

# $\gamma$ ダイバージェンスを用いた非負値行列因子分解法

町田浩平\* 竹之内高志†

(公立はこだて未来大学大学院)‡

## 1 はじめに

世の中には様々な種類のデータが存在する．例えば，パワースペクトル，画像のピクセル値などは非負値を取る行列型データとして表現することができる．このような行列型データを解析して特徴抽出する技術が，テキストマイニング，音響信号処理，画像処理などに応用されている．例えば画像処理では，複数の画像から特徴抽出を行う場合，各々の画像を列ベクトルとして結合して行列型データとして扱う．1メガピクセルの画像をベクトルとして表現した場合，100万次元ベクトルとなるため，画像が  $m$  枚ある場合には， $1000000 \times m$  の巨大な非負値行列を処理することになる．このような，非負値の行列型データの解析手法として，非負値行列因子分解 (Non-negative Matrix Factorization, NMF) が提案された [8]．NMF は，非負値行列を二つの低ランクの非負値行列に分解する多変量解析の手法である．低ランクの非負値行列に分解することにより，観測したデータより少量のパラメータで，データの特徴を表現することができる．また非負値制約を課すことによって，二つの非負値行列はスパースになり，人間が解釈可能な特徴が抽出されやすいという性質がある．従来のNMFでは，コスト関数としてユークリッド距離やKLダイバージェンスといった距離尺度が用いてきた．しかし，しばしば観測データは非観測なノイズを含んでいるため，これらの距離尺度を用いたNMFではノイズの影響を強く受ける場合がある．このようなノイズに対処するために，一般化ダイバージェンスの一つであり，ノイズに対して頑健である  $\beta$ ダイバージェンスをコスト関数として用いたNMFが提案されている．本稿では，ノイズに対してより頑健な性質を持つことで知られる  $\gamma$ ダイバージェンスを用いたNMFアルゴリズムについて検討する．

2章でNMFの定式化について述べ，3章で一般化ダイバージェンスの一例である  $\beta$ ダイバージェンスと  $\gamma$ ダイバージェンスについて解説する．4章において， $\gamma$ ダイバージェンスをコスト関数としたNMFアルゴリズムを2種類の定式化で導出する．5章で，数値実験により導出したアルゴリズムの性質・性能を検証し，考察する．

## 2 設定

NMFでは，解析対象のデータ行列  $V \in \mathbb{R}^{n \times m}$  を二つの非負値行列  $W \in \mathbb{R}^{n \times r}$ ， $H \in \mathbb{R}^{r \times m}$  の行列積を用いて

近似する． $W, H$  は以下の最適化問題を解くことで求めることができる．

$$\begin{aligned} \arg \min_{W, H} \quad & D(V||WH) \\ \text{subject to} \quad & W_{ij} \geq 0, H_{jk} \geq 0 \\ & (i = 1, \dots, n, j = 1, \dots, r, k = 1, \dots, m) \end{aligned} \quad (1)$$

ただし，行列  $M$  に対し， $M_{ij}$  を  $M$  の  $(i, j)$  要素とする． $W$  の各列は  $V$  を表現するための  $r$  個の基底を表し， $H$  の各行は  $W$  の  $r$  個の基底に対応する重みを表す．基底の数  $r$  は  $nm > r(n+m)$  を満たす自然数であるとき， $V$  の持つ特徴を低ランクの  $W$  に縮約することができる． $D(\cdot||\cdot)$  は，二つの確率分布間，もしくは二つの行列間の隔たりを測るコスト関数である．従来，NMFの文脈では，二つの行列間の近似度を測るために，以下に示すユークリッド距離  $D_{EU}$ ，KLダイバージェンス  $D_{KL}$ ，板倉斎藤距離  $D_{IS}$  といった距離尺度が用いられてきた [4, 9, 10, 11]．

$$D_{EU}(A||B) = \sum_{i,t} (A_{it} - B_{it})^2. \quad (2)$$

$$D_{KL}(A||B) = \sum_{i,t} (A_{it} \log \frac{A_{it}}{B_{it}} - A_{it} + B_{it}). \quad (3)$$

$$D_{IS}(A||B) = \sum_{i,t} (\frac{A_{it}}{B_{it}} - \log \frac{A_{it}}{B_{it}} - 1). \quad (4)$$

ただし， $\sum_{i,t}$  は行列の全要素に関する和を表す．KLダイバージェンスや板倉斎藤距離は確率分布間の隔たりを測るために標準的に用いられる距離尺度である．適用する距離尺度に応じて異なる最適化問題を解くことになるため，距離尺度ごとに抽出されるパターンや特徴が異なることに注意する．一般に，観測データにはなんらかの非観測なノイズが加わっており，ノイズの種類を仮定した上で推定を行う．例えば，ユークリッド距離を適用した場合は，観測データに正規分布に従うノイズが加わっていると解釈することができる．データに加わるノイズの分布型は未知であるが，何らかの方法でノイズに適した距離尺度を用いることができれば，より適切な特徴抽出を行うことが可能となる．従来提案されているNMFアルゴリズムは，行列間の隔たりを上述の距離尺度で直接測り，最小化することで  $W, H$  を求めている．その一方で， $V$  と  $WH$  の隔たりを統計モデルを用いて表現・推定することで  $W, H$  を求めることもできる [4]．本稿では，これらの二つの異なる定式化に基づき，これらの  $\gamma$ ダイバージェンスを用いた2種類のNMFのアルゴリズムを導出し，その性質，性能を検証する．

\*g2114029@fun.ac.jp

†ttakashi@fun.ac.jp

‡函館市亀田中野町 116 番地 2

### 3 一般化ダイバージェンス

二つの行列  $A, B$  間の  $\beta$  ダイバージェンス  $D_\beta$ , そして  $\gamma$  ダイバージェンス  $D_\gamma$  は以下のように定義される [1, 3, 6].

$$D_\beta(A||B) = \frac{1}{\beta(\beta-1)} \sum_{i,t} (A_{it}^\beta + (\beta-1)B_{it}^\beta - \beta A_{it} B_{it}^{\beta-1}) \quad (5)$$

$$D_\gamma(A||B) = \frac{1}{\gamma(\gamma+1)} \log \frac{(\sum_{i,t} A_{it}^{\gamma+1})(\sum_{i,t} (B_{it}^{\gamma+1})^\gamma)}{(\sum_{i,t} (A_{it} B_{it}^\gamma))^{\gamma+1}} \quad (6)$$

$\beta, \gamma$  はそれぞれダイバージェンスを規定するパラメータである. 各ダイバージェンスは非負値を取り,  $A = B$  のときのみ 0 となる.  $\beta$  ダイバージェンスは,  $\beta \rightarrow 1$  の極限でユークリッド距離と等価となり,  $\beta \rightarrow 0$  の極限で KL ダイバージェンスと等価となる [1, 2]. また,  $\beta$  ダイバージェンスは, パラメータ  $\beta$  を大きく設定するとノイズに対する頑健性が向上する一方で, 学習効率が低下する. つまり, パラメータ  $\beta$  は頑健性と学習効率を制御しており, 両者はトレードオフの関係にある. これまでの研究では, コスト関数として  $\beta$  ダイバージェンスを用いた NMF アルゴリズムが提案されている [5].  $\gamma$  ダイバージェンスは,  $\gamma \rightarrow 0$  の極限で正規化 KL ダイバージェンスに収束する [3].

$$\lim_{\gamma \rightarrow 0} D_\gamma(A||B) = \sum_{i,t} (\tilde{A}_{it} \log \frac{\tilde{A}_{it}}{\tilde{B}_{it}}) \quad (7)$$

ただし行列  $M$  に対して行列  $\tilde{M}$  を以下で定義する.

$$\tilde{M}_{it} = \frac{M_{it}}{\sum_{kl} M_{kl}} \quad (8)$$

(8) の  $\tilde{M}$  は全要素の総和を取ると 1 となる. また,  $\gamma$  ダイバージェンスは以下の性質を持つ [1].

$$D_\gamma(c_1 A || c_2 B) = D_\gamma(A || B) \quad (9)$$

$c_1, c_2$  は正の任意定数である. この性質は, スケール不変性と呼ばれ,  $c_1, c_2$  はどのような値であっても  $\gamma$  ダイバージェンスの値は変わらない. さらに,  $\gamma$  ダイバージェンスはデータに加わっているノイズを自動的に無視する性質がある [6]. 本稿では, NMF の頑健性を向上させることを目的として,  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムを導出し, その性質について検証する.

### 4 アルゴリズム

最適化問題 (1) を解くために, Lee らはユークリッド距離と KL ダイバージェンスを適用した場合に対して, Multiplicative update rules を提案した [9]. このアルゴリズムの特徴は, 正の値を取るパラメータ  $W, H$  の積の形で更新式が表されることによって, 行列の非負性が保たれることにある. 本章では,  $\gamma$  ダイバージェンスをコスト関数として適用した場合に導かれる二種類の Multiplicative update rules の導出を行う.

#### 4.1 行列間の隔たりに基づく方法

本章では,  $V$  と  $WH$  の隔たりを直接  $\gamma$  ダイバージェンスを用いて測ることを考える.  $\gamma$  ダイバージェンスを用いた NMF のコスト関数は以下で定義される.

$$D_\gamma(V||WH) = \frac{1}{\gamma(\gamma+1)} \log \frac{(\sum_{i,t} V_{it}^{\gamma+1})(\sum_{i,t} ((WH)_{it}^{\gamma+1})^\gamma)}{(\sum_{i,t} (V_{it}(WH)_{it}^\gamma))^{\gamma+1}} \quad (10)$$

(10) を最小化する  $W, H$  を求めるために勾配法を用いる.  $W_{kl}$  に対する (10) 式の勾配は

$$\frac{\partial}{\partial W_{kl}} D_\gamma(V||WH) = \frac{\sum_t (WH)_{kt}^\gamma H_{lt}}{\sum_{i,t} (WH)_{it}^{\gamma+1}} - \frac{\sum_t V_{kt} (WH)_{kt}^{\gamma-1} H_{lt}}{\sum_{i,t} V_{it} (WH)_{it}^\gamma} \quad (11)$$

となり, 勾配法に基づいた  $W_{kl}$  の更新式は,

$$W_{kl} \leftarrow W_{kl} - \eta_{kl} \left[ \frac{\sum_t (WH)_{kt}^\gamma H_{lt}}{\sum_{i,t} (WH)_{it}^{\gamma+1}} - \frac{\sum_t V_{kt} (WH)_{kt}^{\gamma-1} H_{lt}}{\sum_{i,t} V_{it} (WH)_{it}^\gamma} \right] \quad (12)$$

となる.  $\eta_{kl}$  は正の値を取る学習係数である. 学習係数は任意の値に設定できるため,

$$\eta_{kl} = W_{kl} \frac{\sum_{i,t} (WH)_{it}^{\gamma+1}}{\sum_t (WH)_{kt}^\gamma H_{lt}} \quad (13)$$

と設定すると, 更新式 (12) は以下のように書き換えることができる.

$$W_{kl} \leftarrow W_{kl} \frac{\sum_{i,t} (WH)_{it}^{\gamma+1} \sum_t V_{kt} (WH)_{kt}^{\gamma-1} H_{lt}}{\sum_{i,t} V_{it} (WH)_{it}^\gamma \sum_t (WH)_{kt}^\gamma H_{lt}} \quad (14)$$

$W, H$  は非負値であるため, 更新式 (14) により  $W$  の非負性が保たれる. 同様に  $H_{pq}$  の更新式は,

$$H_{pq} \leftarrow H_{pq} \frac{\sum_{i,t} (WH)_{it}^{\gamma+1} \sum_i V_{iq} (WH)_{iq}^{\gamma-1} W_{ip}}{\sum_{i,t} V_{it} (WH)_{it}^\gamma \sum_i (WH)_{iq}^\gamma W_{ip}} \quad (15)$$

となる. まとめると,  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムは以下のように書くことができる.

1.  $W, H$  の要素を適当な方法で初期化する.
2. 収束するまで (15) と (14) を交互に繰り返す.

#### 4.2 統計モデルに基づく方法

4.1 では,  $V$  と  $WH$  の隔たりを  $\gamma$  ダイバージェンスを用いて直接測ることで NMF のアルゴリズムを導出した. 本章では,  $V$  と  $WH$  の隔たりを統計モデルを用いて表現し, データ行列から定義される経験分布と統計モデル間の距離尺度として  $\gamma$  ダイバージェンスを用いることで NMF のアルゴリズムを導出する.

行列  $V$  の各要素を独立かつ同一の分布 (i.i.d) から得られたサンプルとみなし, 以下のような統計モデルを考える.

$$q(V_{it}; (WH)_{it}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(V_{it} - (WH)_{it})^2}{2\sigma^2}\right) \quad (16)$$

データ行列  $V$  から定義される経験分布  $\tilde{p}$  と統計モデル  $q$  間の  $\gamma$  ダイバージェンスは,

$$D_\gamma(\tilde{p}||q) = C - \frac{\gamma - 1}{2(\gamma + 1)} \log(\sigma^2) - \frac{1}{\gamma} \log\left(\sum_{i,t} \exp\left(\frac{-\gamma\delta_{it}^2}{2\sigma^2}\right)\right) \quad (17)$$

となる. ただし,  $C$  は  $W, H$  に依存しない定数であり,  $\delta_{it} = V_{it} - (WH)_{it}$  とする. (17) を  $W, H$  に関して最小化することで, アルゴリズムを導出することができる.  $\gamma = 0$  としたとき,  $\gamma$  ダイバージェンスは正規化 KL ダイバージェンスと一致するため, コスト関数 (17) の最小化は  $D_{EU}$  に基づく通常の NMF アルゴリズムと等価となる. また, 統計モデル (16) において, 正規分布ではなくポアソン分布を仮定し, KL ダイバージェンスを用いた場合は,  $D_{KL}$  を用いた NMF アルゴリズムと等価となる.  $W$  の  $kl$  要素である  $W_{kl}$  に対する (17) の勾配は,

$$\frac{\partial}{\partial W_{kl}} D_\gamma(\tilde{p}||q) \propto \frac{\sum_t \delta_{kt} \exp\left(\frac{-\gamma\delta_{kt}^2}{2\sigma^2}\right) H_{lt}}{\sum_{i,t} \exp\left(\frac{-\gamma\delta_{it}^2}{2\sigma^2}\right)} \quad (18)$$

となる. よって, 勾配法に基づく更新式は,

$$W_{kl} \leftarrow W_{kl} - \eta_{kl} \frac{\sum_t \delta_{kt} \exp\left(\frac{-\gamma\delta_{kt}^2}{2\sigma^2}\right) H_{lt}}{\sum_{i,t} \exp\left(\frac{-\gamma\delta_{it}^2}{2\sigma^2}\right)} \quad (19)$$

となる.  $\eta_{kl}$  は正の値を取る学習係数である. 学習係数は任意の値に設定できるため,

$$\eta_{kl} = W_{kl} \frac{\sum_{i,t} \exp\left(\frac{-\gamma\delta_{it}^2}{2\sigma^2}\right)}{\sum_t \exp\left(\frac{-\gamma\delta_{kt}^2}{2\sigma^2}\right) (WH)_{kt} H_{lt}} \quad (20)$$

と設定すると, 次の更新式が得られる.

$$W_{kl} \leftarrow W_{kl} \frac{\sum_t V_{kt} \exp\left(\frac{-\gamma\delta_{kt}^2}{2\sigma^2}\right) H_{lt}}{\sum_t (WH)_{kt} \exp\left(\frac{-\gamma\delta_{kt}^2}{2\sigma^2}\right) H_{lt}} \quad (21)$$

同様に,  $H_{pq}$  における更新式は,

$$H_{pq} \leftarrow H_{pq} \frac{\sum_i V_{iq} \exp\left(\frac{-\gamma\delta_{iq}^2}{2\sigma^2}\right) W_{ip}}{\sum_i (WH)_{iq} \exp\left(\frac{-\gamma\delta_{iq}^2}{2\sigma^2}\right) W_{ip}} \quad (22)$$

となる. 以上をまとめると,  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムは以下のように書くことができる.

1.  $W, H, \sigma^2$  を適当な方法で初期化する.
2. (22) と (21) に従い  $W, H$  を更新する.
3. (17) を最小化する  $\sigma^2$  を求める.
4. 収束するまで 2. 及び 3. の手順を繰り返し行う.

3 における  $\sigma^2$  に関する最小化は一次元の最小化であるため, 勾配法で比較的簡単に求めることができる.

ここで,  $\gamma$  ダイバージェンスの代わりに,  $\beta$  ダイバージェンスを用いた NMF アルゴリズムについて考える.

経験分布  $\tilde{p}$  と統計モデル  $q$  の間の  $\beta$  ダイバージェンスは,

$$D_\beta(\tilde{p}||q) = C + (2\pi\sigma^2)^{\frac{1-\beta}{2}} \left( \frac{1}{\beta^{\frac{3}{2}}} + \frac{1}{nm(1-\beta)} \sum_{i,t} \exp\left(\frac{(1-\beta)\delta_{it}^2}{2\sigma^2}\right) \right) \quad (23)$$

となる. (23) を  $\gamma$  ダイバージェンスの場合と同様に, 勾配法に基づき Multiplicative update rules を導出すると, 以下の更新式が得られる.

$$W_{kl} \leftarrow W_{kl} \frac{\sum_t V_{kt} \exp\left(\frac{(1-\beta)\delta_{kt}^2}{2\sigma^2}\right) H_{lt}}{\sum_t (WH)_{kt} \exp\left(\frac{(1-\beta)\delta_{kt}^2}{2\sigma^2}\right) H_{lt}} \quad (24)$$

$$H_{pq} \leftarrow H_{pq} \frac{\sum_i V_{iq} \exp\left(\frac{(1-\beta)\delta_{iq}^2}{2\sigma^2}\right) W_{ip}}{\sum_i (WH)_{iq} \exp\left(\frac{(1-\beta)\delta_{iq}^2}{2\sigma^2}\right) W_{ip}} \quad (25)$$

(21) と (24), (22) と (25) はそれぞれ,  $\gamma = \beta - 1$  とすると等価な更新式となる. したがって,  $\beta$  ダイバージェンスを用いた統計モデルに基づく NMF アルゴリズムは,  $\sigma^2$  に関する最適化を除き  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムと等価となる.

## 5 実験

$\gamma$  ダイバージェンスの頑健性を検証するために人工データ実験を行う.  $V$  の各要素を  $[0, 100]$  上の一様分布で生成し, 以下の手続きで実験を行う.

1. 行列  $V \in \mathbb{R}^{30 \times 30}$  を生成する.
2.  $V$  に外れ値ノイズを加え,  $V_{noise}$  とする.
3.  $V_{noise}$  を用いて  $W, H$  を学習する. ただし, 基底数  $r = 15$  とする.
4. 得られた  $W, H$  の行列積  $WH$  と  $V$  の誤差を平均二乗和誤差 (Mean squared error, MSE)  $\frac{1}{nm} D_{EU}(V||WH)$  で測りノイズに対する挙動を検証する.

手順 4 において,  $V_{noise}$  ではなく,  $V$  との誤差を測っていることに注意する.  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムにおける  $W, H$  の各要素の初期値は,

$$\sqrt{\frac{V_{median}}{r}} + \epsilon \quad (26)$$

とする. ただし,  $V_{median}$  は  $V$  の中央値を表し,  $\epsilon$  は正規乱数により生成された微小ノイズである (平均 0, 分散 0.01). 手順 2 におけるノイズは  $V$  の全要素のうち, ランダムに選ばれた 3% の要素をノイズとし, 値を 250 とした. また, 上述の実験を 20 回行うことで, 手法の平均的な性能を検証した.  $\gamma$  の値は  $0 \leq \gamma \leq e$  の範囲で探索して各  $\gamma$  におけるアルゴリズムの頑健性を検証する.

Fig.1 は, 4.1 で導出した行列同士の近似度を測るアルゴリズムにより得られた結果であり, Fig.2 は, 4.2 で導出した統計モデルを用いたアルゴリズムにより得られた結果を示す. 各図において, 横軸は  $\gamma$  を表し, 縦軸は MSE を表す. Fig.1 における  $\gamma = 0$  の点は, 行列間の正規化 KL ダイバージェンスを用いた NMF アルゴリズムに対

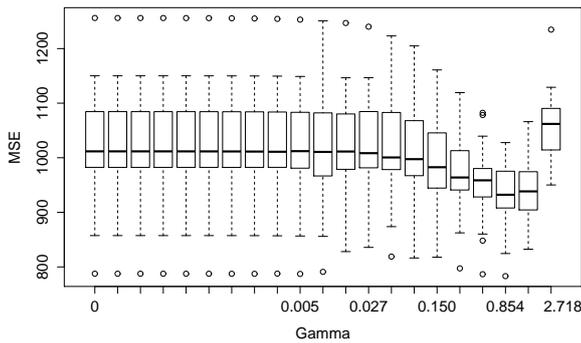


Fig. 1 行列間の隔たりに基づく NMF アルゴリズムによる MSE . 横軸は  $\gamma$  , 縦軸は MSE を表す .

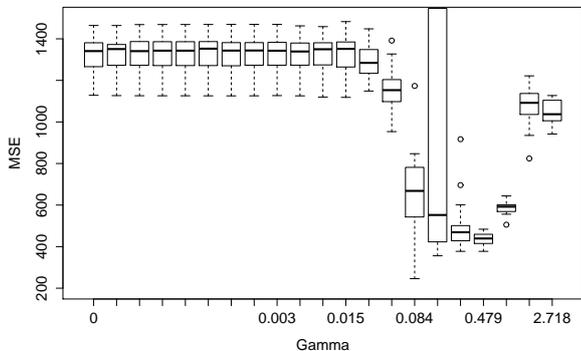


Fig. 2 統計モデルに基づく NMF アルゴリズムによる MSE . 横軸は  $\gamma$  , 縦軸は MSE を表す .

応し, Fig.2 における  $\gamma = 0$  の点は行列間のユークリッド距離  $D_{EU}$  に基づく通常の NMF アルゴリズムに対応することに注意する . Fig.1 , Fig.2 共に ,  $\gamma = 0.1$  付近で MSE が小さくなり始めているが , MSE が最小となる  $\gamma$  の値は二つのアルゴリズムで異なっている . ユークリッド距離 , 正規化 KL ダイバージェンスを用いた NMF アルゴリズムと比較して ,  $\gamma$  ダイバージェンスに基づく NMF アルゴリズムがノイズに対して頑健であることが確認できる . また , 統計モデルに基づく NMF は行列間距離に基づく NMF より , 小さな MSE を達成しており , 適切に  $\gamma$  を選択することができれば , ノイズに対してより頑健となることを示している .

## 6 まとめ

本稿では , ノイズに対する頑健性に注目して , 一般化ダイバージェンスの一つである  $\gamma$  ダイバージェンスをコスト関数として用いた NMF アルゴリズムを , 行列間の隔たりに基づく方法と統計モデルに基づく方法の 2 種類の方法で導出を行った . 導出した NMF アルゴリズムの頑健性を検証するために実験を行い , 外れ値ノイズに対して統計モデルに基づく NMF アルゴリズムが特に頑健

であることを確認した . 本稿では , 観測データの全要素を用いて特徴抽出を行ったが , データに欠測が生じることがしばしばあるため , データの一部を利用して特徴抽出を行うようにアルゴリズムを拡張する必要がある . また , 実験では  $\gamma$  の値を網羅的に探索し  $\gamma$  ダイバージェンスを用いた NMF アルゴリズムのノイズに対する頑健性を調べたが , 実データに適用するには , 交差検証法などを用いてデータから適切な  $\gamma$  を決定する必要がある .

## 参考文献

- [1] A. Cichocki and S. Amari, " Families of Alpha-Beta- and Gamma- Divergence: Flexible and Robust Measures of Similarities, " Entropy, 12(6), pp. 1532-1568, 2010.
- [2] A. Cichocki, C. Sergio and S. Amari, "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization," Entropy, 13(1), pp. 134-170, 2011.
- [3] O. Dikmen, Z. Yang, and E. Oja, "Learning the Information Divergence," arXiv:1406.1385, 2014.
- [4] C. Fevotte, N. Bertin and J. Durrieu, " Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis., " Neural Computation , 21(3), pp. 793-830, 2009.
- [5] C. Fevotte and J. Idier, " Algorithms for non-negative matrix factorization with the beta-divergence," Neural Computation, 23(9), pp.2421-2456, 2010.
- [6] H. Fujisawa, S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," Journal of Multivariate Analysis, 99, pp. 2053-2081, 2008.
- [7] T. Kanamori and H. Fujisawa, "Robust Estimation under Heavy Contamination using Enlarged Models," arXiv:1311.5301, 2013.
- [8] D. D. Lee and H. S. Seung, " Learning the parts of objects by non-negative matrix factorization, " Nature, vol.401, pp. 788-791, 1999.
- [9] D. D. Lee and H. S. Seung, " Algorithms for Non-negative Matrix Factorization, " Neural Information P
- [10] S. Lyu and X. Wang, " On Algorithms for Sparse Multi-factor NMF, " Neural Information Processing Systems, pp. 602-610, 2013.
- [11] K. Takeuchi, K. Ishiguro, A. Kimura and H. Sawada, " Non-negative Multiple Matrix Factorization, " International Joint Conference on Artificial Intelligence, pp. 1713-1720, 2013.