マルチチャネル非負値行列因子分解における初期値依存性の挙動解析 *

☆三浦伊織 (大分大), 太刀岡勇気, 成田知宏, 石井純 (三菱電機),
吉山文教, 上ノ原進吾, 古家賢一 (大分大)

1 はじめに

非負値行列因子分解 (Nonnegative Matrix Factorization: NMF)[1] とは非負値の行列を分解し、解析を行う手法である。行列表現できるデータならば分解可能であるため、音や画像, 文書など多種多様なものに利用できる。音響分野ではマルチチャネル拡張による空間情報を付与することで音源分離を行う手法が提案されている。しかし、従来のマルチチャネル NMF[2][3] は自由度の高いモデルであるため、分離性能に対する初期値依存性が課題となっている [4]。

本稿では、通常ランダムに設定される初期値に対して、いくつかの条件で計算した初期値を設定し分離した際の初期値依存性の挙動解析を行った。

2 マルチチャネル NMF[2][3]

2.1 概要

マルチチャネル NMF とは、NMF をマルチチャネル拡張したものであり、観測行列を 4 つの行列 **H**、**Z**、**T**、**V** に分解する。マルチチャネル NMF では空間情報を用いてスペクトル基底を L 個にクラスタリングすることで事前の学習なしで音源分離を実現する。ただし、位相情報を扱うため複素数を用いる。そこで、複素数における非負性に対応するものとして、エルミート半正定値行列を用いる [2]。

2.2 定式化

M をマイクロホン数として入力ベクトルを $\tilde{\mathbf{x}} = [\tilde{x}_1, \cdots, \tilde{x}_M]^\top$ とする。ただし、$^\top$ は転置を表す。$\tilde{x}_m$ は m 番目のマイクロホンでの Short Time Fourier Transform (STFT) の複素係数であり、スペクトログラムを指す。周波数 i $(1 \le i \le I)$、時間 j $(1 \le j \le J)$ のとき $\tilde{\mathbf{x}}_{ij}$ で表すと行列 $\mathbf{X}$ は $\mathsf{X}_{ij} = \tilde{\mathbf{x}}_{ij}\tilde{\mathbf{x}}_{ij}^H$ もしくは

$$\mathsf{X} = \tilde{\mathbf{x}}_m \tilde{\mathbf{x}}_m^H = \begin{bmatrix} |\tilde{x}_1|^2 & \cdots & \tilde{x}_1\tilde{x}_M^* \\ \vdots & \ddots & \vdots \\ \tilde{x}_M\tilde{x}_1^* & \cdots & |\tilde{x}_M|^2 \end{bmatrix} \tag{1}$$

で表される。ただし、^{H} はエルミート転置を表す。すなわち、I 行 J 列の行列 $\mathbf{X}$ はそれぞれの要素が $M \times M$ の行列を持つ階層的なエルミート半正定値行列となる。この行列 $\mathbf{X}$ をマルチチャネル NMF で分解した結果は、K 個の基底から成る基底行列 **T**、アクティベーション行列 **V**、音源の空間情報を示す空間相関行列 **H** と音源の空間情報と各基底を関連付ける潜在変数行列 **Z** という 4 つの行列に分解され、次式で示

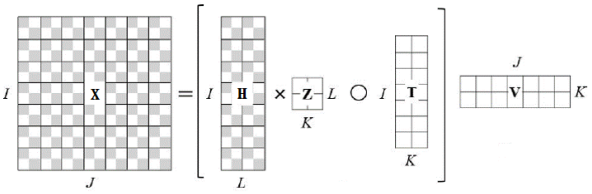

Fig. 1 マルチチャネル NMF で分解された行列の例

される。

$$\mathbf{X} = (\mathbf{HZ} \circ \mathbf{T})\mathbf{V} \tag{2}$$

ただし、$\circ$ はアダマール積を表す。行列 **T** は I 行 K 列の行列、行列 **V** は K 行 J 列の行列、行列 **H** は行列 **X** と同様にそれぞれの要素が $M \times M$ の行列を持つ I 行 L 列の階層的なエルミート半正定値行列、行列 **Z** は L 行 K 列の行列である。L は音源数を表している。Fig. 1 は式 (2) を図式化したものである。このとき、右辺は

$$\hat{\mathsf{X}}_{ij} = \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \mathsf{H}_{il} z_{lk} \right) t_{ik} v_{kj} \tag{3}$$

と表すことができ、理想的には行列 X と $\hat{\mathsf{X}}_{ij}$ を要素に持つ行列 $\hat{\mathsf{X}}$ は等しくなる。しかし、一般的には誤差が生じるため、マルチチャネル NMF では行列 X と行列 $\hat{\mathsf{X}}$ との距離 $D_*(\mathsf{X}, \hat{\mathsf{X}})$ を定義し、この距離を最小化する行列 **T**、 **V**、 **H**、 **Z** を求める。今回はダイナミックレンジが大きい音楽や音声に適している Itakura-Saito (IS) divergence[5] を用いて以下のように定義する。

$$D_{IS}(\mathsf{X}_{ij}, \hat{\mathsf{X}}_{ij}) = tr(\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}) - \log\det \mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1} - M \tag{4}$$

ただし、$tr\,(\cdot)$ は対角要素の和を表している。

2.3 アルゴリズム

Multiplicative update rule[6] と呼ばれる反復アリゴリズムを、ランダムな非負の値で初期化した行列 **T**、**V**、**H**、**Z**に繰り返し適用することで、$D_{IS}(\mathsf{X}, \hat{\mathsf{X}})$ を最小化するような各行列を得る。IS divergence を用いた更新式は以下のようになる。

$$t_{ik} \leftarrow t_{ik} \sqrt{\frac{\sum_l z_{lk} \sum_j v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{X}_{ij}\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}{\sum_l z_{lk} \sum_j v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1}\mathsf{H}_{il})}} \tag{5}$$

---

* Analysis of Initial-value Dependency in Multichannel Nonnegative Matrix Factorization. by Iori Miura (Oita University), Yuuki Tachioka, Tomohiro Narita, Jun Ishii (Mitsubishi Electric), Fuminori Yoshiyama, Shingo Uenohara, and Ken'ichi Furuya (Oita University)

Table 1　実験に用いた音楽データ

ID	Author/Song	Snip	Part
1	Bearlin Roads	85-99 (14 sec)	piano ambient vocals
2	Another Dreamer The Ones We Love	69-94 (25 sec)	drums vocals guitar
3	Fort Minor Remamber The Name	69-94 (24 sec)	drums vocals violin_synth
4	Ultimate Nz Tour	54-78 (18 sec)	drums guitar synth

$$v_{kj} \leftarrow v_{kj} \sqrt{\frac{\sum_l z_{lk} \sum_i t_{ik} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}{\sum_l z_{lk} \sum_i t_{ik} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}} \quad (6)$$

$$z_{lk} \leftarrow z_{lk} \sqrt{\frac{\sum_{i,j} t_{ik} v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}{\sum_{i,j} t_{ik} v_{kj} tr(\hat{\mathsf{X}}_{ij}^{-1} \mathsf{H}_{il})}} \quad (7)$$

H_{il} については次式の A、B を係数に持つ代数リッカチ方程式で求めることができる。

$$A = \sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \quad (8)$$

$$B = \mathsf{H}'_{il} \left(\sum_k z_{lk} t_{ik} \sum_j v_{kj} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \mathsf{X}_{ij}^{-1} \right) \mathsf{H}'_{il} \quad (9)$$

ただし、H'_{il} は更新前の行列 H_{il} を表している。

2.4　正規化

行列 **H** と行列 **Z** については、更新毎に発散を防ぐために正規化を行わなければならない。正規化は以下の式で行った。

$$\mathsf{H}_{il} = \frac{\mathsf{H}_{il}}{tr(\mathsf{H}_{il})} \quad (10)$$

$$z_{lk} = \frac{z_{lk}}{\sum_l z_{lk}} \quad (11)$$

2.5　音源分離

音源分離を行うためにウィナーフィルタを用いる。ウィナーフィルタは一般的には次式で表される。

$$Y = \frac{\hat{S}}{\hat{S} + N} X \quad (12)$$

$Y = \tilde{y}_{ij}^{(l)}$、$\hat{S} = (\sum_{k=1}^{K} z_{lk} t_{ik} v_{kj})\, \mathsf{H}_{il}$、$\hat{S} + N = \hat{\mathsf{X}}_{ij}$、$X = \mathsf{X}_{ij}$ を代入すると、次式のマルチチャネルウィナーフィルタとなり、各クラスタに対応した音源信号を得られる。

$$\tilde{y}_{ij}^{(l)} = \left(\sum_{k=1}^{K} z_{lk} t_{ik} v_{kj} \right) \mathsf{H}_{il} \hat{\mathsf{X}}_{ij}^{-1} \mathsf{X}_{ij} \quad (13)$$

3　初期値設定による挙動解析

ここでは推定が比較的容易と考えられる基底行列 **T** および空間相関行列 **H** に着目し、複数の手法を用いて初期値を設定することで、分離性能がどのように変化するか実験的に分析を行う。

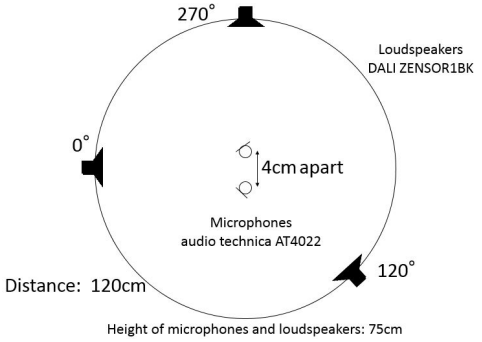

Fig. 2　音源の配置図

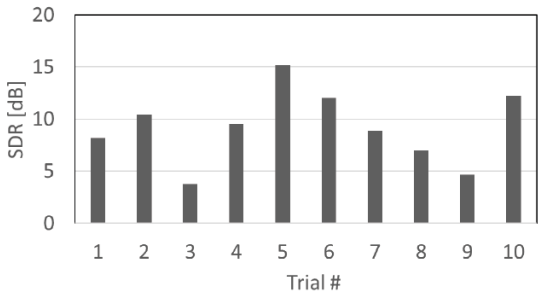

Fig. 3　音源分離性能の初期値依存性

3.1　実験条件

実験に用いた混合信号は Table 1[7] の音楽データに Fig. 2 の環境で測定したインパルス応答 ($M = 2$) を畳み込み作成した。インパルス応答長 300、サンプリング周波数 16 kHz、STFT のフレームサイズ 1024、シフトサイズ 256 とし、基底数 $K = 30$、音源数 $L = 3$、Multiplicative update rule の更新回数 500 とした。また、マルチチャネルでの IS divergence の計算において行列式が 0 になるのを防ぐために X_{ij} の対角要素に 10^{-10} を足している。プログラムは Sawada らのアルゴリズム [2] を MATLAB で実装した。ランダムに生成した 10 個の初期値パターンを用意し、各提案手法によって作成した行列と置き換えて、音源分離を実行する。分離性能の評価基準は次式の Signal-to-Distortion Ratio (SDR)[3] を用いている。

$$\mathrm{SDR} = 10 \log_{10} \frac{\sum_t s^{\mathrm{img}}(t)^2}{\sum_t y^{\mathrm{spat}}(t)^2 + y^{\mathrm{int}}(t)^2 + y^{\mathrm{artif}}(t)^2} \quad (14)$$

ただし、s^{est} は目的音源の推測信号、s^{img} は目的音源の正解信号、y^{spat} は空間（フィルタリング）歪み、y^{int} は目的音源以外の音源の信号、y^{artif} は分離処理による信号の歪みを表す。

3.2　従来法の課題

マルチチャネル NMF は自由度の高いモデルであるため、局所最適解が増え、初期値依存による分離性能のばらつきが問題となる。Fig. 3 はマルチチャネル NMF アルゴリズムにランダムな初期値を 10 回与えて音源分離を行った際の分離性能を示している [4]。この図から、分離性能は初期値ごとに大きく異なっていることがわかる。

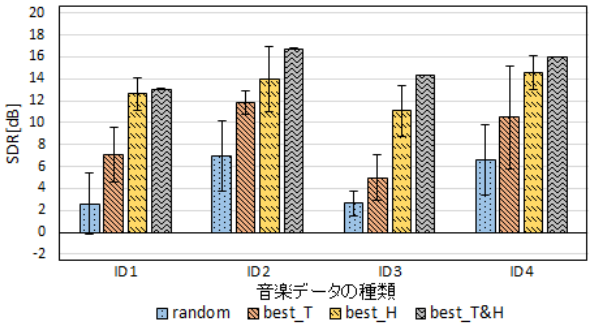

Fig. 4　初期値依存性の実験結果

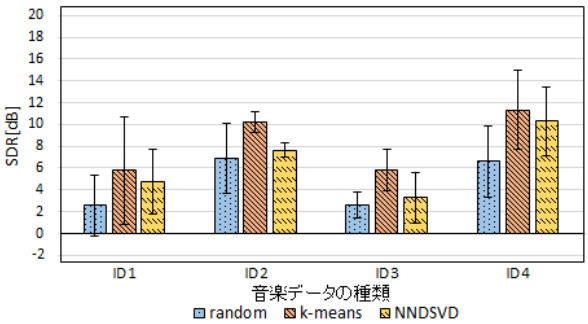

Fig. 5　基底行列に対する実験結果

3.3　最良の T と H を用いた場合

基底行列および空間相関行列に対する初期値依存性が大きいかどうか解析を行う。

ランダムな初期値パターンを 10 個作成し、各パターンで分離を行う (結果を Fig. 4 において “random” として示している)。分離結果が良かったパターンの更新後の各行列は、正しく音源分離が出来ている理想的な値であると仮定する。各パターンの他の初期値は変えずに、更新後の最良の行列を初期値として設定することで実験を行う。ここでは以下のパターンで比較する。ただし、最良の基底行列を設定する場合は基底行列の更新を行わない。

1. 最良の基底行列 (Fig. 4 で “best_T” と示す)
2. 最良の空間相関行列 (“best_H” と示す)
3. best_T と best_H を使用 (“best_T&H” と示す)

Fig. 4 は各音楽データでの分離後における 3 音源の平均 SDR を示したものである。エラーバーは標準偏差を示す。Fig. 4 から、最良の行列を初期値とすることで分離性能が向上することが分かる。また、基底行列と空間相関行列の両方を最良の値にすることで、分離結果のばらつきが大幅に減少している。また、基底行列を固定せずに音源分離する場合も行ったが、固定した場合より性能が下がった。

3.4　基底行列を音源から作成した場合

基底行列はスペクトルパターンを表すので、事前にどのような音源が含まれるのかを知ることが出来れば、初期値として設定することが可能である。本研究では混合前の音源から k-means 法で求める手法と NNDSVD で求める手法で初期値を設定する。

3.4.1　混合前の音源から k-means 法 [8] で作成

k-means 法とはデータ行列 X を任意数のクラスタに分割し、クラスタごとの平均を算出するアルゴリズムである。スペクトログラムに適用する場合、任意の基底数だけスペクトルパターンをクラスタリングし、各クラスタの平均の値を基底として使用することができる。この手法では混合前の 3 音源からそれぞれ 10 個ずつ基底を作成 (計 30 個) し、基底行列の初期値として設定する (Fig. 5 で “k-means” と示す)。

3.4.2　混合前の音源から NNDSVD 法 [9] で作成

NNDSVD (Non-negative Double Singular Value Decomposition) 法と呼ばれる特異値分解を用いた初期化手法を使用して基底を作成する。

特異値分解を使うことで、m 行 n 列の行列 $\mathbf{X}$ を

$$\mathbf{X} = U\Sigma W' \tag{15}$$

のように m 行 k 列の行列 U、k 行 k 列の行列 Σ、k 行 n 行の行列 W' の内積で表すことが出来る。スペクトログラムを特異値分解する場合、基底行列は

$$\mathbf{T} = U\sqrt{\Sigma} \tag{16}$$

となる。しかし行列 $\mathbf{T}$ は負の値を含むので、このまま初期値として扱うことは出来ない。そこで特異値ベクトルの負の成分を正の値に変えて、最後にゼロ成分を分離前の行列の平均値に置き換える NNDSVD 法を適用することで非負の行列となり、初期値として設定することが出来る。k-means 法と同様に、混合前の 3 音源からそれぞれ 10 個ずつの基底を作成し、基底行列の初期値として設定する (Fig. 5 で “NNDSVD” と示す)。

3.4.3　実験結果

Fig. 5 は各音楽データでの分離後における 3 音源の平均 SDR を示したものである。k-means 法もしくは NNDSVD 法を用いて初期値を設定することで、初期値をランダムとして分離を行うよりも SDR が全体的に向上しているが、標準偏差については大きな差は見られなかった。

3.5　空間相関行列をインパルス応答から作成した場合

空間相関行列はマイクロホン間の空間相関を示す行列であり、事前に各音源の位相情報を知ることが出来れば初期値として設定することが可能である。ここでは各音源のインパルス応答から、直接音モデルによる手法と直接音+反射音モデルによる手法で初期値を設定する。

3.5.1　直接音モデルによる作成

音源のインパルス応答をフーリエ変換することで

$$A_i = \begin{bmatrix} a_{i,1} & \cdots & a_{i,M} \end{bmatrix}^\top \tag{17}$$

M 行 1 列のステアリングベクトル A_i が与えられる。A_i と、そのエルミート転置 (1 行 M 列) の積

$$\mathsf{H}_i = A_i A_i^H = \begin{bmatrix} |a_{1,1}|^2 & \cdots & a_{i,1}a_{i,M}^* \\ \vdots & \ddots & \vdots \\ a_{i,M}a_{i,1}^* & \cdots & |a_{i,M}|^2 \end{bmatrix} \tag{18}$$

は周波数ビン i における空間相関を表す。L 個の各音源から H_i を作成することで、マルチチャネル NMF における I 行 L 列の空間相関行列として設定出来る[10]。

本実験では、Fig. 2 の環境で測定したインパルス応答から作成する。Fig. 6 は 240° 方向の音源から左チャネルまでのインパルス応答を示す。ここでは到来方向を示す直接音のみのインパルス応答から作成し、空間相関行列の初期値を設定する (Fig. 7 で “direct” と示す)。

3.5.2 反射音を含めたインパルス応答から作成

直接音モデルでは直接音のみのインパルス応答を使用したが、ここでは反射音を含めて初期値を設定することで、反射音の情報が必要かどうか解析を行う (Fig. 7 で “reflect” と示す)。

3.5.3 実験結果

Fig. 7 は各音楽データでの分離後における 3 音源の平均 SDR を示したものである。反射音を含めたインパルス応答 (reflect) から空間相関行列を設定することで、分離性能が全体的に向上し、ID3 を除く音楽データでは標準偏差が小さくなることが分かった。直接音のみのインパルス応答 (direct) から設定した場合は、ランダムと比べてあまり分離性能は変わらなかったが、標準偏差は小さくなっている。

3.6 考察

Fig. 4 に示されているように、基底行列と空間相関行列の初期値が良ければ、アクティベーション行列と潜在変数行列の値によらず、分離性能が高く、標準偏差が小さい。このことから、基底行列と空間相関行列に対する依存性が大きいということが考えられる。

Fig. 5 では、基底行列を推定することで分離性能は向上したが、分離結果のばらつきは改善しなかった。このことから、スペクトルパターンの推定は成功しているが、クラスタリングの失敗による音源とのミスマッチングで、分離結果のばらつきが発生してしまっていると考えられる。

Fig. 7 では、直接音のみのインパルス応答からの推定では分離性能はあまり変わらなかったが、反射音を含めたインパルス応答から作成することで、分離性能が向上した。このことから, 反射音を含めた情報が必要となるので、空間相関行列を到来方向だけから導くことは難しいと考えられる。

4 まとめと今後の課題

様々な方法でマルチチャネル NMF の初期値を設定することで、どのように分離結果が変化するか解析を行った。実験の結果から、基底行列と空間相関行列に対する初期値依存性が大きいということが確認できた。特に空間相関行列に対する初期値依存性が大きいので、空間相関を正しく推定できるような初期値設定法の提案が課題となる。

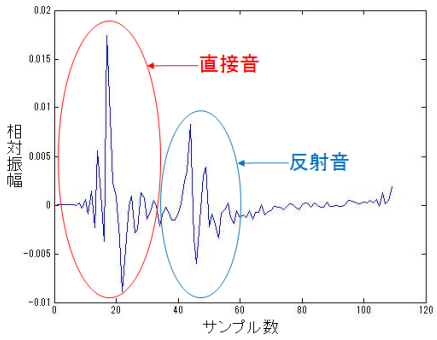

Fig. 6 240° 方向から左チャネルへのインパルス応答

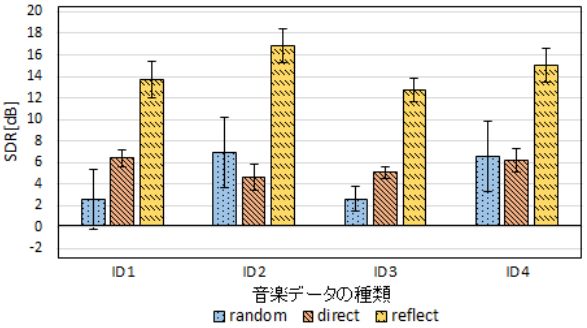

Fig. 7 空間相関行列に対する実験結果

参考文献

[1] D.D. Lee *et al.*, “Learning the parts of objects with nonnegative matrix factorization,” Nature, vol. 401, pp. 788-791,1999.

[2] H. Sawada *et al.*, “Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data,” IEEE Trans. ASLP, vol.21, no.5, pp. 971-982, 2013.

[3] E. Vincent *et al.*, “First stereo audio source separation evaluation campaigh: Data algprithm and results,” Independent Component Analysis and Signal Separation(Springer, Bearlin, 2007), pp. 552-559.

[4] 吉山 文教, 他: “マルチチャネル非負値行列因子分解における分離性能の高い初期値の判別法” 日本音響学会講演論文集, pp. 777-780, 2014.

[5] C. Fvotte, N. Bertin *et al.*, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” Neural Comput., vol. 21, no. 3, pp. 793-830, 2009.

[6] M. Nakano *et al.*, “Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence,” In Proc.MLSP 2010, pp. 283-288, 2010.

[7] S. Araki *et al.*, “The 2011 signal separation evaluation campaign (SiSEC2011): -audio source separation,” Latent Variable Analysis and Signal Separation(Springer, Bearlin, 2012), pp. 414-422.

[8] D. Arthur, S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding,” Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithm, pp. 1027-1035, 2007.

[9] C. Boutsidis, E. Gallopoulos, “SVD based initialization: A head start for nonnegative matrix factorization,” Pattern Recognition letters, vol. 41, pp. 1350-1362, 2008.

[10] 北村 大地, 他: “ランク 1 空間モデルを用いた効率的な多チャネル非負値行列因子分解” 日本音響学会講演論文集, pp. 579-582, 2014.