Vol.2018-MUS-120 No.18
2018/8/23

モノラル音響信号に対する音源分離のための独立低ランクテンソル分析

吉井 和佳[1,2,a)]　北村 昂一[1,b)]　坂東 宜昭[1,3,c)]　中村 栄太[1,d)]　河原 達也[1,e)]

概要：本稿では，独立性と低ランク性に基づく汎用的なテンソル分解技法である独立低ランクテンソル分析 (ILRTA) を提案し，単一チャネル音源分離への応用について述べる．一般に，音響信号の音源分離は，短時間フーリエ変換 (STFT) で得られる時間周波数スペクトログラムを対象として，すべての時間周波数ビンが独立であるという仮定の下で行うことが一般的であった．単一チャネル音源分離においては，音源スペクトログラムの低ランク性に基づく非負値行列分解 (NMF) が代表的である．一方，複数チャネル音源分離においては，音源スペクトログラムの独立性に着目した独立成分分析 (ICA) やその多変量拡張である独立ベクトル分析 (IVA) が利用でき，最近では，NMF と IVA を統合した独立低ランク行列分析 (ILRMA) が提案されている．ILRMA および ILRTA はともに，低ランクな音源スペクトログラムを推定する点で共通しているが，ILRMA は複数チャネル信号に対して，チャネル間を無相関化する線形分離フィルタを推定するのに対し，ILRTA は単一チャネル信号に対して，時刻間および周波数間を無相関化する線形変換を推定する点で異なる．我々は以前，NMF を拡張し，すべてのビン間の相関を考慮できる相関テンソル分解 (CTF) を提案した．ILRMA が複数チャネル NMF (MNMF) の特殊形であるのと同様，ILRTA は CTF の特殊形であり，CTF の莫大な計算量を削減することができる．また，ILRTA は，任意の階数のテンソルデータの各軸を同時無相関化する世界初の枠組みであり，ILRMA を特殊形に含む，時間軸・周波数軸・チャネル軸の同時無相関化に基づく複数チャネル音源分離への展開も可能になる．

1. はじめに

音響信号の音源分離は，音響イベント検出 [1]，実環境下での音声認識 [2]，音楽の自動採譜 [3] などにおける基礎技術となっている．これまで，単一・複数チャネル音源分離は，短時間フーリエ変換 (STFT) 領域で行われることが一般的であった．単一チャネル音源分離は，原理的に不良設定問題であり，解の曖昧性を解消するには，音源スペクトログラムが満たすべき性質を仮定する必要がある．一方，マルチチャネル音源分離においては，音源数とマイク数が同じである決定条件であれば，音源に関する事前知識を用いなくても (Blind Source Separation, BSS)，音源スペクトログラムの空間的な性質に着目することで，良い分離結果が得られることが知られている．

[1] 京都大学 大学院情報学研究科 知能情報学専攻
Yoshida-honmachi, Sakyo, Kyoto, Kyoto 606-8501, Japan
[2] 理化学研究所 革新知能統合研究センター (AIP)
15F, 1-4-1 Nihonbashi, Chuo, Tokyo 103-0027, Japan
[3] 産業技術総合研究所 (AIST) 知能システム研究部門
Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
[a)] yoshii@kuis.kyoto-u.ac.jp
[b)] kitamura@sap.ist.i.kyoto-u.ac.jp
[c)] y.bando@aist.go.jp
[d)] enakamura@sap.ist.i.kyoto-u.ac.jp
[e)] kawahara@i.kyoto-u.ac.jp

単一チャネル音響信号に対する音源分離を行うには，非負値行列分解 (Nonnegative Matrix Factorization, NMF) [4] がしばしば利用される．NMF は，入力となる非負値行列（パワースペクトログラム）を二つの非負値行列（基底スペクトルの集合と対応する音量ベクトルとの集合）の積で近似する．NMF には多くの変種が存在するが，混合音の複素スペクトログラム中の時間周波数ビンはすべて独立であり（現実には成立しない），それぞれが異なる複素ガウス分布に従うという仮定のもとでは，混合音のパワースペクトログラムに対して Itakura-Saito (IS) ダイバージェンスに基づく NMF (IS-NMF) [5] を適用することが理論的に妥当である．IS-NMF の結果に基づくウィナーフィルタを用いると，時間周波数ビンごとに独立に，混合音の複素成分を音源信号の複素成分の和に分解することができる．このとき，混合音と音源信号の複素スペクトログラムの位相は同一にならざるを得ず，復元される時間領域の音源信号の品質には限界があった．時間信号と対応する位相を復元する方法 [6, 7] も提案されているが，必ずしも分離音の品質が向上するわけではなかった．

NMF において位相の不整合が起こる本質的な原因は，全ての時間周波数ビンが独立であると仮定することにある．理論上は，無限の長さを持つ定常信号をフーリエ変換すれ

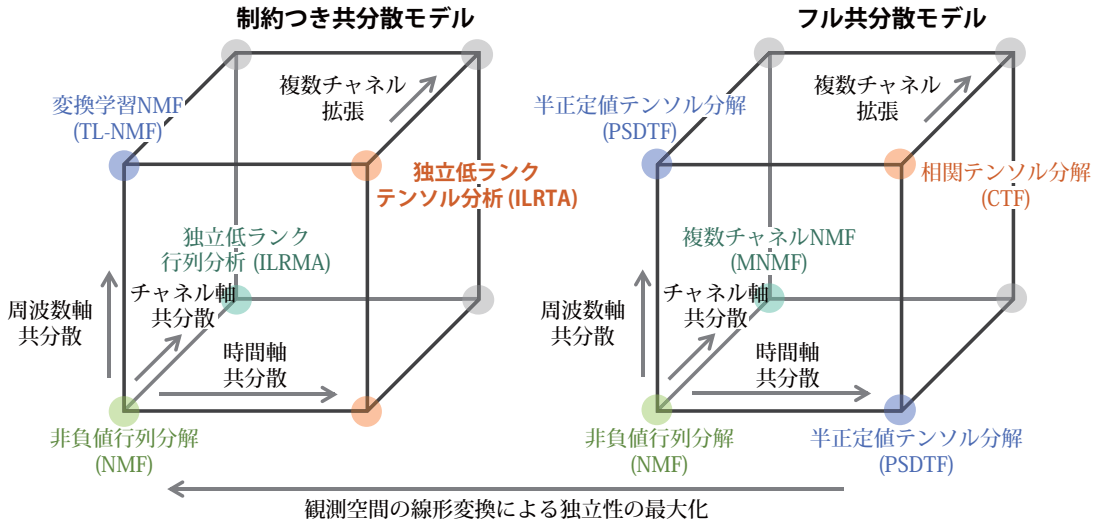

図 1 共分散モデリングに基づく単一・複数チャネル音源分離手法

ば，周波数軸は独立になる．しかし，有限の長さを持つ非定常信号に対して STFT を適用した場合，時間周波数ビン間には相関が発生することが避けられない．

我々は最近，全ての時間周波数ビン間の共分散を取り扱うことができる相関テンソル分解 (Correlated Tensor Factorizaion, CTF) [8] を提案し，単一チャネル音源分離に適用した．CTF では，与えられた半正定値行列（混合音スペクトログラムのすべての時間周波数ビン上の巨大な共分散行列）を，少数の半正定値行列（周波数軸上の共分散行列）群と対応する別の半正定値行列（時間軸上の共分散行列）群とのクロネッカー積の和で近似する．CTF の結果に基づくウィナーフィルタリングを行うと，時間周波数領域の音源スペクトログラムを一挙に推定することができる．CTF は，NMF の拡張となっており，半正定値テンソル分解 (Positive Semidefinite Tensor Factorization, PSDTF) [9, 10] や非負値テンソル分解 (Nonnegative Tensor Factorization, NTF) [11] をその特殊形として包含する．しかし，CTF は計算量が莫大で，現実的には実行が困難であった．具体的には，F 個の周波数ビンと T フレームからなる混合音のスペクトログラムを分解するには，NMF は $\mathcal{O}(KFT)$ であるが，CTF は $\mathcal{O}(KF^3T^3)$ であった．

音源の独立性に基づく複数チャネル音源分離においては，チャネル間の共分散構造が重要な役割を果たす．例えば，周波数領域における独立成分分析 (Independent Component Analysis, ICA) [12] では，チャネル間を無相関化することで，周波数ビン域ごとに音源成分を分離することができる．異なる周波数ビン間で同じ音源を対応付けるパーミュテーション問題を回避するため，音源スペクトルが多変量分布に従うとして，すべての周波数ビンを一挙に取り扱う独立ベクトル分析 (Independent Vector Analysis, IVA) [13, 14] が提案されている．さらに，音源スペクトログラムの低ランク性を導入することにより，IVA と NMF を統合した独立低ランク行列分析 (Independent Low-Rank Matrix Analysis, ILRMA) が提案されている．これら一連の手法は，音源スペクトログラムに関する仮定（優ガウス性や低ランク性）を満たしつつ，チャネル間を独立にする分離行列を推定する点で共通している．別の分離行列の推定方法として，異なる時刻における空間相関行列を同時対角化する手法も提案されている [15–17].

複数チャネル音源分離に着想を得て，本稿では，独立性と低ランク性に基づく単一チャネル音源分離のための独立低ランクテンソル分析 (Independent Low-Rank Tensor Analysis, ILRTA) を提案する．その核心部は，周波数軸と時間軸をそれぞれ無相関化する変換行列を推定することで，時間周波数領域での CTF を，変換後の空間での NMF として高速実行することにある．ILRMA が，分離行列の推定と分離スペクトログラムの NMF を反復するのに対し，ILRTA では，変換行列の推定と変換スペクトログラムの NMF を反復する．この結果，CTF の計算量は $\mathcal{O}(F^2T) + \mathcal{O}(FT^2) + \mathcal{O}(F^4) + \mathcal{O}(T^4) + \mathcal{O}(KFT)$ となる．

本研究の主な貢献は，時間・周波数・チャネル軸の共分散モデリングという観点から，従来の音源分離手法を統一的に記述する統一理論の構築にある（図 1）．ILRMA が，複数チャネル NMF [18] における空間相関行列（チャネル軸上の共分散行列）をランク 1 に制限したものであるのに対して，ILRTA は，CTF における周波数軸上の共分散行列および時間軸上の共分散行列をそれぞれ同時対角化できるよう制限したものである．また，ILRTA は，離散フーリエ変換に代わる最適な変換を NMF と同時に学習する変換学習 NMF (Transform-Learning NMF, TL-NMF) の拡張とみなせる．ILRTA は，任意の階数のテンソルデータの各軸を同時無相関化する世界初の枠組みであり，ILRMA を特殊形に含む，時間・周波数・チャネル軸の同時無相関化に基づく複数チャネル音源分離への展開も可能になる．

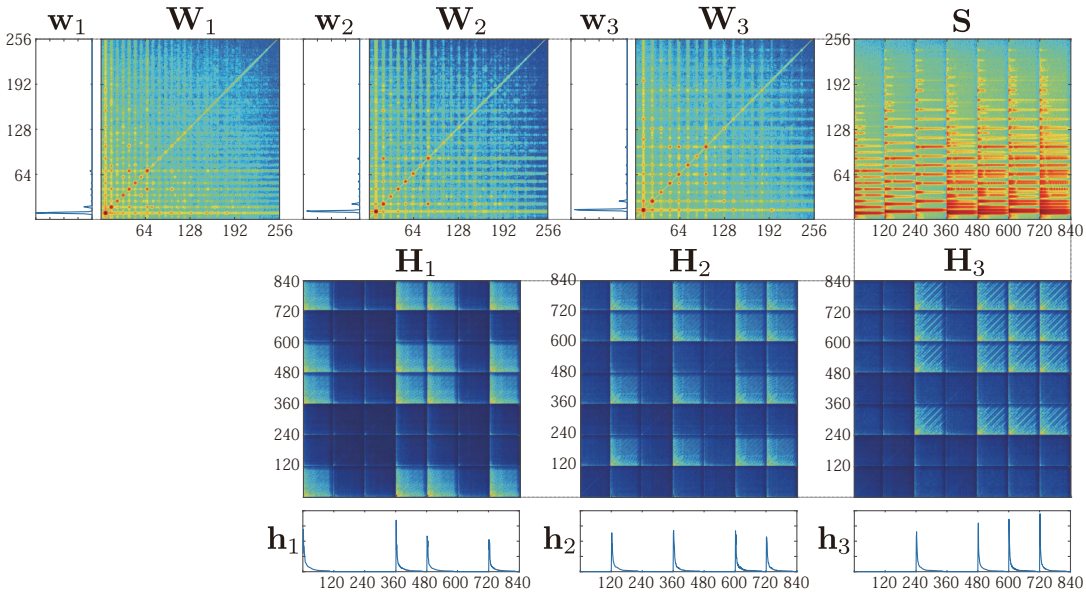

図 2 非負値行列分解，半正定値テンソル分解および相関テンソル分解の比較

2. 相関テンソル分解

本章では，本研究の基礎となる相関テンソル分解 (CTF) [8] と単一チャネル音源分離への応用について述べる．

2.1 定式化

$\mathbf{X} \in \mathbb{S}_+^{D_1 D_2 \cdots D_M}$ を半正定値行列とする．ここで，$\mathbb{S}_+^D$ は $D \times D$ の半正定値行列を表す [19]．$\mathbf{X}$ の各次元は，M 個の正整数 $\{D_m\}_{m=1}^M$ の積に分解できるとする．CTF では，与えられた $\mathbf{X}$ を，K 個の基底から構成される半正定値行列 $\mathbf{Y}$ で近似する．

$$\mathbf{X} \approx \mathbf{Y} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \bigotimes_{m=1}^{M} \mathbf{V}_{km} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \mathbf{V}_{k1} \otimes \cdots \otimes \mathbf{V}_{kM} \quad (1)$$

ここで，$\{\mathbf{V}_{km} \in \mathbb{S}_+^{D_m}\}_{m=1}^M$ は，基底 k に対応する半正定値行列の集合であり $\otimes$ はクロネッカー積を表す．

半正定値行列 $\mathbf{X}$ と $\mathbf{Y}$ の近似誤差を評価するには，Bregman 行列ダイバージェンス [20] が便利である．

$$\mathcal{D}_\phi(\mathbf{X}|\mathbf{Y}) = \phi(\mathbf{X}) - \phi(\mathbf{Y}) - \mathrm{tr}\big(\nabla\phi(\mathbf{Y})^{\mathrm{T}}(\mathbf{X} - \mathbf{Y})\big) \quad (2)$$

ここで，ϕ は，$\mathbb{S}_+^{D_1 D_2 \cdots D_M}$ 上の厳密に凸な関数である．音響信号の音源分離では，複素スペクトルが多変量ガウス分布に従うという仮定すれば，$\phi(\mathbf{Z}) = -\log|\mathbf{Z}|$ となる log-det (LD) ダイバージェンス [21] を用いるのが適切である．

$$\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y}) = -\log\left|\mathbf{X}\mathbf{Y}^{-1}\right| + \mathrm{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - D_1 \cdots D_M \quad (3)$$

式 (1) は $\mathbf{X}^{(m)} \approx \sum_{k=1}^K \mathbf{V}_{km} \otimes (\bigotimes_{m' \neq m} \mathbf{V}_{km'})$ と書き直せるため，$\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y})$ を最小化する $\mathbf{V}_{km}$ を推定するうえでは，$M = 2$ の場合のみを考えれば十分である．ここで，$\mathbf{X}^{(m)}$ は $\mathbf{X}$ の成分を入れ替えたものを表す．

2.2 位置づけ

半正定値行列 $\mathbf{X} \in \mathbb{S}_+^{FT}$ が与えられたときに，次式を満たす二つの半正定値行列群 $\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ および $\{\mathbf{H}_k \in \mathbb{S}_+^{\mathrm{T}}\}_{k=1}^K$ を推定することを考える．

$$\mathbf{X} \approx \mathbf{Y} \stackrel{\text{def}}{=} \sum_{k=1}^{K} \mathbf{W}_k \otimes \mathbf{H}_k \quad (4)$$

ここで，F および T は正整数である（例：周波数ビン数とフレーム数）．いま，$\mathbf{Y}_k = \mathbf{W}_k \otimes \mathbf{H}_k$ とすると，$\mathbf{Y} = \sum_k \mathbf{Y}_k$ が成立する．また，$[\mathbf{z}]$ を，対角成分にベクトル $\mathbf{z}$ をもつ対角行列を表すものとする．図 2 で示される通り，もしすべての半正定値行列が対角行列であれば，すなわち，$\mathbf{X} = [\mathbf{x}]$，$\mathbf{W}_k = [\mathbf{w}_k]$，$\mathbf{H}_k = [\mathbf{h}_k]$ が成立すれば，LD-CTF は IS-NMF [5] に帰着する．

$$x_{ft} \approx y_{ft} \stackrel{\text{def}}{=} \sum_{k=1}^{K} w_{kf} h_{kt} \quad (5)$$

ここで，x_{ft} および y_{ft} はそれぞれ，非負値ベクトル $\mathbf{x} \in \mathbb{R}_+^{FT}$ および $\mathbf{y} \in \mathbb{R}_+^{FT}$ の要素を表すとする．もし，$\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ および $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ のいずれかが対角行列である場合，LD-CTF は LD-PSDTF [9, 10] に帰着する．

$$\hat{\mathbf{X}}_f \approx \sum_{k=1}^{K} w_{kf} \mathbf{H}_k \quad \text{or} \quad \check{\mathbf{X}}_t \approx \sum_{k=1}^{K} \mathbf{W}_k h_{kt} \quad (6)$$

ここで，$\hat{\mathbf{X}}_f \in \mathbb{S}_+^T$ は，$\mathbf{X}$ から f に関連する行と列を抽出することで得られる半正定値行列であり，$\check{\mathbf{X}}_t \in \mathbb{S}_+^F$ も同様に定義されるものとする．LD-PSDTF は，ある特定の次元（例えば周波数軸あるいは時間軸）の共分散構造をとらえることができるのに対し，LD-CTF は全ての次元の共分散構造を同時にとらえることができる．したがって，LD-PSDTF および LD-CTF では，位相情報を用いた高品質な分離が達成できる．

2.3 パラメータ推定

与えられた $\mathbf{X}$ に対して，$\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ および $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ を推定するため，収束保証付きの反復アルゴリズムが提案されている [8]．まず，二つの半正定値行列 $\mathbf{A}$ および $\mathbf{B}$ の幾何平均 $\mathbf{A}\#\mathbf{B}$ は次式で定義される [22–24]．

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}^{\frac{1}{2}}\left(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}\right)^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}(\mathbf{A}^{-1}\mathbf{B})^{\frac{1}{2}} \quad (7)$$

このとき，$\mathbf{W}_k$ および $\mathbf{H}_k$ の更新式は次式で与えられる．

$$\mathbf{W}_k \leftarrow \mathbf{A}_k^{-1}\#(\mathbf{W}_k\mathbf{B}_k\mathbf{W}_k) \quad (8)$$

$$\mathbf{H}_k \leftarrow \mathbf{C}_k^{-1}\#(\mathbf{H}_k\mathbf{D}_k\mathbf{H}_k) \quad (9)$$

ここで，$\mathbf{A}_k \in \mathbb{S}_+^F$，$\mathbf{B}_k \in \mathbb{S}_+^F$，$\mathbf{C}_k \in \mathbb{S}_+^T$ および $\mathbf{D}_k \in \mathbb{S}_+^T$ は半正定値行列であり，次式で与えられる．

$$\mathbf{A}_k = (\mathbf{I}_F \otimes \mathbf{1}_T^{\mathrm{T}})\left((\mathbf{1}_F \otimes \mathbf{H}_k^{\mathrm{T}}) \odot \mathbf{Y}^{-1}\right)(\mathbf{I}_F \otimes \mathbf{1}_T)$$

$$\mathbf{B}_k = (\mathbf{I}_F \otimes \mathbf{1}_T^{\mathrm{T}})\left((\mathbf{1}_F \otimes \mathbf{H}_k^{\mathrm{T}}) \odot \mathbf{Y}^{-1}\mathbf{X}\mathbf{Y}^{-1}\right)(\mathbf{I}_F \otimes \mathbf{1}_T)$$

$$\mathbf{C}_k = (\mathbf{1}_F^{\mathrm{T}} \otimes \mathbf{I}_T)\left((\mathbf{W}_k^{\mathrm{T}} \otimes \mathbf{1}_T) \odot \mathbf{Y}^{-1}\right)(\mathbf{1}_F \otimes \mathbf{I}_T)$$

$$\mathbf{D}_k = (\mathbf{1}_F^{\mathrm{T}} \otimes \mathbf{I}_T)\left((\mathbf{W}_k^{\mathrm{T}} \otimes \mathbf{1}_T) \odot \mathbf{Y}^{-1}\mathbf{X}\mathbf{Y}^{-1}\right)(\mathbf{1}_F \otimes \mathbf{I}_T)$$

ここで，$\mathbf{I}_D$ および $\mathbf{1}_D$ はそれぞれ，サイズ D の単位行列および全要素が 1 の長さ D のベクトルを表し，$\odot$ は要素積を表す．このアルゴリズムの計算量は $\mathcal{O}(KF^3T^3)$ である．

2.4 単一チャネル音源分離

単一チャネル音源分離における LD-CTF の確率的な解釈を説明する．まず，F 周波数ビンと T フレームからなる混合音の複素スペクトログラム $\mathbf{S} \in \mathbb{C}^{F\times T}$ のすべての時間周波数ビンを行優先で直列化したベクトルを $\mathbf{s} \in \mathbb{C}^{FT}$ とする．共分散行列 $\mathbf{X} = \mathbf{s}\mathbf{s}^{\mathrm{H}}$ はランク 1 となる．同様に，音源 k の複素スペクトログラム $\mathbf{S}_k \in \mathbb{C}^{F\times T}$ を直列化したベクトルを $\mathbf{s}_k \in \mathbb{C}^{FT}$ とし，共分散行列 $\mathbf{Y}_k \in \mathbb{S}_+^{FT}$ をもつ多変量複素ガウス分布に従うと仮定する．

$$\mathbf{s}_k \mid \mathbf{Y}_k \sim \mathcal{N}_c(\mathbf{s}_k \mid \mathbf{0}, \mathbf{Y}_k) \quad (10)$$

ここで，$\mathbf{Y}_k$ の制約はないので，IS-NMF や LD-PSDTF と異なり，すべての時間周波数ビン間の完全な共分散構造が考慮できることに注意する，ガウス分布には再生性があることから，複素スペクトルの加法性 $\mathbf{s} = \sum_k \mathbf{s}_k$ を仮定すれば，次式が成り立つ．

$$\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_c(\mathbf{s} \mid \mathbf{0}, \mathbf{Y}) \quad (11)$$

したがって，観測データ $\mathbf{s}$ に対する対数尤度が計算できる．

$$\begin{aligned}\log p(\mathbf{s}|\mathbf{Y}) &\overset{c}{=} -\log|\mathbf{Y}| - \mathrm{tr}(\mathbf{X}\mathbf{Y}^{-1}) \\ &\overset{c}{=} -\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y})\end{aligned} \quad (12)$$

したがって，LD-CTF は、式 (11) で与えられる確率モデルの最尤推定と等価である．

$\mathbf{W}_k$ および $\mathbf{H}_k$ が求まれば，ウィナーフィルタを用いて，観測変数 $\mathbf{s}$ から潜在変数 $\mathbf{s}_k$ を事後推論できる．

$$\begin{aligned}&p(\mathbf{s}_k \mid \mathbf{s}, \mathbf{W}, \mathbf{H}) \\ &= \mathcal{N}_c\big(\mathbf{s}_k \mid \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}, \mathbf{Y} - \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{Y}_k\big)\end{aligned} \quad (13)$$

音源 k の時間信号は，位相復元手法に頼らずに，$\mathbb{E}[\mathbf{s}_k] = \mathbf{Y}_k\mathbf{Y}^{-1}\mathbf{s}$ に逆 STFT を直接適用すれば求められる．

3. 独立低ランクテンソル分析

提案する独立低ランクテンソル分析 (ILRTA) は，LD-CTF に対して，周波数軸上の K 個のフル共分散行列および時間軸上の K 個のフル共分散行列が，それぞれ同時対角化できる場合の特殊形である．このとき，同時対角化に用いる行列は，周波数領域あるいは時間領域を別の領域へ線形変換する行列となっている．変換後の領域では共分散行列が対角行列となる，すなわち，その領域を構成するビン間は無相関となる．ここで，ガウス分布を仮定していることから，無相関は独立であることと等価であることに注意されたい．その結果，時間周波数領域での LD-CTF は，二つの軸をともに線形変換した領域での IS-NMF と等価となり，大幅な計算量の削減が可能になる．

3.1 定式化

式 (4) で与えられる LD-CTF の定式化において，$\{\mathbf{W}_k \in \mathbb{S}_+^F\}_{k=1}^K$ および $\{\mathbf{H}_k \in \mathbb{S}_+^T\}_{k=1}^K$ が，それぞれ同時対角化可能であると仮定する．

$$\forall k\ \mathbf{W}_k = \mathbf{P}^{-1}[\tilde{\mathbf{w}}_k]\mathbf{P}^{-\mathrm{H}} \quad (14)$$

$$\forall k\ \mathbf{H}_k = \mathbf{Q}^{-1}[\tilde{\mathbf{h}}_k]\mathbf{Q}^{-\mathrm{H}} \quad (15)$$

ここで，$\tilde{\mathbf{w}}_k \in \mathbb{R}_+^F$ および $\tilde{\mathbf{h}}_k \in \mathbb{R}_+^T$ は非負値ベクトルであり，$\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_F]^{\mathrm{H}} \in \mathbb{C}^{F\times F}$ および $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_T]^{\mathrm{H}} \in \mathbb{C}^{T\times T}$ は正則行列である．ここでは， TL-NMF [25] のように，ユニタリ行列には限定しない．もし，$\mathbf{P}$ および $\mathbf{Q}$ が単位行列であれば，ILRTA は IS-NMF に帰着する．もし，$\mathbf{P}$ あるいは $\mathbf{Q}$ が単位行列であれば，ILRTA は同時対角化制約付き LD-PSDTF に帰着する（図 1）．一方，$\mathbf{X}$ を近似する行列 $\mathbf{Y}$ は，次式で与えられる．

$$\mathbf{Y} = \sum_{k=1}^K \mathbf{W}_k \otimes \mathbf{H}_k = \mathbf{R}^{-1}\left(\sum_{k=1}^K [\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]\right)\mathbf{R}^{-\mathrm{H}} \quad (16)$$

ここで，$\mathbf{R} = \mathbf{P} \otimes \mathbf{Q}$ とした．読みやすさのため，観測値 $\tilde{x}_{ft}$ および近似値 $\tilde{y}_{ft}$ を次式で定義しておく．

$$\begin{aligned}\tilde{x}_{ft} &= (\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{p}_f \otimes \mathbf{q}_t) \\ &= \mathbf{p}_f^{\mathrm{H}}(\mathbf{I}_F \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{I}_F \otimes \mathbf{q}_t)\mathbf{p}_f \\ &= \mathbf{q}_t^{\mathrm{H}}(\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{I}_T)\mathbf{X}(\mathbf{p}_f \otimes \mathbf{I}_T)\mathbf{q}_t\end{aligned} \quad (17)$$

$$\tilde{y}_{ft} = \sum_{k=1}^K \tilde{w}_{kf}\tilde{h}_{kt} \quad (18)$$

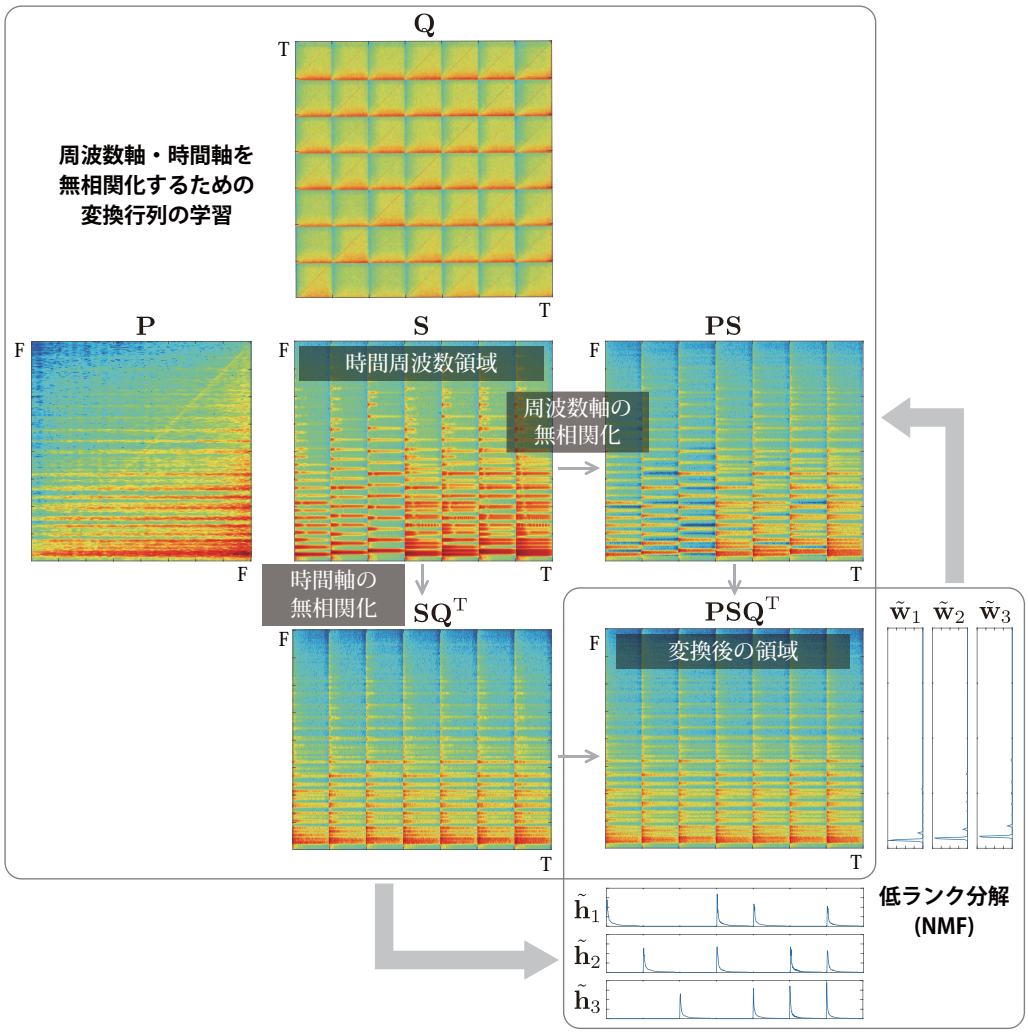

図 3 変換行列の学習と NMF の反復に基づく ILRTA のパラメータ推定

また，式 (12) で与えられるコスト関数に対して，式 (14) および式 (14) を代入すると次式が得られる．

$$\begin{aligned}\mathcal{D}_{\mathrm{LD}}(\mathbf{X}|\mathbf{Y}) &= -\log|\mathbf{X}\mathbf{Y}^{-1}| + \mathrm{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - FT \\ &\overset{c}{=} -T\log|\mathbf{P}\mathbf{P}^{\mathrm{H}}| - F\log|\mathbf{Q}\mathbf{Q}^{\mathrm{H}}| \\ &\quad + \sum_{f=1}^{F}\sum_{t=1}^{T}\log\tilde{y}_{ft} + \sum_{f=1}^{F}\sum_{t=1}^{T}\tilde{x}_{ft}\tilde{y}_{ft}^{-1} \quad (19)\end{aligned}$$

3.2 パラメータ推定

与えられた $\mathbf{X}$ に対して，基底ベクトル $\{\tilde{\mathbf{w}}_k\}_{k=1}^K, \{\tilde{\mathbf{h}}_k\}_{k=1}^K$，および変換行列 $\mathbf{P}, \mathbf{Q}$ を求める反復アルゴズムを導出する．図 3 に示す通り，変換行列の更新と，変換後の空間における NMF による基底ベクトルの更新を交互に行う．

3.2.1 基底ベクトルの更新

式 (19) を $\tilde{\mathbf{w}}_k$ および $\tilde{\mathbf{h}}_k$ に関して最小化するうえでは，第三項および第四項のみに着目すればよい．x と y の IS ダイバージェンスを $\mathcal{D}_{\mathrm{IS}}(x|y) = -\log x/y + x/y - 1$ とすると，この二項の和は，$\sum_{ft}\mathcal{D}_{\mathrm{IS}}(\tilde{x}_{ft}|\tilde{y}_{ft})$ と定数を除いて同じである．したがって，$\tilde{\mathbf{w}}_k$ および $\tilde{\mathbf{h}}_k$ に対しては，IS-NMF の乗法更新則 [26] がそのまま適用できる．

$$\tilde{w}_{kf} \leftarrow \tilde{a}_{kf}^{-1}\#(\tilde{w}_{kf}\tilde{b}_{kf}\tilde{w}_{kf}) = \tilde{a}_{kf}^{-\frac{1}{2}}\tilde{b}_{kf}^{\frac{1}{2}}\tilde{w}_{kf} \quad (20)$$

$$\tilde{h}_{kt} \leftarrow \tilde{c}_{kt}^{-1}\#(\tilde{h}_{kt}\tilde{d}_{kt}\tilde{h}_{kt}) = \tilde{c}_{kt}^{-\frac{1}{2}}\tilde{d}_{kt}^{\frac{1}{2}}\tilde{h}_{kt} \quad (21)$$

ここで，$\tilde{a}_{kf}$，$\tilde{b}_{kf}$，$\tilde{c}_{kt}$ および $\tilde{d}_{kt}$ は次式で与えられる．

$$\tilde{a}_{kf} = \sum_{t=1}^{T}\tilde{h}_{kt}\tilde{y}_{ft}^{-1} \quad \tilde{b}_{kf} = \sum_{t=1}^{T}\tilde{h}_{kt}\tilde{x}_{ft}\tilde{y}_{ft}^{-2} \quad (22)$$

$$\tilde{c}_{kt} = \sum_{f=1}^{F}\tilde{w}_{kf}\tilde{y}_{ft}^{-1} \quad \tilde{d}_{kt} = \sum_{f=1}^{F}\tilde{w}_{kf}\tilde{x}_{ft}\tilde{y}_{ft}^{-2} \quad (23)$$

ここで，LD-CTF は半正定値行列の幾何平均（式 (8) および式 (9)）を，IS-NMF は非負値の幾何平均（式 (20) および式 (21)）を計算するという興味深い対応関係がある．

3.2.2 変換行列の更新

式 (19) を $\mathbf{P}$ に関して最小化するうえでは，第一項および第四項のみに着目すればよい．式 (17) を用いると，この二項の和は，Majorization-Minimization (MM) 原理に基づく IVA [14] のコスト関数と同じ形式をしている．したがって，$\mathbf{P}$ を更新するには，反復射影法 (Iterative Projection, IP) と呼ばれる反復アルゴリズムを適用できる．

$$\text{方向の更新：}\ \mathbf{p}_f \leftarrow (\mathbf{P}\mathbf{U}_f)^{-1}\mathbf{e}_f \tag{24}$$

$$\text{ノルムの更新：}\ \mathbf{p}_f \leftarrow (\mathbf{p}_f^{\mathrm{H}}\mathbf{U}_f\mathbf{p}_f)^{-\frac{1}{2}}\mathbf{p}_f \tag{25}$$

ここで，$\mathbf{e}_f \in \mathbb{R}^F$ は f 番目の要素のみが 1 の単位ベクトルであり，$\mathbf{U}_f \in \mathbb{S}_+^F$ は次式で与えられる．

$$\mathbf{U}_f = \sum_{t=1}^{T}(\mathbf{I}_F \otimes \mathbf{q}_t^{\mathrm{H}})\mathbf{X}(\mathbf{I}_F \otimes \mathbf{q}_t)\tilde{y}_{ft}^{-1} \tag{26}$$

同様に，$\mathbf{Q}$ の更新則も導出できる．

$$\text{方向の更新：}\ \mathbf{q}_t \leftarrow (\mathbf{Q}\mathbf{V}_t)^{-1}\mathbf{e}_t \tag{27}$$

$$\text{ノルムの更新：}\ \mathbf{q}_t \leftarrow (\mathbf{q}_t^{\mathrm{H}}\mathbf{V}_t\mathbf{q}_t)^{-\frac{1}{2}}\mathbf{q}_t \tag{28}$$

ここで，$\mathbf{e}_t \in \mathbb{R}^{\mathrm{T}}$ は t 番目の要素のみが 1 の単位ベクトルであり，$\mathbf{V}_t \in \mathbb{S}_+^T$ は次式で与えられる．

$$\mathbf{V}_t = \sum_{f=1}^{F}(\mathbf{p}_f^{\mathrm{H}} \otimes \mathbf{I}_T)\mathbf{X}(\mathbf{p}_f \otimes \mathbf{I}_T)\tilde{y}_{ft}^{-1} \tag{29}$$

3.3 単一チャネル音源分離

単一チャネル音源分離における ILRTA の働きについて考察する．まず，式 (16) を式 (11) に代入すると，ILRTA の確率モデルが得られる．

$$\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_c\left(\mathbf{s} \;\middle|\; \mathbf{0}, \mathbf{R}^{-1}\left(\sum_{k=1}^{K}[\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]\right)\mathbf{R}^{-\mathrm{H}}\right) \tag{30}$$

このとき，$\mathbf{R} = \mathbf{P} \otimes \mathbf{Q}$ を変換行列とする $\mathbf{s} \in \mathbb{C}^{FT}$ の線形変換も多変量複素ガウス分布に従う．

$$\mathbf{R}\mathbf{s} \mid \mathbf{Y} \sim \mathcal{N}_c\left(\mathbf{R}\mathbf{s} \;\middle|\; \mathbf{0}, \sum_{k=1}^{K}[\tilde{\mathbf{w}}_k] \otimes [\tilde{\mathbf{h}}_k]\right) \tag{31}$$

ここで，$\mathbf{R}\mathbf{s} \in \mathbb{C}^{FT}$ は，空間変換後のスペクトログラム $\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}} \in \mathbb{C}^{F\times T}$ を直列化して得られる複素ベクトルである．式 (31) で，多変量複素ガウス分布の共分散行列が対角行列であることに着目すると，$\mathbf{S}$ に含まれる時間周波数ビンは相関を持っていても，$\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ に含まれるビンは無相関であることが分かる．また，$\mathbf{R}\mathbf{s}$，すなわち $\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ が与えられたもとで，式 (31) を最大化する $\tilde{\mathbf{w}}_k$ および $\tilde{\mathbf{h}}_k$ を求める問題は，$\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ に対する IS-NMF と等価である．IS-NMF はすべてのビンの独立性を仮定していることから，$\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ は，分解対象として好ましい性質を持っている．変換行列 $\mathbf{P}$ および $\mathbf{Q}$ は，$\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ の独立性および低ランク性ができる限り満たされるように学習される．

ILRTA では，混合音の複素スペクトログラム $\mathbf{S}$ の周波数軸および時間軸をそれぞれ無相関化する変換行列 $\mathbf{P}$ および $\mathbf{Q}$ の推定と，無相関化したスペクトログラム $\mathbf{P}\mathbf{S}\mathbf{Q}^{\mathrm{T}}$ に対する IS-NMF という互いに依存した二つのタスクを，収束するまで交互に反復する．これは，ILRMA における，混合音の複素スペクトログラムのチャネル軸を無相関化する分離行列の推定と，分離したスペクトログラムに対する IS-NMF という二つのタスクの反復と同型である．

3.4 残された課題

ILRTA を安定的に実行するには，いくつか技術的な課題が残されている．ILRTA は過剰パラメータモデルであるので，初期値依存性が極めて高く，ランダムな初期値では動作しない．ILRTA の自由度は $K(F+T)+F^2+T^2$ であり，LD-CTF の自由度である $K(F^2+T^2)$ のおよそ K 分の 1 ではあるが，依然として，解析対象となる複素スペクトログラム $\mathbf{S}$ の自由度 FT よりもはるかに大きい．そのため，まず，IS-NMF を実行して，基底ベクトル $\mathbf{w}_k$ およびアクティベーション $\mathbf{h}_k$ を求めたのち，$\tilde{\mathbf{w}}_k \leftarrow \mathbf{w}_k$，$\tilde{\mathbf{h}}_k \leftarrow \mathbf{h}_k$，$\mathbf{P} \leftarrow \mathbf{I}_F$，$\mathbf{Q} \leftarrow \mathbf{I}_T$ と初期化するとよい．これは，収束に要する反復回数の削減にも効果的である．

変換行列を求めるのに IP 法を利用する際にも問題がある（3.2.2 節）．音源分離において，$\mathbf{X} = \mathbf{s}\mathbf{s}^{\mathrm{H}}$ はランク 1 の行列であるため，式 (26) および式 (29) は効率的に計算することができる．

$$\mathbf{U}_f = \underbrace{(\mathbf{S}\mathbf{Q}^{\mathrm{T}})}_{F\times T}\underbrace{[[\tilde{y}_{f1},\cdots,\tilde{y}_{fT}]^{\mathrm{T}}]}_{T\times T}\underbrace{(\mathbf{S}\mathbf{Q}^{\mathrm{T}})^{\mathrm{H}}}_{T\times F} \tag{32}$$

$$\mathbf{V}_t = \underbrace{(\mathbf{P}\mathbf{S})^{\mathrm{H}}}_{T\times F}\underbrace{[[\tilde{y}_{1t},\cdots,\tilde{y}_{Ft}]^{\mathrm{T}}]}_{F\times F}\underbrace{(\mathbf{P}\mathbf{S})}_{F\times T} \tag{33}$$

ここで，一般的の条件である $F < T$ である場合を考えると，$\mathbf{V}_t$ は $T\times T$ の行列ではあるが，そのランクは F となり，逆行列が計算できない．この場合，主成分分析 (PCA) などの次元圧縮法が有効であると考えられる．また，LD-PSDTF で得られた $\{\mathbf{H}_k\}_{k=1}^K$ に対して近似的な同時対角化を行うことで $\mathbf{Q}$ を求める方法も考えられる．大規模な逆行列計算に伴う数値的な不安定さを解決するには，$\mathbf{P}$ の更新は数回にとどめておくことや，TL-NMF [25] と同様に，$\mathbf{P}$ をユニタリ行列に限定するなどが考えられる．

4. 評価

本章では，ILRTA の性能を評価するため，その特殊形である IS-NMF と LD-PSDTF の性能を比較した予備実験の結果について報告する．

4.1 実験条件

実験には，MIDI のピアノ音を用いた．三つの音高 (C4, E4, G4) をもつ 1.2 秒間の音響信号を準備し，それらを 7 つの異なる組み合わせで重畳したもの (C4, E4, G4, C4+E4, C4+G4, E4+G4, C4+E4+G4) を連結して 8.4 秒の音響信号を合成した．サンプリング周波数は 16[kHz] とした．窓幅 512 点ガウス窓を用いて，窓シフト長 160 点の STFT を行うことで，複素スペクトログラム $\mathbf{S} \in \mathbb{C}^{F\times T}$ を得た（$F = 256$，$T = 840$）．3.4 節で議論した通り，この条件では $\mathbf{Q}$ の更新に問題があるので，$\mathbf{Q} = \mathbf{I}_T$ として，周波数軸変換行列 $\mathbf{P}$ と基底ベクトル $\tilde{\mathbf{w}}_k$ および $\tilde{\mathbf{h}}_k$ を推定した．この手法は，周波数軸上の共分散行列が同時対角化可能であ

表 1 音源分離精度 [dB]

Method	SDR	SIR	SAR
IS-NMF	18.9	24.2	20.4
LD-PSDTF	22.8	28.5	24.2
ILRTA (高速近似 LD-PSDTF)	24.3	31.4	25.2

るように制限した LD-PSDTF と等価であり，LD-PSDTF の高速近似解法とみることができる．ILRTA および LD-PSDTF ともに，IS-NMF の結果を用いて初期化を行った．BSS Eval Toolbox [27] を用いて，Source-to-Distortion Ratio (SDR)，Source-to-Interferences Ratio (SIR) および Sources-to-Artifacts Ratio (SAR) で評価した．

4.2 実験結果

表 1 に実験結果を示す．すべての評価基準において，ILRTA は IS-NMF および LD-PSDTF より優れた性能を示した．興味深いことに，$\mathbf{Q} = \mathbf{I}_T$ に制限した ILRTA は，LD-PSDTF の近似であるにもかかわらず，LD-PSDTF より優れた性能を示した．このことは，LD-PSDTF や LD-CTF のような過剰パラメータモデルでは，適切な制約を導入することでパラメータ数を削減し，より良い局所解を見つけやすくすることが効果的であることを示唆している．実際に，ILRTA によって得られた同時対角化可能な共分散行列は，LD-PSDTF で得られた共分散行列と同様であることを確認している（図 2）

5. おわりに

本稿では，独立低ランクテンソル分析 (ILRTA) と呼ぶ新しい低ランク分解手法を提案し，単一チャネル音源分離への応用について述べた．ILRTA は，混合音の複素スペクトログラムの性質に合わせて，すべての時間周波数ビンを無相関化するべく，周波数軸の変換行列および時間軸の変換行列を推定すると同時に，変換されたスペクトログラムに対して低ランク分解を行う．小規模な実験では，相関テンソル分解 (LD-CTF) の特殊形である半正定値テンソル分解 (LD-PSDTF) と比較して，その高速近似解法となるよう設定した ILRTA は優れた性能を示した．本来 LD-CTF の高速近似解法である ILRTA の能力を引き出すには，初期値依存性の問題と，理論的・数値的の両面で，取り扱う行列が非正則になる問題に対処する必要がある．

ILRTA は複数の軸の共分散構造を同時にモデル化ができる汎用的な枠組みのため，ICA, IVA, ILRMA と同様に，チャネル間を無相関化する機構を取り込むことにより，複数チャネル音源分離への拡張を行う予定である（図 1）．また，音源分離以外にも，推薦システムなど，低ランク近似が有効なタスクへの応用も検討していきたい．

謝辞: 本研究の一部は，JSPS 科研費 No. 26700020, No. 16H01744，JSPS 特別研究員奨励費 No. 16J05486, および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511, 2015.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: Challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[4] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems (NIPS)*, pages 556–562, 2000.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[6] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[7] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In *Workshop on Statistical and Perceptual Audition (SAPA)*, pages 23–28, 2008.

[8] K. Yoshii. Correlated tensor factorization for audio source separation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 731–735, 2018.

[9] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto. Infinite positive semidefinite tensor factorization for source separation of mixture signals. In *International Conference on Machine Learning (ICML)*, pages 576–584, 2013.

[10] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto. Beyond NMF: Time-domain audio source separation without phase reconstruction. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 369–374, 2013.

[11] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2004.

[13] T. Kim, T. Eltoft, and T.-W. Lee. Independent vector analysis: An extension of ICA to multivariate components. In *International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 165–172, 2006.

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

[15] E. Weinstein, M. Feder, and A. V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Transactions on Speech and Audio Processing*, 1(4):405–413, 1993.

[16] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations.

Physical Review Letters, 72(23):3634–3636, 1994.

[17] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[18] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982, 2013.

[19] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[20] L. M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[21] B. Kulis, M. Sustik, and I. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research (JMLR)*, 10:341–376, 2009.

[22] T. Ando. Topics on operator inequalities. Technical report, Division of Applied Mathematics, Research Institute of Applied Electricity, Hokkaido University, Japan, 1974.

[23] T. Andoa, C.-K. Li, and R. Mathias. Geometric means. *Linear Algebra and its Applications*, 385(1):305–334, 2004.

[24] M. Congedo, B. Afsari, A. Barachant, and M. Moakher. Approximate joint diagonalization and geometric mean of symmetric positive definite matrices. *PLoS ONE*, 10(4):1–25, 2015.

[25] D. Fagot, H. Wendt, and C. Févotte. Nonnegative matrix factorization for transform learning. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2431–2435, 2018.

[26] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama. Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta divergence. In *International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 283–288, 2010.

[27] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.