4Q-01

マイクロホンアレイ音源分離のための複素 t 分布に基づく多チャネル非負値行列因子分解

北村昂一[1]　坂東宜昭[2]　糸山克寿[2]　吉井和佳[2]

[1]京都大学 工学部 情報学科　[2]京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

混合音の音源分離は，実環境下で音声認識や音環境認識を行う上で必要不可欠な技術である．特に，音源やマイクロホンアレイに関する事前情報を用いないブラインド音源分離技術が盛んに研究されており，代表的なものに独立ベクトル解析 (IVA) [1] や Second-order Blind Identification (SOBI) [2] が挙げられる．しかし，このような周波数領域における音源分離では，各周波数ビンごとに得られた音源の並び順を揃えるパーミュテーション問題を解くことが困難であった．この問題を解決するため，大塚らは，各周波数ビンごとに音源を分離すると同時に，方向情報を用いてクラスタリングを行う潜在的ディリクレ配分法 (LDA) の拡張を提案している．一方，澤田らは，音源の方向情報だけではなく，音色情報を同時に利用する多チャネル非負値行列因子分解 (MNMF) [3] を提案している．具体的には，多チャネルの混合音スペクトログラムが与えられると，音源スペクトルの音色のテンプレート（基底スペクトル），その励起度合い（アクティベーション），さらに音源方向に対する空間相関行列とを一挙に推定可能である．

従来の MNMF では，各音源の基底スペクトルが複素ガウス分布に従うという仮定をおくことで，複素ガウス分布の加法性から，混合音スペクトログラムも複素ガウス分布に従うことになる．したがって，MNMF は，複素ガウス尤度を持つ確率モデルの最尤推定と解釈することができる．しかし，実際の音源は複素ガウス分布より裾が重いことが一般的であるため，初期値依存性が高く，局所解に陥りやすい傾向があった．

本研究では，尤度関数に複素 t 分布を用いることで，より広いクラスの MNMF である t-MNMF を提案する．単チャネル NMF では，複素ガウス尤度に基づく（板倉斉藤距離基準の）IS-NMF や複素コーシー尤度に基づく Cauchy NMF [4] を特殊形に含む複素 t 分布を尤度関数に持つ t-NMF が提案され，t 分布の自由度 ν を調節することで（$\nu=1$：コーシー分布, $\nu\to\infty$：ガウス分布），初期値依存性が軽減することが報告されている．MNMF では，多くのパラメータ数からなる空間相関行列の推定が必要であるため，初期値や外れ値に対する頑健性が期待できる確率モデルの定式化には有益である．

2. t-MNMF

本研究では，下記に定める音源分離問題を扱う．

入力	M チャネル同期マイクロホンアレイで観測される混合音スペクトル $\boldsymbol{x}_{ij}\in\mathbb{C}^M$
出力	L 個の音源スペクトル $\{\boldsymbol{y}_{ij}^{(1)},\cdots,\boldsymbol{y}_{ij}^{(L)}\}\in\mathbb{C}^M$

ここで，$1\le i\le I$, $1\le j\le J$ および $1\le l\le L$ はそれぞれ，周波数，時刻，音源を表す．また，$\boldsymbol{x}_{ij}$ は，各チャネルごとに観測信号を短時間フーリエ変換し，周波数 i，時刻 j における値を M 個並べたものである．

Multichannel Nonnegative Matrix Factorization based on the Complex t Distribution for Sound Source Separation: *K. Kitamura, Y. Bando, K. Itoyama, and K. Yoshii (Kyoto Univ.)*

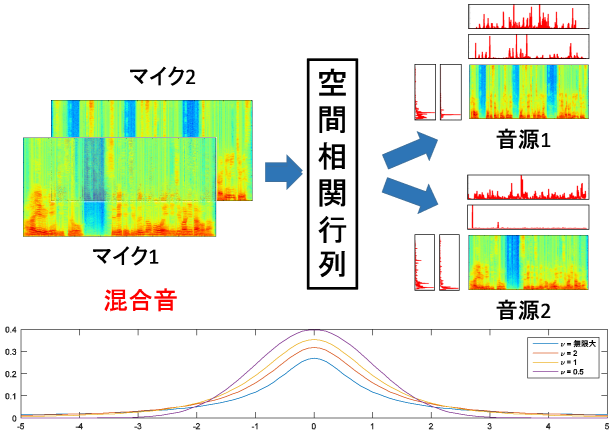

図 1: MNMF と t 分布の確率密度関数

2.1 多チャネル非負値行列因子分解

MNMF の目標は，混合音スペクトルの自己相関行列 $\mathbf{X}_{ij}=\boldsymbol{x}\boldsymbol{x}^H\in\mathbb{C}^{M\times M}\succeq\mathbf{0}$ を，音源ごとの空間相関行列の線形和に分解することである（詳細は [3] を参照）．

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

ここで，$1\le k\le K$ は基底，$\mathbf{H}_{il}\in\mathbb{C}^{M\times M}\succeq\mathbf{0}$ は周波数 i における音源 l の空間相関行列，$\{t_{1k},\cdots,t_{Ik}\}\in\mathbb{R}_+^I$ は k 番目の基底スペクトル，$\{v_{k1},\cdots,v_{kJ}\}\in\mathbb{R}_+^J$ は対応するアクティベーションを表す．$z_{lk}\in\mathbb{R}$ は各音源 l に対する基底 k の寄与を示す潜在変数である．MNMF は，$\mathbf{X}_{ij}$ と $\hat{\mathbf{X}}_{ij}$ との距離を最小化する問題であるが，本稿では統計的な見地から対数尤度を最大化する問題として議論する．MNMF 結果が得られれば，各音源信号は多チャネルウィナーフィルタを用いて得ることができる．

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

2.2 複素 t 分布に基づく確率モデル

本稿で提案する t-MNMF の対数尤度関数は，自由度 $\nu>0$ を持つ多変量複素 t 分布を用いて定義する．

$$\begin{aligned}&\log p(\mathbf{X}_{ij}|\hat{\mathbf{X}}_{ij})\\&\overset{c}{=}-\log|\hat{\mathbf{X}}_{ij}|+\tfrac{2M+\nu}{2}\log\left(1+\tfrac{2}{\nu}\mathrm{tr}(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij})\right)\end{aligned}$$

ここで，自由度 ν が $\nu\to\infty$ のとき多変量複素ガウス分布に，$\nu=1$ のとき多変量複素コーシー分布に帰着する．従来の複素ガウス尤度に基づく MNMF [3] では，対数尤度関数は多変量複素ガウス分布で与えられていた．

$$\log p(\mathbf{X}_{ij}|\hat{\mathbf{X}}_{ij})\overset{c}{=}-\log|\hat{\mathbf{X}}_{ij}|+\mathrm{tr}(\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1})$$

t-MNMF のパラメータを最適化するには，補助関数法に基づく乗法更新アルゴリズムが利用できる．

$$t_{ik}\leftarrow t_{ik}\sqrt{\frac{\sum_j v_{kj}\frac{2M+\nu}{\nu+2\mathrm{tr}(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij})}\sum_l z_{lk}\alpha_{ijl}}{\sum_j v_{kj}\sum_l z_{lk}\beta_{ijl}}}$$

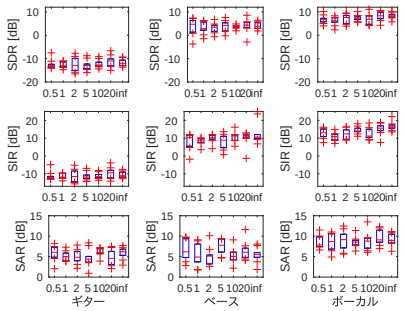

図 2: ID1 の音源分離精度

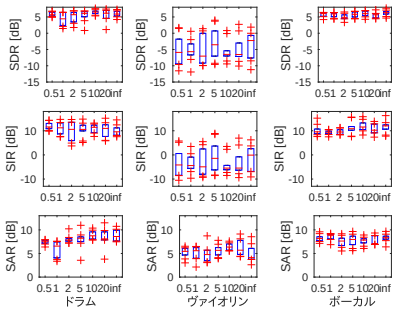

図 3: ID2 の音源分離精度

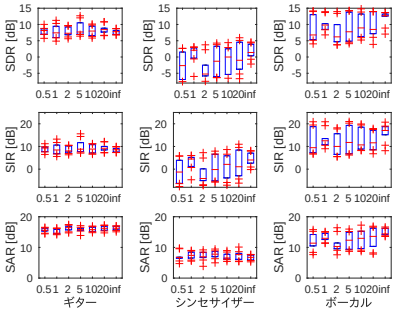

図 4: ID3 の音源分離精度

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\sum_j t_{ik}\frac{2M+\nu}{\nu+2\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\right)}\sum_l z_{lk}\alpha_{ijl}}{\sum_j t_{ik}\sum_l z_{lk}\beta_{ijl}}}$$

$$z_{lk} \leftarrow z_{lk}\sqrt{\frac{\sum_j t_{ik}v_{kj}\frac{2M+\nu}{\nu+2\mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\right)}\sum_l \alpha_{ijl}}{\sum_j t_{ik}v_{kj}\sum_l \beta_{ijl}}}$$

$\alpha_{ijl} = \mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\mathbf{H}_{il}\right)$, $\beta_{ijl} = \mathrm{tr}\left(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{H}_{il}\right)$ とした．また，$\mathbf{H}_{il}$ は次の方程式の解として与えられる．

$$\mathbf{H}_{il}\mathbf{A}\mathbf{H}_{il} = \mathbf{H}'_{il}\mathbf{B}\mathbf{H}'_{il}$$

ただし，$\mathbf{H}'_{il}$ は一つ前の更新時の相関行列とし，行列 $\mathbf{A}$ と $\mathbf{B}$ は以下で与えられる:

$$\mathbf{A} = \sum_{i,j} z_{lk}t_{ik}v_{kj}\hat{\mathbf{X}}_{ij}^{-1}$$

$$\mathbf{B} = \left(\sum_{j,k}\frac{2M+\nu}{\nu+2\mathrm{tr}(\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij})}z_{lk}t_{ik}v_{kj}\hat{\mathbf{X}}_{ij}^{-1}\mathbf{X}_{ij}\hat{\mathbf{X}}_{ij}^{-1}\right)$$

2.3 複素ガウス尤度と複素コーシー尤度の性質

従来の MNMF では，各音源スペクトルが複素ガウス分布に従うことを仮定しているため，複素ガウス分布の再生性から，パワースペクトルの加法性が保証される．単チャネルの NMF では，各音源スペクトルが複素コーシー分布に従うを仮定することにより，複素コーシー分布の再生性から，振幅スペクトルの加法性が保証される [4]．一方で，多チャネル NMF の場合には，「多変量」複素コーシー分布は再生性を持たないため，厳密には振幅スペクトルの加法性は保証されないが，単チャネル NMF と同様好ましい性質をもつことが期待される．

3. 評価実験

3.1 実験条件

音楽信号を用いた分離精度の評価実験を行った．音源数 $L = 3$，観測チャネル数 $M = 3$ のシミュレーション混合音を用いて評価した．観測信号は，RWCP [5] のインパルス応答 (E2A) を各音源に畳みこみ生成した．音源信号は表 1 に示した SiSEC [6] の 3 種の音楽データのそれぞれ 3 楽器を用いた．実験条件はサンプリングレートが 16kHz，窓長が 512 サンプル，シフト長が 160 サンプル，基底数 60，反復回数 520 である．H は対角要素を 1/M で非対角要素を 0 で初期化した．Z，T，V は非負値でランダムに初期化した．更新方法はまず T，V のみを 20 回更新し，その後 T，V，Z，H を 500 回更新した．音源分離精度は，BSS Eval Toolbox [7] を用いて，source-to-distortion ratio (SDR)，source-to-interferences ratio (SIR)，および source-to-artifacts ratio (SAR) で評価をした．

表 1: 音源信号

ID	SiSEC データベースの楽曲条件名	各音源の楽器
1	bearlin_roads_snip_85_99	ギター/ベース/ボーカル
2	for_minor_remember_the_name_snip_54_78	ドラム/ヴァイオリン/ボーカル
3	ultimate_nz_tour_snip_43_61	ギター/シンセサイザー/ボーカル

3.2 実験結果

図 2-4 はそれぞれの楽曲に対し自由度 $\nu = 0.5, 1, 2, 5, 10, 20, \infty$ に変えて 10 回試行した際の SDR, SIR，SAR である．またそれぞれの試行において変数の初期化はランダムに行っている．

4. おわりに

本稿では，多チャネル信号に対して複素 t 分布に基づく多チャネル非負値行列因子分解を用いて分離を行い，その精度を比較した．空間相関行列のランクを 1 に制約した MNMF が提案されており，音源分離精度の向上が確認されている [8]．将来的には，空間相関行列にランク 1 制約を取り入れた複素 t 分布に基づく MNMF を導出し，音源分離精度の比較を行う．

謝辞 本研究の一部は，JSPS 科研費 24220006, 15K12063 の支援を受けた．

参考文献

[1] Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 189–192. IEEE, 2011.

[2] Belouchrani, et al. A blind source separation technique using second-order statistics. *IEEE Transactions on,Signal Processing*, Vol. 45, No. 2, pp. 434–444, 1997.

[3] Sawada, et al. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on, Audio, Speech, and Language Processing*, Vol. 21, No. 5, pp. 971–982, 2013.

[4] Liutkus, et al. Cauchy nonnegative matrix factorization. In *(WASPAA), 2015 IEEE Workshop on*, pp. 1–5.

[5] Nakamura, et al. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *LREC*, 2000.

[6] Araki, et al. The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation. In *Latent Variable Analysis and Signal Separation*, pp. 414–422. 2012.

[7] Vincent, et al. Performance measurement in blind audio source separation. *IEEE Transactions on, Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469, 2006.

[8] Kitamura, et al. Relaxation of rank-1 spatial constraint in overdetermined blind source separation. In *EUSIPCO 2015*, pp. 1271–1275.