音楽音響信号解析のためのスパース学習

Sparse Learning for Music Signal Analysis

吉井和佳　糸山克寿

Abstract

本稿では，統計的音響信号処理に興味を持つ研究者向けに，モノラル音響信号の音源分離のための非負値行列分解（NMF）について解説する．従来，音源信号のパワースペクトルの加法性が仮定できる場合には，板倉・斎藤ダイバージェンスをコスト関数に持つ NMF（IS-NMF）が適切であり，複素ガウス分布をゆう度関数に持つ確率モデルとしての解釈が可能であることが知られていた．最近，音源信号の振幅スペクトルの加法性を仮定できる場合には，複素コーシー分布をゆう度関数に持つ NMF（Cauchy NMF）が適切であることが発見されている．本稿ではこれらに加えて，両者を特殊形に含む複素スチューデント t 分布に基づく NMF を紹介する．

キーワード：音源分離，非負値行列分解，最ゆう推定

1. はじめに

音楽音響信号の音源分離は，種々のアプリケーションの基礎を成す重要な課題である．例えば，音楽内容に基づく類似楽曲検索においては，混合音をそのまま取り扱うのではなく，各楽器パートを個別に解析できれば，より適切な類似度計算が可能となる．また，ユーザが楽曲を鑑賞する際には，あたかも楽曲制作におけるミクシングエンジニアのように，混合音中の楽器パートの音量を個別に調節することができれば，音楽の楽しみが広がる．このように，ユーザが既存の楽曲をカスタマイズしながら鑑賞することを可能にする能動的音楽鑑賞システムが盛んに研究されている．

現在，市販の楽曲は，ステレオや 5.1 チャネルなどマルチチャネルで録音されていることが一般的である．しかし，ビームフォーミングや独立成分分析などのマルチチャネル音源分離技術は利用できない．楽曲制作においては，あらかじめ楽器パートやボーカルを個別にモノラル録音しておき，チャネル間の音量バランスを変えてミクシングすることで定位感を演出するのが一般的である．したがって，マイクロホンアレーを用いて直接収録されたマルチチャネル音響信号を対象とする場合と異なり，チャネル間の位相情報に着目した音源分離は適切ではない．

モノラルの音楽音響信号を楽器音に分離するという不良設定問題を解く上で，信号に内在するスパース性に着目するアプローチは有望である．一見複雑に思える音楽音響信号も，高々有限個の楽器音が重畳しているにすぎない．例えば，各ピアノ曲においては，出現する音高は調に依存して偏っていることから，ピアノ 88 鍵の使われやすさにスパース性が存在する．すなわち，時間・周波数平面上の混合音スペクトログラムは高次元データであるが，それは高々有限個の音高のスペクトルが様々な音量で重畳することで生成されている．本稿では，この考え方に基づく非負値行列分解（NMF : Nonnegative Matrix Factorization）[(1), (2)] と確率モデルとしての解釈について解説する．

2. 統計的音響信号処理

本章では，統計的音響信号処理の基本的なアプローチを整理しておく．

吉井和佳　正員　京都大学大学院情報学研究科知能情報学専攻
E-mail　yoshii@kuis.kyoto-u.ac.jp
糸山克寿　京都大学大学院情報学研究科知能情報学専攻
E-mail　itoyama@kuis.kyoto-u.ac.jp
Kazuyoshi YOSHII, Member and Katsutoshi ITOYAMA, Nonmember (Graduate School of Informatics, Kyoto University, Kyoto-shi, 606-8501 Japan).
電子情報通信学会誌　Vol.99 No.5 pp.456-460 2016 年 5 月
©電子情報通信学会 2016

2.1 確率モデルの定式化

統計的音響信号処理では，潜在信号が変化したり，重畳したりして観測信号が得られる生成過程を確率モデルとして定式化し，逆に，観測信号が与えられたときに，未知の潜在信号を推定する問題を考える．音源分離問題においては，潜在信号は各楽器音の音源信号であり，観測信号は混合音である．このとき，観測信号が生成される確率（ゆう度）を最大化するような潜在信号を求める問題（最ゆう推定）を解けばよい．

確率モデルを定式化する際には，物理的な現象を十分に反映させることが重要である．具体的には，まず，潜在信号を入力とし，観測信号を出力する何らかの「関数」を定義する．音源分離問題では，潜在信号の和を出力する関数を考える．次に，潜在信号が何らかの確率分布に従うことを仮定すれば，関数による変数変換の結果，観測信号が従う確率分布が導ける．

2.2 非負値行列分解の統計的な解釈

混合音のスペクトログラムが与えられると，各時刻の混合音スペクトルを少数の音源スペクトルのスパースな線形和として低ランク近似するのがNMFである．具体的には，混合音スペクトルと近似スペクトルとのかい離度を最小化する問題を解く．これは，ある確率モデルの最ゆう推定としての解釈が可能である．

もし，物理的にパワースペクトルの加法性が成立していれば，板倉・斎藤（IS）ダイバージェンス(用語)に基づくNMF（IS-NMF）が理論的に妥当である[2]．具体的には，音源スペクトルが複素ガウス分布に従うと仮定すると，それらが重畳して得られる観測スペクトルも複素ガウス分布に従う．このとき，観測スペクトルに対する複素ガウスゆう度の最大化が，ISダイバージェンスの最小化と等価となる．

実際には，振幅スペクトルの加法性が成立するという仮定の下で，カルバックライブラ（KL）ダイバージェンス(用語)に基づくNMF（KL-NMF）を用いて音源分離を行うことが一般的である[1]．具体的には，音源スペクトルがポアソン分布に従うと仮定すると，観測スペクトルもポアソン分布に従う．このとき，観測スペクトルに対するポアソンゆう度の最大化が，KLダイバージェンスの最小化と等価となる．しかし，振幅スペクトルは音粒子のヒストグラムであるという解釈に基づき，振幅値が整数値に限定されている点で物理的に不自然である．にもかかわらず，実験的にはIS-NMFよりも良い結果を与えることが多かった[3]．

最近，振幅スペクトルの加法性の下では，複素コーシー分布に基づくNMF（Cauchy NMF）が理論的に妥当であることが報告されている[4]．具体的には，音源スペクトルが複素コーシー分布に従うと仮定すると，観測スペクトルも複素コーシー分布に従う．複素コーシー分布は裾が重く（平均・分散共に無限大に発散），外れ値に頑健であるという性質を持っており，実在する音源信号をモデル化するのに都合が良い．

本稿では，IS-NMFやCauchy NMFを特殊形に含む複素スチューデントt分布に基づくNMF（t-NMF）[5]について解説する．複素t分布は自由度パラメータνを持ち，$\nu=1,\infty$のときそれぞれ複素コーシー分布及び複素ガウス分布に帰着する．これら以外では，スペクトルの加法性は成立しないが，Cauchy NMFとIS-NMFの中間的な性質を持つNMFを実現することは実用面と学術面で有意義である．

3. 非負値行列分解の確率モデル

本章では，NMFの確率モデルとモノラル音響信号の音源分離への応用について述べる．

3.1 コスト関数最小化としての定式化

NMFの目的は，入力として与えられる非負値行列$\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N]\in\mathbb{R}_+^{M\times N}$に対し，

$$\boldsymbol{X}\approx\boldsymbol{W}\boldsymbol{H}^T\overset{\text{def}}{=}\boldsymbol{Y} \tag{1}$$

となる二つの非負値行列$\boldsymbol{W}=[\boldsymbol{w}_1,\cdots,\boldsymbol{w}_K]\in\mathbb{R}_+^{M\times K}$と$\boldsymbol{H}=[\boldsymbol{h}_1,\cdots,\boldsymbol{h}_K]\in\mathbb{R}_+^{N\times K}$への低ランク分解を行うことである．ただし，$\boldsymbol{w}_k\in\mathbb{R}_+^M$及び$\boldsymbol{h}_k\in\mathbb{R}_+^N$はそれぞれ基底ベクトル及び対応する重みベクトルであり，基底数は$K\ll\min(M,N)$とする．ここで，再構成行列を$\boldsymbol{Y}=[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N]\in\mathbb{R}_+^{M\times N}$とすると，

$$\boldsymbol{x}_n\approx\sum_{k=1}^{K}h_{kn}\boldsymbol{w}_k\overset{\text{def}}{=}\boldsymbol{y}_n \tag{2}$$

と書ける．すなわち，各非負値ベクトル$\boldsymbol{x}_n$は基底ベクトル群$\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_K\}$の線形結合$\boldsymbol{y}_n$で近似される．

観測値x_{nm}とNMFによる近似値y_{nm}との間の誤差$\mathcal{D}(x_{nm}|y_{nm})$を評価する尺度には，次式で定義される（一般化）KLダイバージェンス[1]及びISダイバージェンス[2]がよく利用される．

■ 用　語　解　説

ISダイバージェンス・KLダイバージェンス　確率分布間のかい離度を測るために用いる尺度．通常の距離とは異なり，対称性を持たないことに注意．

再生性　同じ分布族に含まれる確率分布を持つ独立な確率変数に対して，その和の確率分布もまた同じ族に含まれる性質のこと．

$$\mathcal{D}_{\mathrm{KL}}(x_{nm}|y_{nm})=x_{nm}\log\frac{x_{nm}}{y_{nm}}-x_{nm}+y_{nm} \quad (3)$$

$$\mathcal{D}_{\mathrm{IS}}(x_{nm}|y_{nm})=\frac{x_{nm}}{y_{nm}}-\log\frac{x_{nm}}{y_{nm}}-1 \quad (4)$$

これらは常に非負値をとり，$x_{nm}=y_{nm}$ のときのみ 0 となる．また，通常の距離尺度と異なり，非対称性 $\mathcal{D}(x_{nm}|y_{nm})\neq\mathcal{D}(y_{nm}|x_{nm})$ を持つ．最終的に，再構成行列 $\boldsymbol{Y}$ に対するコスト関数は次式で与えられる．

$$\mathcal{D}(\boldsymbol{X}|\boldsymbol{Y})=\sum_{m=1}^{M}\sum_{n=1}^{N}\mathcal{D}(x_{nm}|y_{nm}) \quad (5)$$

式(5)を最小化する $\boldsymbol{W}$ 及び $\boldsymbol{H}$ を求めるには，乗法更新アルゴリズム[6]が利用できる．

3.2 音源分離への応用

音楽音響信号を対象とする場合，楽曲は複数トラックをミクシングして制作されることが多いため，混合音の生成過程として，遅延や残響がない瞬時混合過程を考えることは妥当である．このとき，混合音の複素スペクトログラム $\widehat{\boldsymbol{X}}=[\hat{\boldsymbol{x}}_1,\cdots,\hat{\boldsymbol{x}}_N]\in\mathbb{C}^{M\times N}$ は K 個の音源信号の複素スペクトログラムの和で表せる．

$$\widehat{\boldsymbol{X}}=\sum_{k=1}^{K}\widehat{\boldsymbol{X}}_k \quad (6)$$

ここで，M は周波数ビン数，N はフレーム数であり，$\widehat{\boldsymbol{X}}_k=[\hat{\boldsymbol{x}}_{k1},\cdots,\hat{\boldsymbol{x}}_{kN}]\in\mathbb{C}^{M\times N}$ は k 番目の音源信号の複素スペクトログラムである．

混合音の音源分離とは，観測変数 $\widehat{\boldsymbol{X}}$ が与えられたときに，潜在変数 $\widehat{\boldsymbol{X}}_k$ を求める逆問題である．そこで，各音源信号の「非負値」スペクトログラム $\boldsymbol{X}_k=[\boldsymbol{x}_{k1},\cdots,\boldsymbol{x}_{kN}]\in\mathbb{R}_+^{M\times N}$ は，ランク 1 の非負値行列 $\boldsymbol{Y}_k=[\boldsymbol{y}_{k1},\cdots,\boldsymbol{y}_{kN}]\in\mathbb{R}_+^{M\times N}$ で近似できると仮定する(図 1)．

$$\boldsymbol{X}_k\approx\boldsymbol{w}_k\boldsymbol{h}_k^T\overset{\mathrm{def}}{=}\boldsymbol{Y}_k \quad (7)$$

すなわち，非負値行列 $\boldsymbol{Y}_k$ における各フレーム n の非負値スペクトル $\boldsymbol{y}_{kn}$ は，基底スペクトル $\boldsymbol{w}_k\in\mathbb{R}_+^M$ を重み h_{kn} でスケーリングするだけで得られる相似形であると仮定している($\boldsymbol{y}_{kn}=h_{kn}\boldsymbol{w}_k$)．ここで，$\boldsymbol{x}_n=\sum_k\boldsymbol{x}_{kn}$ かつ $\boldsymbol{y}_n=\sum_k\boldsymbol{y}_{kn}$ としておく．

NMF を適用することで $\boldsymbol{W}$ 及び $\boldsymbol{H}$ が推定されたとすると，観測変数 $\hat{x}_{nm}$ が与えられたときの潜在変数 $\hat{x}_{knm}$ の推定値はウィーナーフィルタリングで得られる．

$$\mathbb{E}[\hat{x}_{knm}|\hat{x}_{nm}]=y_{knm}y_{nm}^{-1}\hat{x}_{nm} \quad (8)$$

ただし，音源スペクトル $\hat{x}_{knm}$ の位相は混合音スペクトル $\hat{x}_{nm}$ と同一となっている．最後に，逆フーリエ変換を用いれば，時間領域の音源信号を復元できる．

音源分離結果の評価は，人工的に合成した混合音を用いて，もとの音源信号が利用可能な環境下で行うことが一般的である．評価尺度としては，分離信号と音源信号との近さや，含まれる雑音レベル，音源信号間の干渉の度合いなどが挙げられる．

3.2.1 KL-NMF に基づく音源分離[1]

KL-NMF に基づく音源分離では，混合音の複素スペクトル $\hat{\boldsymbol{x}}_n$ ではなく，混合音の「振幅」スペクトル $\boldsymbol{x}_n$ の生成モデルを考える．まず，各音源信号の振幅スペクトル $\boldsymbol{x}_{kn}$ の各要素 x_{knm} が y_{knm} をパラメータに持つポアソン分布に従うことを仮定する．

$$x_{knm}\sim\mathrm{Poisson}(y_{knm}) \quad (9)$$

ここで，本来は非負の実数をとるはずの振幅値 x_{knm} は非負の整数値をとることを仮定した．更に，便宜的に振

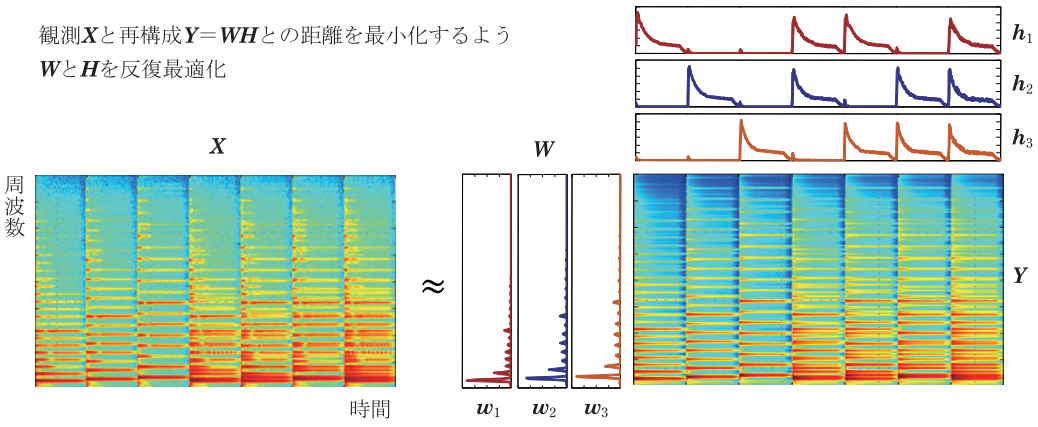

図 1　非負値スペクトログラムに対する NMF の適用結果　各音源信号の非負値スペクトログラム $\boldsymbol{X}_k$ は，ランク 1 のスペクトログラム $\boldsymbol{Y}_k=\boldsymbol{w}_k\boldsymbol{h}_k^T$ で近似される．

幅スペクトルの加法性を仮定する．このとき，ポアソン分布の再生性(用語)から，混合音の振幅スペクトル $\boldsymbol{x}_n$ の各要素 x_{nm} もポアソン分布に従う．

$$x_{nm} \sim \mathrm{Poisson}(y_{nm}) \tag{10}$$

ここで，$x_{nm}=\sum_k x_{knm}$ かつ $y_{nm}=\sum_k y_{knm}$ である．したがって，KL-NMF に対応する確率モデルの対数ゆう度は次式で与えられる．

$$\log p(x_{nm}|y_{nm}) \overset{c}{=} x_{nm}\log y_{nm} - y_{nm} \tag{11}$$

式(11)の符号を反転させると，式(3)と定数項（観測 x_{nm} を含む）を除いて等しい．したがって，式(11)の最大化（最ゆう推定）は式(3)の最小化と等価である．

3.2.2 IS-NMF に基づく音源分離(2)

IS-NMF に基づく音源分離では，混合音の複素スペクトル $\tilde{\boldsymbol{x}}_n$ の生成モデルを考える．まず，各音源信号の複素スペクトル $\tilde{\boldsymbol{x}}_{kn}$ の各要素 $\tilde{x}_{knm}$ が，y_{knm} を分散パラメータに持つ複素ガウス分布に従うことを仮定する．

$$\tilde{x}_{knm} \sim \mathcal{N}_c(0, y_{knm}) \tag{12}$$

式(6)で与えられる複素スペクトルの加法性の下では，複素ガウス分布の再生性から，混合音の複素スペクトル $\tilde{\boldsymbol{x}}_n$ の各要素 $\tilde{x}_{nm}$ も複素ガウス分布に従う．

$$\tilde{x}_{nm} \sim \mathcal{N}_c(0, y_{nm}) \tag{13}$$

したがって，パワースペクトルの加法性が成立している．いま，パワー値を $x_{nm}=\tilde{x}_{nm}\tilde{x}_{nm}^*$ とすると，IS-NMF に対応する確率モデルの対数ゆう度は次式で与えられる．

$$\log p(x_{nm}|y_{nm}) \overset{c}{=} -\log y_{nm} - \frac{x_{nm}}{y_{nm}} \tag{14}$$

式(14)の符号を反転させると，式(4)と定数項を除いて等しい．したがって，式(14)の最大化（最ゆう推定）は式(4)の最小化と等価である．

3.2.3 Cauchy NMF に基づく音源分離(4)

Cauchy NMF に基づく音源分離では，IS-NMF における複素ガウス分布を，やはり再生性を持つ複素コーシー分布に置き換えた生成モデルを考える．

$$\tilde{x}_{nm} \sim \mathcal{C}_c(0, y_{nm}) \tag{15}$$

このとき，振幅スペクトルの加法性が成立することが知られている(7)．したがって，混合音のパワー値 $x_{nm}=\tilde{x}_{nm}\tilde{x}_{nm}^*$ に対して，Cauchy NMF に対応する確率モデルの対数ゆう度は次式で与えられる．

$$\log p(x_{nm}|y_{nm}) \overset{c}{=} -\log y_{nm} - \frac{3}{2}\log\left(1+\frac{2x_{nm}}{y_{nm}}\right) \tag{16}$$

したがって，Cauchy NMF は，式(16)の最大化（最ゆう推定）として定式化でき，式(16)の符号を反転させたものをコスト関数とみなした NMF と等価である．

4. 非負値行列分解の新展開

本章では，最新の話題である IS-NMF 及び Cauchy NMF を特殊形に含む t-NMF(5)について解説する．t-NMF は，一般に IS-NMF より初期値依存性に強く，データに合わせて IS-NMF と Cauchy NMF との中間的な性質を持つ NMF を構成することができる．

4.1 確率モデルの定式化

t-NMF に基づく音源分離では，式(13)で与えられる IS-NMF のゆう度関数において，複素ガウス分布を複素 t 分布に置き換えた生成モデルを考える．

$$\tilde{x}_{nm} \sim \mathcal{T}_c^{\nu}(0, y_{nm}) \tag{17}$$

ここで，ν は自由度パラメータであり，$\nu=1$ とすると式(15)に，$\nu\to\infty$ とすると式(13)に帰着する．ただし，ν がこれらの値以外をとる場合には，確率分布の再生性，すなわち非負値（振幅あるいはパワー）スペクトルの加法性が成立しないことに注意する．混合音のパワー値 $x_{nm}=\tilde{x}_{nm}\tilde{x}_{nm}^*$ に対して，t-NMF に対応する確率モデルの対数ゆう度は次式で与えられる．

$$\log p(x_{nm}|y_{nm}) \overset{c}{=} -\log y_{nm} - \left(1+\frac{\nu}{2}\right)\log\left(1+\frac{2x_{nm}}{\nu y_{nm}}\right) \tag{18}$$

4.2 乗法更新アルゴリズム

式(18)を最大化する $\boldsymbol{W}$ 及び $\boldsymbol{H}$ を求めるため，補助関数法に基づく乗法更新アルゴリズム（表1）が提案されている(5)．ただし，スケールの任意性を解消するため，$\sum_m w_{km}=1$ を満たすよう $\boldsymbol{w}_k$ 及び $\boldsymbol{h}_k$ をスケーリングしておく．$\nu\to\infty$ のとき $z_{nm}\to x_{nm}$ であるので，従来知られていた IS-NMF の乗法更新アルゴリズムと一致する．$\nu=1$ のときは，Cauchy NMF の新たな乗法更新アルゴリズムを与える．

表1　t-NMF の最ゆう推定

Require：非負値行列 $\boldsymbol{X}\in\mathbb{R}_{+}^{M\times N}$，基底数 K
1：非負値行列 $\boldsymbol{W}\in\mathbb{R}_{+}^{M\times K}$ をランダムに初期化
2：非負値行列 $\boldsymbol{H}\in\mathbb{R}_{+}^{N\times K}$ をランダムに初期化
3：While not converged do
4：$z_{nm}=\left(\frac{\nu}{2+\nu}x_{nm}^{-1}+\frac{2}{2+\nu}y_{nm}^{-1}\right)^{-1}$
5：$w_{km}\leftarrow w_{km}\left(\frac{\sum_n z_{nm}h_{kn}/y_{nm}^2}{\sum_n h_{kn}/y_{nm}}\right)^{\frac{1}{2}}$
6：$h_{kn}\leftarrow h_{kn}\left(\frac{\sum_m z_{nm}w_{km}/y_{nm}^2}{\sum_m w_{km}/y_{nm}}\right)^{\frac{1}{2}}$
7：end while
8：return 非負値行列 $\boldsymbol{W}$, $\boldsymbol{H}$

t-NMF は，z_{nm} を仮想的な観測データとみなした IS-NMF に相当する．ただし，z_{nm} は観測データ x_{nm} と再構成データ y_{nm} に対する重み ν:2 の調和平均であり，自由度 ν が大きくなるほど x_{nm} に近づく．IS-NMF は $\nu\to\infty$ であるから，観測データ x_{nm} のみに着目する．一方，Cauchy NMF は $\nu=1$ であり，観測データ x_{nm} と再構成データ y_{nm} を 1:2 の重みで考慮することから，観測データに対して過学習しないことが分かる．

5. おわりに

NMF に基づく音楽音響信号の音源分離においては，音源スペクトルの加法性，すなわち確率分布の再生性が重要な性質である．これを念頭に，ある仮定の下では，複素ガウスゆう度に基づく IS-NMF と複素コーシーゆう度に基づく Cauchy NMF が理論的に妥当な確率モデルであることを示した．更に，これらを特殊形に含む複素 t ゆう度に基づく NMF を解説した．

現在，音楽音響信号を楽器音ではなく，音色に基づいて楽器パートに分離する技術の開発が進みつつある．更に，NMF に基づく音源分離技術は，音楽に限らず，マルチチャネル環境下において音声や環境音を分離するなど，産業上への応用が期待されている．

文　献

（1） P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics（WASPAA）, pp. 177-180, 2003.
（2） C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis,” Neural Comput., vol. 21, no. 3, pp. 793-830, 2009.
（3） B. King, C. Févotte, and P. Smaragdis, “Optimal cost function and magnitude power for NMF-based speech separation and music interpolation,” International Workshop on Machine Learning for Signal Processing（MLSP）, pp. 1-6, 2012.
（4） A. Liutkus, D. Fitzgerald, and R. Badeau, “Cauchy nonnegative matrix factorization,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics（WASPAA）, 2015.
（5） 吉井和佳，糸山克寿，後藤真孝，“音楽音響信号解析のためのステューデント t 分布に基づく非負値行列分解と半正定値テンソル分解，”信学技報，IBISML2015-70, pp. 131-138, Nov. 2015.
（6） 亀岡弘和，“非負値行列因子分解の音響信号処理への応用，”音響誌，vol. 68, no. 11, pp. 559-565, 2012.
（7） A. Liutkus and R. Badeau, “Generalized Wiener filtering with fractional power spectrograms,” International Conference on Acoustics, Speech, and Signal Processing（ICASSP）, pp. 266-270, 2015.

（平成 27 年 12 月 7 日受付　平成 28 年 1 月 23 日最終受付）

吉井　和佳（正員）
2008 京大大学院情報学研究科博士後期課程了．同年，産業技術総合研究所情報技術研究部門に入所．2014 京大大学院情報学研究科講師に着任．音楽情報処理，統計的音響信号処理の研究に従事．博士（情報学）．

糸山　克寿
2011 京大大学院情報学研究科博士後期課程了．同年，同大学院情報学研究科助教に着任．音楽情報処理，音楽鑑賞インタフェース等の研究に従事．博士（情報学）．