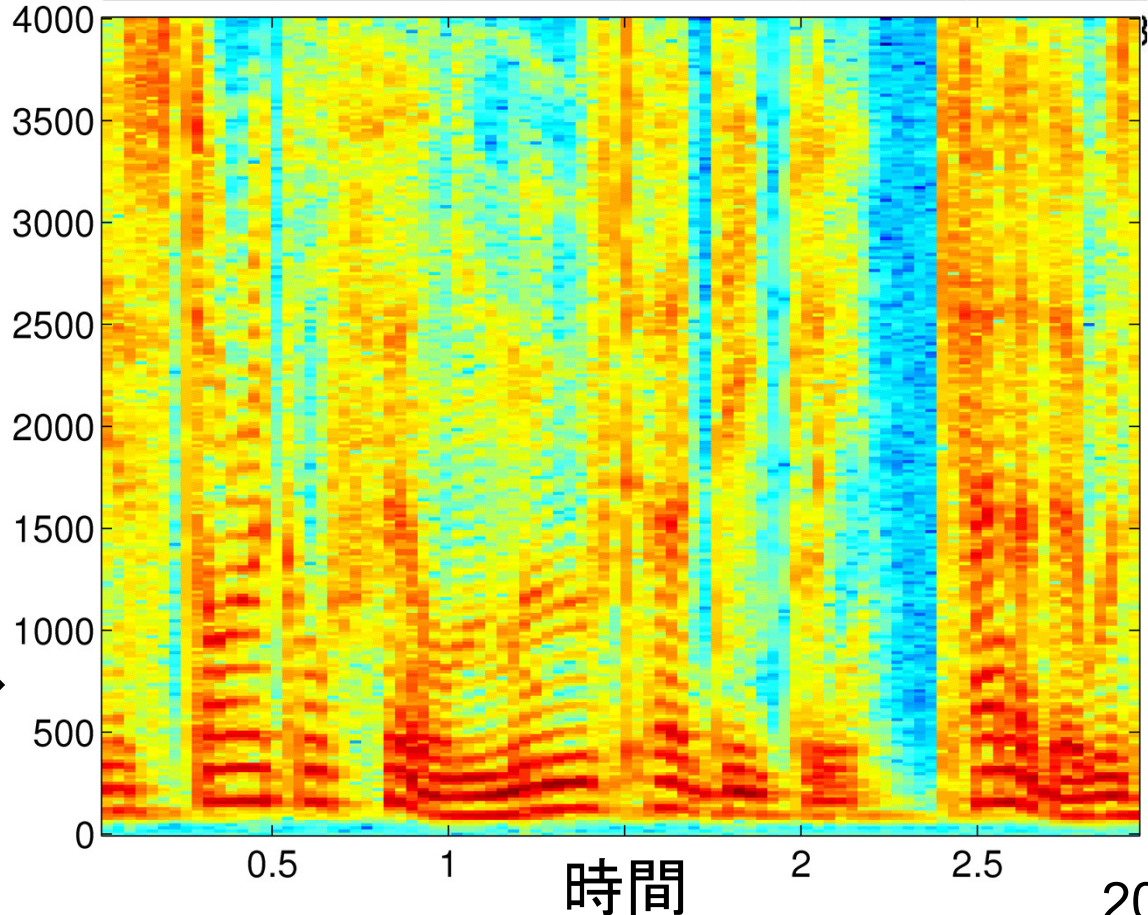
スペクトログラム

$t$  時間

$f$  周波数

$x(t, f)$  複素数

周波数



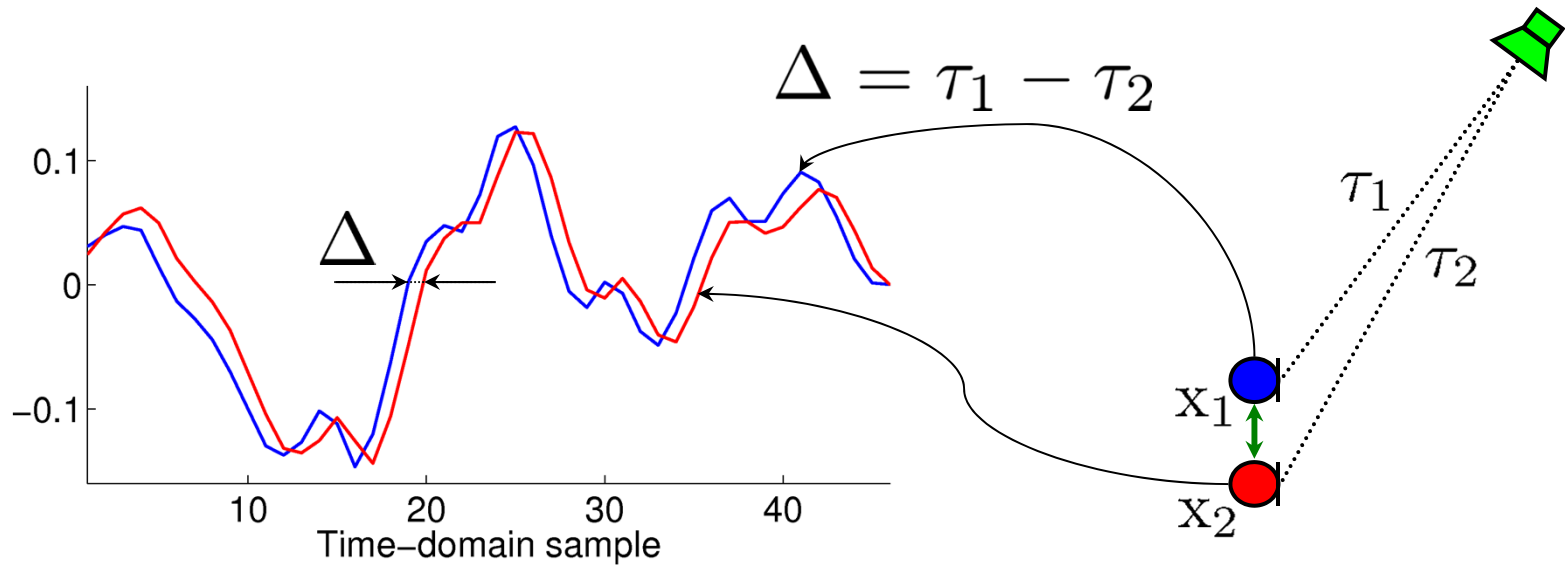
絶対値の対数を表示



赤いほど大きな値  
青は小さな値

# 音の到達時間差

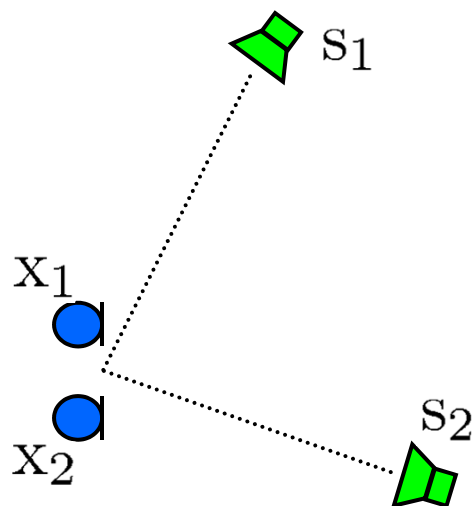
- 音源からマイクロホンまでの距離が異なるため到達時間差が発生する



- 時間周波数毎に計算できる

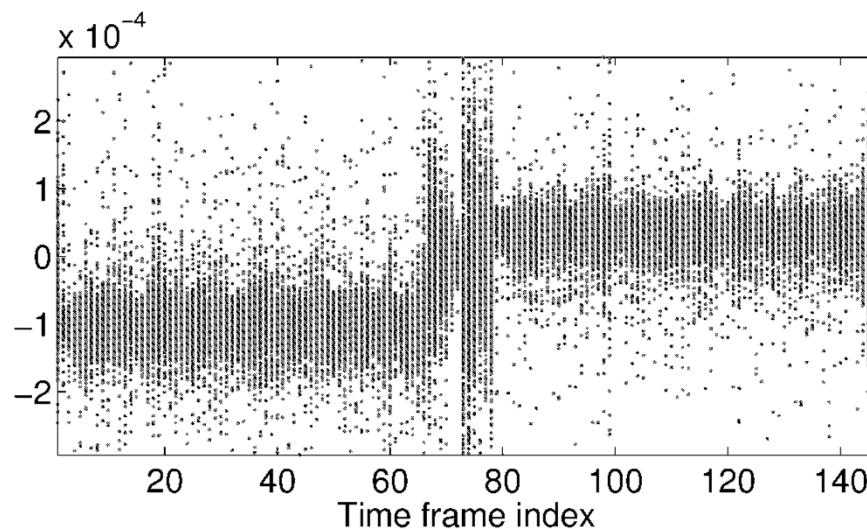
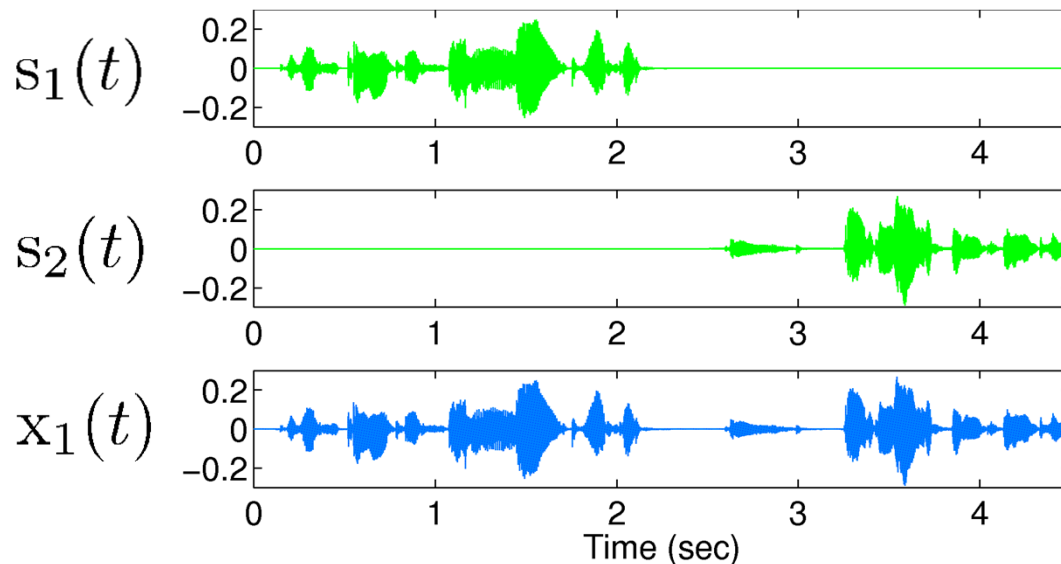
$$\Delta(t, f) = \frac{\arg[x_1(t, f)/x_2(t, f)]}{-2\pi f}$$

# 時間周波数毎の到達時間差推定



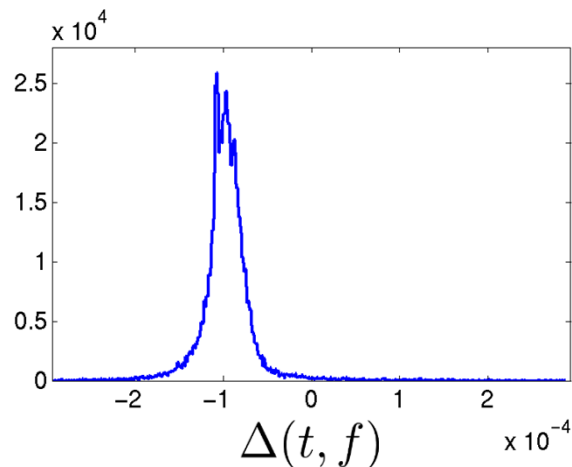
順番に話している

$$\Delta(t, f)$$

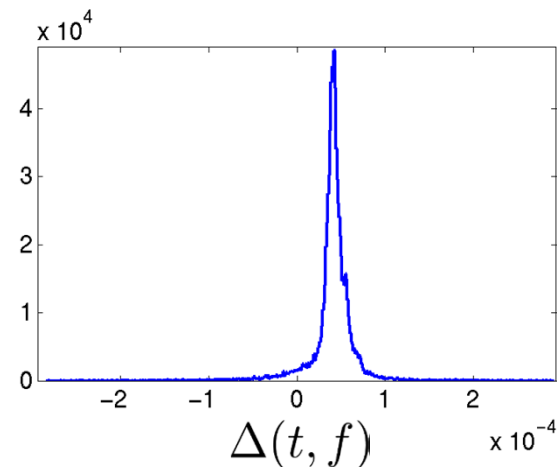


# 各音の時間差のヒストグラム

$$p(\Delta|z = c_1)$$



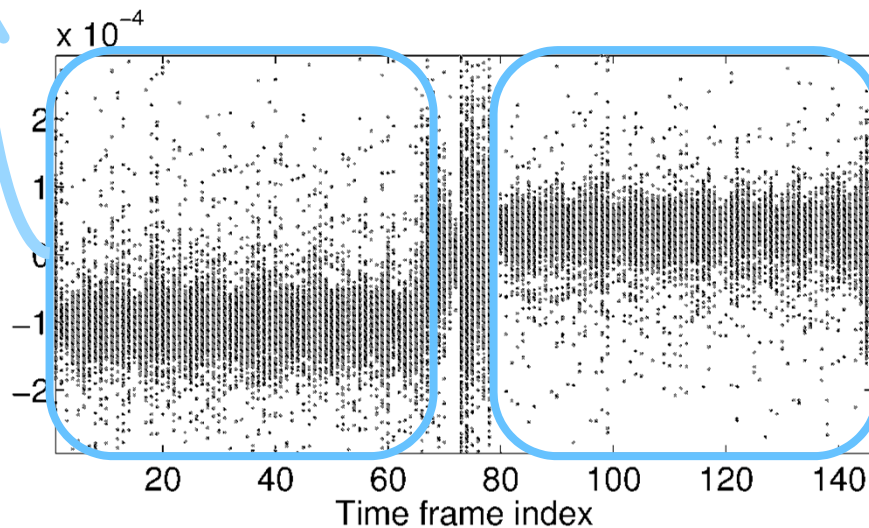
$$p(\Delta|z = c_2)$$



$z$ はどちらの  
人が話して  
いるかを  
示す変数

$s_1$ のみ

$\Delta(t, f)$



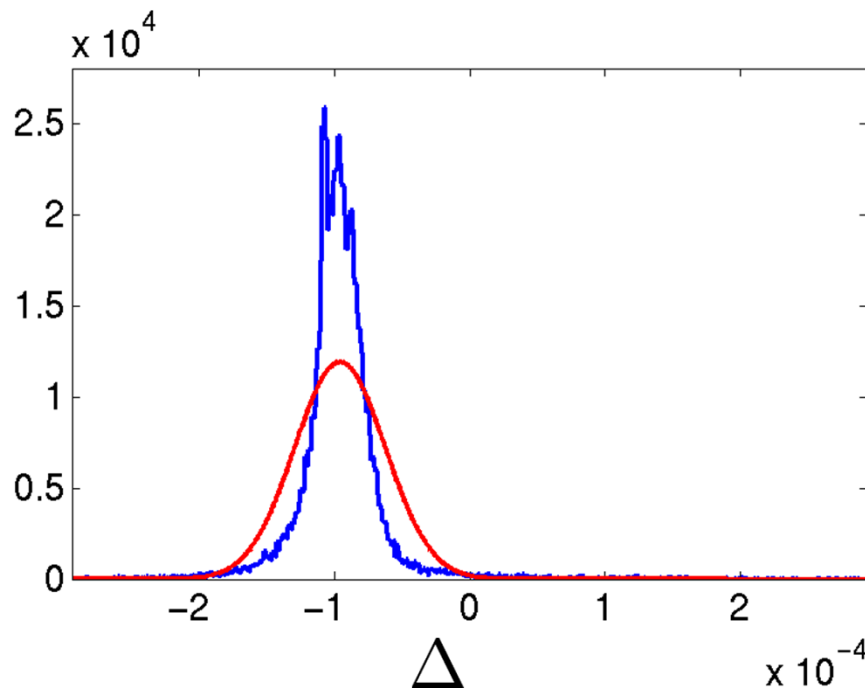
$s_2$ のみ

# 一つの音の時間差をモデル化

$s_1$ のみが話している場合の条件付き確率

$$p(\Delta|z = c_1)$$

青色ヒストグラム



ガウス分布でモデル化

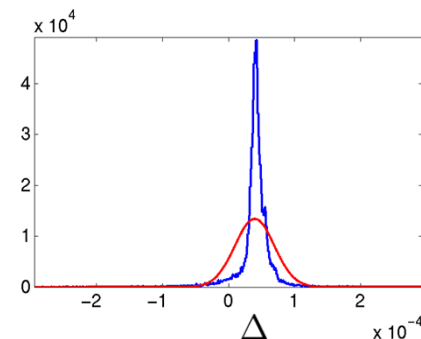
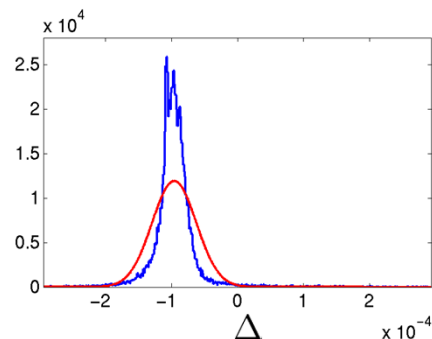
$$p(\Delta|z = c_1) = \mathcal{N}(\Delta|\mu_1, \sigma_1^2)$$

赤色曲線

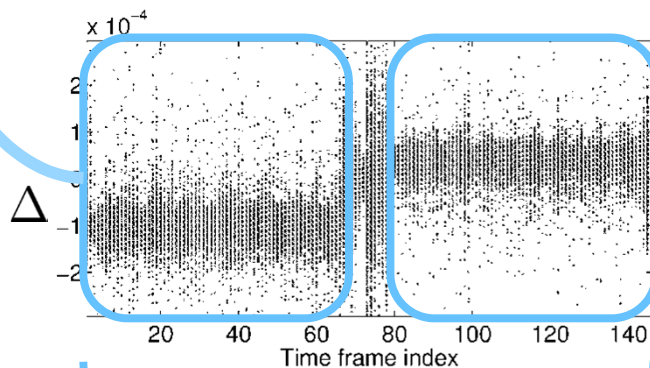
# 各音の時間差, 全体の時間差

$$p(\Delta|z = c_1) = \mathcal{N}(\Delta|\mu_1, \sigma_1^2)$$

$$p(\Delta|z = c_2) = \mathcal{N}(\Delta|\mu_2, \sigma_2^2)$$

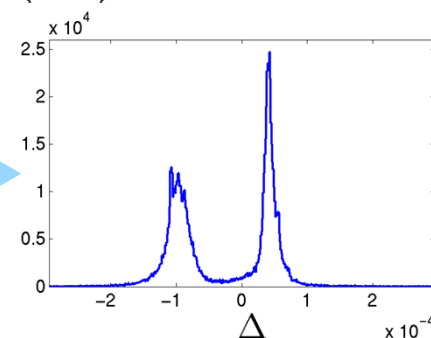


$s_1$ のみ



$s_2$ のみ

$p(\Delta)$



全体

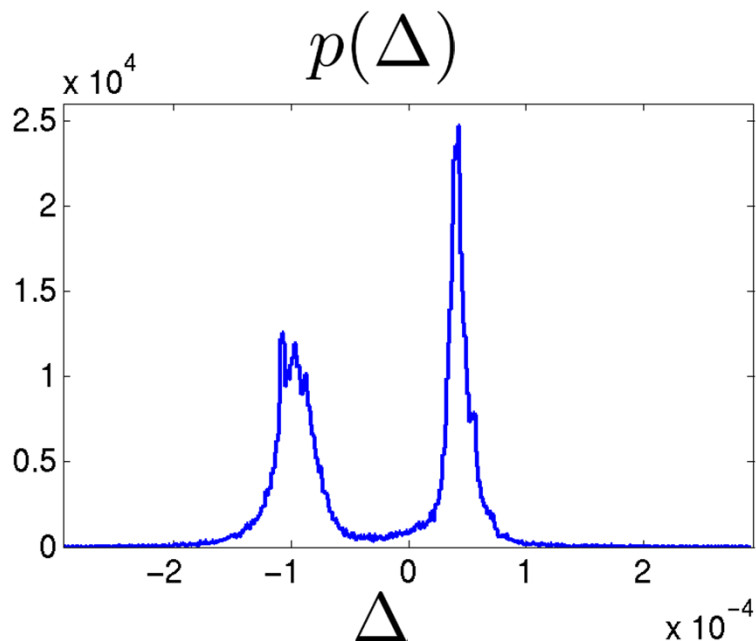
# ガウス混合モデル(GMM)

## 混合モデル

$$p(\Delta) = \sum_{z=c_1}^{c_2} p(z)p(\Delta|z)$$

$$\left\{ \begin{array}{l} \text{和の法則} \quad p(\Delta) = \sum_{z=c_1}^{c_2} p(\Delta, z) \\ \text{積の法則} \quad p(\Delta, z) = p(z)p(\Delta|z) \end{array} \right.$$

$p(\Delta|z) = \mathcal{N}(\Delta|\mu, \sigma^2)$  条件付き確率が**ガウス分布**

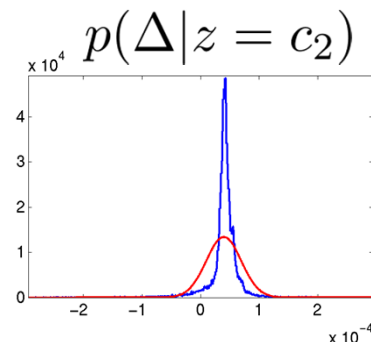
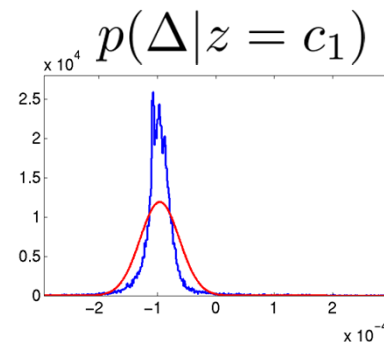


=

$$p(z = c_1) \times$$

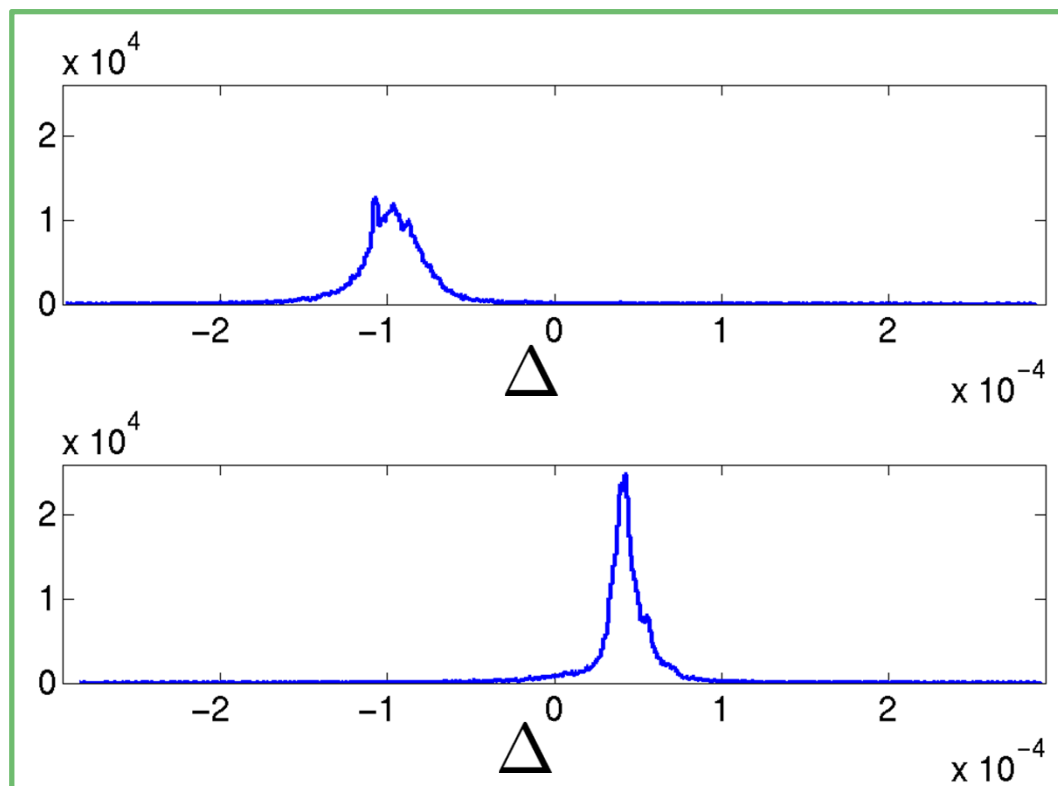
+

$$p(z = c_2) \times$$



# 同時確率と周辺確率の関係

$p(\Delta, z)$

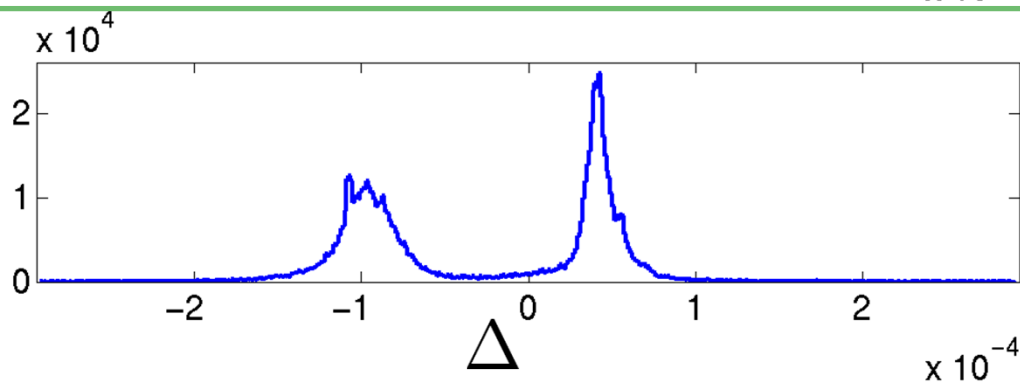


$p(z)$

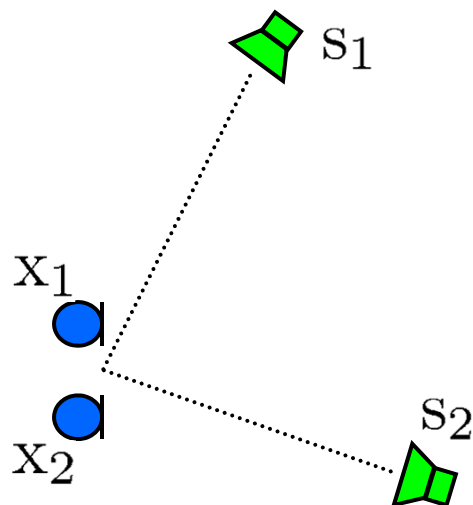
0.4889

0.5111

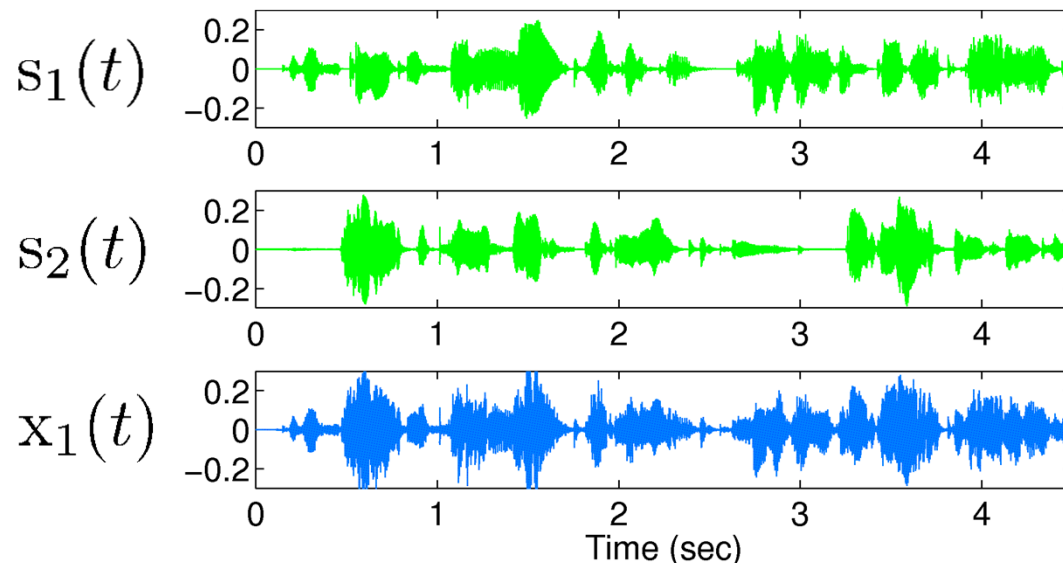
$p(\Delta)$



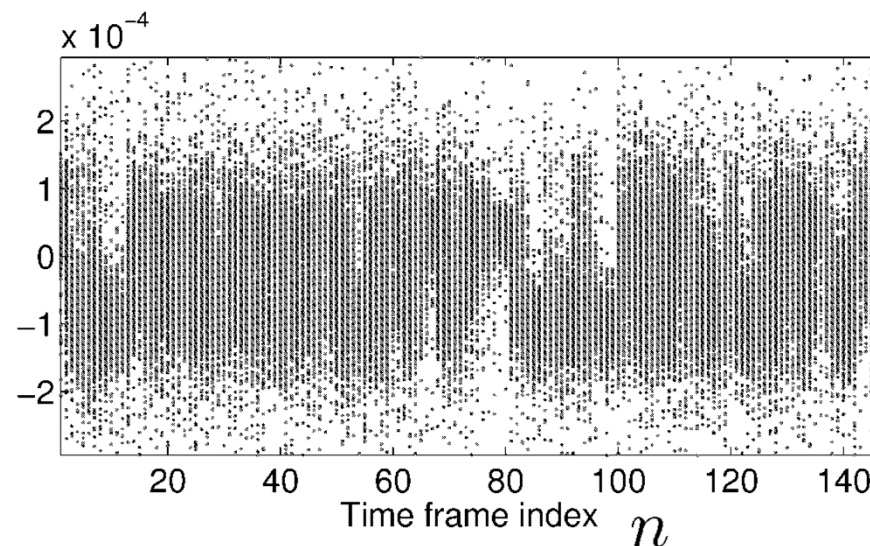
# 時間周波数毎の到達時間差推定



同時に話している



$$\Delta(t, f) = \frac{\arg[x_1(t, f)/x_2(t, f)]}{-2\pi f}$$



# 今度は音が混ざってしまった

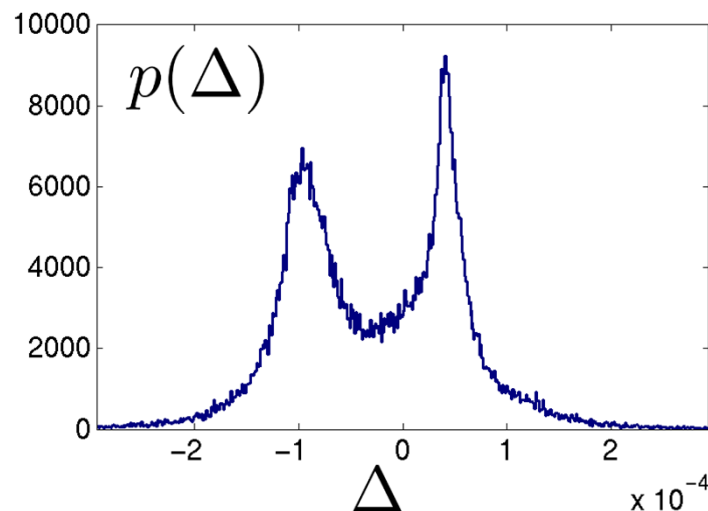
時間差に関する周辺確率しか得られない

変数  $z$  が隠れ変数に



隠れ変数モデル

$$p(\Delta) = \sum_{z=c_1}^{c_2} p(z)p(\Delta|z)$$



各観測サンプル  $\Delta$  に対して,  $z$  の値を推定する  
時間差が  $s_1$  のものと  $s_2$  のものに分離される

# 隠れ変数モデル＝混合モデルを詳細化

$$p(\Delta) = \sum_{z=c_1}^{c_2} p(z)p(\Delta|z)$$

各音源の表れやすさ  
(混合重み)

$$p(z = c_k) = \pi_k$$
$$\sum_{k=1}^2 \pi_k = 1$$

各音源による時間差を  
ガウス分布で表現

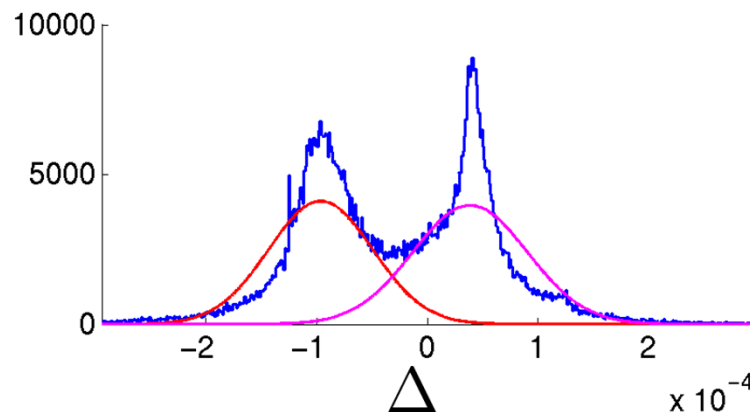
$$p(\Delta|z = c_k) = \mathcal{N}(\Delta|\mu_k, \sigma_k^2)$$

## ガウス混合モデル(GMM)

$$p(\Delta|\theta) = \sum_{k=1}^2 \pi_k \mathcal{N}(\Delta|\mu_k, \sigma_k^2)$$

推定すべきパラメータ

$$\theta = \{\mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2\}$$



# GMMのパラメータ推定

---

## 定式化

時間差のサンプル  $\Delta_1, \dots, \Delta_N$  に対して

**対数尤度**  $\sum_{n=1}^N \log p(\Delta_n | \theta)$  を最大化する

**パラメータ集合**  $\theta = \{\mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2\}$  を求める

➡ 単一のガウス分布とは異なり, 閉じた式では解けない

## EMアルゴリズムで効率的に解ける

E-step: 現在のパラメータによる**隠れ変数**の**負担率**を計算

M-step: 現在の負担率に従って**パラメータ**を最適化

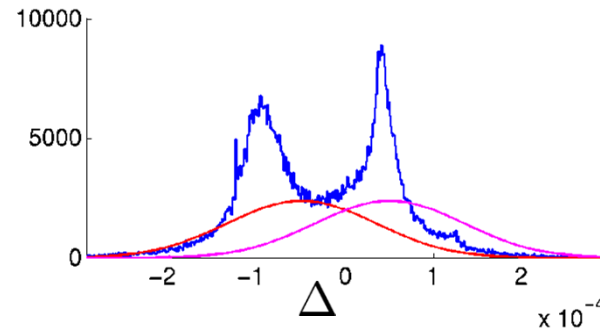
# EMアルゴリズム

## パラメータ初期化

$$\mu_1, \mu_2$$

$$\sigma_1, \sigma_2$$

$$\pi_1 = 0.5, \pi_2 = 0.5$$

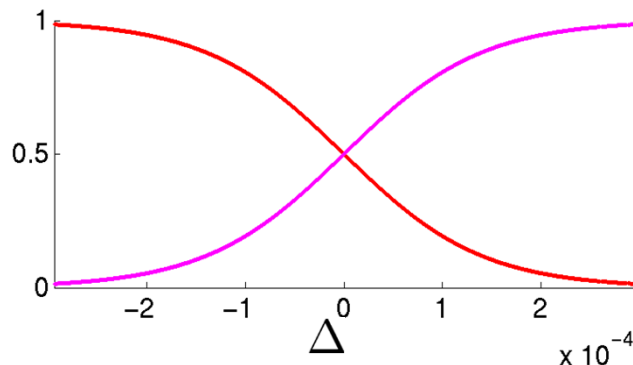


## E-step

各サンプル  $n$  がどのガウス分布  $k$  に属しているかを計算

$$\gamma_{nk} = p(z_n = c_k | \Delta_n, \theta)$$

負担率

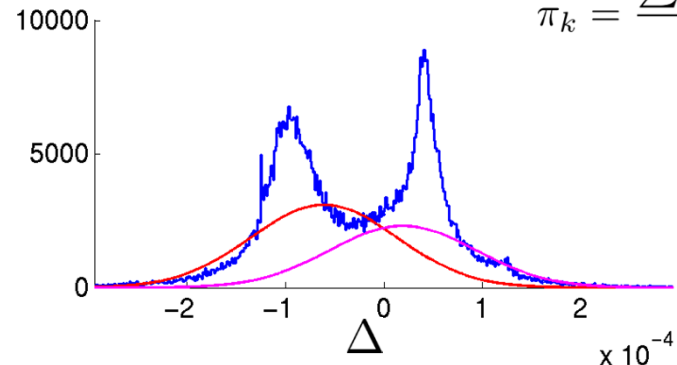


## M-step

負担率に従って  
パラメータを最適化

$$\mu_k = \frac{\sum_n \gamma_{nk} \Delta_n}{\sum_n \gamma_{nk}} \quad \sigma_k^2 = \frac{\sum_n \gamma_{nk} (\Delta_n - \mu_k)^2}{\sum_n \gamma_{nk}}$$

$$\pi_k = \frac{\sum_n \gamma_{nk}}{N}$$



変化がなくなる  
(収束)まで繰り返す

# 負担率の計算

---

ベイズの定理から

$$p(z|\Delta) = \frac{p(\Delta|z)p(z)}{p(\Delta)} = \frac{p(\Delta|z)p(z)}{\sum_z p(\Delta|z)p(z)}$$

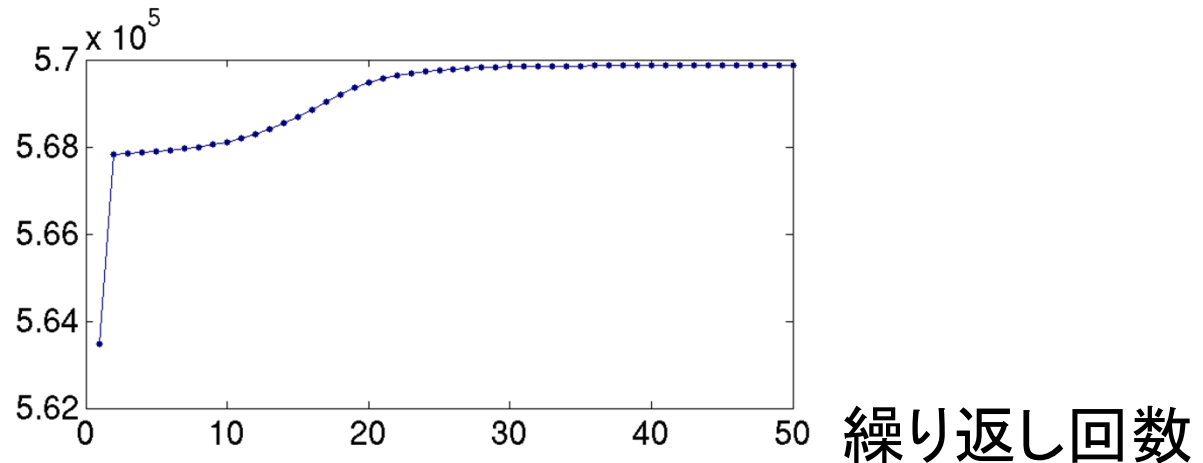
今回の定式化に沿って書くと

$$p(z_n = c_k | \Delta_n, \theta) = \frac{\mathcal{N}(\Delta_n | \mu_k, \sigma_k^2) \pi_k}{\sum_{k=1}^2 \mathcal{N}(\Delta_n | \mu_k, \sigma_k^2) \pi_k}$$

$$\mathcal{N}(\Delta_n | \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left[ -\frac{(\Delta_n - \mu_k)^2}{2\sigma_k^2} \right]$$

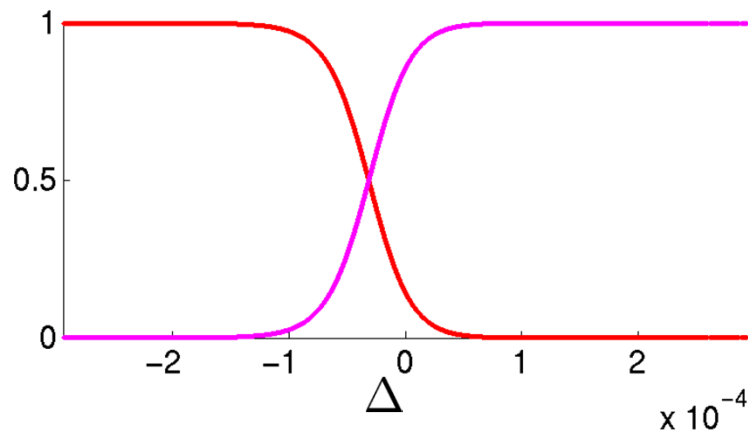
# 収束の様子

対数尤度



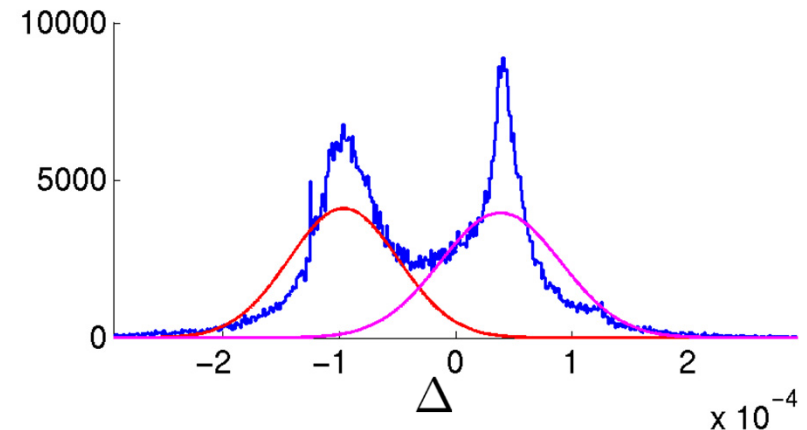
収束時の負担率

$$\gamma_{nk} = p(z_n = c_k | \Delta_n, \theta)$$



収束時のパラメータ

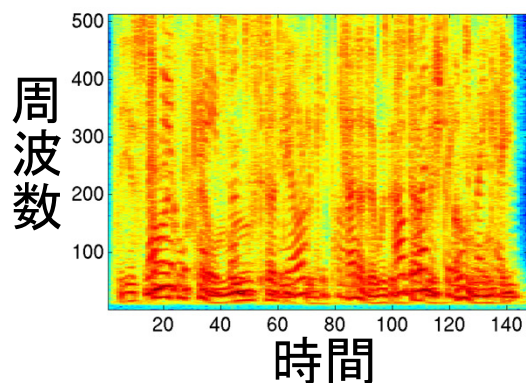
$$\theta = \{\mu_1, \sigma_1, \pi_1, \mu_2, \sigma_2, \pi_2\}$$



# 時間周波数マスキング

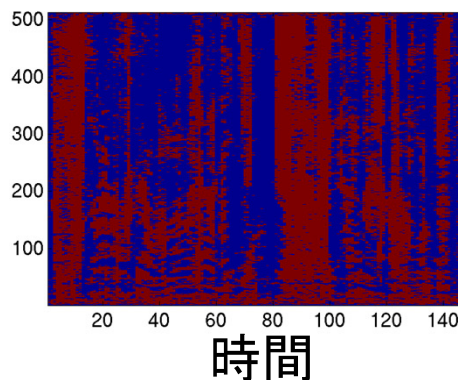


混合信号

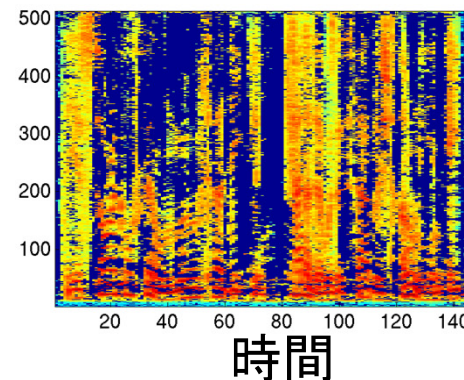


マスク

$$\gamma_{n1} > 0.6$$



分離信号

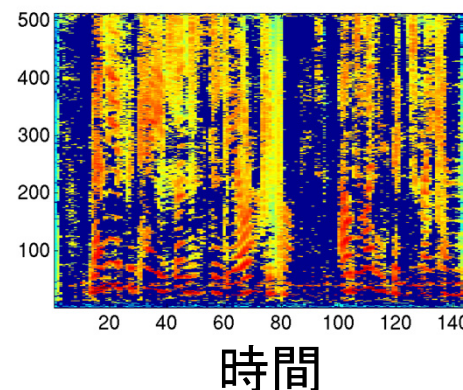
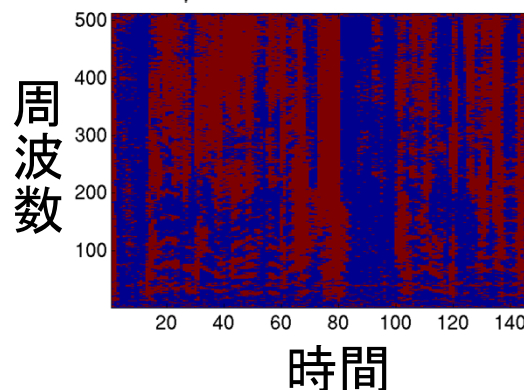


収束時の負担率  
からマスクを作成

■ 1 ■ 0

混合モデルにより  
音源分離を達成

$$\gamma_{n2} > 0.6$$



# 前半(GMM)のまとめ

---

- 確率の基礎
  - » 同時確率, 周辺確率, 条件付き確率
- ガウス分布
  - » 一つの塊データをモデル化
- 混合モデル
  - » 二つ以上の塊データをモデル化
  - » 各塊はガウス分布  $\Rightarrow$  GMM
- 最適化手法
  - » 混合モデルはEMアルゴリズムで効率的に解ける

# 非負值行列因子分解

NMF: Non-negative Matrix Factorization

$$\mathbf{X} = \mathbf{T} \mathbf{V}$$

# アウトライン(NMF)

---

- データの行列表現
  - » 購買履歴, 文書群, 音
- NMF(非負値行列因子分解)とは
  - » NMF: Non-negative Matrix Factorization
- NMFの応用例
  - » 楽器音の分析, 曲からのヴォーカル分離
  - » 文書群の分析
- 基底の数と近似誤差
- NMFアルゴリズム

# データの行列表現

- 購買履歴

|        |                                                                                   | 商品(群)                                                                             |                                                                                   |                                                                                   |
|--------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
|        |                                                                                   |  |  |  |
| ユーザ(群) |  | 2                                                                                 | 0                                                                                 | 0                                                                                 |
|        |  | 0                                                                                 | 3                                                                                 | 1                                                                                 |
|        |  | 1                                                                                 | 1                                                                                 | 2                                                                                 |

多種多様なデータが  
行列形式で表現できる

- 文書

- » Bag-of-words

- 音

- » スペクトログラム (spectrogram)

# Bag-of-words 表現

ビッグデータ時代を迎え、データへの教師ラベル付けが不要な「教師なし機械学習」の重要性が増している。本講演では、教師なし学習の中でもクラスタリング技術に着目する。特に、ガウス混合モデル(GMM: Gaussian Mixture Model)と、非負値行列因子分解(NMF: Non-negative Matrix Factorization)の基礎を説明し、実データへの適用例をいくつか示す。

|     |   |
|-----|---|
| 教師  | 3 |
| データ | 3 |
| 学習  | 2 |
| モデル | 1 |
| :   |   |

# 文書集合の行列表現

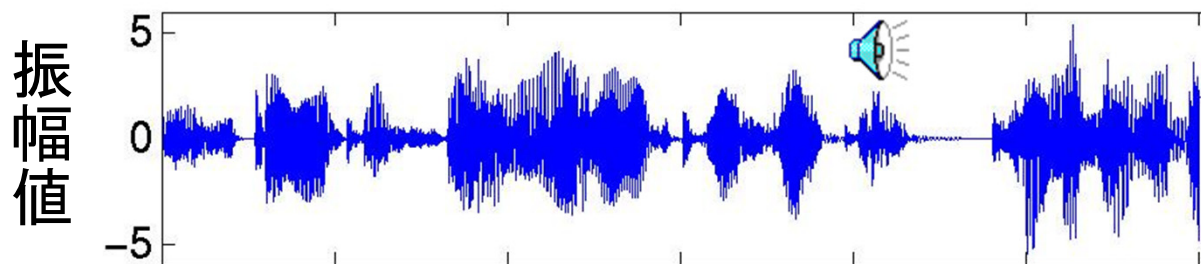
- 1つの文書をBag-of-words表現
- 文書集合を行列表現

|    |      | 文書 |   |   |   |   |   |   |   |   |   |   |   |
|----|------|----|---|---|---|---|---|---|---|---|---|---|---|
| 単語 | 教師   | 3  | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|    | データ  | 3  | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
|    | 学習   | 2  | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|    | モデル  | 1  | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|    | 円    | 0  | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
|    | 産業   | 0  | 0 | 2 | 0 | 3 | 3 | 2 | 0 | 0 | 1 | 0 | 0 |
|    | アフリカ | 0  | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |
|    | 天然   | 0  | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 0 | 0 |
|    | 資源   | 0  | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 2 | 0 | 0 |
|    | :    |    |   |   |   |   |   |   |   |   |   |   |   |

# 音の時間周波数表現

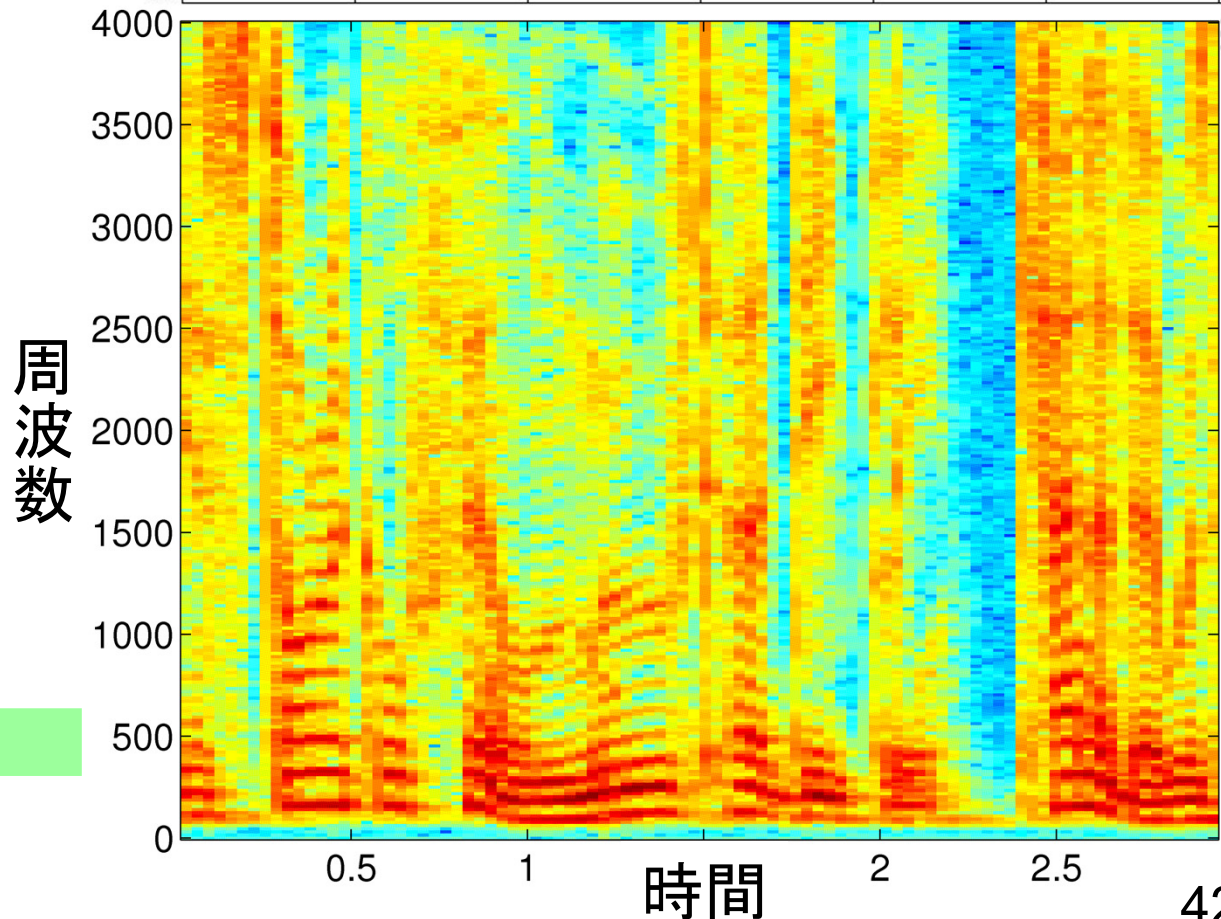
波形

時間  
振幅値



スペクトログラム

時間  
周波数  
複素数



これを行列表現  
とみなす

# アウトライン(NMF)

---

- データの行列表現
  - » 購買履歴, 文書群, 音
- NMF(非負値行列因子分解)とは
  - » NMF: Non-negative Matrix Factorization
- NMFの応用例
  - » 楽器音の分析, 曲からのヴォーカル分離
  - » 文書群の分析
- 基底の数と近似誤差
- NMFアルゴリズム

# NMF: non-negative matrix factorization

- 非負値行列因子分解

»  $I \times J$  行列を  $I \times K$  行列と  $K \times J$  行列の積に分解

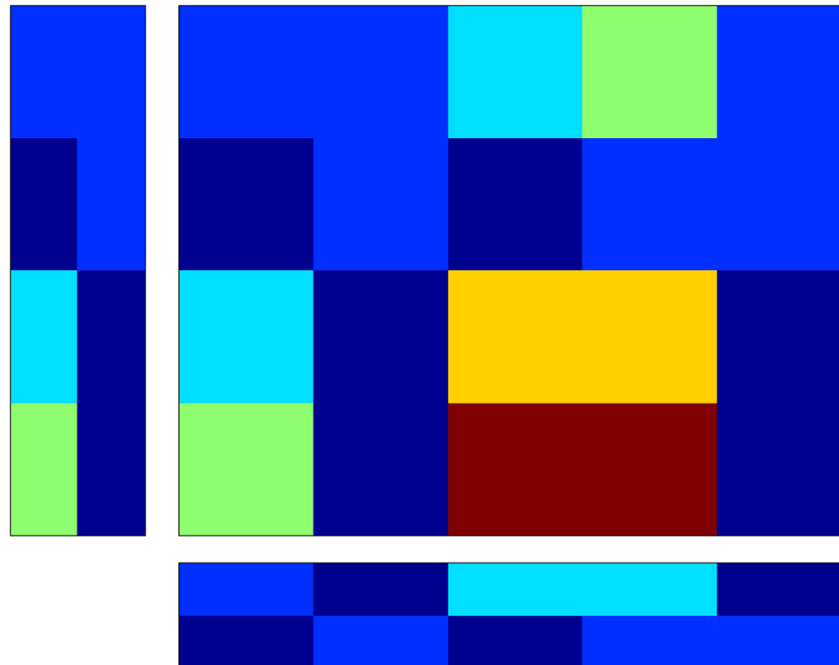
$$\begin{array}{c} J \\ \boxed{\mathbf{X}} \\ I \end{array} = \begin{array}{c} K \\ \boxed{\mathbf{T}} \\ I \end{array} \begin{array}{c} J \\ \boxed{\mathbf{V}} \\ K \end{array}$$

» 要素はすべて非負 (0 か 正の値)

$$\begin{bmatrix} 5 & 4 & 10 & 3 & 12 \\ 2 & 2 & 4 & 1 & 5 \\ 1 & 2 & 2 & 0 & 3 \\ 3 & 2 & 6 & 2 & 7 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 & 0 & 3 \\ 1 & 0 & 2 & 1 & 2 \end{bmatrix}$$

# 図による行列分解表示

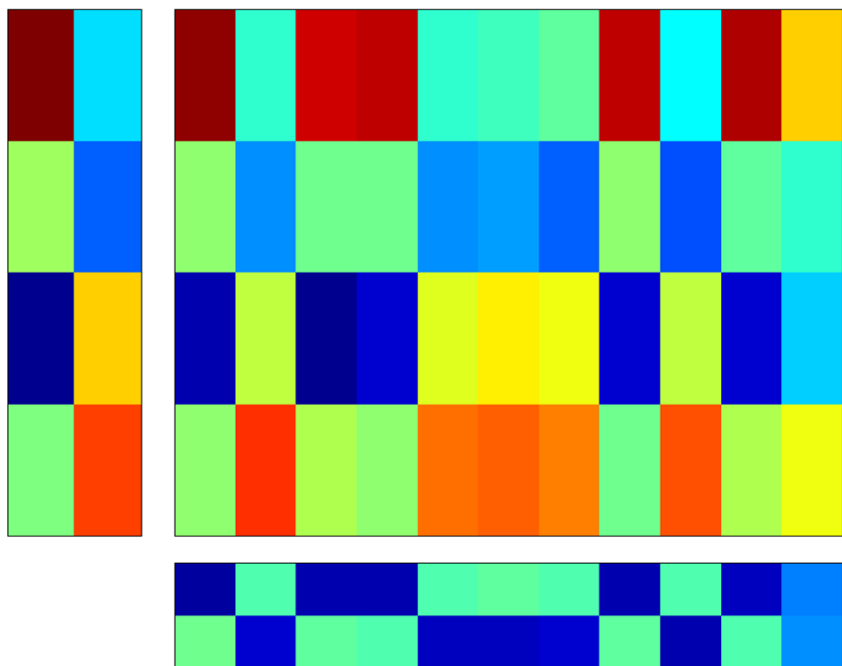
$$\begin{bmatrix} 1 & 1 & 2 & 3 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 2 & 0 & 4 & 4 & 0 \\ 3 & 0 & 6 & 6 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 & 2 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$



# 特徴的な例

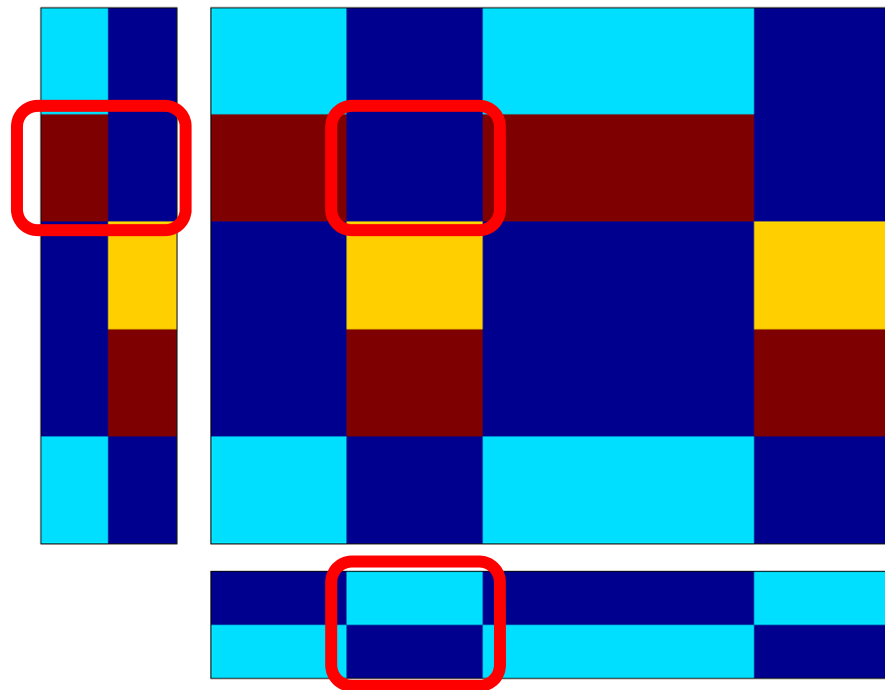
## クラスタリング

頻出パターンで基底を構成することでXとTVの誤差が小さくなる



## 双方向クラスタリング

行方向, 列方向, それぞれに頻出パターンがあれば, 同時にクラスタリングされる



# 非負性により, 0が0を生み出す

因子分解前の  
行列  $\mathbf{X}$  の要素



因子分解後はK次元  
ベクトルの内積で表現

**0要素**に対して, 対応するベクトルは直交

負の値を使える場合

$$0 = \begin{vmatrix} 1 & -1 \end{vmatrix} \begin{vmatrix} 1 \\ 1 \end{vmatrix}$$

0以外の値で表現できる

全ての要素が**非負値**の場合

$$0 = \begin{vmatrix} 0 & 1 \end{vmatrix} \begin{vmatrix} 1 \\ 0 \end{vmatrix}$$

**0**を使う必要あり

# アウトライン

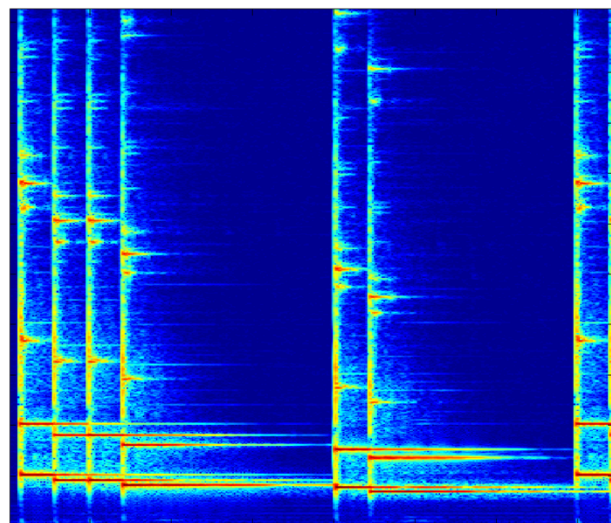
---

- データの行列表現
  - » 購買履歴, 文書群, 音
- NMF (非負値行列因子分解) とは
  - » NMF: Non-negative Matrix Factorization
- NMFの応用例
  - » 楽器音の分析, 曲からのヴォーカル分離
  - » 文書群の分析
- 基底の数と近似誤差
- NMFアルゴリズム

# NMFを音響信号に適用

$$\begin{matrix} & J \\ I & \boxed{X} \end{matrix} = \begin{matrix} & K \\ I & \boxed{T} \end{matrix} \begin{matrix} & J \\ & \boxed{V} \end{matrix} \begin{matrix} K \\ \end{matrix}$$

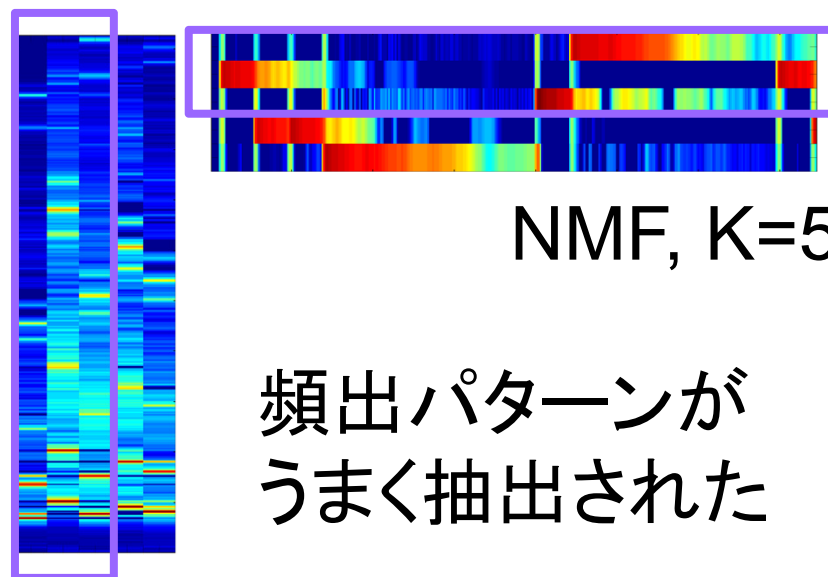
周波数



時間



=

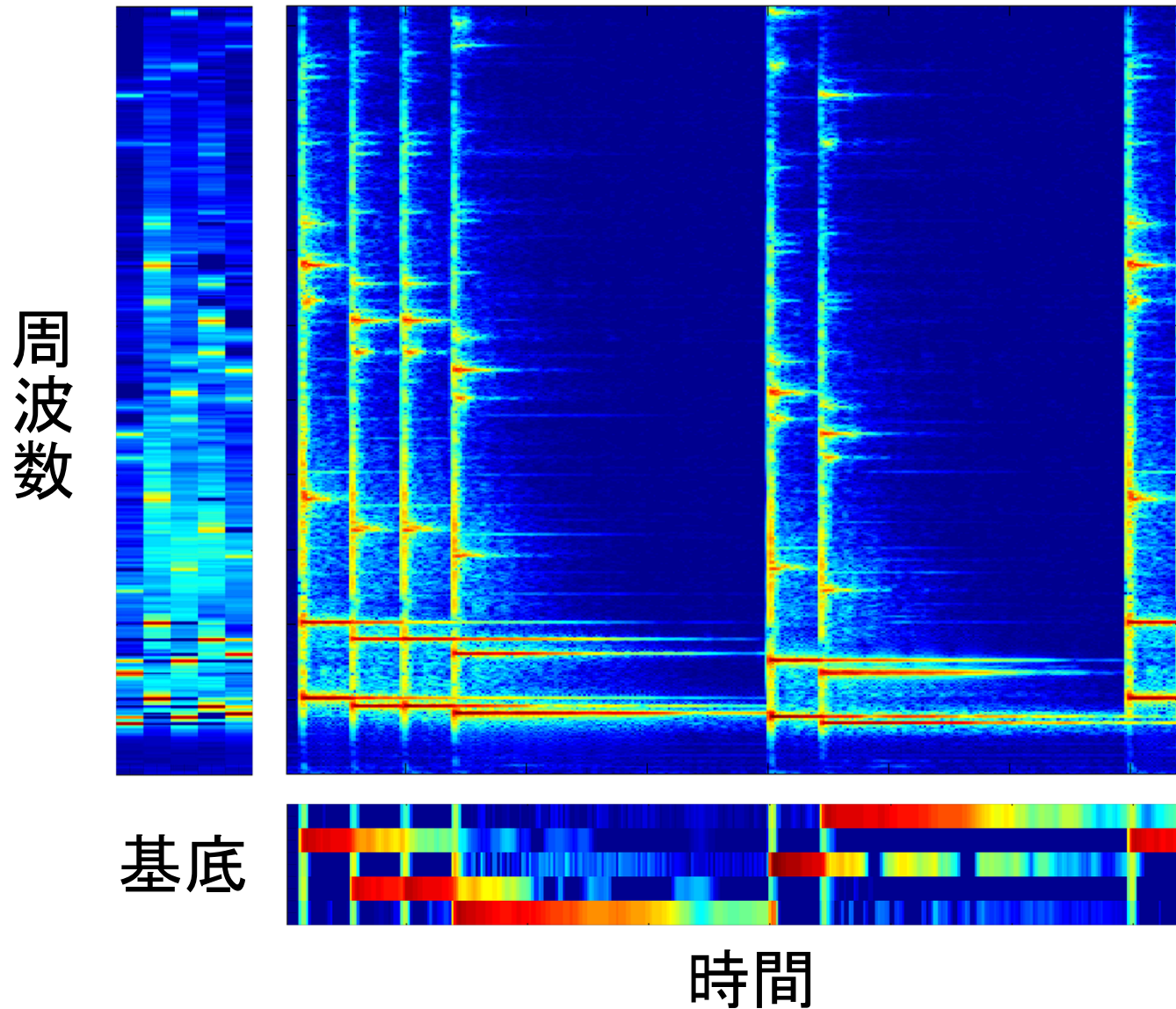


頻出パターンが  
うまく抽出された

NMF基底3つの音

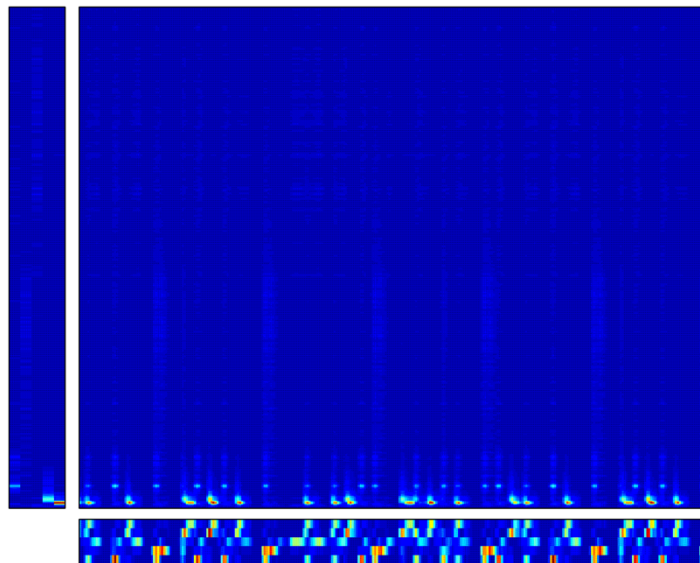


# NMF結果の表示

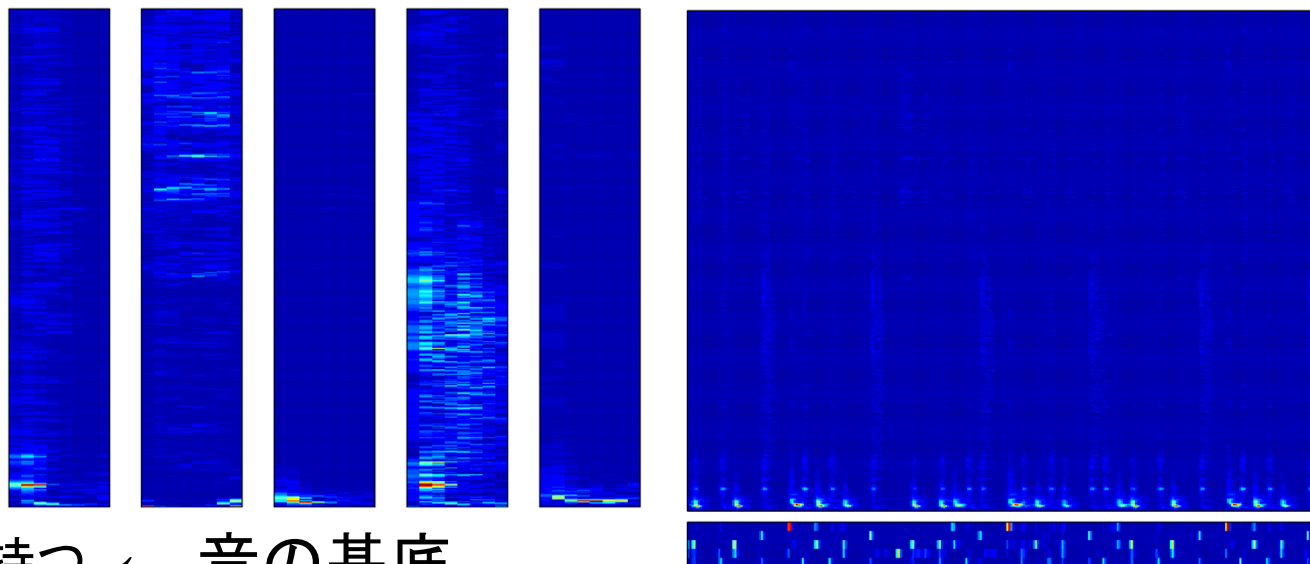


# NMFの拡張: Convolutional-NMF

NMF



Convolutional-  
NMF

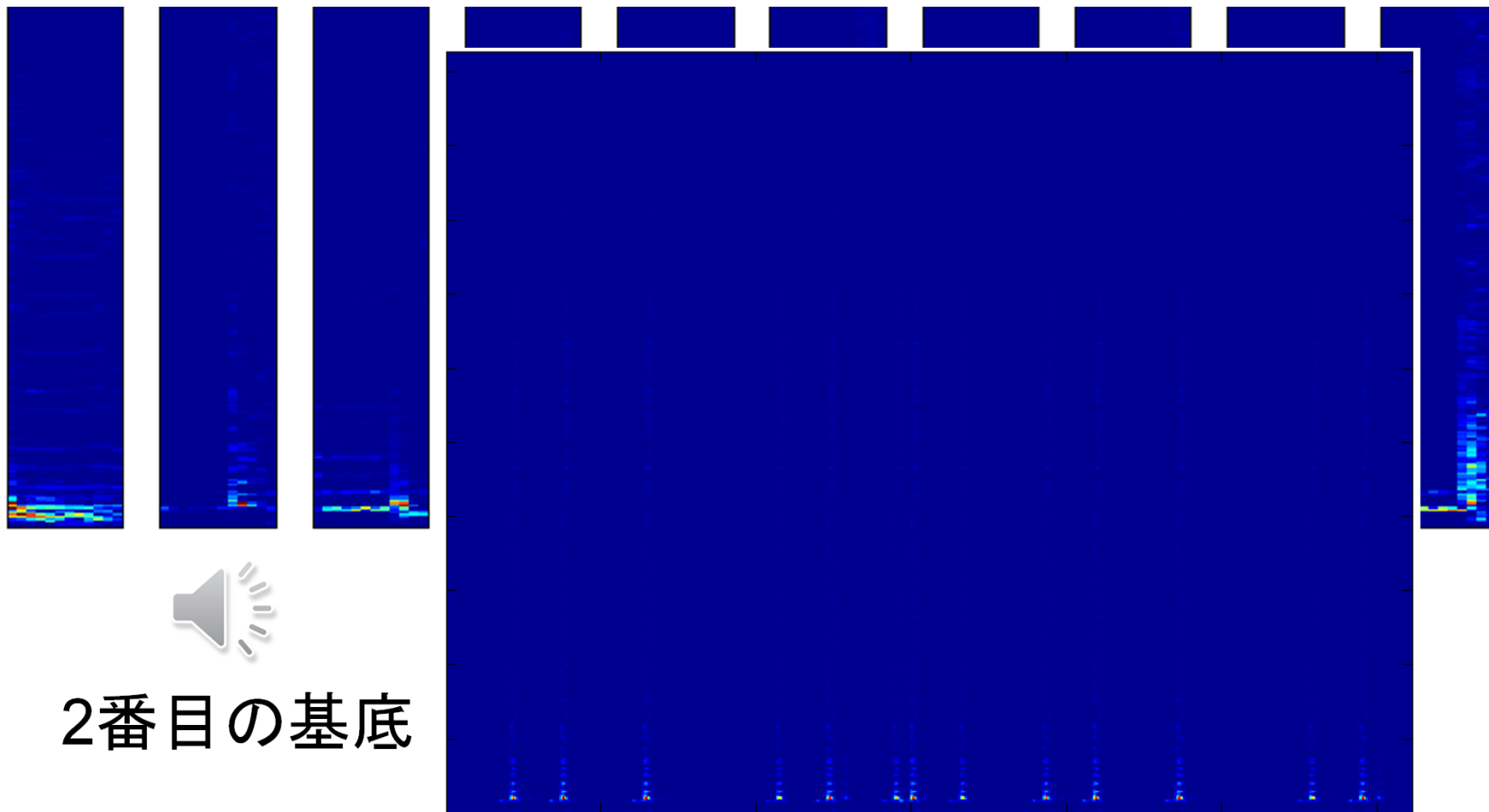


時間方向の幅を持つ ← 音の基底

# NMFによるヴォーカル分離



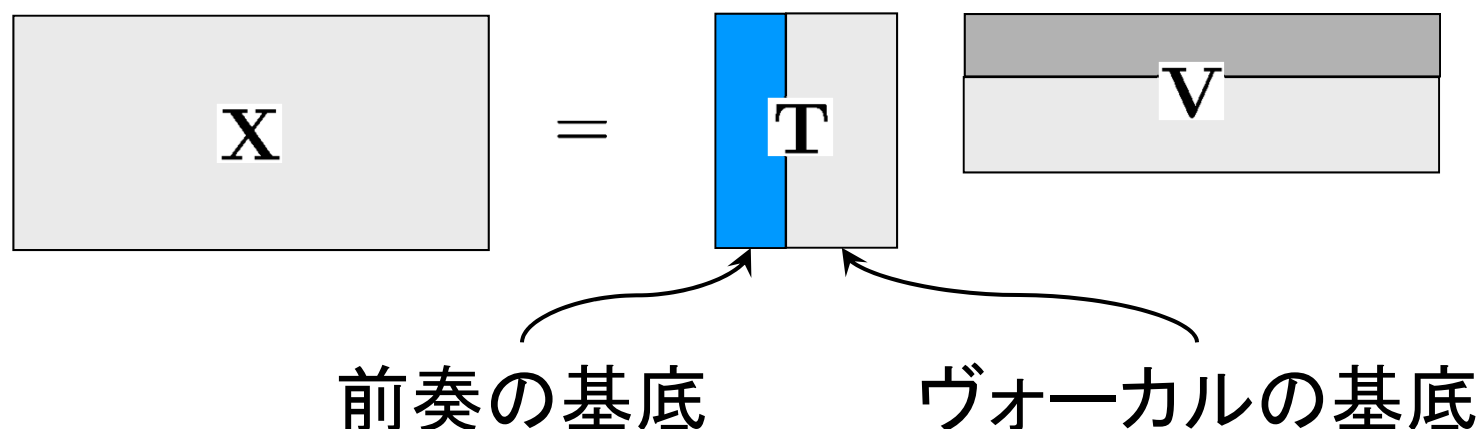
前奏10秒で学習した基底



2番目の基底

# NMFによるヴォーカル分離

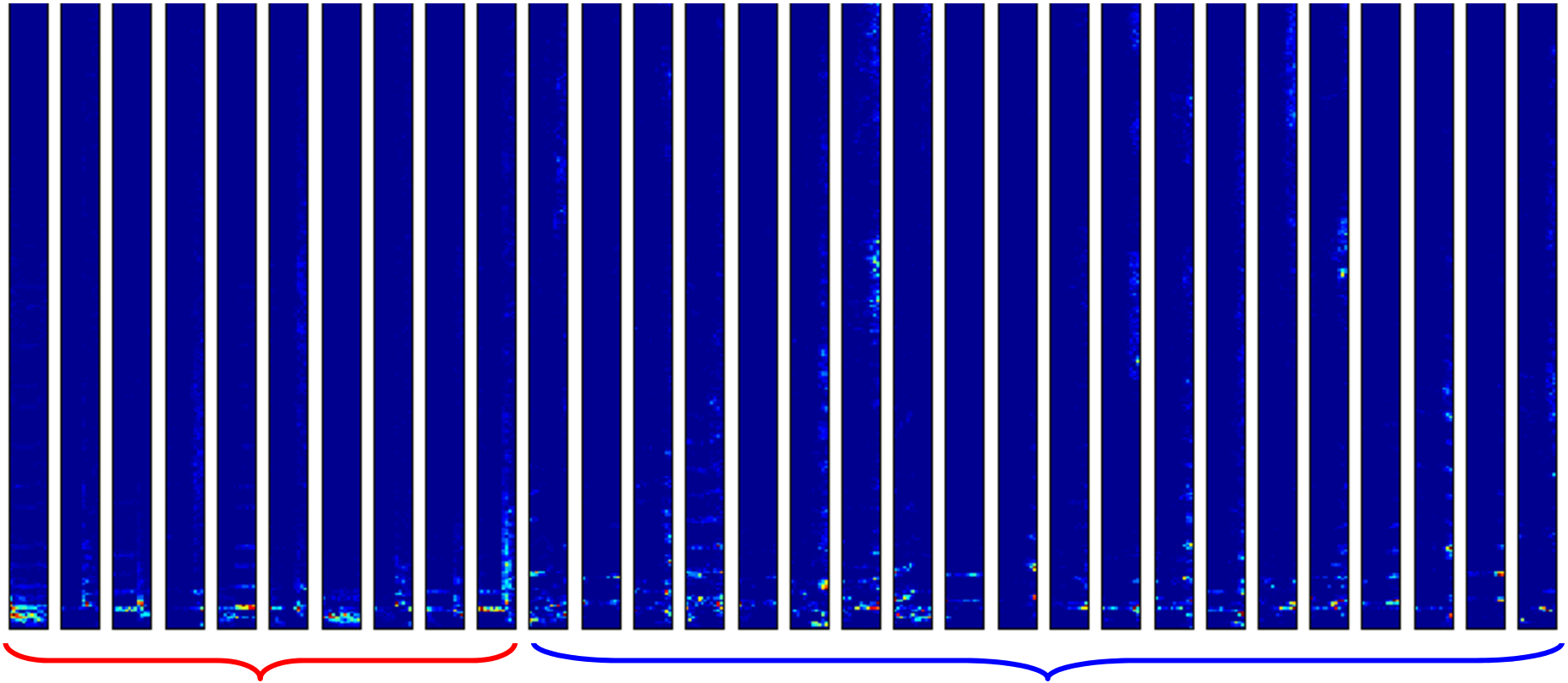
🔊 今度は、全体20秒に対してNMFを適用



- NMFの学習アルゴリズムの際
  - » 前奏の基底  $T$  は更新しない
  - » ヴォーカルの  $T$  と  $V$  , および前奏の  $V$  を更新する

# NMFで学習した30個の基底

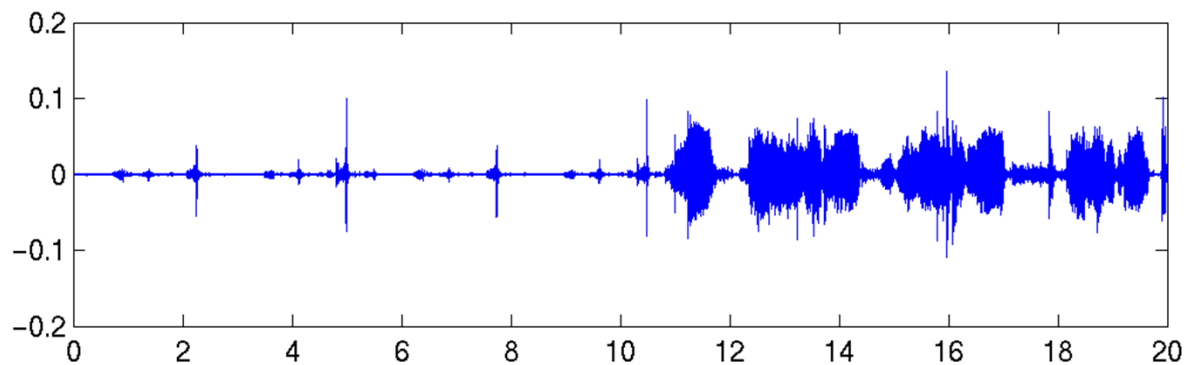
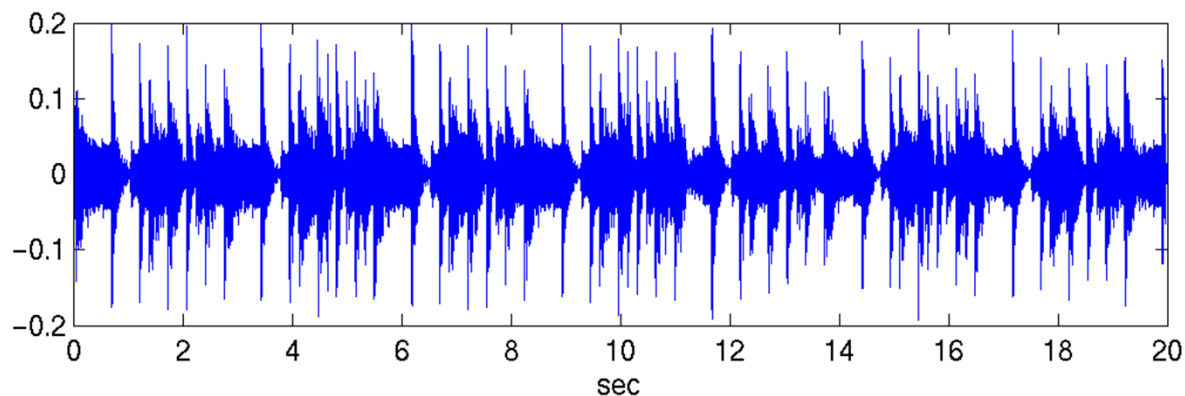
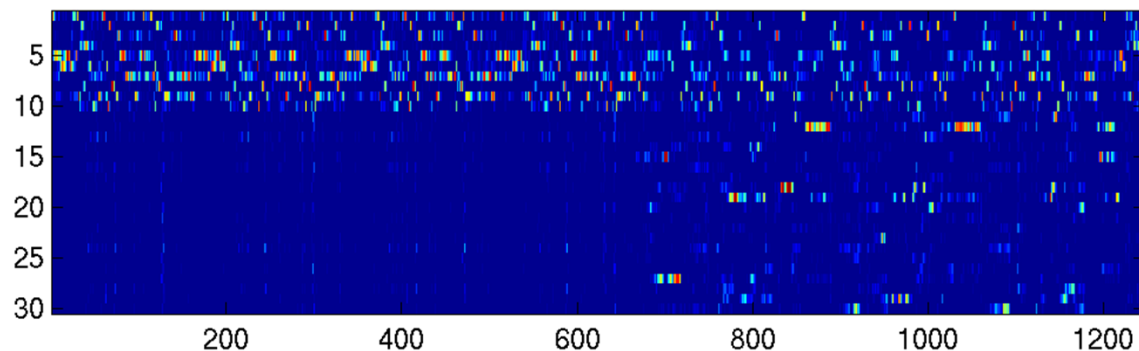
---



前奏の基底

ヴォーカルの基底

# ヴォーカル分離の結果



# NMFを用いた文書群の分析

- 文章をBag-of-words表現
- 文書集合を行列表現
- NMFにより頻出単語パターンを抽出

|    |      | 文書 |   |   |   |   |   |   |   |   |   |   |   |
|----|------|----|---|---|---|---|---|---|---|---|---|---|---|
| 単語 | 教師   | 3  | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|    | データ  | 3  | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
|    | 学習   | 2  | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
|    | モデル  | 1  | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|    | 円    | 0  | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
|    | 産業   | 0  | 0 | 2 | 0 | 3 | 3 | 2 | 0 | 0 | 1 | 0 | 0 |
|    | アフリカ | 0  | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |
|    | 天然   | 0  | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 0 | 0 |
|    | 資源   | 0  | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 2 | 0 | 0 |
|    | :    |    |   |   |   |   |   |   |   |   |   |   |   |

# 20 Newsgroups data set の解析

## • 20個のNewsgroupに投稿された文書集合

From: xxx@yyy.zzz.edu  
Subject: Re: Windows gripe...

> There's one thing about Windows that really frosts me.  
> I have 20MB of RAM installed in my system. I use a 5MB (2.5MB  
> under Windows) disk-cache, and a 4MB permanent swap file.  
>  
> While I can never fill the memory up, I still have problems  
> sometimes because I run out of GDI resources. What gives?  
> I think Windows could manage these resources a little better.  
>  
Are you using Windows 3.0 or 3.1? If you're still on 3.0, 3.1 devotes about twice as much memory to these and runs out much less frequently. If 3.1, you might use one of the resource monitors (such as the one that comes with the Windows 3.1 Resource Kit or one of the many shareware ones available) to see which programs are hogging the resources (every icon, internal graphics brush, etc. in every program running uses a certain amount of this limited memory area. Also, some don't give it back when they're finished).

comp.graphics  
comp.os.ms-windows.misc  
comp.sys.ibm.pc.hardware  
comp.sys.mac.hardware  
comp.windows.x  
rec.autos  
rec.motorcycles  
rec.sport.baseball  
rec.sport.hockey  
sci.crypt  
sci.electronics  
sci.med  
sci.space  
misc.forsale  
talk.politics.misc  
talk.politics.guns  
talk.politics.mideast  
talk.religion.misc  
alt.atheism  
soc.religion.christian

# NMFによる解析結果(T)

$J = 18828$

$K = 20$

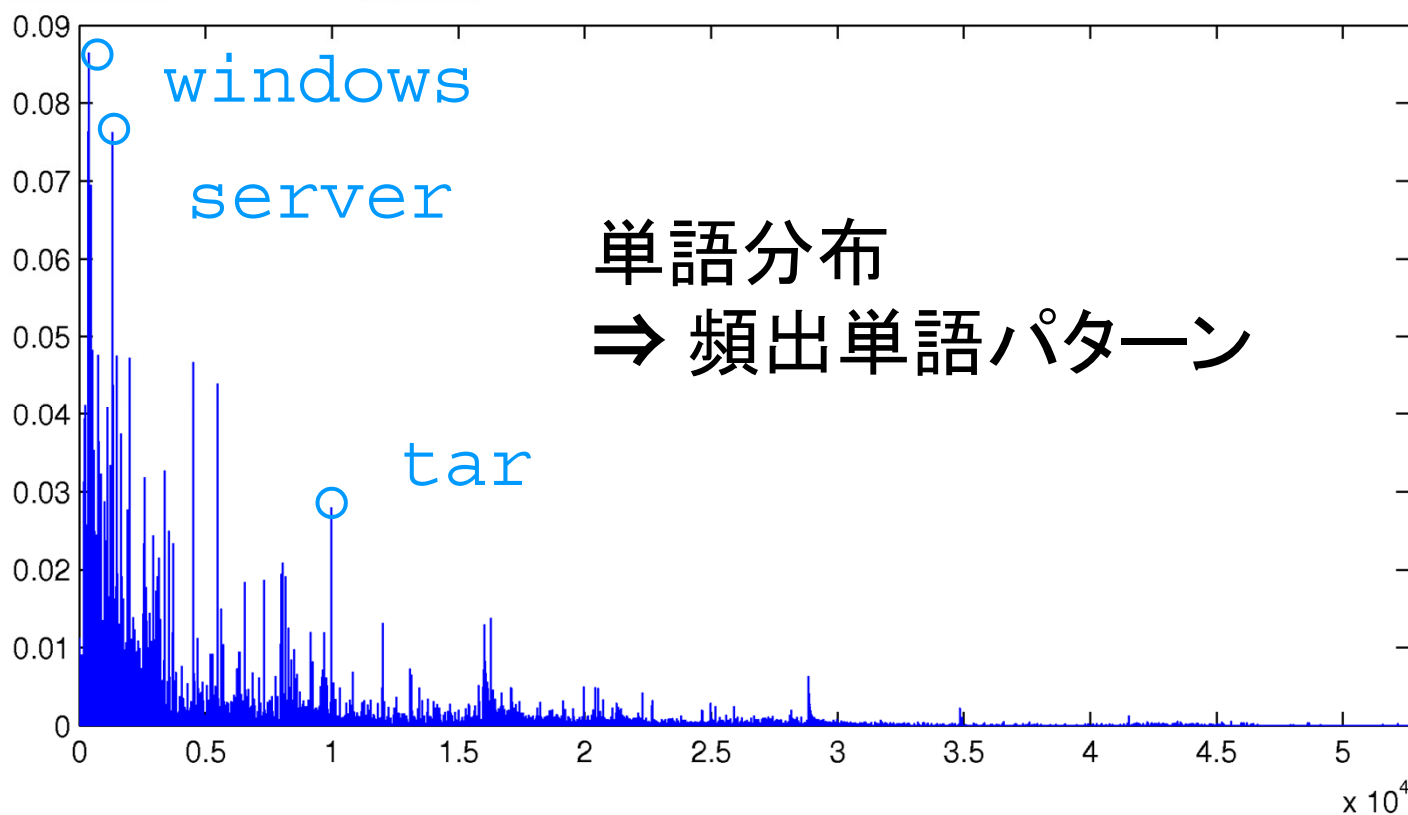
$J$

$V$

$K$

$I = 52647$   
単語

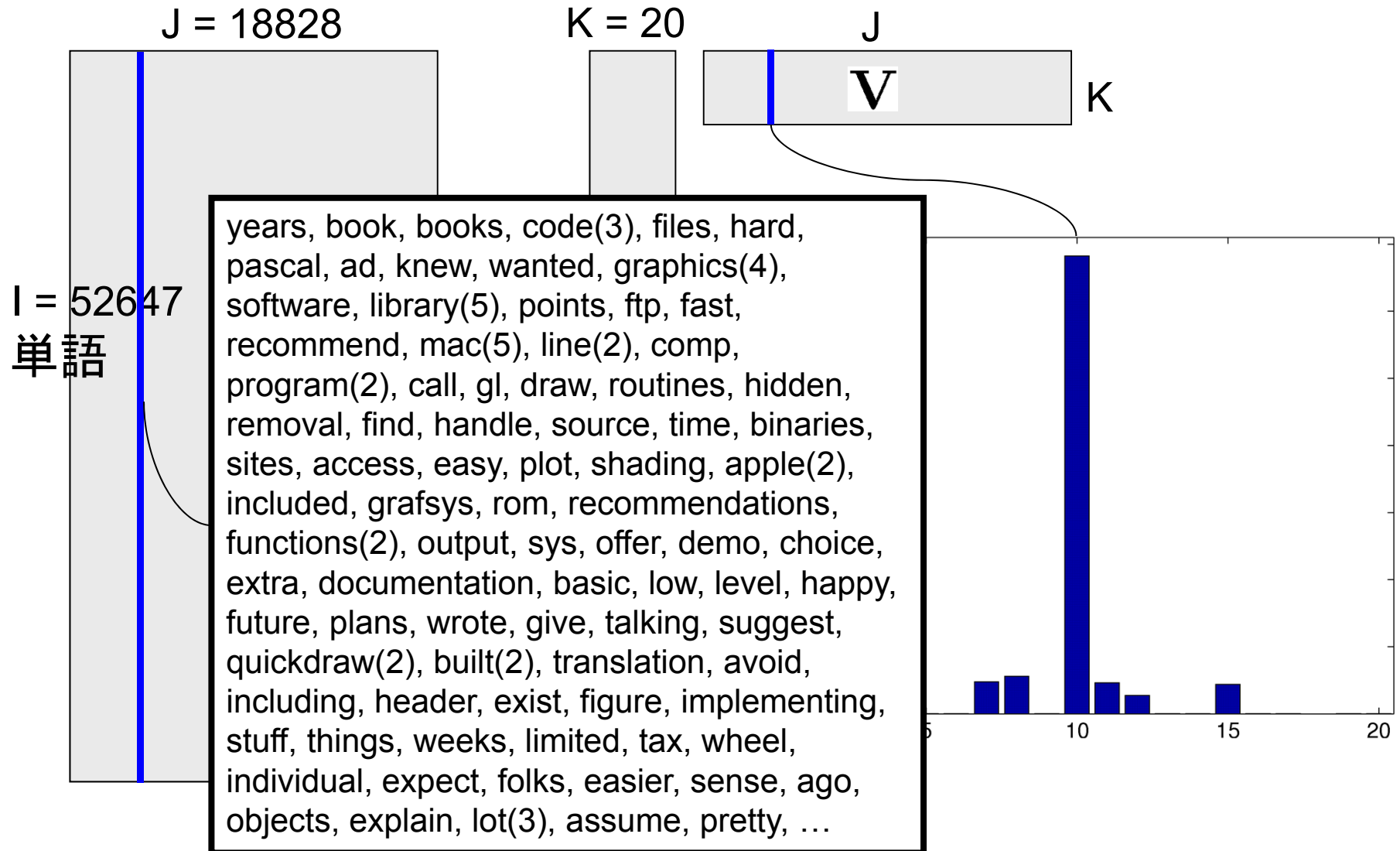
$X$



# 各基底 T の上位10個の単語

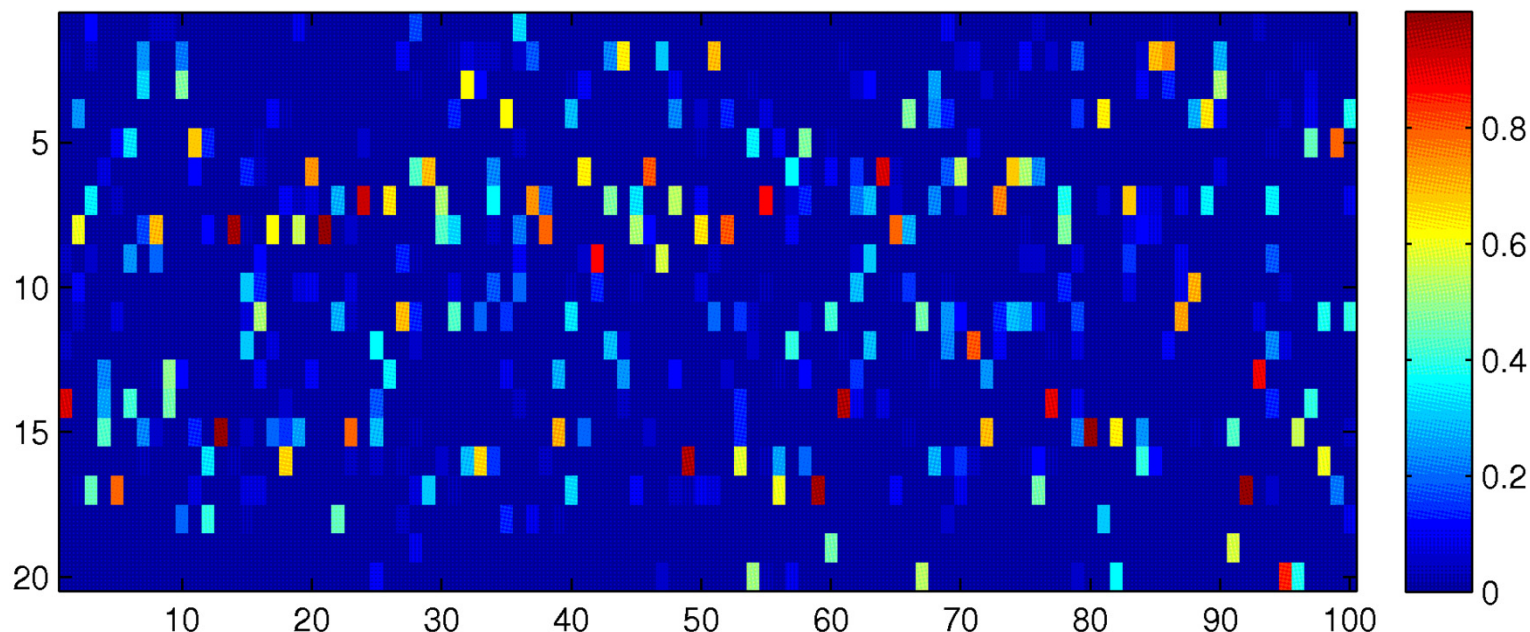
|                                                                                              |                             |
|----------------------------------------------------------------------------------------------|-----------------------------|
| 1. db, cx, ah, mc, mv, chz, lk, sc, hz, bh                                                   | a. comp.graphics            |
| 2. people, fire, fbi, gun, koresh, children, time, batf, wa                                  | b. comp.os.ms-windows.misc  |
| 3. game, team, hockey, games, play, season, nhl, la, y                                       | c. comp.sys.ibm.pc.hardware |
| 4. windows, file, graphics, server, sun, motif, ftp, versi                                   | d. comp.sys.mac.hardware    |
| 5. earth, space, entry, mission, output, orbit, program,                                     | e. comp.windows.x           |
| 6. power, car, light, water, good, oil, high, time, ground                                   | f. rec.autos                |
| 7. mail, key, list, information, send, email, internet, sof                                  | g. rec.motorcycles          |
| 8. drive, dos, card, windows, scsi, system, pc, disk, m                                      | h. rec.sport.baseball       |
| 9. clipper, chip, key, gun, encryption, law, government                                      | i. rec.sport.hockey         |
| 10. image, file, files, program, data, color, bit, images, j                                 | j. sci.crypt                |
| 11. car, bike, good, time, back, dod, ride, day, left, mak                                   | k. sci.electronics          |
| 12. government, president, people, clinton, states, state                                    | l. sci.med                  |
| 13. god, jesus, bible, christian, church, christ, people, c                                  | m. sci.space                |
| 14. people, israel, jews, israeli, jewish, religion, war, ara                                | n. misc.forsale             |
| 15. people, evidence, science, question, make, things,                                       | o. talk.politics.misc       |
| 16. year, good, time, game, team, baseball, players, hit                                     | p. talk.politics.guns       |
| 17. space, nasa, launch, cost, gov, henry, dc, station, s                                    | q. talk.politics.mideast    |
| 18. max, window, pl, giz, wm, bhj, widget, bxn, gk, sl                                       | r. talk.religion.misc       |
| 19. armenian, turkish, armenians, people, armenia, turk                                      | s. alt.atheism              |
| 20. medical, health, disease, study, cancer, patients, doctor, treatment, medicine, research | t. soc.religion.christian   |

# NMFによる解析結果 (V)



# クラスタリングとしての解釈

ランダムに100個選んだ  $V$  の列

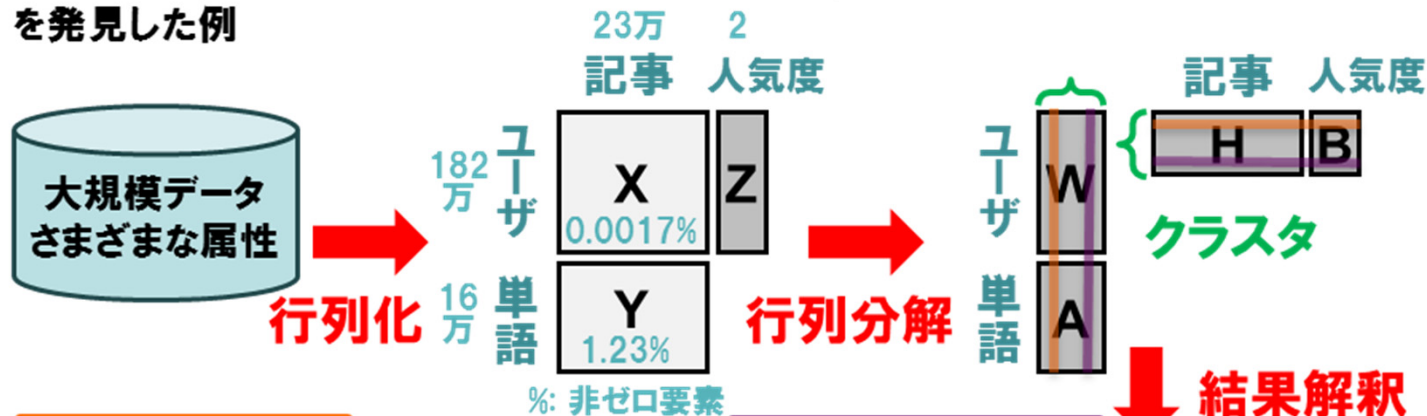


- 多くの文書が1個の基底で表現されている
  - » クラスタリングとしての解釈
  - » 基底数 $K=20 \Rightarrow$  クラスタ数20
- NMFにより, 大量の文書のクラスタリングが達成される

# 複合データの解析 ～ある現象を様々な側面で観測～

## NMMF: Non-negative Multiple Matrix Factorization

Twitterまとめ記事(Togetter)を解析し、トピック単語群(話題)とオーソリティ(その話題によく発言し、よくフォローされているユーザ)の組からなるいくつかのクラスタを発見した例



### クラスタ1「宇宙」

#### トピック単語群(話題)

宇宙 衛星 地球 ロケット 軌道 jaxa はやぶさ 探査 打ち上げ  
飛行 移籍 あかつき 機 nasa 打ち上げる プレイブ  
カンパニー 土 iss

| オーソリティ          | プロフィール             | フォロー数  |
|-----------------|--------------------|--------|
| hadukino        | 中国の宇宙・天文情報         | 1,321  |
| ohnuki_tsuyoshi | 宇宙開発が仕事            | 5,477  |
| 5thstar         | 飛行士選抜ファイナリスト       | 2,386  |
| ShinyaMatsuura  | 宇宙開発・科学関連物書        | 18,609 |
| sctracker       | spacecraft tracker | 847    |

### クラスタ2「Hadoop」

#### トピック単語群(話題)

使う Hadoop クラウド さん AWS ない 話 データこれいい 開発  
セッション そう PHP 的 処理 DB サーバ ヒップホップ 人 MySQL  
会場 Mongo DB 見る Azure 今日 すごい 資料 書く 勉強

| オーソリティ     | プロフィール                  | フォロー数 |
|------------|-------------------------|-------|
| understeer | Solutions Architect     | 2,368 |
| yutuki_r   | Cloud, Hadoop, NoSQL    | 800   |
| kimtea     | NoSQL, Google AppEngine | 2,742 |
| repeatedly | D, Ruby, Python, Opera  | 4,081 |
| siena      | ウェブ技術全般, 情報設計           | 1,065 |

# アウトライン

---

- データの行列表現
  - » 購買履歴, 文書群, 音
- NMF (非負値行列因子分解) とは
  - » NMF: Non-negative Matrix Factorization
- NMF の応用例
  - » 楽器音の分析, 曲からのヴォーカル分離
  - » 文書群の分析
- 基底の数と近似誤差
- NMF アルゴリズム

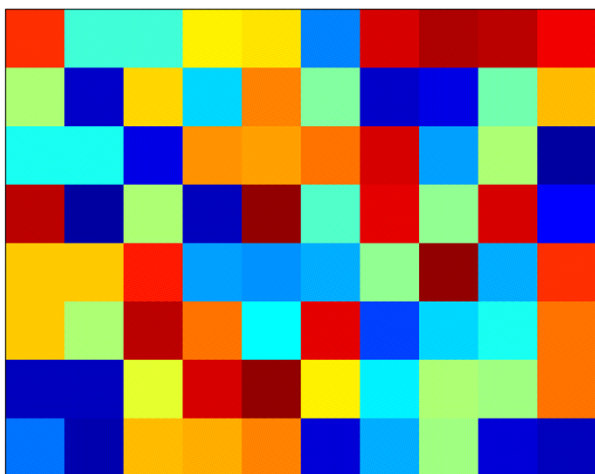
# 基底の数 $K$

---

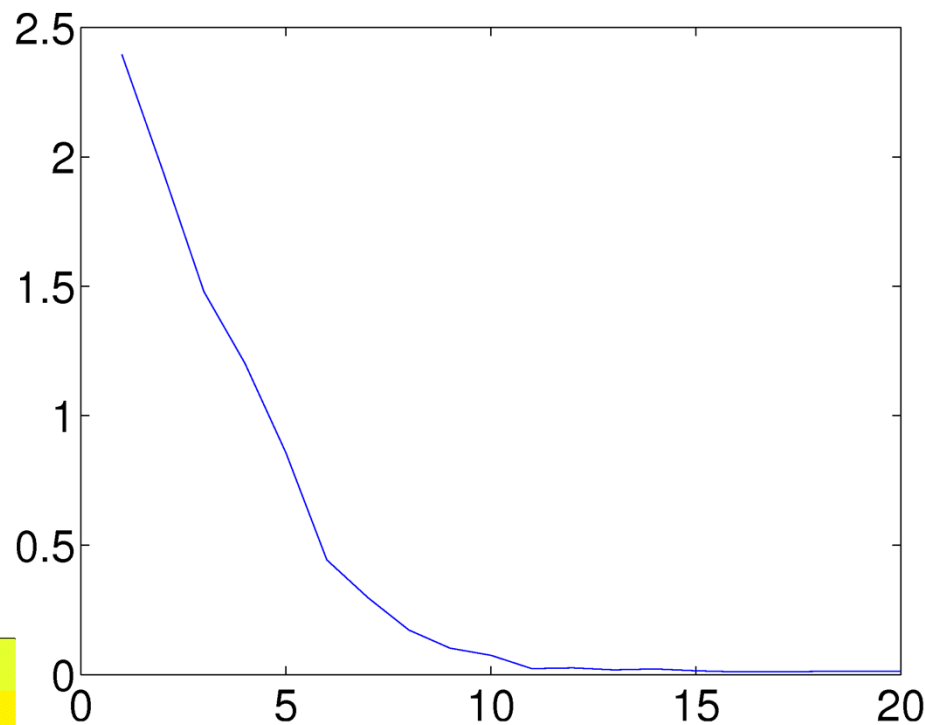
- 解析の際: クラスタの数を指定
  - 最適な $K$ の値の決定は自明ではない
    - » 何個の音パターンを取り出すか
    - » 文書集合を何個にクラスタリングするか
- $K$ を増やすことで因子分解後の表現能力が高まる
  - » 近似誤差が減少する
  - » 増やし過ぎるとOver-fittingし, 汎化性能が下がる
- $K$ の決定の一つの方法
  - » 近似誤差の減少が飽和した辺り

# 基底の数と近似誤差

$X$

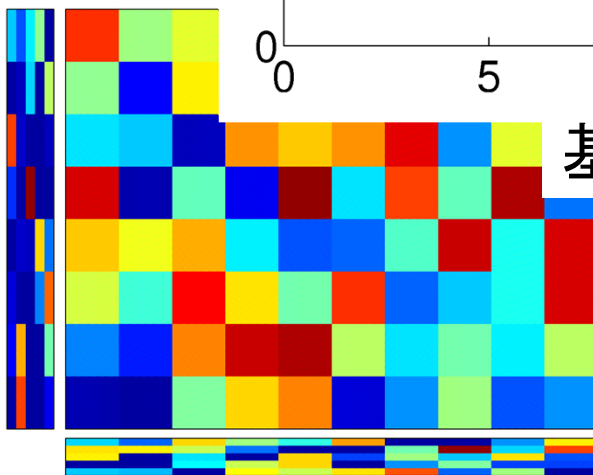
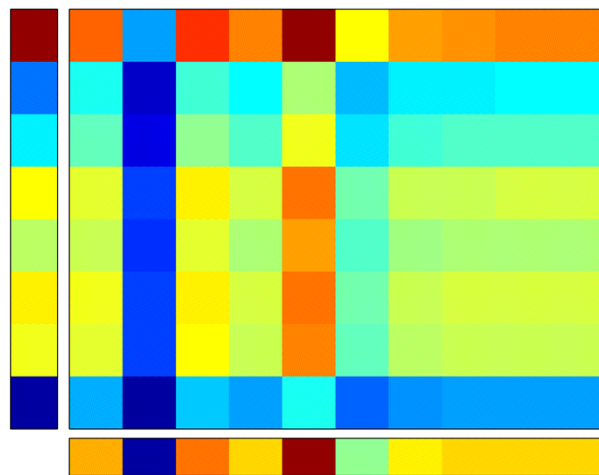


近似誤差

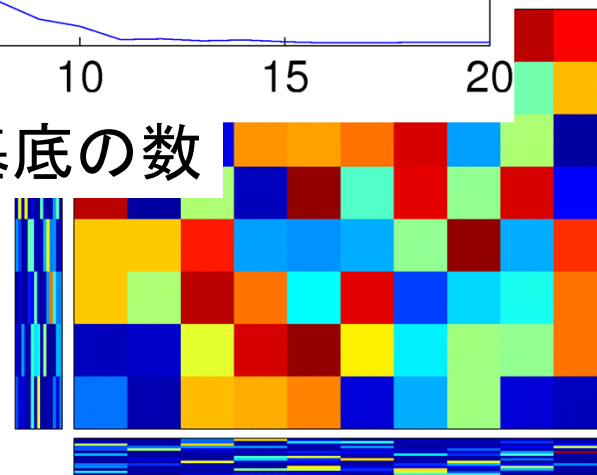


$TV$

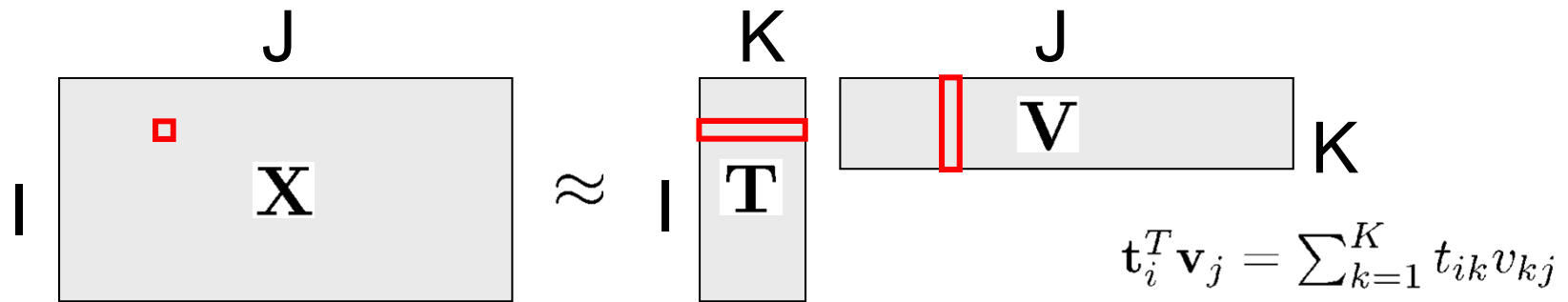
$K=1$



基底の数



# 近似誤差の定義



$$D(\mathbf{X} || \mathbf{T}\mathbf{V}) = \sum_{i=1}^I \sum_{j=1}^J d(\underline{x_{ij}} || \underline{\mathbf{t}_i^T \mathbf{v}_j})$$

- Euclid距離の二乗

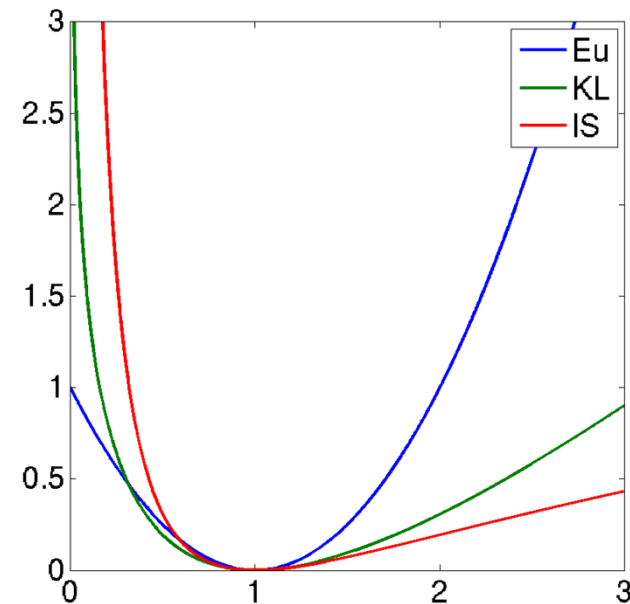
$$d(x_{ij} || \mathbf{t}_i^T \mathbf{v}_j) = (x_{ij} - \mathbf{t}_i^T \mathbf{v}_j)^2$$

- 一般化 KL divergence

$$d(x_{ij} || \mathbf{t}_i^T \mathbf{v}_j) = x_{ij} \log \frac{x_{ij}}{\mathbf{t}_i^T \mathbf{v}_j} - x_{ij} + \mathbf{t}_i^T \mathbf{v}_j$$

- IS divergence (板倉齊藤距離)

$$d(x_{ij} || \mathbf{t}_i^T \mathbf{v}_j) = \frac{x_{ij}}{\mathbf{t}_i^T \mathbf{v}_j} - \log \frac{x_{ij}}{\mathbf{t}_i^T \mathbf{v}_j} - 1$$



# NMF アルゴリズム

---

- 事前に決めておくこと: 近似誤差の定義
  - » さきほど紹介した3種類, あるいはその他
- アルゴリズムへの入力
  - » 非負値行列  $\mathbf{X}$ , 基底の数  $K$
- アルゴリズムの出力
  - » 非負値行列  $\mathbf{T}$ ,  $\mathbf{V}$
- アルゴリズムの目的
  - » 近似誤差の最小化

$$F(\mathbf{T}, \mathbf{V}) = D(\mathbf{X} || \mathbf{TV}) = \sum_{i=1}^I \sum_{j=1}^J d(x_{ij} || \mathbf{t}_i^T \mathbf{v}_j)$$

# Multiplicative update rules

ランダムな非負値で初期化した行列 $\mathbf{T}$ ,  $\mathbf{V}$ に下記の更新式を何回か繰り返し適用することで, 分解後の行列 $\mathbf{T}$  と $\mathbf{V}$  が得られる

$\hat{x}_{ij} = \mathbf{t}_i^T \mathbf{v}_j$  は, 元の行列の要素  $x_{ij}$  と等しくなるべきもの

Euclid 距離

$$t_{ik} \leftarrow t_{ik} \frac{\sum_j x_{ij} v_{kj}}{\sum_j \hat{x}_{ij} v_{kj}}, \quad v_{kj} \leftarrow v_{kj} \frac{\sum_i x_{ij} t_{ik}}{\sum_i \hat{x}_{ij} t_{ik}}$$

KL divergence

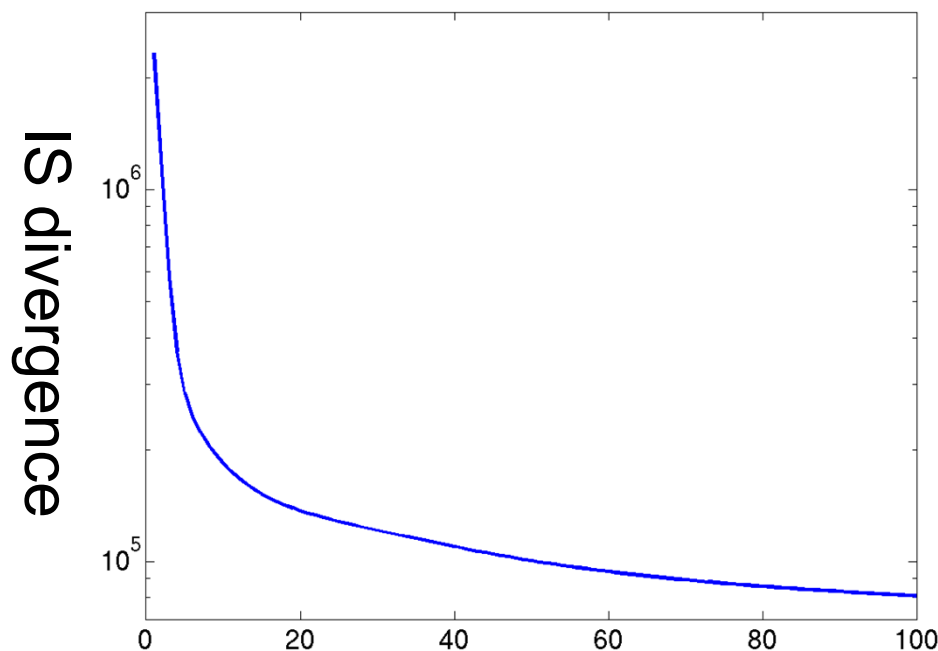
$$t_{ik} \leftarrow t_{ik} \frac{\sum_j \frac{x_{ij}}{\hat{x}_{ij}} v_{kj}}{\sum_j v_{kj}}, \quad v_{kj} \leftarrow v_{kj} \frac{\sum_i \frac{x_{ij}}{\hat{x}_{ij}} t_{ik}}{\sum_i t_{ik}}$$

IS divergence

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_j \frac{x_{ij}}{\hat{x}_{ij}} \frac{v_{kj}}{\hat{x}_{ij}}}{\sum_j \frac{v_{kj}}{\hat{x}_{ij}}}}, \quad v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_i \frac{x_{ij}}{\hat{x}_{ij}} \frac{t_{ik}}{\hat{x}_{ij}}}{\sum_i \frac{t_{ik}}{\hat{x}_{ij}}}}$$

# NMFアルゴリズム 実行の様子(1)

- 音響信号(ヴォーカル分離)の例, 20秒全体  
  »  $513 \times 1247$ の行列(密行列)
- Convolutional-NMF, 基底数  $K = 30$ , 時間幅  $L = 12$



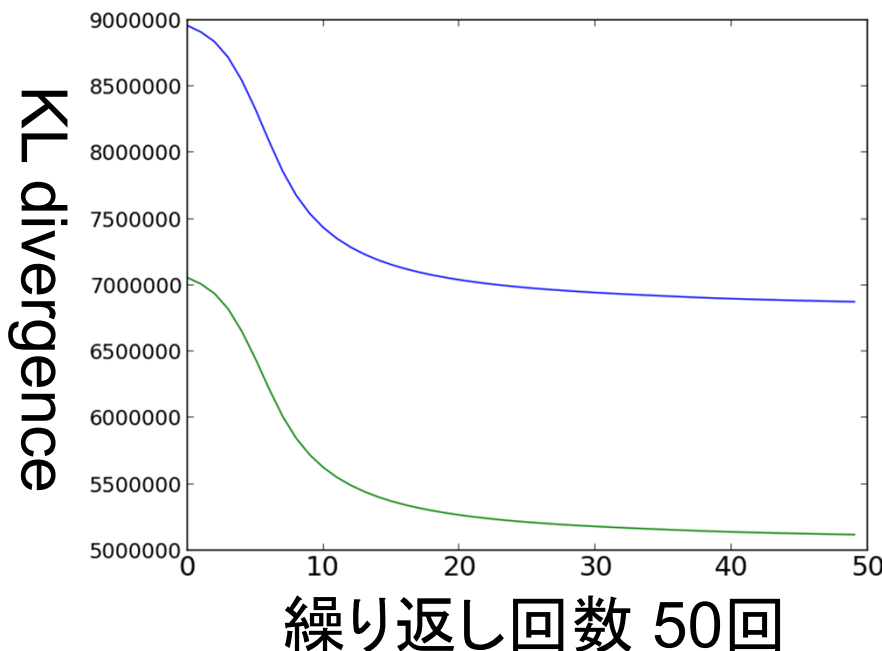
実行時間63秒

繰り返し回数 100回

(Matlab, Xeon W3680 3.33GHz)

# NMFアルゴリズム 実行の様子(2)

- 20 Newsgroups Bag-of-words
  - » 約5万×約2万のスパース行列, 非ゼロ要素数: 約138万
- スパース行列表現(ゼロ要素はメモリを使用しない)でのNMF, 基底数  $K = 20$



全要素の誤差を計算  
実行時間464秒

非ゼロ要素との誤差のみ計算  
実行時間50秒

(Python, Xeon W3680 3.33GHz)

# 後半(NMF)のまとめ

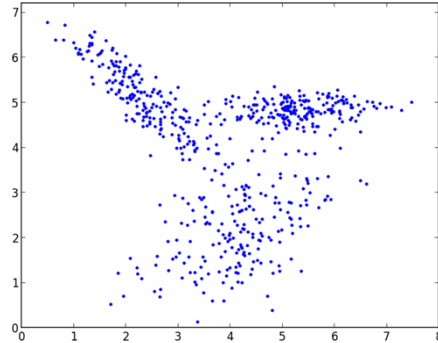
---

- 多種多様なデータの行列表現
  - » 文書をBag-of-words, 音をスペクトログラム
  - » 複合データは複数の行列で表現
- 頻出パターン抽出法
  - » 多くの状況でクラスタリング的解釈が可能
- 応用例
  - » ヴォーカル分離, 文書集合の自動分類, 他多数
- 簡単に実装／実行できるアルゴリズム

# 全体のまとめ

## データ そのもの

実際には高次元過ぎて  
図示できない



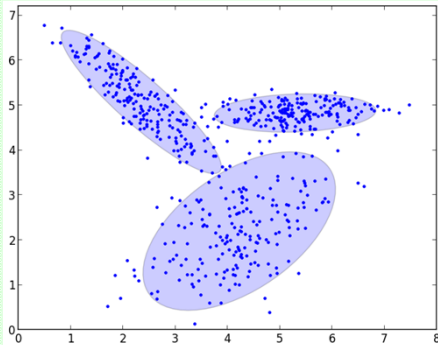
## 行列 そのもの

実際には行数／列数が膨大で  
把握が困難

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 2 | 0 | 3 | 3 | 2 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 2 |
| 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 2 | 0 | 0 |

## GMM

3つの塊から成る  
構造が把握できる



## NMF

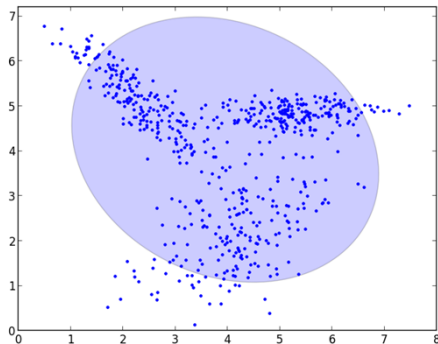
3つのグループが抽出され  
各グループの傾向が見える

|     |     |     |
|-----|-----|-----|
| 0.4 | 0.  | 0.  |
| 0.4 | 0.  | 0.  |
| 0.3 | 0.  | 0.  |
| 0.2 | 0.  | 0.  |
| 0.  | 0.  | 0.2 |
| 0.  | 0.  | 0.9 |
| 0.  | 0.6 | 0.  |
| 0.  | 0.5 | 0.2 |
| 0.  | 0.4 | 0.2 |

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| 0 | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 3 | 2 | 1 | 2 |
| 0 | 0 | 2 | 0 | 3 | 3 | 2 | 0 | 0 | 1 | 0 | 0 |

## 単一 ガウス分布

全体を捉えているが  
構造は表現されていない



## 累計値

全体の傾向は見えるが  
特徴的なグループは見えない

5  
5  
4  
2  
4  
11  
8  
6  
7

9 4 4 4 5 6 3 2 4 5 3 3