Mémoire d'habilitation à diriger des recherches (HdR)
École Doctorale en Sciences Fondamentales et Appliquées
Université Nice Sophia Antipolis

Contributions à la factorisation en matrices non-négatives

Cédric Févotte

Habilitation préparée au Laboratoire Lagrange (UMR CNRS, Observatoire de la Côte d'Azur & Université Nice Sophia Antipolis) et soutenue le 13 octobre 2014 à la Faculté des Sciences de Nice, devant le jury constitué de :

Patrick Flandrin, directeur de recherche CNRS à l'ENS Lyon	**président**
Pierre Comon, directeur de recherche CNRS à Gipsa-Lab, Grenoble	**rapporteur**
Jean-Christophe Pesquet, professeur à l'Université Paris-Est, Marne-la-Vallée	**rapporteur**
Arshia Cont, chercheur à l'IRCAM, Paris	**rapporteur**
Jean-François Cardoso, directeur de recherche CNRS au LTCI, Paris	**examinateur**
André Ferrari, professeur à l'Université Nice Sophia Antipolis	**examinateur**

Table des matières

Remerciements

Je remercie chaleureusement les membres de mon jury, Jean-François Cardoso, Pierre Comon, Arshia Cont, André Ferrari, Patrick Flandrin et Jean-Christophe Pesquet, que je tiens tous fort en estime, pour m'avoir fait l'honneur d'examiner mes travaux.

Pour le soutien qu'ils m'ont apporté jusqu'ici à divers moments de ma carrière, je souhaite très particulièrement remercier Raphaël Blouet, Olivier Cappé, Nicolas Dobigeon, Christian Doncarli, Arnaud Doucet, Slim Essid, Simon Godsill, Jérôme Idier, Christian Jutten, Marie-Françoise Lucas, Éric Moulines, Gaël Richard et Bruno Torrésani.

Pour son soutien de tout premier plan, j'embrasse fougueusement ma compagne Belén.

Ce mémoire est dédié à mes grands-pères Jean et Roger, dont chacun de mes titres universitaires a fait la plus grande fierté.

Préambule

Ce document est organisé en trois parties. La première partie est une notice individuelle décrivant mon parcours et rapportant l'ensemble de mes activités professionnelles depuis le début de ma thèse en 2000. La seconde partie rapporte mes activités de recherche consacrées depuis 2007 à un thème particulier, la factorisation en matrices non-négatives. Cette partie unifie dans un même formalisme un ensemble de mes publications, et en particulier six publications représentatives de mes recherches sur ce sujet, annexées dans la troisième partie.

Dans tout le document, les citations faites à des articles se réfèrent à la bibliographie commune commençant à la page 69.

Première partie

Notice individuelle

Chapitre 1

Curriculum vitæ

Cédric Févotte

Né en 1977 en Meurthe-et-Moselle.
Docteur-ingénieur en traitement du signal.

Laboratoire Lagrange
Université Nice Sophia Antipolis
Parc Valrose, 06000 Nice, France
+33 4 92 07 63 80
cfevotte@unice.fr
http://www.unice.fr/cfevotte/

Carrière

depuis 2007 **Chargé de recherche CNRS**

depuis 2013 *Laboratoire Lagrange - UMR 7293 CNRS, Observatoire de la Côte d'Azur & Université Nice Sophia Antipolis*, Nice, France.

2007–2012 *Laboratoire Traitement et Communication de l'Information (LTCI) - UMR 5141 CNRS & Télécom ParisTech*, Paris, France.

2007 **Chercheur contractuel** à ***Télécom ParisTech*** (*8 mois*)
(anciennement École Nationale Supérieure des Télécommunications). Au sein du réseau européen d'excellence K-Space (Knowledge space of semantic inference for automatic annotation and retrieval of multimedia content).

2006–2007 **Ingénieur de recherche** à ***Mist-Technologies*** (*10 mois*)
(startup devenue *Audionamix*, Paris). Développement d'outils de décomposition des signaux musicaux, pour le remastering mono/stéréo en son 5.1 home cinéma.

2003–2006 **Chercheur post-doctorant** à l'***Université de Cambridge***
Dans le *Signal Processing Laboratory* et au sein du réseau européen HASSIP (Harmonic analysis and statistics for signal and image processing).

Formation

2000–2003 **Thèse de doctorat** en traitement du signal
Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), UMR 6597 CNRS, École Centrale, École des Mines & Université de Nantes.
"Approche temps-fréquence pour la séparation aveugle de sources non-stationnaires". Encadrée par Christian Doncarli.

1999–2000 **DEA** en automatique et informatique appliquée. *École Centrale de Nantes.*

1997–2000 **Études d'ingénieur généraliste**. *École Centrale de Nantes.*

1995–1997 **Classes préparatoires**. *Lycée Henri Poincaré, Nancy.*

Mobilité court-terme

2012 **Visiting scientist** à ***Mitsubishi Electric Research Laboratories*** *(3 mois)*
Cambridge MA, États-Unis. Dans l'équipe *Speech & Audio*, en disponibilité du CNRS.

2010, 2012 Séjours à l'*Université du Bosphore*, Istanbul *(2 × 2 semaines)*

2002 Séjour doctoral à l'*Université de Maribor*, Slovénie *(1 mois)*

Thèmes de recherche

Traitement du signal et des images & *apprentissage statistique*, pour le débruitage, la séparation de sources, les problèmes inverses, notamment en audio. Estimation statistique, inférence bayésienne, optimisation, simulation stochastique, approximations variationnelles, modèles à variables latentes, analyse en composantes indépendantes, analyse en composantes parcimonieuses, factorisation en matrices non-négatives, représentations temps-fréquence.

1.1 Encadrement, enseignement et diffusion de la science

1.1.1 Encadrement doctoral

2012– Nicolas Seichepine. *Méthodes de co-factorisation douce pour la segmentation audiovisuelle*. Co-encadrée à 40% avec Slim Essid et Olivier Cappé (Télécom ParisTech), École doctorale EDITE de Paris.

2009–2012 Augustin Lefèvre. *Dictionary learning methods for single-channel audio source separation*. Co-encadrée à 50% avec Francis Bach (INRIA). Soutenue le 3 octobre 2012. École doctorale EDSP de Cachan.

2007–2010 Laurent Oudre. *Reconnaissance d'accords à partir de signaux audio par l'utilisation de gabarits théoriques*. Soutenue le 3 novembre 2010. Co-encadrée à 50% avec Yves Grenier (Télécom ParisTech). École doctorale EDITE de Paris.

1.1.2 Encadrement post-doctoral

2010–2012 Onur Dikmen, sur le projet ANR-TANGERINE. (*2 ans*)

2008–2009 Alexey Ozerov, sur le projet ANR-SARAH, avec Maurice Charbit. (*18 mois*)

1.1.3 Encadrement de visiteurs

2013 Dennis Sun, doctorant de l'Université de Stanford, États-Unis. *(3 mois)*

2012 Brian King, doctorant de l'Université de Washington, États-Unis. *(3 mois)*

1.1.4 Encadrement de stages de master

2012 Nicolas Seichepine. *Factorisation multimodale pour la segmentation audiovisuelle*. Co-encadrement avec Slim Essid (Télécom ParisTech), stage de Master MVA.

2009 Augustin Lefèvre. *Décompositions du signal sonore par factorisation en matrices non-négatives pénalisée*. Co-encadrement avec Francis Bach (INRIA), stage de Master MVA.

1.1.5 Participation à l'enseignement

Depuis le début de ma thèse en 2000 et jusqu'à aujourd'hui, j'ai assuré des enseignements en automatique & traitement du signal, restauration des signaux audio et séparation de sources dans diverses institutions (École des Mines de Nantes, École Centrale de Nantes, Télécom ParisTech, Université de Cambridge) et à des niveaux élève-ingénieur et master. Le détail de ces enseignement est présenté à la table 1.1.

1.1.6 Participation à l'organisation de conférences et *workshops*

International

Workshops

- Workshops *Methodological aspects of hyperspectral imaging* (MAHI), Nice, octobre 2012 & 2013 ($\sim$ 30 participants).
- Workshop *Music signal processing* en prélude à la conférence Acoustics'08, Télécom ParisTech, 27 juin 2008 ($\sim$ 30 participants).
- Workshop *Harmonic analysis and statistics for signal and image processing*, Cambridge UK, 13-17 septembre 2004 ($\sim$ 20 participants).

Sessions spéciales

- Session spéciale *Music structure analysis & sound source separation*, international symposium on Computer Music Modeling and Retrieval (CMMR), Málaga, juin 2010.

	Débruitage des signaux audio (décliquage par interpolation temporelle, réduction du bruit de fond)	Décompositions adaptatives du signal et des images (analyse en composantes indépendantes, NMF, séparation de sources)	Automatique & traitement du signal (systèmes linéaires invariants, transformées de Fourier, échantillonnage, processus aléatoires)	Total (équiv. TD)
2000 - 2003			– Cursus ingénieur EMN (135h TD/TP) – Cursus ingénieur ECN (90h TD/TP)	225h
2003 - 2005			– Cursus ingénieur Univ. Cambridge (12h TD)	12h
2007-2008	– Formation continue TPT (2h C) – Master ATIAM (2h C, 3h TP, 3h EX)			12h
2008-2009	– Formation continue TPT (2h C) – Master ATIAM (2h C, 3h TP, 3h EX) – Cursus ingénieur 3^e année ECN (2h C, 3h TP)			18h
2009-2010	– Master ATIAM (2h C, 3h TP, 3h EX) – Cursus ingénieur 3^e année ECN (2h C, 3h TP)	– Master ATIAM (1h C) – Cursus ingénieur 3^e année TPT (3h C, 3h TP) – Cursus ingénieur 1^{ere} année TPT (1h C)		25,5h
2010-2011	– Master ATIAM (2h C, 3h TP, 3h EX) – Cursus ingénieur 3^e année ECN (2h C, 3h TP)	– Master ATIAM (1h C) – Cursus ingénieur 3^e année TPT (3h C, 3h TP) – Cursus ingénieur 1^{ere} année TPT (3h C) – Formation spécifique des ingénieurs du Corps des Mines (3h C)		33h
2011-2012	– Master ATIAM (2h C, 3h TP, 3h EX) – Cursus ingénieur 3^e année ECN (2h C, 3h TP)	– Master ATIAM (2h C, 3h TP) – Cursus ingénieur 3^e année TPT (2h C, 3h TP)	– Cursus ingénieur 1^{ere} année TPT (15h C, 9h TD, 15h TP)	73,5h
2012-2013	– Cursus ingénieur 3^e année ECN (2h C, 3h TP)			6h
2013-2014	– Cursus ingénieur 3^e année ECN (2h C, 3h TP)			6h

TABLE 1.1 – Récapitulatif des enseignements. C = cours ; TD = travaux dirigés ; TP = travaux pratiques ; EX = oraux d'examen. EMN = École des Mines de Nantes ; ECN = École Centrale de Nantes ; TPT = Télécom ParisTech ; ATIAM = Acoustique, Traitement du signal et Informatique Appliqués à la Musique (master Paris 6, IRCAM & Télécom ParisTech).

– Session spéciale *Nonnegative matrix and tensor factorisations : Statistical methods and applications*, European Signal Processing Conference (EUSIPCO), avec Taylan Cemgil (Université du Bosphore, Istanbul), Glasgow, août 2009.

Président invité de sessions (chairman)
Conférences SPARS 2009, IEEE SSP 2011, ICASSP 2013-2014.

National

École d'été
– Direction scientifique de la 5^e École d'Été de Peyresq GRETSI & GdR ISIS, sur le thème *Apprentissage en traitement du signal et des images*, juin 2010.

Journées thématiques
– Journée GdR ISIS *Carrières en Signal, Image & Vision*, avec avec Nicolas Dobigeon (IRIT, Toulouse), 11 octobre 2013. ($\sim$ 120 participants)
– Journée GdR ISIS *Avancées récentes en traitement du signal audio*, avec Valentin Emiya (LIF, Marseille), 18 octobre 2012. ($\sim$ 60 participants)
– Journée GdR ISIS *Contraintes de non-négativité en traitement du signal et des images*, avec Nicolas Dobigeon (IRIT, Toulouse), 1er février 2011. ($\sim$ 60 participants)

Sessions spéciales
– Session spéciale *Applications de la séparation de sources*, avec Rémi Gribonval et Emmanuel Vincent, GRETSI, Paris, septembre 2003.

1.2 Projets, relations industrielles et valorisation

1.2.1 Financements par projet

Les montants indiqués correspondent à la subvention allouée à l'ensemble des partenaires.

Financements ANR

2009–2013 Financement ANR "Jeunes chercheurs" pour le projet TANGERINE (*Theory and applications of nonnegative matrix factorization*), 140 k€. Porteur.

2006–2009 Financement ANR "Recherche innovation en audiovisuel et multimédia (RIAM)" pour le projet SARAH (*Standardisation of high-definition audio remastering*), avec Mist-Technologies et les studios Copra. 361 k€. Co-porteur.

Financements CNRS

2013–2015 Financement "Grandes masses de données scientifiques (MASTODONS)" pour le projet DISPLAY (*Distributed processing for very large arrays in radioastronomy*), avec le LTCI et le laboratoire SATIE. 50 k€. Participant.

2011–2012 Financement "Projet exploratoire premier soutien (PEPS)" pour le projet ESTOMAT (*Estimation de l'ordre dans les modèles à factorisation matricielle*), avec Nicolas Dobigeon (IRIT, Toulouse). 10 k€. Co-porteur.

2011–2012 Financement "Échange de chercheurs" pour le projet BAYTEN (*Fast Bayesian matrix and tensor factorisation methods for nonstationary multivariate time series analysis*) avec Taylan Cemgil (Université du Bosphore, Istanbul). 4 k€. Co-porteur.

2001–2003 Projet jeunes chercheurs GdR ISIS *Ressources pour la séparation de signaux audiophoniques* avec l'IRISA et l'IRCAM. 9 k€. Co-porteur.

Financement Observatoire de la Côte d'Azur

2013 Financement "Bonus Qualité Recherche (BQR)" pour le projet PHASE (*The estimation of phase in latent factor models*). 6 k€. Porteur.

Financement Institut Mines-Télécom

2013 Financement Futur & Ruptures pour la thèse de Nicolas Seichepine. 112 k€. Co-porteur.

1.2.2 Mise en disponibilité, consultance

2012 Visiting scientist à Mitsubishi Electric Research Laboratories (Cambridge MA, États-Unis), au sein de l'équipe "Speech & Audio", en disponibilité du CNRS *(3 mois)*.

2009 Consultant en traitement du signal pour l'entreprise de géophysique CGGVeritas *(6 jours)*.

1.2.3 Brevets

2013 Dépôt d'un brevet US pour un procédé de modélisation et débruitage de la parole, issu de mon séjour à Mitsibushi Electric Research Laboratories [OFB11].

2011 Dépôt d'un brevet US pour un procédé de séparation de sources audio, issu du projet ANR SARAH [HFL12].

1.2.4 Valorisation

L'une des méthodes de reconnaissance d'accord développée avec Laurent Oudre durant sa thèse est utilisée dans le moteur de recherche d'information musicale MuMa, développé par Exalead au sein du projet Quaero. [1]

1.3 Animation et management de la recherche

1.3.1 Pilotage

International

depuis 2012	Membre du comité technique IEEE Machine Learning for Signal Processing (MLSP). Chairman du sous-commité "Elections and nominations" depuis 2013.
depuis 2012	Administrateur de la liste de diffusion LVA/ICA sur la séparation de sources ($\sim$ 650 membres).

National

depuis 2009	Membre du comité de direction du GdR ISIS (Information, Signal, Images et Vision), co-responsable du réseau des doctorants.
2011–2012	Membre élu du conseil de laboratoire du LTCI.
2008	Participation à un rapport de l'IGAENR sur la simplification des procédures administratives (achats, missions, recrutements) dans les unités de recherche [AF08], suite à une réponse à une consultation publique organisée par le ministère de l'enseignement supérieur et de la recherche.

1.3.2 Expertise

International

Rapports d'habilitation

2012	Rapporteur des travaux de Dr Zhirong Yang pour l'obtention du titre de *Docent*, Aalto University, Finlande.

Participation à des jurys de thèse à l'étranger

2013	*Opponent* de la thèse de Nasser Mohammadiha (encadrée par Arne Leijon), Royal Institute of Technology (KTH), Stockholm.
2012	Membre du jury de thèse de Kenan Yılmaz (encadrée par Taylan Cemgil), Université du Bosphore (Istanbul).
2011	*Opponent* de la thèse de Zhanyu Ma (encadrée par Arne Leijon), Royal Institute of Technology (KTH), Stockholm.
2010	*External examiner* de la thèse de Paul Peeling (encadrée par Simon Godsill), Université de Cambridge.

Expertise pour des programmes de financement

2014	Rapporteur d'une demande de financement EPSRC (UK).
2011	Rapporteur d'une demande de financement NSF (USA).

1. `http://labs.exalead.com/project/muma`

National

Participation à des comités de recrutement de maîtres de conférence

2014 Poste en section 61 à l'ENSEEIHT, Toulouse.
2012 Poste en section 61 à l'Université de Nice.
2011 Poste en section 61 à l'École Centrale de Nantes.
Poste en section 26/27 à l'Université de Provence.

Participation à des jurys de thèse en France

2013 Examinateur de la thèse de François Rigaud (encadrée par Bertrand David et Laurent Daudet), Télécom ParisTech.

Autres

2013– Membre du jury du prix de thèse "Signal, Image & Vision" du GRETSI, GdR ISIS & Club EEA.

1.3.3 Activité éditoriale

Journaux

- Éditeur associé de la revue IEEE Trans. Signal Processing, depuis 2014.
- Rapports de lecture principalement pour les revues IEEE Trans. Signal Processing, IEEE Trans. Audio, Speech and Language Processing, IEEE Trans. Pattern Analysis and Machine Intelligence, IEEE Trans. Neural Networks.

Conférences

- *Area Chair* (en *Machine Learning*) pour la conférence EUSIPCO 2014.
- Membre des comités de programme des conférences EUSIPCO (*European Signal Processing Conference*) 2010–2012 ; CMMR (*Computer Music Modeling and Retrieval*), 2010, 2012 ; IEEE WASPAA (*Workshop on Applications of Signal Processing to Audio and Acoustics*) 2011, 2013 ; LVA/ICA (*Latent Variable Analysis and Signal Separation*) 2012 ; IEEE MLSP (*Machine Learning for Signal Processing*) 2012–2014, IEEE ICASSP 2012–2014.

1.4 Prix & distinctions

- *ICASSP 2014 Best student paper award* (attribué à Dennis Sun, Stanford University), pour l'article "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence" [SF14], issu de son séjour à Nice et dont je suis co-auteur.
- Prime d'excellence scientifique du CNRS, attribuée en 2012 et pour 4 ans.

Chapitre 2

Activités de recherche

Une très large partie de mes activités de recherche concerne la séparation aveugle de sources (SAS). La SAS consiste à estimer n signaux inconnus (les *sources*) d'après la seule observation de m mélanges d'entre eux (les *observations* ou *canaux*). Le problème de la SAS comporte différents niveaux de difficulté, selon principalement trois caractéristiques.

La première de ces caractéristiques est la nature du mélange. Le problème le plus simple est le cas des mélanges linéaires instantanés invariants dans le temps, lorsque les observations à l'instant t sont une combinaison linéaire des sources au même instant t. Cependant, en pratique, les mélanges sont souvent convolutifs, éventuellement variants dans le temps et parfois non-linéaires. Le problème standard de la "cocktail party" consiste par exemple à séparer des enregistrements de locuteurs parlant simultanément dans le brouhaha d'une pièce. La réverbération crée un effet de convolution et le déplacement des locuteurs produit un mélange variant dans le temps. Des effets de non-linéarité peuvent être observés, par exemple si les microphones saturent.

Une autre caractéristique des problèmes de SAS est le nombre de sources n par rapport au nombre d'observations m. Dans le cas (sur-)déterminé ($m \geq n$), il suffit d'estimer les paramètres de mélange et d'en appliquer l'inverse aux observations pour estimer les sources. Au contraire, le cas sous-déterminé ($m < n$) est un problème mal posé car le mélange n'est plus inversible et une information *a priori* sur les sources (typiquement, un modèle) est nécessaire à leur reconstruction.

Une troisième caractéristique est la nature des sources (stationnaire, non-stationnaire, parcimonieuse, non-négative, etc.) qui conditionne généralement le type de méthode de séparation à employer.

Durant ma thèse (2000–2003), j'ai contribué à une méthode de séparation de mélanges (sur-)déterminés de sources non-stationnaires, basée sur la diagonalisation simultanée de matrices de représentation temps-fréquence spatiales de la classe de Cohen. Je me suis également intéressé à l'évaluation des méthodes de séparation de sources audio (SSA). Puis, lorsque j'étais chercheur post-doctorant à Cambridge (2003–2006), j'ai commencé à étudier des approches bayésiennes pour la séparation de mélanges sous-déterminés bruités, en particulier pour des problèmes audio. Ce travail m'a conduit plus généralement à développer des modèles probabilistes de signaux audio exploitant des principes de parcimonie structurée dans des bases temps-fréquence. Lors de ce post-doctorat je me suis également ouvert aux méthodes de simulation Monte-Carlo par chaînes de Markov. En 2006 j'ai rejoint la startup Mist-Technologies à Paris. Au sein de cette entreprise j'ai commencé à m'intéresser à la séparation de sources en contexte mono-capteur par factorisation non-négative du spectrogramme.

Depuis mon entrée au CNRS en 2007, je me suis quasiment entièrement consacré au problème de factorisation en matrices non-négatives (dont j'utiliserai l'acronyme anglais NMF pour *nonnegative matrix factorization*), avec des contributions à la fois générales sur le plan méthodologique (construction d'algorithmes de descente par majoration-minimisation, méthodes d'estimation de l'ordre, approches probabilistes) et des contributions applicatives dans le domaine de la séparation

de sources audio. Ces contributions à la NMF feront l'objet de la partie II de ce mémoire. Au préalable, je présente succinctement dans les paragraphes qui suivent les recherches que j'ai réalisées préalablement à mon entrée au CNRS.

2.1 Recherche doctorale (sept. 2000 – oct. 2003)

Laboratoire	Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN) UMR 6597 CNRS, École Centrale, École des Mines & Université de Nantes.
Sujet	Approche temps-fréquence pour la séparation aveugle de sources non-stationnaires
Encadrant	Christian Doncarli
Financement	Bourse ministérielle

Séparation aveugle de sources avec des représentations temps-fréquence bilinéaires

Dans le cas (sur-)déterminé, une procédure classique de SAS consiste à *blanchir* les observations (décorrélation spatiale) puis à diagonaliser simultanément un ensemble de matrices correspondant à une représentation conjointe des observations, afin d'estimer la matrice de mélange. Dans le cas de signaux non-stationnaires, Belouchrani & Amin [BA98] ont montré qu'il est est pertinent d'utiliser la classe de Cohen des représentations temps-fréquence spatiales (RTFS). Pour chaque point temps-fréquence, la RTFS des observations est une matrice dont les termes diagonaux contiennent les RTF de chacun des canaux observés et dont les termes non-diagonaux contiennent les RTF conjointes de deux canaux.

Ma principale contribution a été l'élaboration d'un critère permettant d'optimiser la sélection des matrices à diagonaliser simultanément. Ce critère permet de détecter les points temps-fréquence correspondants à des auto-termes simples des sources (i.e., des points où seule une source est présente). Ce travail a été publié dans [FD04, HFDZ02]. Il a été appliqué à la séparation de signaux électromyographiques dans [FFDM04], dans le cadre du projet européen *NEW* (Neuromuscular assessment in the Elderly Worker).

Bloc-diagonalisation simultanée pour la séparation de mélanges convolutifs

La méthode de séparation précédente peut être généralisée à la séparation de mélanges convolutifs à réponse impulsionnelle finie. Le principe consiste à ajouter des versions retardées des sources au vecteur contenant les sources initiales, ce qui permet de reformuler le mélange convolutif en un mélange instantané. Toutefois les nouvelles sources ne sont plus toutes mutuellement indépendantes et il apparaît que l'étape de diagonalisation simultanée doit être remplacée par une étape de *bloc*-diagonalisation simultanée. Comme de nombreuses méthodes de séparation de mélanges convolutifs, cette méthode ne produit pas les sources originales mais plusieurs versions filtrées de chaque source. Une étape d'identification aveugle SIMO est donc nécessaire pour retrouver les sources originales. Ce travail est présenté dans [FD03, FDD03]. Un algorithme de bloc-diagonalisation simultanée de type Jacobi est décrit dans [LFMV02, FT07].

Évaluation des méthodes de séparation de sources audio

Durant mon doctorat, j'ai été avec Emmanuel Vincent (alors à l'IRCAM) et Rémi Gribonval (IRISA) l'un des initiateurs du projet Jeunes Chercheurs "Ressources pour la séparation de sources audio" financé par le GdR ISIS. Les objectifs de ce projet ont été :

- d'établir une classification des problèmes variés de la SSA en tâches générales, suivant l'objectif à accomplir (qualité audio des sources, indexation, compression, etc.),
- de définir la structure d'une base de signaux test pour l'évaluation de chaque tâche,
- de cataloguer les algorithmes de séparation existants et d'identifier à quelle tâche ils peuvent s'appliquer,

– de définir des critères d'évaluation pertinents pour chaque tâche afin de pouvoir évaluer *numériquement* la qualité de séparation (rapport signal à bruit, taux d'erreur de classification, rapport de compression, etc.),

Le dernier point en particulier, concernant l'évaluation, a donné lieu à un article [VGF06] à fort impact (environ 650 citations recensées sur Google Scholar au moment de la rédaction de ce mémoire) et à la boîte à outils MATLAB `BSS_EVAL` pour l'évaluation de méthodes de SAS. Cette dernière est devenue un outil d'évaluation standard. [1]

2.2 Recherche post-doctorale (nov. 2003 – mars 2006)

Laboratoire	Signal Processing Laboratory, Cambridge University Engineering Department
Thème	Traitement statistique du signal audio
Responsable	Simon J. Godsill
Financement	projet européen HASSIP
	(Harmonic analysis and statistics for signal and image processing)

Approches bayésiennes pour la séparation de sources

En pratique les mélanges de sources sont souvent sous-déterminés ($m < n$). Par exemple, en audio, les enregistrements stéréophoniques sont sous-déterminés dès lors que plus de deux sources sont simultanément en présence. Le problème sous-déterminé est un problème mal posé qui requiert une information *a priori* sur les sources, typiquement, un *modèle*.

Un modèle simple consiste à exploiter des hypothèses de parcimonie des sources. Les signaux audio en particulier adoptent une représentation parcimonieuse sur des *dictionnaires* adaptés (tels que des atomes de Fourier court-terme, des cosinus locaux, des ondelettes), c'est-à-dire que la majorité de leurs coefficients de décomposition sont proches de zéro. En modélisant le signal audio comme une combinaison linéaire d'atomes MDCT (une base orthonormale de cosinus locaux), dans laquelle les coefficients de décomposition sont modélisés par un processus aléatoire identiquement et indépendamment distribué (i.i.d) Student t, nous avons élaboré une approche bayésienne basée sur un échantillonneur de Gibbs (une méthode classique de simulation Monte Carlo par chaînes de Markov) pour la séparation de mélanges linéaires instantanés sous-déterminés et bruités [FG06a]. Une approche variationnelle de type *mean field*, plus rapide mais moins robuste, est également considérée dans [CFG07]. Le modèle utilisé dans ces dernières publications est relativement générique dans la mesure où la méthodologie utilisée peut être directement appliquée à des sources non-audio pour peu qu'elles admettent une représentation parcimonieuse dans un dictionnaire donné. Pour améliorer la qualité de la séparation de sources dans le cas audio nous avons élaboré des modèles de source plus complexes, prenant en compte les spécificités de l'audio. Ainsi, des modèles dépendants de la fréquence sont étudiés dans [FG05] et des modèles prenant en compte des contraintes structurelles (modélisant des phénomènes physiques tel que la persistance temporelle des coefficients temps-fréquence) sont étudiés dans [Fév06].

Des modèles temps-fréquence gaussiens, permettant une inférence rapide avec l'algorithme EM, sont également considérés dans [FC05]. Lorsque des techniques de type EM ou variationnel sont utilisées, l'initialisation de la matrice de mélange joue un rôle important dans la convergence de l'algorithme vers un minimum global. À cette fin, nous avons proposé dans [DF06] une méthode de type *kernel PCA* pour calculer une estimation grossière de la matrice de mélange dans les mélanges de sources parcimonieuses.

1. The `BSS_EVAL` MATLAB toolbox : `bass-db.gforge.inria.fr/bss_eval`

Représentations parcimonieuses adaptatives et structurées

Le travail ci-dessus mélange des techniques de séparation de sources et de régression linéaire parcimonieuse, qui est le problème consistant à trouver une représentation parcimonieuse d'un signal dans un dictionnaire donné (généralement redondant). Si les modèles de source audio que nous avons utilisés dans le cas de la séparation de sources sont construits sur des bases (principalement pour limiter la complexité calculatoire), nous avons également étudié des techniques de régression linéaire parcimonieuse dans des dictionnaires formés par l'union de bases (et donc redondants). Une méthode générique inspirée de la sélection bayésienne de variables est proposée dans [FG06b]. Cette approche est appliquée à l'extraction des parties tonales et transitoires dans la musique dans [FTDG08] : le signal musical est décomposé sur l'union d'une base MDCT avec courte résolution temporelle pour les transitoires (avec contrainte structurelle selon l'axe des fréquences) et d'une base MDCT avec longue résolution temporelle pour les parties tonales (avec contrainte structurelle selon l'axe temporel).

2.3 Recherche en entreprise (mai 2006 – fév. 2007)

Entreprise	Mist-Technologies, Paris (devenue Audionamix en 2008)
Thème	Séparation de sources en mono-capteur, remastering.

Durant mon passage au sein de l'entreprise Mist-Technologies je me suis intéressé au problème de *remastering* d'enregistrements dits du "back catalog". Ce dernier terme concerne tous les enregistrements audio produits pour l'industrie de la musique ou du cinéma avant l'ère du son dit "3D". Le remastering consiste à produire à partir d'un mélange mono ou stéréo un enregistrement multicanal (de plus de 2 canaux) destiné à être "spatialisé" au moyen d'autant d'enceintes que de canaux (par exemple 5 petites enceintes et 1 caisson de basse pour le format "5.1"). Il est souvent impossible ou trop coûteux de créer un enregistrement 3D sur la base des sources originales. Dans bien des cas les pistes audio auront été perdues ou détruites, et si elles existent, il coûtera très cher au studio qui souhaite ressortir un film ancien en son 3D de solliciter les services d'un ingénieur du son qui saura préserver l'intégrité artistique du mix mono ou stéréo. Ainsi, au contraire du *remixing*, le *remastering* décompose astucieusement le mélange original sans perte et le redistribue sur l'ensemble d'enceintes.

La méthode à laquelle j'ai contribué au sein de Mist-Technologie s'est construite sur une approche séparation de sources. Sans être directement liée à la NMF elle m'a permis de me familiariser avec cette thématique qui est devenue mon principal sujet de recherche à partir de 2007.

2.4 Recherche au CNRS (depuis 2007)

Affiliations :

depuis 2013	Laboratoire Lagrange, Nice UMR 7293 CNRS, Observatoire de la Côte d'Azur & Univ. Nice Sophia Antipolis
2007–2012	Laboratoire Traitement et Communication de l'Information (LTCI), Paris UMR 5141 CNRS & Télécom ParisTech

De manière générale, les données à traiter sont souvent non-négatives par nature, comme par exemple les intensités de pixels, les amplitudes spectrales, les mesures de quantité ou de prix (consommations alimentaires, valeurs boursières) ou encore les avis quantitatifs (notes de clients sur internet). Le traitement optimal de ces données requiert parfois un traitement sous contrainte de non-négativité. La NMF est une technique de régression linéaire apparue à la fin des années 90 qui est devenue un sujet de recherche important dans les domaines de l'apprentissage statistique et du traitement du signal et de l'image. Elle consiste simplement à approcher une matrice à coefficients non-négatifs par un produit de deux autres matrices non-négatives, où l'une des matrices représente un dictionnaire de motifs élémentaires caractéristiques des données et l'autre matrice contient les

coefficients d'activation de ces motifs.

La NMF a été appliquée à de nombreux problèmes (reconnaissance de formes, clustering, fouille de données, séparation de sources, filtrage collaboratif, etc.) dans de nombreux domaines (traitement de données textuelles, traitement du signal audio, bioinformatique, etc.). La NMF, et sa généralisation à la factorisation de tenseurs non-négatifs (NTF), sont des sujets qui nécessitent des réponses à de nombreux problèmes ouverts. Mes recherches sur ce thème ont été soutenues par un projet jeunes chercheurs de l'ANR (projet TANGERINE, 2009–2012). La méthodologie développée au sein de ce projet a notamment été appliquée à des problèmes spécifiques relevant du traitement du signal musical et de l'analyse de document multimédia. Le descriptif de ces recherches est développée dans la partie II de ce manuscrit.

Chapitre 3

Production scientifique

3.1 Bibliométrie

Données au 14 mai 2014.

***h*-index de Google Scholar : 24**
(nb. total de citations : **2753**)

***h*-index de Thompson Reuters Web of Science : 11**
(nb. total de citations : **657** ; 628 sans auto-citation)

17 articles dans des revues de premier rang (1 IEEE Trans. Pattern Analysis and Machine Intelligence, 1 IEEE Signal Processing Magazine, 2 Neural Computation, 1 IEEE Trans. Signal Processing, 7 IEEE Trans. Audio, Speech and Langage Processing, 1 IEEE Trans. Multimedia, 1 IEEE Trans. Biomedical Engineering, 2 IEEE Signal Processing Letters, 1 Digital Signal Processing).
48 articles dans des actes de colloques internationaux à comité de lecture (notamment 1 NIPS, 23 IEEE conferences and workshop, 8 EUSIPCO).
2 chapitres d'ouvrage

56 co-auteurs de **25** institutions différentes (15 internationales, 10 nationales).

Profil Google Scholar : `http://scholar.google.com/citations?user=kAWEUqOAAAAJ&hl=en`

3.2 Liste complète des publications

Articles de journaux à comité de lecture

[J1] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using non-negative factorizations. *IEEE Signal Processing Magazine*, 31(3) :66–75, May 2014.

[J2] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7) :1592–1605, July 2013.

[J3] S. Essid and Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Transactions on Multimedia*, 15(2) :415–425, Feb. 2013.

[J4] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10) :5163–5175, Oct. 2012.

[J5] L. Oudre, C. Févotte, and Y. Grenier. Probabilistic template-based chord recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(8) :2249 – 2259, Nov. 2011.

[J6] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9) :2421–2456, Sep. 2011.

[J7] L. Oudre, Y. Grenier, and C. Févotte. Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7) :2222 – 2233, Sep. 2011.

[J8] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :564–575, Mar. 2010.

[J9] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :550–563, Mar. 2010.

[J10] 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, Mar. 2009.

[J11] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill. Sparse linear regression with structured priors and application to denoising of musical audio. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1) :174–185, Jan. 2008.

[J12] A. T. Cemgil, C. Févotte, and S. J. Godsill. Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing*, 17(5) :891–913, Sep. 2007. Special issue *Bayesian source separation*, ed. E. E. Kuruoğlu and K. H. Knuth.

[J13] C. Févotte and S. J. Godsill. A Bayesian approach to blind separation of sparse sources. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6) :2174–2188, Nov. 2006.

[J14] 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, Jul. 2006.

[J15] C. Févotte and S. J. Godsill. Sparse linear regression in unions of bases via Bayesian variable selection. *IEEE Signal Processing Letters*, 13(7) :441–444, Jul. 2006.

[J16] D. Farina, C. Févotte, C. Doncarli, and R. Merletti. Blind separation of linear instantaneous mixtures of non-stationary surface myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 51(9) :1555 –1567, Sep. 2004.

[J17] C. Févotte and C. Doncarli. Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Processing Letters*, 11(3) :386–389, Mar. 2004.

Articles de conférences à comité de lecture

[C1] D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[C2] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Piecewise constant nonnegative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[C3] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Co-factorisation douce en matrices non-négatives. application au regroupement multimodal de locuteurs. In *Proc. Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France, Sep. 2013.

[C4] N. Dobigeon and C. Févotte. Robust nonnegative matrix factorization for nonlinear unmixing of hyperspectral images. In *Proc. IEEE Workshop Hyperspectral image and signal processing : Evolution in remote sensing (WHISPERS)*, Gainesville, FL, June 2013.

[C5] C. Févotte, J. Le Roux, and J. R. Hershey. Non-negative dynamical system with application to speech and audio. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[C6] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[C7] J. Le Roux, C. Févotte, and J. R. Hershey. A new non-negative dynamical system for speech and audio modeling. In *Proc. Acoustical Society of Japan Spring Meeting*, Mar. 2013.

[C8] A. Lefèvre, F. Bach, and C. Févotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, Oct. 2012.

[C9] T. Gerber, M. Dutasta, L. Girin, and C. Févotte. Professionally-produced music separation guided by covers. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, Oct. 2012.

[C10] S. Essid and C. Févotte. Decomposing the video editing structure of a talk-show using nonnegative matrix factorization. In *Proc. IEEE International Conference on Image Processing (ICIP)*, Orlando, Florida, Sep. 2012.

[C11] B. King, C. Févotte, and P. Smaragdis. Optimal cost function and magnitude power for NMF-based speech separation and music interpolation. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sep. 2012.

[C12] O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2267–2275, Granada, Spain, Dec. 2011. MIT Press.

[C13] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. In *Proc. NIPS workshop on Sparse Representation and Low-rank Approximation*, Sierra Nevada, Spain, Dec. 2011.

[C14] A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, Oct. 2011.

[C15] C. Févotte and J. Idier. Algorithmes de factorisation en matrices non-négatives fondée sur la beta-divergence. In *Proc. 23e colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, Sep. 2011.

[C16] O. Cappé, C. Févotte, and D. Rhodes. Algorithme EM en ligne simulé pour la factorisation non-négative probabiliste. In *Proc. 23e colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, Sep. 2011.

[C17] A. Lefèvre, F. Bach, and C. Févotte. Factorisation de matrices structurée en groupes avec la divergence d'Itakura-Saito. In *Proc. 23e colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, Sep. 2011.

[C18] C. Févotte, O. Cappé, and A. T. Cemgil. Efficient Markov chain Monte Carlo inference in composite models with space alternating data augmentation. In *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, pages 221 – 224, Nice, France, June 2011.

[C19] C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[C20] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[C21] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[C22] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[C23] L. Oudre, C. Févotte, and Y. Grenier. Probabilistic framework for template-based chord recognition. In *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, St-Malo, France, Oct. 2010.

[C24] H. Lantéri, C. Theys, C. Richard, and C. Févotte. Split gradient method for nonnegative matrix factorization. In *Proc. 18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, Aug. 2010.

[C25] C. Févotte and A. Ozerov. Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues. In S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, editors, *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, volume 6684 of *Lecture Notes in Computer Science*, pages 102–115, Málaga, Spain, 2010., June 2010. Springer. Long paper.

[C26] L. Oudre, Y. Grenier, and C. Févotte. Template-based chord recognition : influence of the chord types. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, Oct. 2009.

[C27] A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, USA, Oct. 2009.

[C28] L. Oudre, Y. Grenier, and C. Févotte. Chord recognition using measures of fit, chord templates and filtering methods. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, Oct. 2009.

[C29] C. Févotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009.

[C30] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 15–19, Glasgow, Scotland, Aug. 2009.

[C31] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3137–3140, Taipei, Taiwan, Apr. 2009.

[C32] N. Bertin, C. Févotte, and R. Badeau. A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1545 –1548, Taipei, Taiwan, Apr. 2009.

[C33] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, St-Malo, France, Apr. 2009.

[C34] R. Blouet, G. Rapaport, I. Cohen, and C. Févotte. Evaluation of several strategies for single sensor speech/music separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 37–40, Las Vegas, USA, Apr. 2008.

[C35] S. J. Godsill, A. T. Cemgil, C. Févotte, and P. J. Wolfe. Bayesian computational methods for sparse audio and music processing. In *Proc. 15th European Signal Processing Conference (EUSIPCO)*, Poznań, Poland, Sep. 2007.

[C36] C. Févotte and F. Theis. Pivot selection strategies in Jacobi joint block-diagonalization. In *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 177–187, London, UK, Sep. 2007.

[C37] C. Févotte. Bayesian blind separation of audio mixtures with structured priors. In *Proc. 14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep. 2006. Special session *Undetermined sparse audio source separation* (invited paper).

[C38] L. Benaroya, R. Blouet, C Févotte, and I. Cohen. Single sensor source separation using multiple-window STFT representation. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[C39] F. Desobry and C. Févotte. Kernel PCA based estimation of the mixingmmatrix in linear instantaneous mixtures of sparse sources. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[C40] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani. Sparse regression with structured priors : application to audio denoising. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[C41] C. Févotte and S. J. Godsill. Blind separation of sparse sources using Jeffrey's inverse prior and the EM algorithm. In *Proc. 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Charleston, SC, USA, Mar. 2006.

[C42] V. Y. F. Tan and C. Févotte. A study of the effect of source sparsity for various transforms on blind audio source separation performance. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, Rennes, France, Nov. 2005.

[C43] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 78–81, Mohonk, NY, USA, Oct. 2005.

[C44] C. Févotte and S. J. Godsill. A Bayesian approach to time-frequency based blind source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, USA, Oct. 2005.

[C45] A. T. Cemgil, C. Févotte, and S. J. Godsill. Blind separation of sparse sources using variational EM. In *Proc. 13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, Sep. 2005.

[C46] C. Févotte, S. J. Godsill, and P. J. Wolfe. Bayesian approach for blind separation of underdetermined mixtures of sparse sources. In *Proc. 5th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Granada, Spain, Sep. 2004.

[C47] D. Farina, F. Lebrun, C. Févotte, C. Doncarli, and R. Merletti. Blind source separation of linear mixtures of non-stationary surface EMG signals. In *Proc. 19e colloque GRETSI sur le Traitement du Signal et des Images*, Paris, France, Sep. 2003.

[C48] E. Vincent, C. Févotte, and R. Gribonval. Comment évaluer les algorithmes de séparation audio ? In *Proc. 19e colloque GRETSI sur le Traitement du Signal et des Images*, Paris, France, Sep. 2003.

[C49] C. Févotte, A. Debiolles, and C. Doncarli. Blind separation of FIR convolutive mixtures : application to speech signals. In *Proc. 1st ISCA Workshop on Non-Linear Speech Processing*, Le Croisic, France, May 2003.

[C50] C. Févotte and C. Doncarli. A unified presentation of blind source separation methods for convolutive mixtures using block-diagonalization. In *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, Apr. 2003.

[C51] E. Vincent, C. Févotte, R. Gribonval, and al. A tentative typology of audio source separation tasks. In *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, Apr. 2003.

[C52] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte. Proposals for performance measurement in source separation. In *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, Apr. 2003.

[C53] A. Holobar, C. Févotte, C. Doncarli, and D. Zazula. Single autoterms selection for blind source separation in time-frequenccy plane. In *Proc. 11th European Signal Processing Conference (EUSIPCO)*, Toulouse, France, Sep. 2002.

[C54] L. De Lathauwer, C. Févotte, B. De Moor, and J. Vandewalle. Jacobi algorithm for joint block diagonalization in blind identification. In *Proc. 23rd Symposium on Information Theory in the Benelux*, pages 155–162, Louvain-la-Neuve, Belgium, May 2002.

[C55] E. Le Carpentier and C. Févotte. Séparation de sources autorégressives gaussiennes par maximum de vraisemblance et filtrage de Kalman. In *Proc. 18e colloque GRETSI sur le Traitement du Signal et des Images*, Toulouse, France, Sep. 2001.

Thèses

[T1] C. Févotte. *Approche temps-fréquence pour la séparation aveugle de sources non-stationnaires (A time-frequency approach for blind separation of non-stationary sources).* PhD thesis, École Centrale de Nantes et Université de Nantes, 2003.

[T2] C. Févotte. *Acoustique des salles : modélisation de l'environnement sonore (Room acoustics : identification of room transfer functions).* Master's thesis, École Centrale de Nantes, 2000.

Brevets

[B1] J. Hershey, C. Févotte, and J. Le Roux. Method for transforming non-stationary signals using a dynamic model, Oct. 2012. US Patent 13657077, filed.

[B2] A. Ozerov, C. Févotte, and R. Blouet. Automatic source separation via joint use of segmental information and spatial diversity, Feb. 2011. US Patent 13021692, filed.

Divers

[D1] P. Aimé and C. Févotte. La simplification administrative de la gestion des unités de recherche. Rapport de l'Inspection Générale de l'Éducation Nationale et de la Recherche (IGAENR), no 2008-089, Oct. 2008. `http://media.enseignementsup-recherche.gouv.fr/file/Concours_2008/22/3/2008-089simplification_44223.pdf`.

3.3 Séminaires et présentations invités

Séminaires invités

International

2012 Erwin Schrödinger Institute, Vienna (*Modern methods of time-frequency analysis semester*)

2011 Mitsubishi Electrics Research Laboratories, Cambridge MA (*Mini-symposium on audio and music signal processing*)
Massachusetts Institute of Technology, Cambridge MA (*Stochastic systems group seminar*)

2010 Université du Bosphore, Istanbul
Cambridge University Engineering Dept, UK

2009 Telefonica R&D, Madrid

2007 Dublin City University

2005 Queen Mary University of London

2003 Cambridge University Engineering Dept, UK

National

2014 Xerox Research Center Europe, Grenoble
2013 I3S, Nice (*Workshop on tensors and application*)
2012 LATP, Marseille (*Séminaire Signal-Apprentissage*)
GIPSA-lab, Grenoble (*Séminaire Images-Signal*)
2010 LITIS, Rouen
2009 IRCCyN, Nantes
Laboratoire Fizeau, Nice
IRISA, Rennes
2007 Télécom Paris
2004 IRISA, Rennes
2003 IRISA, Rennes

Présentations invitées

International

2013 Séminaire Dagstuhl *Computational audio analysis*
2010 Session spéciale *Music structure analysis & sound source separation*, CMMR (également organisateur de la session)
2009 Session spéciale *Nonnegative matrix and tensor factorisations : statistical methods and applications*, EUSIPCO (également organisateur de la session)
2006 Session spéciale *Undetermined sparse audio source separation*, EUSIPCO

National

2013 Journée GdR ISIS *Traitement de données à valeurs complexes*, Marseille

Deuxième partie

Contributions à la factorisation en matrices non-négatives

Introduction

Étant donnée une matrice $\mathbf{V}$ de dimension $F \times N$ à coefficients non-négatifs, la factorisation en matrices non-négatives (NMF) consiste à trouver une approximation

$$\mathbf{V} \approx \mathbf{WH}, \tag{3.1}$$

telle que les matrices $\mathbf{W}$ et $\mathbf{H}$ soient à coefficients non-négatifs et de dimensions $F \times K$ et $K \times N$, respectivement. [1]

L'ordre K de la factorisation est souvent choisi tel que $K < F$, de sorte que la matrice $\hat{\mathbf{V}} = \mathbf{WH}$ forme une approximation de rang faible de $\mathbf{V}$. L'autre cas, $K \geq F$, est envisageable mais nécessite alors une forme de régularisation des facteurs (e.g., parcimonie de $\mathbf{H}$) afin d'éviter des solutions exactes triviales.

C'est un article de Lee & Seung paru dans la revue *Nature* en 1999 [LS99] qui a lancé la NMF, en dépit de travaux préliminaires tels que ceux de Paatero [PT94, Paa97]. Dans cet article phare, la NMF est présentée comme une méthode d'apprentissage de dictionnaire, à mettre aux côtés par exemple de la quantification vectorielle (VQ), de l'analyse en composantes principales (PCA) [Bur09] ou de l'analyse en composantes indépendantes (ICA) [Com94]. En effet, la NMF permet d'approcher chaque vecteur colonne de $\mathbf{V}$ (représentant dans ce formalisme un échantillon de données observées) sous la forme

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
donnée		"variables explicatives"	"régresseurs"
		"dictionnaire"	"coefficients de décomposition"
		"motifs"	"activations"

La particularité de la NMF comparée aux autres méthodes d'apprentissage de dictionnaire est d'imposer la non-négativité de $\mathbf{W}$ et $\mathbf{H}$. La non-négativité de $\mathbf{W}$ assure l'interprétabilité du dictionnaire (dans le sens où les motifs et les données appartiennent au même espace non-négatif). La non-négativité de $\mathbf{H}$ assure la non-négativité de $\mathbf{WH}$ (et donc que $\mathbf{V}$ est bien approchée par une autre matrice non-négative) et induit en corollaire une représentation dite "par partie". En effet, du fait de cette contrainte, l'approximation de $\mathbf{V}$ ne peut être que constructive : il est interdit de retrancher des éléments du dictionnaire dans l'approximation d'une donnée. Cela implique que $\mathbf{W}$ aura tendance à contenir des "briques élémentaires" caractéristiques des données.

Ce principe est illustrée dans l'article de Lee & Seung par une expérience que nous reproduisons ici. Nous avons appliqué la PCA et la NMF à un jeu de 2429 imagettes représentant des visages,

1. L'adjectif "non-négatif" est une traduction de l'anglais *nonnegative*. Je l'utilise pour qualifier un nombre réel tel que $x \geq 0$. En d'autres termes, il est synonyme de l'adjectif français "positif". L'adjectif anglais *positive* qualifie quant à lui un réel *strictement* supérieur à 0. L'usage de la terminologie "matrice non-négative" (une matrice à coefficients non-négatifs donc) permet d'éviter toute confusion avec "matrice positive", qui peut désigner une matrice symétrique dont les valeurs propres sont positives (appelée aussi plus précisément "matrice semi-définie positive"). La terminologie utilisée permet en outre de rester proche de l'anglais, et en particulier de l'acronyme NMF qui, plutôt que par exemple "FMP", s'est imposé en français... Par ailleurs, la question de comment traduire *nonnegative matrix factorization* peut se poser. Factorisation non-négative de matrice ? Factorisation de matrice non-négative ? Factorisation en matrices non-négatives ? Le troisième choix m'a paru le mieux traduire le principe fondamental de la NMF qui est de produire une approximation de $\mathbf{V}$ telle que $\mathbf{W}$, $\mathbf{H}$ et donc $\mathbf{WH}$ sont toutes trois à coefficients non-négatifs.

issues du "CBCL face dataset".[2] Un ensemble de 49 de ces imagettes est représenté en figure 3.1 (a). Chaque imagette, de dimension 19×19, est vectorisée puis insérée dans l'une des colonnes de $\mathbf{V}$. Une PCA et une NMF d'ordre $K = 25$ sont appliquées à ce jeu de données. Les dictionnaires appris sont représentés en figure 3.1 (b) & (c). La PCA fournit une représentation holistique des données alors que la NMF en fournit une décomposition par parties. La PCA soulève en outre un problème d'interprétabilité des résultats dans la mesure où le dictionnaire retourné contient des valeurs négatives (en rouge).

Depuis l'article de Lee & Seung en 1999, la NMF a connu essor important dans les domaines du traitement du signal et de l'apprentissage statistique. Elle a par exemple été appliquée à l'analyse de données environnementales [PT94], la création automatique de résumés vidéo [CF02], la fouille de données textuelles [LS99, XLG03], la transcription musicale et la séparation de sources [SB03, Vir07], l'analyse de marqueurs génétiques [BTGM04], la classification de whiskies écossais [YFH06], l'imagerie hyperspectrale [BBL$^+$07], la gestion de portefeuille [DRdFC07], le filtrage collaboratif [Wu07], l'analyse de consommations alimentaires [ZFVC11], le débruitage et l'*inpainting* d'images [MBPS10].

Cette deuxième partie du document présente mes contributions principales à la NMF, sujet de mes recherches depuis mon entrée au CNRS en 2007. La présentation est organisée en quatre chapitres.

Le chapitre 4 décrit des algorithmes de type majoration-minimisation (MM) pour la NMF fondée sur la β-divergence, une famille de mesures de dissemblance qui comprend l'erreur quadratique, la divergence de Kullback-Leibler généralisée et la divergence d'Itakura-Saito. Ce chapitre traite également de l'estimation de l'ordre de la factorisation. Il repose sur les deux publications suivantes, reproduites en partie III :

[FI11] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9) :2421–2456, Sep. 2011. `http://www.unice.fr/cfevotte/publications/journals/neco11.pdf`

[TF13] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7) :1592 – 1605, July 2013. `http://www.unice.fr/cfevotte/publications/journals/pami13_ardnmf.pdf`

Le chapitre 5 traite du cas particulier de la NMF fondée sur la divergence d'Itakura-Saito (IS-NMF) et de ses applications en traitement du signal audio. Il est montré que IS-NMF sous-tend un modèle probabiliste de la transformée de Fourier court-terme (TFCT) pertinent pour les signaux non-stationnaires composites, tels que les signaux audio. Un modèle généralisé aux signaux multicanaux avec application à la séparation de sources est présenté. Ce chapitre repose sur les deux publications suivantes, reproduites en partie III :

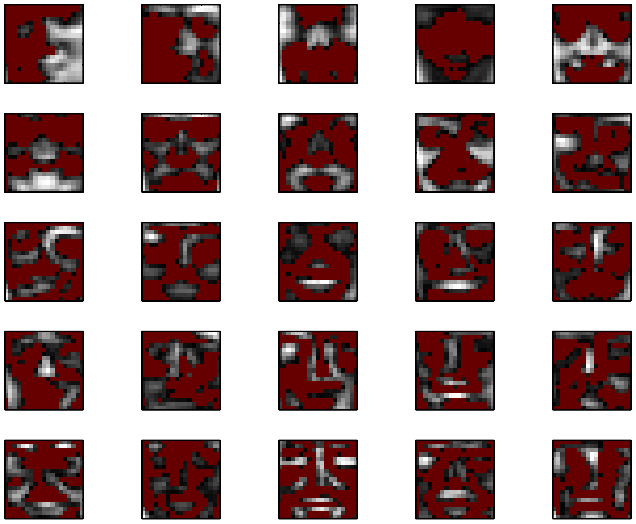
[FBD09] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3) :793–830, Mar. 2009. `http://www.unice.fr/cfevotte/publications/journals/neco09_is-nmf.pdf`

[OF10] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :550– 563, Mar. 2010. `http://www.unice.fr/cfevotte/publications/journals/ieee_asl_multinmf.pdf`

2. CBCL Face database #1, MIT Center for Biological and Computation Learning, `http://www.ai.mit.edu/projects/cbcl`

(a) 49 imagettes issues du “CBCL face dataset”

(b) Dictionnaire appris par PCA ($K = 25$)



(c) Dictionnaire appris par NMF ($K = 25$)

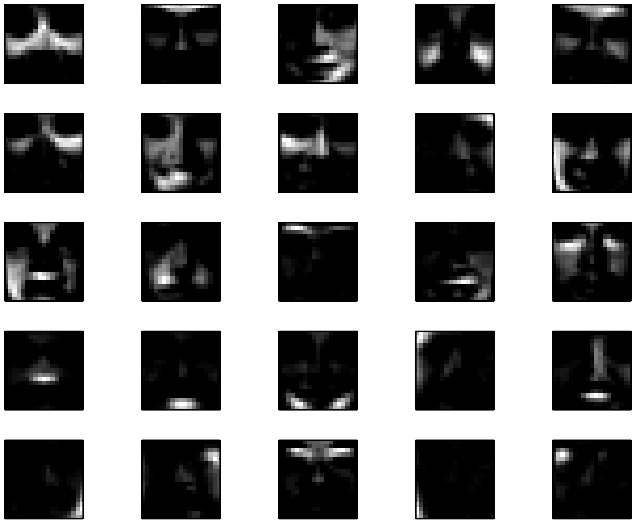


FIGURE 3.1 – Résultats de PCA et NMF appliquée au “CBCL face dataset”. La PCA fournit une représentation holistique des données alors que la NMF en fournit une décomposition par parties. La PCA soulève en outre un problème d’interprétabilité des résultats dans la mesure où le dictionnaire retourné contient des valeurs négatives (en rouge).

Le chapitre 6 présente un critère d'estimation alternatif aux approches présentées dans les chapitres 4 et 5, basé sur un traitement probabiliste approfondi de la NMF. Alors que les approches usuelles pour la NMF peuvent s'interpréter comme une estimation au sens du maximum de la vraisemblance conjointe de $\mathbf{V}$, $\mathbf{W}$ et $\mathbf{H}$, nous présentons une approche de type vraisemblance marginalisée dans laquelle la vraisemblance conjointe est intégrée selon le facteur $\mathbf{H}$. Ce dernier, dont le nombre de colonnes croît avec le nombre de données, est ainsi traité comme un paramètre de "nuisance". Cette approche se révèle avoir des propriétés intéressantes que nous expliciterons, telle que la régularisation automatique du rang de la factorisation. Ce chapitre repose sur les deux publications suivantes, reproduites en partie III :

[DF12] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10) :5163–5175, Oct. 2012.
`http://www.unice.fr/cfevotte/publications/journals/ieee_sp_mmle.pdf`

[DF11] O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.
`http://www.unice.fr/cfevotte/publications/proceedings/nips11.pdf`

Enfin, le chapitre 7 décrit succinctement les travaux réalisés dans le cadre des trois thèses que j'ai co-encadrées.

Chapitre 4

Algorithmes pour la NMF fondée sur la β-divergence

Dans la littérature sur la NMF, la factorisation (3.1) est très généralement obtenue par résolution du problème de minimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) \quad \text{s.c.} \quad \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{4.1}$$

où la notation $\mathbf{A} \geq 0$ exprime la non-négativité des coefficients de $\mathbf{A}$ (et non celle des valeurs propres) et où $D(\mathbf{V}|\mathbf{WH})$ est une mesure de dissemblance telle que

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}). \tag{4.2}$$

La fonction $d(x|y)$ est une mesure de dissemblance entre scalaires (parfois appelée fonction de coût), i.e., une fonction de $\mathbb{R}_+ \times \mathbb{R}_+$ dans $\mathbb{R}_+$ avec un unique minimum égal à zéro en $x = y$. Une fonction de coût souvent considérée pour la NMF est la β-divergence. Elle apparaît par exemple dans les publications [OP08, FCC09, FBD09, CZPA09, VBB10, DCL10, HBD10, LYO12, LDP+13, RDD13, SCY13, HBFR14]. Nous décrivons dans ce chapitre des algorithmes pour la minimisation de cette divergence.

4.1 Présentation de la β-divergence

Ce paragraphe introduit un ensemble de propriétés de la β-divergence auxquelles il sera fait référence dans ce chapitre et ceux qui suivent.

4.1.1 Définition

La β-divergence [BHHJ98, CA10] est une famille continue de divergences dont l'expression est donnée par

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log\frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{4.3}$$

La β-divergence prend comme cas particuliers l'erreur quadratique, la divergence de Kullback-Leibler (KL) généralisée et la divergence d'Itakura-Saito (IS), pour $\beta = 2, 1$ et 0, respectivement. Les cas $\beta = 1$ et $\beta = 0$ sont des cas obtenus par limite et la β-divergence offre donc un continuum de mesures de dissemblance représenté en Figure 4.1. Le paramètre β est un degré de liberté propre à la modélisation des données et sa valeur peut être ou bien fixée arbitrairement ou bien apprise sur

un jeu d'apprentissage pour un contexte et une application donnés. Pour illustration, la Figure 4.2 présente des résultats d'interpolation obtenus par NMF avec la β-divergence (nous utiliserons par la suite l'abréviation β-NMF) pour différentes valeurs de β et K. La valeur de β peut être ajustée pour obtenir la qualité visuelle désirée.

4.1.2 Comportement par rapport à l'échelle

Pour tout scalaire λ positif on a

$$d_\beta(\lambda x|\lambda y) = \lambda^\beta d_\beta(x|y). \tag{4.4}$$

Il apparaît donc que le critère $D(\mathbf{V}|\mathbf{WH})$ défini à l'équation 4.2 pénalisera davantage les écarts d'approximation sur les coefficients de large amplitude pour $\beta > 0$ que les écarts sur les coefficients de faible amplitude, tandis que pour $\beta < 0$, la factorisation reposera inversement davantage sur les coefficients de faible amplitude de $\mathbf{V}$. Cette propriété constitue un premier critère pour le choix de β, selon l'importance donnée aux faibles coefficients dans le contexte applicatif considéré. On notera que la divergence d'IS est invariante à l'échelle, i.e., $D_{\mathrm{IS}}(\lambda x|\lambda y) = D_{\mathrm{IS}}(x|y)$. Comme nous le verrons au chapitre 5, cette propriété est bienvenue pour la factorisation de spectrogrammes audio, dont l'échelle des valeurs est très contrastée, et dans lesquels faibles et grandes énergies ont une importance relativement égale du point de vue perceptif.

4.1.3 Interprétation probabiliste

La β-divergence est liée à la distribution dite de Tweedie [Twe84]. Ce lien apparaît dans [Jør87] et est rapporté dans [CZPA09]. La distribution de Tweedie est un cas particulier de la famille exponentielle à dispersion (EDM, pour *exponential dispersion model*) [Jør87], elle-même une extension de la famille exponentielle. La distribution de Tweedie est caractérisée par une relation polynomiale entre la moyenne et la variance, telle que

$$\mathrm{var}[x] = \phi\, \mu^{2-\beta}, \tag{4.5}$$

où $\mu = \mathrm{E}[x]$ est la moyenne, β est le paramètre de forme and ϕ est le paramètre de dispersion. La distribution de Tweedie n'est définie que pour $\beta \leq 1$ et $\beta \geq 2$. Pour $\beta \neq 0, 1$, sa densité de probabilité peut s'écrire

$$\mathrm{T}(x|\mu, \phi, \beta) = h(x, \phi) \exp\left[\frac{1}{\phi}\left(\frac{1}{\beta - 1} x \mu^{\beta-1} - \frac{1}{\beta}\mu^\beta\right)\right] \tag{4.6}$$

où $h(x, \phi)$ est la fonction de base. Pour $\beta \in \{0, 1\}$, la densité prend une forme limite de (4.6). Le support de $T(x|\mu, \phi, \beta)$ varie avec la valeur de β, mais le domaine de définition de μ est généralement $\mathbb{R}_+$, excepté pour $\beta = 2$, valeur pour laquelle le support est $\mathbb{R}$ et la distribution de Tweedie coïncide avec la distribution gaussienne de moyenne μ et de variance ϕ. Pour $\beta = 1$ (et $\phi = 1$), la distribution de Tweedie coïncide avec la distribution de Poisson. Pour $\beta = 0$, elle coïncide avec la distribution Gamma avec paramètre de forme $\alpha = 1/\phi$ et paramètre d'échelle μ/α. À noter que la fonction de base h n'admet une forme analytique que pour $\beta \in \{-1, 0, 1, 2\}$. Finalement, la fonction dite de *déviance* de la distribution de Tweedie, i.e., le log-ratio des vraisemblances du modèle saturé ($\mu = x$) et du modèle général est proportionnel à la β-divergence. Plus précisément, on a

$$\log \frac{\mathrm{T}(x|\mu = x, \phi, \beta)}{\mathrm{T}(x|\mu, \phi, \beta)} = \frac{1}{\phi} d_\beta(x|\mu). \tag{4.7}$$

De ce fait, la β-divergence sous-tend une log-vraisemblance pour la distribution de Tweedie, lorsque celle-ci est définie. Avec l'hypothèse que les coefficients $\{v_{fn}\}$ de $\mathbf{V}$ soient indépendants conditionnellement à $\mathbf{WH}$, la log-vraisemblance oppposée s'écrit

$$-\log p(\mathbf{V}|\mathbf{W}, \mathbf{H}) = \frac{1}{\phi} D_\beta(\mathbf{V}|\mathbf{WH}) + \mathrm{cst}. \tag{4.8}$$

En définitive, la β-NMF sous-tend un modèle probabiliste des données $\mathbf{V}$, paramétré par $\mathbf{W}$, $\mathbf{H}$, β et ϕ. Le modèle est tel que $v_{fn} \sim \mathrm{T}([\mathbf{WH}]_{fn}, \phi, \beta)$ et vérifie en particulier $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$.

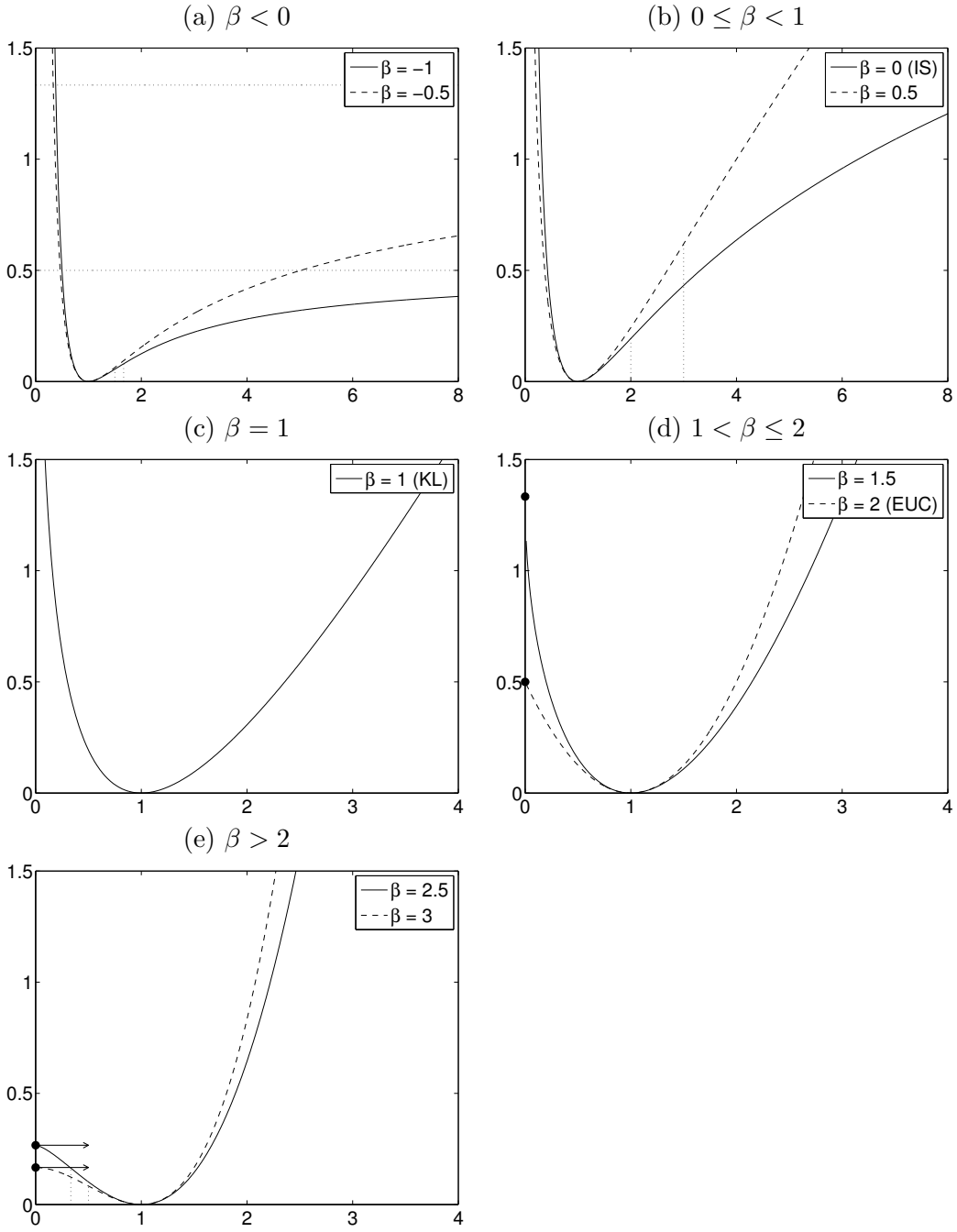


FIGURE 4.1 – β-divergence $d_\beta(x|y)$ comme fonction de y (avec $x = 1$). Les sous-figures illustrent le comportement de la β-divergence pour ses 5 régimes. La divergence est convexe pour $1 \leq \beta \leq 2$ comme le montre les sous-figures (c) et (d). Sur les autres figures, les points d'inflexion sont indiqués par les traits pointillés verticaux. Pour $\beta < 0$, la divergence possède une asymptote horizontale en $y \to \infty$ de coordonnée $x^\beta/(\beta(\beta-1))$. Pour $\beta > 1$, la divergence prend la valeur finie $x^\beta/(\beta(\beta-1))$ en $y = 0$, où la dérivée est nulle pour $\beta > 2$. (Figures reproduites de [FI11])

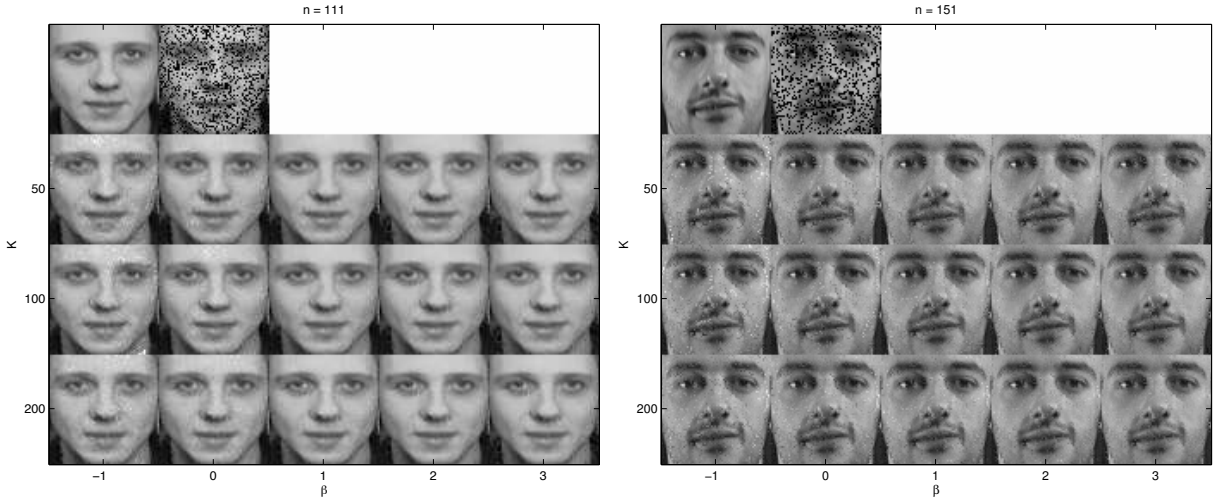


FIGURE 4.2 – Résultats d'interpolation sur le jeu de données Olivetti, constitué de 400 images de visages (40 personnes, 10 prises de vue par personne), de dimensions 64×64 et codées en 8 bits. Les images forment les colonnes de la matrice $\mathbf{V}$, dont 25% des pixels ont été éliminés au hasard. Une β-NMF est appliquée sur la base des données restantes et les pixels manquants sont reconstruits en utilisant les valeurs données par la reconstruction $\mathbf{WH}$. Les reconstructions présentées sont obtenues pour $K = \{50, 100, 200\}$ et $\beta = \{-1, 0, 1, 2, 3\}$. Perceptiblement, les reconstructions obtenus avec $\beta = 3$ s'avèrent de meilleure qualité. (Figures reproduites de [FI11])

4.2 Algorithmes

Un algorithme heuristique pour la minimisation de (4.2) a été proposé dans [CZA06]. En collaboration avec Jérôme Idier (IRCCyN, Nantes), nous avons construits de nouveaux algorithmes, garantissant la décroissance de la fonction objectif à chaque itération. Ces travaux ont abouti à la publication de l'article [FI11] reproduit en partie III. Les algorithmes proposés reposent sur la construction (locale) d'une fonction majorante (globale) de la fonction objectif. Un premier type d'algorithme, dit de majoration-minimisation, repose sur la minimisation itérative de cette fonction, donnant lieu à des mises à jour multiplicatives. Nous avons ensuite proposé un nouveau type d'algorithme, dit de majoration-égalisation, forme sur-relaxée du précédent qui produit en pratique une convergence plus rapide. Ces algorithmes sont présentés ci-après.

4.2.1 Préliminaires

Avant de s'intéresser au cas spécifique de la β-NMF, il convient de rapporter que l'essentiel des algorithmes de NMF existant dans la littérature, quelque soit la mesure de dissemblance utilisée, reposent sur l'architecture suivante. Des mises à jours itératives par bloc de $\mathbf{W}$ et $\mathbf{H}$ sont faites séparément, conditionnellement à la valeur courante de l'autre paramètre, selon le principe de l'algorithme 1.

On notera que sans contraintes supplémentaires, les mises à jours de $\mathbf{W}$ et $\mathbf{H}$ sont équivalentes, par transposition :

$$\mathbf{V} \approx \mathbf{WH} \iff \mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T. \tag{4.9}$$

Ainsi, on pourra se concentrer sur la mise à jour d'un des deux facteurs, par exemple $\mathbf{H}$ étant donnée la valeur courante de $\mathbf{W}$. Enfin, sous la forme de la fonction objectif donnée par l'équation (4.2), l'optimisation de la fonction objectif selon $\mathbf{H}$ pour $\mathbf{W}$ fixé est séparable en les colonnes de $\mathbf{H}$ (et respectivement en les lignes de $\mathbf{W}$ s'agissant de l'optimisation selon ce paramètre pour $\mathbf{H}$ fixé) :

$$D(\mathbf{V}|\mathbf{WH}) = \sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n). \tag{4.10}$$

Algorithm 1 Schéma standard de descente par bloc pour la NMF.

Initialiser $\mathbf{W}^{(0)}$
for $i = 1, \ldots, N_{iter}$ **do**
  % Descente sur $\mathbf{H}$
  Choisir $\mathbf{H}^{(i)}$ tel que $D(\mathbf{V}|\mathbf{W}^{(i-1)}\mathbf{H}^{(i)}) \leq D(\mathbf{V}|\mathbf{W}^{(i-1)}\mathbf{H}^{(i-1)})$
  % Descente sur $\mathbf{W}$
  Choisir $\mathbf{W}^{(i)}$ tel que $D(\mathbf{V}|\mathbf{W}^{(i)}\mathbf{H}^{(i)}) \leq D(\mathbf{V}|\mathbf{W}^{(i-1)}\mathbf{H}^{(i)})$
end for

Finalement, dans le schéma standard décrit par l'algorithme 1, on pourra se concentrer sur le problème d'optimisation réduit donné par

$$\min_{\mathbf{h}\geq\mathbf{0}} C(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h}), \tag{4.11}$$

dans lequel $\mathbf{v}$ est un vecteur non-négatif de dimension F, $\mathbf{W}$ est une matrice non-négative de dimensions $F \times K$ et $\mathbf{h}$ est la variable à optimiser, de dimension K. Ce dernier problème d'optimisation s'apparente à un problème de *régression linéaire non-négative*. Il a reçu une attention considérable dans le domaine de la restauration d'images, où $\mathbf{v}$ joue le rôle d'une image vectorisée bruitée, $\mathbf{W}$ celui d'un opérateur de convolution et $\mathbf{h}$ celui d'une image originale. On pourra notamment citer l'algorithme de Lucy-Richardson pour la divergence de KL généralisée (vraisemblance Poissonienne) [Ric72, Luc74], l'algorithme ISRA [DWM86, De 93] pour l'erreur quadratique et les travaux d'Eggermont *et al.* pour la divergence d'Itakura-Saito [CET99]. Certains de ces algorithmes coïncident avec des algorithmes de NMF devenus standard sans que cette filiation soit bien établie dans la littérature NMF. Nous dédions un paragraphe à ce sujet dans notre article [FI11].

4.2.2 Algorithme heuristique

Un algorithme de descente de gradient à pas adaptatif a été proposé pour la β-NMF dans [CZA06]. Le pas est réglé selon une heuristique qui permet de garantir la non-négativité des mises à jour, qui deviennent "multiplicatives". La décroissance de la fonction objectif est observée expérimentalement à chaque itération. Le même algorithme peut être obtenu par une autre heuristique, présentée dans notre article [FBD09], et décrite ici. On peut facilement montrer que la dérivée $\nabla_{h_k} C(\mathbf{h})$ de la fonction objectif (4.11) peut s'exprimer comme la différence de deux fonctions non-négatives, i.e., telle que $\nabla_{h_k} C(\mathbf{h}) = \nabla^+_{h_k} C(\mathbf{h}) - \nabla^-_{h_k} C(\mathbf{h})$. La règle de mise à jour heuristique s'écrit simplement

$$h_k^{\mathrm{H}} = \tilde{h}_k \frac{\nabla^-_{h_k} C(\mathbf{h})|_{h_k=\tilde{h}_k}}{\nabla^+_{h_k} C(\mathbf{h})|_{h_k=\tilde{h}_k}}, \tag{4.12}$$

où $\tilde{h}_k$ désigne l'itéré courant. La forme multiplicative de la mise à jour assure la non-négativité des itérées, étant données des initialisations positives. La règle de mise à jour peut être interprétée comme un algorithme de descente : h_k est mis à jour vers la gauche quand le gradient est positif et respectivement vers la droite lorsque le gradient est négatif. Un point fixe $h_k^\star$ de l'algorithme implique ou bien que $\nabla_{h_k} C(\mathbf{h})|_{h_k=h_k^\star} = 0$ ou que $h_k^\star = 0$.

Nous avons montré dans nos travaux [FI11] que cette forme multiplicative obtenue de manière heuristique (et donc sans preuve de convergence tant pour la suite de valeurs prises par la fonction objectif que pour la suite des itérées) peut être obtenue par algorithme de majoration-minimisation, parfois à un exposant près, comme expliqué dans le paragraphe suivant.

4.2.3 Algorithme de majoration-minimisation

Le principe de majoration-minimisation (MM) consiste à majorer localement à l'itéré courant $\tilde{\mathbf{h}}$ la fonction $C(\mathbf{h})$ par une fonction $G(\mathbf{h}|\tilde{\mathbf{h}})$ dont la minimisation par rapport à $\mathbf{h}$ est explicite.

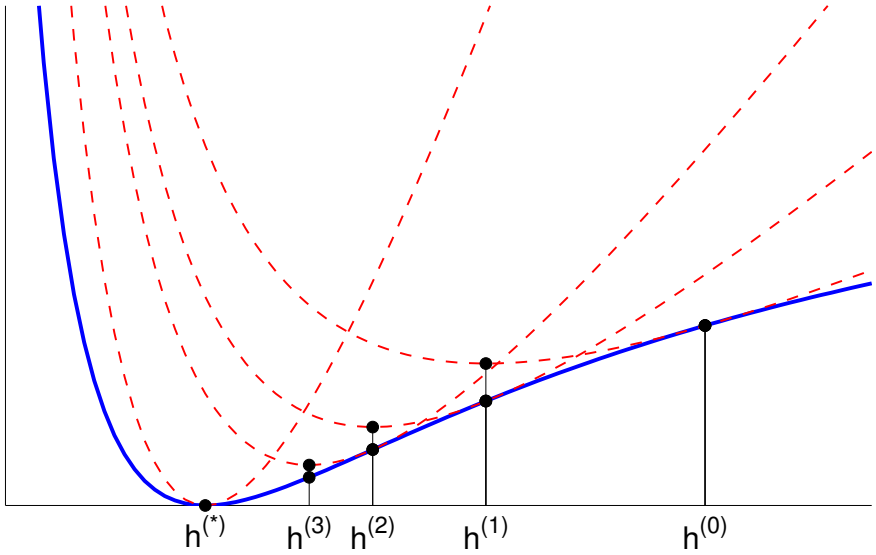


FIGURE 4.3 – Illustration du principe de majoration-minimisation (MM) sur un exemple scalaire. L'algorithme est initialisé en $h^{(0)}$. Une fonction majorante est construite en ce point. Sa minimisation produit le nouvel itéré, et ainsi de suite jusqu'à convergence. (Figure reproduite de [SFM+14])

Mathématiquement, la fonction $G(\mathbf{h}|\tilde{\mathbf{h}})$ doit vérifier $G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}) = C(\tilde{\mathbf{h}})$ et $\forall \mathbf{h}$, $G(\mathbf{h}|\tilde{\mathbf{h}}) \geq C(\mathbf{h})$. La minimisation de $G(\mathbf{h}|\tilde{\mathbf{h}})$ produit un algorithme de descente pour la minimisation de $C(\mathbf{h})$. En effet, à l'itération $(i+1)$ on a

$$C(\mathbf{h}^{i+1}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{i}). \tag{4.13}$$

Le principe MM est illustré en figure 4.3 ; un bon tutoriel de cette classe d'algorithme est donné dans [HL04].

Afin de construire une majorante pour notre problème, nous exploitons le fait que la β-divergence peut toujours s'écrire comme la somme d'un terme convexe et d'un terme concave, tel que $d(x|y) = \breve{d}(x|y) + \hat{d}(x|y) + \text{cst}$. Par exemple, dans le cas de la divergence d'IS, une décomposition naturelle (non-unique) est $\breve{d}(x|y) = x/y$ et $\hat{d}(x|y) = \log y$. Des décompositions convexe-concaves pour tous les autres cas sont donnés dans [FI11]. À noter que pour $1 \leq \beta \leq 2$, la divergence est convexe par rapport à son deuxième argument et que la partie concave est donc nulle. Il en résulte que la fonction objectif $C(\mathbf{h})$ peut elle-même s'écrire comme la somme d'un terme convexe et d'un terme concave, tel que $C(\mathbf{h}) = \breve{C}(\mathbf{h}) + \hat{C}(\mathbf{h}) + \text{cst}$, avec $\breve{C}(\mathbf{h}) = \sum_f \breve{d}(v_f|\sum_k w_{fk}h_k)$ et $\hat{C}(\mathbf{h}) = \sum_f \hat{d}(v_f|\sum_k w_{fk}h_k)$.

Étant donnée cette décomposition notre approche consiste à majorer séparément les parties convexes et concaves et à additionner les majorantes. En notant $\tilde{\mathbf{h}}$ l'itéré courant, la partie convexe peut être majorée localement en $\tilde{\mathbf{h}}$ au moyen d'une inégalité de type Jensen. En notant $\tilde{v}_f = [\mathbf{W}\tilde{\mathbf{h}}]_f$ et $\tilde{\lambda}_{fk} = (w_{fk}\tilde{h}_k)/\tilde{v}_f$, vérifiant par construction $\sum_k \tilde{\lambda}_{fk} = 1$, on a

$$\breve{C}(\mathbf{h}) = \sum_f \breve{d}(v_f|\sum_k w_{fk}h_k) \tag{4.14}$$

$$\leq \sum_f \sum_k \tilde{\lambda}_{fk} \breve{d}(v_f|\frac{w_{fk}h_k}{\tilde{\lambda}_{fk}}) \tag{4.15}$$

$$= \sum_{fk} \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \breve{d}\left(v_f|\tilde{v}_f\frac{h_k}{\tilde{h}_k}\right) \tag{4.16}$$

$$\stackrel{\text{def}}{=} \breve{G}(\mathbf{h}|\tilde{\mathbf{h}}) \tag{4.17}$$

	$\beta < 1$	$1 \leq \beta \leq 2$	$\beta > 2$
$\gamma(\beta)$	$\frac{1}{2-\beta}$	1	$\frac{1}{\beta-1}$

TABLE 4.1 – Exposant apparaissant dans les mises à jour MM.

La partie concave $\widehat{C}(\mathbf{h})$ peut quant à elle être majorée par sa tangente en $\tilde{\mathbf{h}}$, ce qui s'écrit

$$\widehat{C}(\mathbf{h}) \leq \widehat{C}(\tilde{\mathbf{h}}) + \nabla_{\mathbf{h}}^T C(\tilde{\mathbf{h}})(\mathbf{h} - \tilde{\mathbf{h}}) \tag{4.18}$$

$$= \sum_f \left[\widehat{d}(v_f|\tilde{v}_f) + \widehat{d}'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k) \right] \tag{4.19}$$

$$\stackrel{\text{def}}{=} \widehat{G}(\mathbf{h}|\tilde{\mathbf{h}}) \tag{4.20}$$

On notera que dans les deux cas, la majoration découple l'optimisation par rapport à $\mathbf{h}$ en la ramenant à une optimisation séparée par rapport à chacune de ses variables. En d'autres termes, dans les deux cas la majorante s'exprime comme une somme sur k de fonctions de h_k (et de h_k seulement).

À ce stade, on se limitera à dire que, dans le cas de la β-divergence, la majorante $G(\mathbf{h}|\tilde{\mathbf{h}}) = \breve{G}(\mathbf{h}|\tilde{\mathbf{h}}) + \widehat{G}(\mathbf{h}|\tilde{\mathbf{h}})$ résultante des opérations précédentes peut être minimisée analytiquement, voir les détails dans [FI11], produisant la mise à jour suivante

$$h_k^{\text{MM}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right)^{\gamma(\beta)}, \tag{4.21}$$

où l'exposant $\gamma(\beta)$ est donné en Table 4.1. En développant l'équation (4.12), on s'apercevrait que la mise à jour heuristique et la mise à jour MM coïncident à l'exposant $\gamma(\beta)$ près, qui vaut toujours 1 dans le cas heuristique (et donc que les mises à jour heuristiques et MM coïncident complètement pour $1 \leq \beta \leq 2$). En pratique on s'aperçoit que, lorsqu'elles ne coïncident pas, la mise à jour heuristique produit une convergence plus rapide que la mise à jour MM. Cela peut s'expliquer par le fait que la mise à jour heuristique produit des pas toujours plus grands que la mise à jour MM. En effet on peut montrer que

$$\forall \beta, \forall k, \quad |h_k^{\text{H}} - \tilde{h}_k| \geq |h_k^{\text{MM}} - \tilde{h}_k|. \tag{4.22}$$

En revanche, on ne sait toujours pas montrer que l'algorithme heuristique produit bien un algorithme de descente (au sens de la descente monotone de la fonction objectif à chaque itération), hormis pour l'intervalle $0 \leq \beta \leq 2$. La démonstration de ce résultat pour le cas $0 \leq \beta \leq 1$ est disponible dans [FI11].

4.2.4 Algorithme de majoration-égalisation

La stratégie de MM peut être poussée un pas plus loin en remplaçant la minimisation par une "égalisation". En effet, la relation (4.13) permet de construire $\mathbf{h}^{(i+1)}$ tel que $G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)})$, sans que $\mathbf{h}^{(i+1)}$ soit nécessairement le minimiseur de $G(\mathbf{h}|\mathbf{h}^{(i)})$. En particulier, on peut choisir $\mathbf{h}^{(i+1)}$ tel que $G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) = G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^i)$, lorsqu'un tel point existe (et tel que $\mathbf{h}^{(i+1)} \neq \mathbf{h}^{(i)}$). Ce cas est illustré en figure 4.4. Nous avons donné à cette approche le nom de "majoration-égalisation" (ME). L'estimateur ME, lorsqu'il existe, n'est pas facile à calculer dans le cas général. Quelques cas particuliers de valeurs de β pour lequelles l'estimateur ME peut être calculé sont donnés dans [FI11]. Pour ces cas-là, l'algorithme ME produit une convergence plus rapide (en itérations et en temps CPU) que l'algorithme MM, comme en témoignent les expériences rapportées dans [FI11].

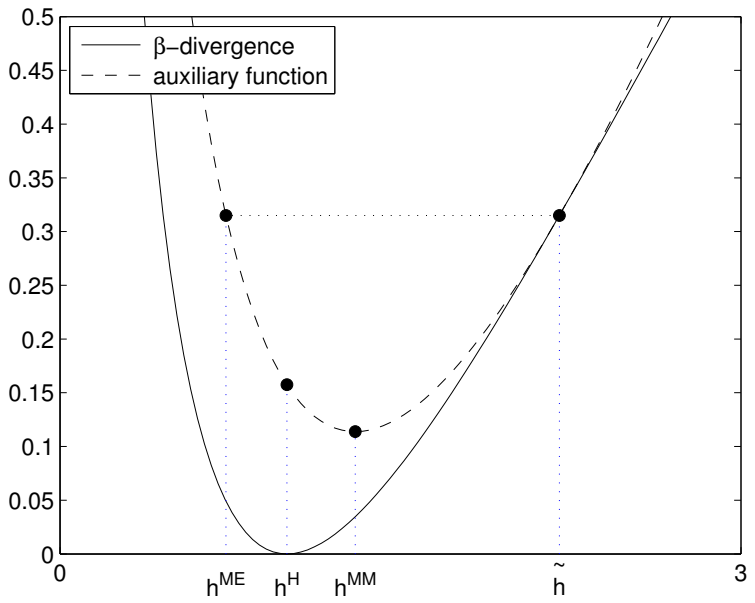


FIGURE 4.4 – β-divergence $d_\beta(x|y)$ pour $\beta = 0.5$ (avec $x = 1$) et sa fonction majorante en dimension 1 (avec $\tilde{h} = 2.2$). La mise à jour h^{MM} correspond au minimum de la fonction majorante. La mise à jour h^{ME} se trouve de l'autre côté de la "vallée" définie par la fonction auxiliaire. Dans ce cas précis la mise à jour heuristique h^{H} se situe entre les deux autres mises à jour, sans que cela soit vrai pour tout β. En outre, il apparaît que la mise à jour heuristique minimise la fonction objectif dans le cas scalaire, sans que cela soit vrai dans le cas multidimensionnel. (Figure reproduite de [FI11])

4.3 Approche pénalisée pour la détermination de l'ordre

Le problème de la NMF, tel qu'il est posé à l'équation (4.1), nécessite de spécifier l'ordre de la factorisation K. La détermination de ce paramètre est un problème important en NMF, qui n'a encore fait l'objet que d'assez peu de travaux. Avec Vincent Y. F. Tan (National University of Singapore), nous avons proposé une approche originale pour la détermination automatique de l'ordre, basée sur l'ajout d'une fonction de pénalisation à la mesure de dissemblance (4.2), et construite comme suit. Ces travaux ont fait l'objet de l'article [TF13] reproduit en partie III.

La NMF peut être envisagée comme l'approximation d'une matrice non-négative par une somme de K matrices non-négatives de rang 1, i.e.,

$$\mathbf{V} \approx \mathbf{w}_1 \underline{h}_1 + \ldots + \mathbf{w}_K \underline{h}_K, \tag{4.23}$$

où $\mathbf{w}_k$ désigne la $k^{\text{ème}}$ colonne de $\mathbf{W}$ and $\underline{h}_k$ la $k^{\text{ème}}$ ligne de $\mathbf{H}$. Nous avons proposé de pénaliser (4.2) par une fonction de la norme des vecteurs $\mathbf{w}_k$ et $\underline{h}_k$. Plus précisément, notre approche consiste à minimiser une fonctionnelle de la forme

$$C(\mathbf{W}, \mathbf{H}) = D_\beta(\mathbf{V}|\mathbf{WH}) + \lambda \sum_{k=1}^{K} \log\left(f(\mathbf{w}_k) + f(\underline{h}_k) + a\right), \tag{4.24}$$

où $f(\mathbf{x}) \geq 0$ et $f(\mathbf{x}) = 0$ si et seulement si $\mathbf{x} = 0$. Cette fonction peut être interprétée comme un critère d'estimation au sens du maximum a posteriori dans un modèle bayésien hierarchique décrit dans l'article [TF13]. La fonction $\log(x + a)$ est une pénalité concave qui favorise la parcimonie des normes des composantes matricielles de rang 1 dans la décomposition (4.23). Elle a donc pour effet d'éliminer (en mettant $\mathbf{w}_k$ et $\underline{h}_k$ à zéro) un sous-ensemble des K composantes initiales. En pratique on pourra se donner une valeur de K assez grande et laisser l'algorithme éliminer les

(a) Échantillon de données

(b) Dictionnaire **W** appris par NMF régularisée

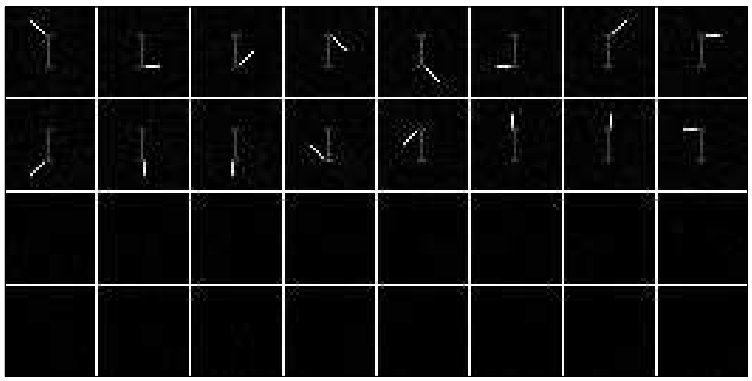


FIGURE 4.5 – (a) Échantillon d'images bruitées du jeu de données *swimmer* [DS04]. Les données sont constituées d'un nageur formé d'un tronc fixe et de quatre membres pouvant chacun prendre quatre positions. Chaque colonne de **V** est constituée des pixels d'une image. (b) Colonnes de **W** (sous forme matricielle), après décomposition par NMF régularisée de **V**, avec $K = 32$ composantes. Seize composantes sont mises exactement à zéro. Les seize autres représentent les membres du nageur, dans chacune des quatre positions possibles. Le tronc est une composante invariante qui, sans contrainte supplémentaire, ne peut être isolée. (Figures reproduites de [TF13])

composantes "inutiles" au sens du critère choisi. Ce nombre de composantes éliminées repose bien sûr sur le choix des hyperparamètres λ et a. Nous proposons une heuristique pour l'estimation de ces paramètres, basée sur l'application de la méthode des moments au modèle statistique qui sous-tend la fonctionnelle (4.24).

Nous avons construit des algorithmes de majoration-minimisation pour la résolution du problème (4.24) pour deux choix possible de $f(\mathbf{x})$: le cas $f(\mathbf{x}) = \|\mathbf{x}\|_1$, correspondant à un choix de prior exponentiel pour $\mathbf{w}_k$ et $\underline{h}_k$, et le cas $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$, correspondant à un choix de prior demi-normal (Gaussienne tronquée) pour $\mathbf{w}_k$ et $\underline{h}_k$. La figure 4.5 présente un résultat de factorisation sur un jeu de données synthétiques couramment utilisé en NMF.

Chapitre 5

Cas particulier de la NMF avec la divergence d'Itakura-Saito

Nous décrivons dans ce chapitre des applications de la NMF à un contexte spécifique : le traitement du signal audio. Le paragraphe 5.1 décrit le principe général de cette approche tandis que les paragraphes 5.2 et 5.3 décrivent mes propres contributions à cette application. Plus précisément, le paragraphe 5.2 décrit l'intérêt de l'utilisation d'une divergence particulière, la divergence d'Itakura-Saito (IS), pour la décomposition de spectrogrammes audio. Dans un second temps, le paragraphe 5.3 décrit une généralisation multicapteur de la NMF avec la divergence d'IS pour la séparation d'enregistrements musicaux multicanaux.

5.1 NMF et traitement du signal audio

De nombreuses recherches en NMF ont été guidées par des applications en traitement du signal audio telles que la transcription automatique et la séparation de sources musicales. Dans ce contexte la matrice de données $\mathbf{V}$ correspond au spectrogramme d'un signal audio monocanal que l'on cherche analyser. Appliquée à ce spectrogramme, la NMF peut extraire dans $\mathbf{W}$ les motifs spectraux des éléments se répétant dans le signal, tels que des spectres de notes, accords ou éléments percussifs. Ce principe est illustré à la figure 5.1. L'idée est due à Smaragdis & Brown dans un article de 2003 qui a connu un fort impact [SB03].

Le principe général de la décomposition de spectrogrammes par NMF décrit en figure 5.1 soulève quelques questions. Comment choisir la représentation temps-fréquence ? Quelle mesure de dissemblance choisir pour la factorisation ? Comment reconstruire les composantes séparées dans le domaine temporel, et notamment comment traiter la question de la phase ? Ces questions n'avaient été que peu traitées lorsque j'ai commencé à m'intéresser à la NMF en contexte audio et les choix relevant de ces interrogations étaient souvent faits de façon ad-hoc. Un article de Virtanen avait mis en lumière les meilleures performance de la divergence de Kullback-Leibler généralisée comparée à l'erreur quadratique pour un problème de séparation de sources monocapteur [Vir07] mais sans donner de réponse satisfaisante au problème du choix de la représentation ni à celui de la reconstruction de la phase. J'ai essayé d'apporter une réponse à toutes ces questions en adoptant une approche "générative" à la modélisation factorielle de la transformée de Fourier court-terme (TFCT).

5.2 Un modèle probabiliste à facteurs latents pour la TFCT

Cette partie rapporte des résultats de notre article [FBD09] reproduit en partie III.

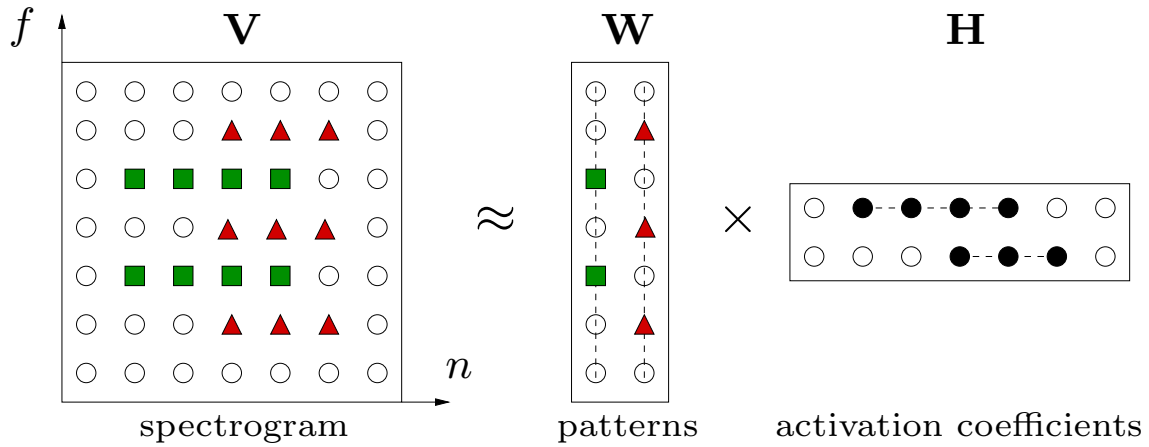


FIGURE 5.1 – Principe de la NMF appliquée à un spectrogramme schématique constitué de deux objets sonores. Les colonnes de $\mathbf{W}$ contiennent les motifs spectraux caractérisant chaque objet tandis que la matrice $\mathbf{H}$ contient les activations de ces motifs dans le spectrogramme.

5.2.1 Modèle composite Gaussien (GCM)

Nous avons proposé un modèle *composite* de la TFCT, appelé *Gaussian composite model* (GCM), construit comme suit. Soit $\mathbf{X} \in \mathbb{C}^{F \times N}$ la TFCT – à valeurs complexes – d'un signal donné $x \in \mathbb{R}^T$. Le modèle GCM est défini par

$$x_{fn} = \sum_k c_{fkn} \tag{5.1}$$

$$c_{fkn} \sim \mathrm{N}_c(0, w_{fk} h_{kn}). \tag{5.2}$$

Les variables $\{c_{fkn}\}$ définissent des composantes *latentes*. Dans la suite on désignera comme $k^{\text{ème}}$ composante la matrice temps-fréquence complexe $\mathbf{C}_k = \{c_{fkn}\}_{fn}$. Les composantes sont modélisées comme mutuellement indépendantes.

Le modèle proposé s'interprète de la façon suivante. L'équation (5.1) définit le caractère composite du modèle : la TFCT du signal observé est exprimée comme une somme exacte de composantes, i.e., $\mathbf{X} = \sum_k \mathbf{C}_k$. L'équation (5.2) définit le modèle probabiliste caractérisant les composantes, supposées distribuées selon un Gaussienne complexe circulaire. [1] Les composantes sont supposées centrées car les signaux audio sont communément supposés l'être dans le domaine temporel et cette hypothèse reste vraie dans le domaine de Fourier. [2] Elles sont modélisées par leur variance qui prend une structure de rang 1. Plus précisement, chaque composante est caractérisée par une variance spectrale $\mathbf{w}_k$ modulée en amplitude par les coefficients $\{h_{kn}\}_n$. En notation matricielle, on pourra écrire $\mathbf{C}_k \sim \mathrm{N}_c(\mathbf{0}, \mathbf{w}_k \underline{h}_k)$. Indépendamment de sa pertinence, le modèle est bien posé dans la mesure où il repose sur des hypothèses physiques réalistes : additivité des composantes dans le domaine complexe, conservation de la phase dans le modèle (bien que modélisée de manière non-informative), hypothèse de centrage des composantes, modélisation Gaussienne peu contraignante.

Venons-en maintenant au lien avec la NMF. Par conservation de la modélisation Gaussienne par additivité, les composantes latentes peuvent être facilement "marginalisées" de la distribution des

1. Une variable aléatoire complexe a pour distribution $\mathrm{N}_c(\mu, \lambda)$ si et seulement si ses parties réelles et imaginaires sont indépendantes et ont pour distribution $\mathrm{N}(\mathrm{Re}[\mu], \lambda/2)$ et $\mathrm{N}(\mathrm{Im}[\mu], \lambda/2)$, respectivement, où N désigne la loi normale réelle.

2. Si l'espérance de x_t est nulle alors l'espérance de sa transformée de Fourier $X_f = \sum_t x_t \, \mathrm{e}^{-j2\pi ft/F}$ l'est également, par linéarité de l'espérance.

observations, conduisant simplement à

$$x_{fn} \sim \mathrm{N}_c(0, [\mathbf{WH}]_{fn}) \tag{5.3}$$

La log-vraisemblance des données et des paramètres $\mathbf{W}$ et $\mathbf{H}$ s'écrit alors

$$-\log p(\mathbf{X}|\mathbf{WH}) = \sum_{fn} \left(\frac{|x_{fn}|^2}{[\mathbf{WH}]_{fn}} + \log[\mathbf{WH}]_{fn} \right) + FN \log \pi \tag{5.4}$$

$$= D_{\mathrm{IS}}(|\mathbf{X}|^2|\mathbf{WH}) + \mathrm{cst} \tag{5.5}$$

où $D_{\mathrm{IS}}(\cdot|\cdot)$ désigne la divergence d'IS définie au paragraphe 4.1, $|\mathbf{X}|^2$ désigne le spectrogramme de puissance, ayant pour coefficients $|x_{fn}|^2$, et 'cst' est un terme constant par rapport aux paramètres $\mathbf{W}$ et $\mathbf{H}$. Il apparaît que l'estimation au sens du maximum de vraisemblance des paramètres $\mathbf{W}$ et $\mathbf{H}$ dans le modèle GCM conduit à un problème de factorisation non-négative du spectrogramme de puissance $\mathbf{V} = |\mathbf{X}|^2$ avec la divergence d'IS. Réciproquement, IS-NMF sous-tend un modèle composite Gaussien de la TFCT.

Notre approche générative répond donc déjà à deux des questions posées dans le paragraphe 5.1, concernant le choix de la représentation temps-fréquence et de la mesure de dissemblance. Nous préconisons de factoriser le spectrogramme de puissance en utilisant la divergence d'IS. Par ailleurs, concernant la reconstruction des composantes séparées, notre modèle probabiliste permet de choisir un estimateur a posteriori des composantes une fois les paramètres $\mathbf{W}$ et $\mathbf{H}$ estimés. Un choix naturel pourra être la moyenne a posteriori $\mathrm{E}[c_{fkn}|\mathbf{X}, \mathbf{W}, \mathbf{H}]$ qui, sous nos hypothèses Gaussiennes, est simplement donnée par l'équation de filtrage dite de Wiener :

$$\hat{c}_{fkn} = \mathrm{E}[c_{fkn}|\mathbf{X}, \mathbf{W}, \mathbf{H}] = \frac{w_{fk}h_{kn}}{[\mathbf{WH}]_{fn}} x_{fn}. \tag{5.6}$$

Des composantes temporelles $\hat{c}_{kt}$ peuvent être ensuite reconstruites par inversion *overlap-add* des composantes temps-fréquence $\hat{\mathbf{C}}_k$ estimées. À noter que la décomposition obtenue est sans perte, à la fois dans le domaine TFCT, i.e., $\mathbf{X} = \sum_k \hat{\mathbf{C}}_k$, et dans le domaine temporel, i.e., $x_t = \sum_k \hat{c}_{kt}$.

5.2.2 Modèle de bruit multiplicatif

Le modèle composite Gaussien offre un support probabiliste clair à la décomposition de signaux par factorisation non-négative du spectrogramme. Une autre interprétation probabiliste peut être donnée à IS-NMF, dénuée de tout rapport à la factorisation de spectrogrammes. Soit $\mathbf{V}$ une matrice non-négative arbitraire et soit le modèle à bruit multiplicatif donné par

$$v_{fn} = [\mathbf{WH}] \, . \, \varepsilon_{fn} \tag{5.7}$$

$$\varepsilon_{fn} \sim \mathrm{G}(\alpha, \alpha) \tag{5.8}$$

où $\mathrm{G}(\alpha, \alpha)$ désigne la loi Gamma d'espérance égale à 1 (obtenu lorsque le paramètre de forme est égal au paramètre d'échelle). Alors on peut facilement obtenir que la log-vraisemblance s'écrit comme

$$-\log p(\mathbf{V}|\mathbf{WH}) = \alpha \, D_{\mathrm{IS}}(\mathbf{V}|\mathbf{WH}) + \mathrm{cst}, \tag{5.9}$$

où 'cst' est ici un terme constant par rapport à $\mathbf{W}$ et $\mathbf{H}$. IS-NMF sous-tend donc plus généralement un problème d'estimation au sens du maximum de vraisemblance de $\mathbf{W}$ et $\mathbf{H}$ dans un bruit multiplicatif Gamma. À noter que dans le modèle définit par (5.7)-(5.8), l'espérance conditionnelle des données est donnée par $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$. Dans le cas particulier où $\alpha = 1$ et $v_{fn} = |x_{fn}|^2$, le modèle multiplicatif se déduit du modèle GCM. Ce résultat est obtenu en utilisant la propriété que le module au carré d'une variable aléatoire Gaussienne complexe est distribué selon une loi exponentielle, cas particulier de la distribution Gamma correspondant à $\alpha = 1$.

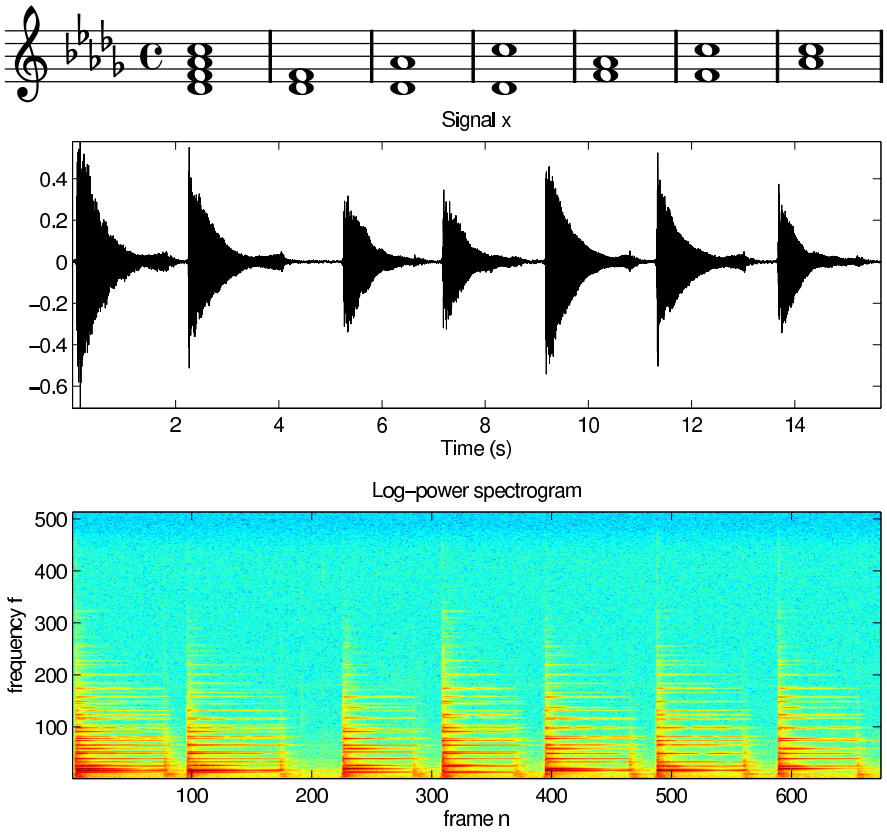


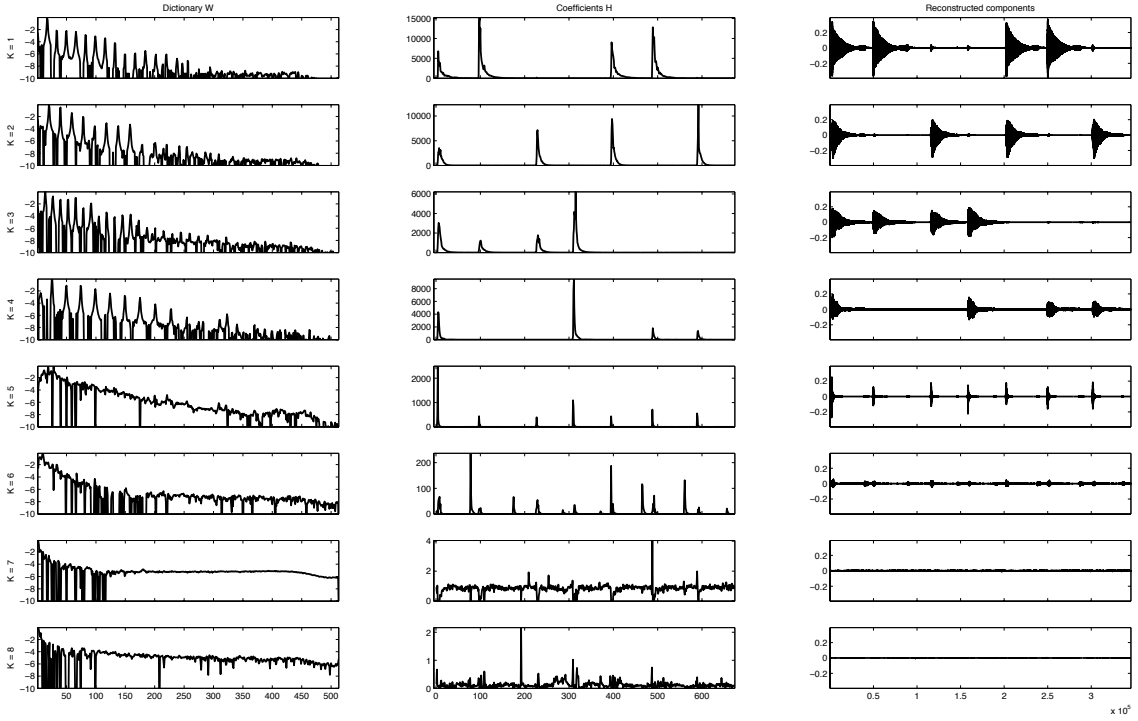
FIGURE 5.2 – Trois représentations des données : (haut) partition musicale, (centre) séquence temporelle, (bas) log-spectrogramme de puissance. En notation MIDI (protocole de communication en informatique musicale), les notes ont pour fréquence fondamentale 61, 65, 68 et 72. (Figures reproduites de [FBD09])

5.2.3 Exemple de décomposition audio

Nous avons appliqué IS-NMF ainsi qu'une approche concurrente au spectrogramme d'une séquence de piano (enregistrée en conditions réelles) composée de 4 notes jouées ensemble dans la première mesure puis par paires dans toutes les combinaisons possibles dans les mesures suivantes, voir figure 5.2. L'approche concurrente est la NMF avec la divergence de KL généralisée appliquée au spectrogramme de magnitude, tel que préconisé dans [Vir07]. IS-NMF est appliquée quant à elle au spectrogramme de puissance, tel que préconisé au paragraphe 5.2.1. Les résultats de décompositions, obtenus avec $K = 8$, sont donnés en figure 5.3. Globalement, les deux méthodes permettent d'extraire les quatre notes mélangées dans quatre composantes séparées. Cependant, la décomposition obtenue avec IS-NMF est plus précise. Les valeurs MIDI des notes séparées sont rigoureusement exactes aux valeurs réelles. En outre, une écoute des composantes estimées révèle que les $5^{\text{ème}}$ et $6^{\text{ème}}$ capturent précisément le son du marteau frappant les cordes du piano (transitoire) ainsi que le bruit mécanique de la pédale douce du piano lorsqu'elle est activée. Les $7^{\text{ème}}$ et $8^{\text{ème}}$ sont des composantes inaudibles. Un tel niveau de précision n'est pas obtenu avec la méthode KL-NMF.

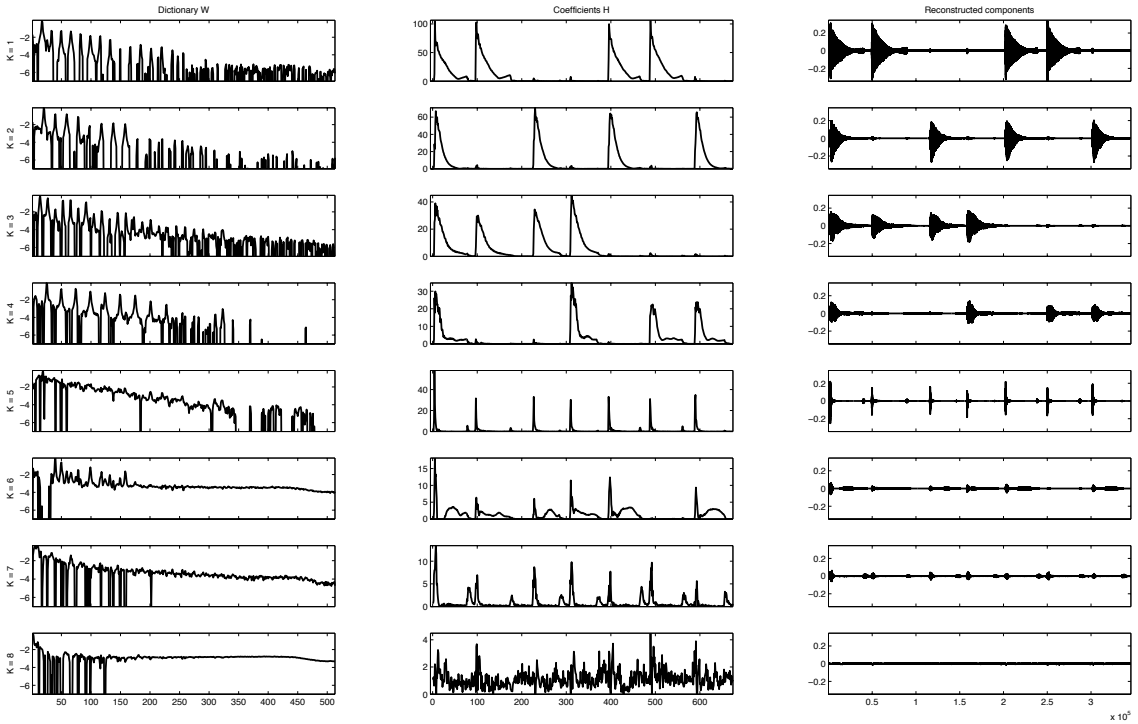
Dans l'article [KFS12], nous avons rapporté des résultats de simulations plus approfondies sur un problème de séparation de locuteurs. Nous avons dans cet article utilisé la NMF avec la β-divergence et comparé les performances de séparation obtenues pour plusieurs valeurs de β et plusieurs valeurs de l'exposant appliqué au spectrogramme (magnitude, puissance et valeur intermédiaire). Il en ressort que les meilleurs résultats sont obtenus pour des valeurs de β dans le voisinage de zero, i.e., pour des divergences proches de la divergence d'IS. En particulier, parmi les différents critères

(a) IS-NMF appliquée au spectrogramme de puissance



Fréquences fondamentales MIDI estimées : 65.0 68.0 61.0 72.0 0 0 0 0

(b) KL-NMF appliquée au spectrogramme de magnitude



Fréquences fondamentales MIDI estimées : 65.2 68.2 61.0 72.2 0 56.2 0 0

FIGURE 5.3 – Résultats de décompositions avec (a) IS-NMF et (b) KL-NMF ; (gauche) colonnes de **W** en échelle logarithmique, (centre) lignes de **H** en échelle linéaire, (droite) composantes temporelles reconstruites avec le filtre de Wiener défini à l'équation (5.6). (Figures reproduites de [Fév10])

d'évaluation utilisés, la NMF avec la divergence IS appliqué au spectrogramme de puissance (correspondant donc au modèle GCM) donne en moyenne les meilleurs résultats en terme de réjection des sources interférentes (critère SIR introduit dans [VGF06]). D'autres auteurs ont fait des études expérimentales similaires pour d'autres applications, par exemple en transcription [VBB10, DCL10], et il en ressort qu'il est généralement souhaitable d'utiliser en contexte audio des valeurs de β entre 0 et 0,5. D'autres auteurs encore ont mis en œuvre un test statistique de sélection de modèle en utilisant l'interprétation probabiliste de la β-divergence pour des spectrogrammes audio et ont rapporté que le modèle GCM obtenait le meilleur score [LYO12].

Notre article [FBD09] décrivant nos travaux sur IS-NMF, tels que rapportés dans cette partie, connaît un bon écho dans la communauté audio, avec plus de 300 citations 5 ans après sa publication. À noter que dans le contexte audio, une forme de corrélation entre les colonnes successives de $\mathbf{H}$ est attendue, en raison de la forte corrélation temporelle qui existe dans pour ces signaux. Ainsi, une part de mon activité s'est aussi consacrée à des formes de IS-NMF contrainte par un modèle dynamique des coefficients d'activation. Un paragraphe y est consacré dans l'article [FBD09] reproduit en partie III. D'autres variantes sont publiées dans [Fév11, FLH13]. Enfin, j'ai co-écrit avec Paris Smaragdis et d'autres auteurs un article donnant une vue d'ensemble des travaux dédiés aux modèles de NMF dynamique, paru dans *IEEE Signal Processing Magazine* [SFM+14].

5.3 IS-NMF multicanal pour la séparation de sources musicales

Les techniques d'analyse audio basées sur la factorisation du spectrogramme concernent les signaux monocanal. Or, les enregistrements audio sont souvent disponibles dans des formats multicanaux, en particulier stéréo. En collaboration avec Alexey Ozerov, post-doc à Télécom ParisTech sur le projet ANR SARAH en 2008–2009, nous avons généralisé le modèle GCM au contexte multicanal. Ces travaux ont été publiés dans l'article [OF10] reproduit en partie III. Dans le modèle proposé, les signaux observés sont un mélange convolutif de signaux sources, dont le spectrogramme est modélisé par un modèle GCM. Le mélange convolutif est approché par un modèle linéaire instantané dans chaque canal de fréquence. Mathématiquement, en notant $x_{i,fn}$ la TFCT de la $i^{\text{ème}}$ observation et $s_{j,fn}$ la TFCT de la $j^{\text{ème}}$ source, notre modèle s'écrit

$$x_{i,fn} = \sum_j a_{ij,f} s_{j,fn}, \tag{5.10}$$

$$s_{j,fn} \sim \mathrm{N_c}(0, [\mathbf{W}_j \mathbf{H}_j]_{fn}), \tag{5.11}$$

où les coefficients de mélange $\{a_{ij,f}\}$, à valeurs complexes et dépendant de la fréquence f, modélisent une convolution de type bande étroite (durée de réverbération bien inférieure à la taille de la fenêtre d'analyse). Le modèle est illustré en figure 5.4. Nous avons proposé un algorithme de type espérance-minimisation (EM) permettant l'estimation des filtres de mélange et des paramètres NMF de chaque source. L'algorithme repose sur une architecture dans laquelle les sources $\{s_{j,fn}\}$ sont traitées comme données manquantes, à la manière de [MCG97, CSDP02, FC05]. La méthode est fonctionnelle aussi bien en contexte sous-déterminé qu'en contexte sur-déterminé, bien qu'évidemment plus performante dans le second cas. Notre méthode a obtenu les meilleurs scores parmi 14 autres méthodes pour la tâche "underdetermined speech and music mixtures" lors de la campagne d'évaluation internationale SiSEC 2008 (Signal separation evaluation campaign). [3]

Une variante semi-supervisée de cette approche, dans laquelle l'utilisateur peut "informer" le système de séparation des plages d'activation de chacune des sources, a fait l'objet d'un dépôt de brevet US en 2011 avec la société Audionamix, ainsi que d'une publication [OFBD11]. Au sein du projet ANR SARAH, cette technique a été utilisée par des ingénieurs du son des studios Copra pour le remélange en son 3D d'enregistrements réels issus de CDs commerciaux. C'est à ma connaissance l'une des premières fois qu'une technique de séparation de sources musicales a produit des résultats exploitables dans le cadre d'une application réelle.

3. `http://www.irisa.fr/metiss/SiSEC08/SiSEC_underdetermined/test_eval.html`

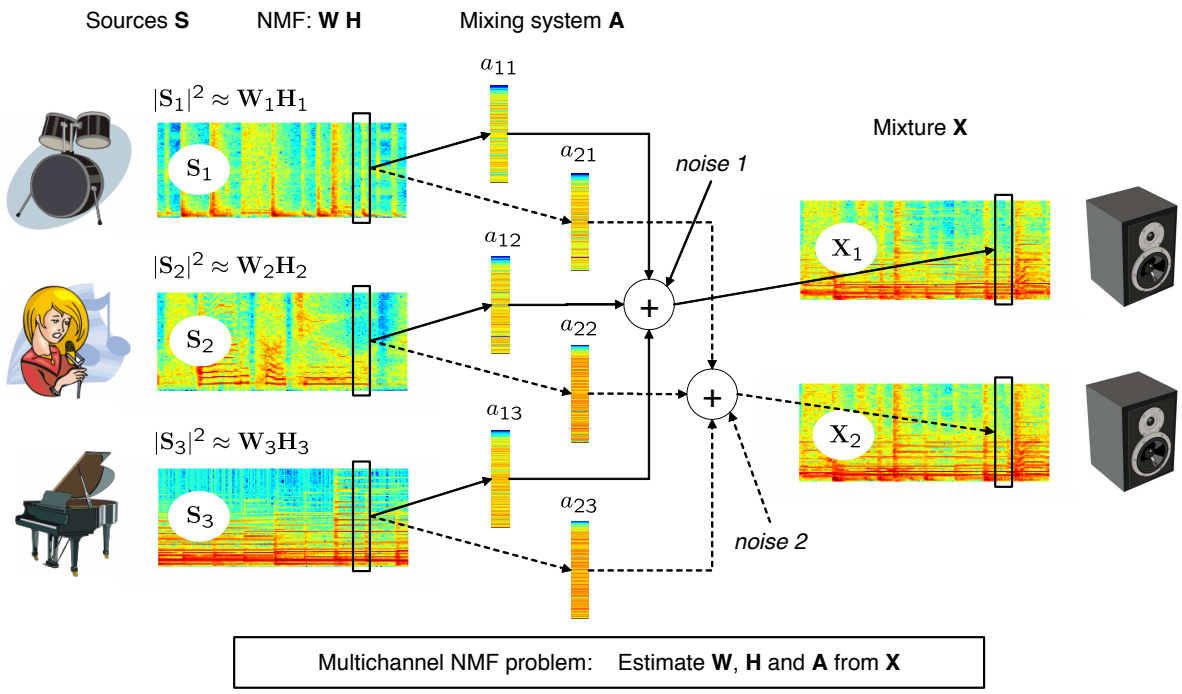


FIGURE 5.4 – Modèle IS-NMF multicanal. (Figure reproduite de [OF10])

On notera pour conclure ce chapitre, que les équations (5.10)–(5.11) définissent un modèle pour le tenseur formé des TFCTs des différents canaux. À ce titre nous avons également étudié les résultats de décompositions tensorielles canoniques appliquées aux spectrogrammes de puissance des canaux [FO10], comme d'autres auteurs l'avait aussi fait avant nous [FCC05]. L'architecture de descente par bloc et mise à jour MM décrite au chapitre 4 peut facilement se généraliser à la décomposition CANDECOMP/PARAFAC non-négative, présentée par exemple dans [LC09]. Comme on pourra s'en douter les résultats ne sont pas aussi satisfaisants qu'avec l'approche NMF multicanal proposée, qui repose sur une modélisation plus adéquate des données audio à analyser.

Chapitre 6

Estimation par maximum de vraisemblance marginalisée

Nous avons vu au paragraphe 4.1.3 que les mesures de dissemblance utilisées en NMF peuvent souvent être ramenées à une log-vraisemblance des données $\mathbf{V}$ et des paramètres $\mathbf{W}$, $\mathbf{H}$, tel que

$$\underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}}\ D(\mathbf{V}|\mathbf{WH}) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}}\ -\log p(\mathbf{V}|\mathbf{WH}). \tag{6.1}$$

Ainsi l'erreur quadratique sous-tend un bruit additif gaussien. Une divergence de Kullback-Leibler généralisée sous-tend un bruit Poisson. Une divergence d'Itakura-Saito sous-tend un bruit multiplicatif Gamma. Le problème de la NMF posé selon (4.1) peut donc être ramené à une estimation conjointe au sens du maximum de vraisemblance des paramètres $\mathbf{W}$ et $\mathbf{H}$ sous certaines hypothèses probabilistes. Cependant, l'estimation est mal posée car le nombre de paramètres croît avec le nombre d'observations ; en effet chaque colonne $\mathbf{v}_n$ implique un paramètre $\mathbf{h}_n$. Nous avons ainsi cherché à poser différemment le problème de la NMF en traitant $\mathbf{H}$ comme une variable latente, associée à un prior, et "marginalisée" de la fonction de vraisemblance. On cherche ainsi à optimiser selon $\mathbf{W}$ la vraisemblance marginale définie par

$$p(\mathbf{V}|\mathbf{W}) = \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{WH})p(\mathbf{H})d\mathbf{H}. \tag{6.2}$$

La vraisemblance marginale n'est en général pas facile à optimiser, ni même disponible en forme close. En collaboration avec Onur Dikmen, post-doctorant à Télécom ParisTech de 2010 à 2012 recruté sur le projet ANR TANGERINE, nous avons proposé pour deux modèles probabilistes (les modèles Gamma-Poisson et Gamma-Exponentiel décrits au paragraphe 6.1) des algorithmes d'inférence variationnelle et stochastique pour l'évaluation et l'optimisation de la vraisemblance marginale [DF11, DF12]. Ces deux articles sont reproduits en partie III. Comparée à l'approche NMF standard (équivalente à l'optimisation de la vraisemblance conjointe de $\mathbf{W}$ et $\mathbf{H}$), nous avons constaté que l'estimé de $\mathbf{W}$ par cette méthode possède de meilleures capacités de description et de prédiction des données. En particulier, une forme d'auto-régularisation des colonnes de $\mathbf{W}$ a été observée : les colonnes "superflues" de $\mathbf{W}$ sont automatiquement mises à zéro, de manière similaire à l'approche décrite au paragraphe 4.3, mais sans nécessité de pénaliser explicitement le terme d'attache aux données (et en particulier de choisir λ).

Nous décrivons au paragraphe 6.1 les modèles probabilistes auxquels nous avons appliqué notre étude. Puis, nous proposons au paragraphe 6.2 une comparaison conceptuelle et empirique de l'estimateur de maximum de vraisemblance marginalisée et de l'estimateur plus habituel de maximum de vraisemblance conjointe.

6.1 Modèles considérés

Le premier modèle considéré pour notre étude est le modèle dit "Gamma-Poisson" (GaP) [Can04] défini par

$$v_{fn} \sim \text{Pois}([\mathbf{WH}]_{fn}), \tag{6.3}$$
$$h_{kn} \sim \text{G}(\alpha_k, \beta_k). \tag{6.4}$$

Il constitue un modèle de données entières. Il a en particulier été beaucoup utilisé pour l'analyse sémantique de documents textuels [Can04, BJ06]. Dans ce contexte v_{fn} correspond au nombre d'occurrences du mot f dans le document n. Les colonnes du dictionnaire $\mathbf{W}$ peuvent s'interprèter comme des sujets thématiques (e.g., "politique", "sport", "culture", etc.), caractérisés chacun par la cooccurrence de certains mots. Les coefficients de $\mathbf{h}_n$ donnent les proportions de chacun des sujets dans le document n.

Le deuxième modèle considéré, que nous avons appelé "Gamma-Exponentiel" (GEx), correspond au modèle de bruit multiplicatif exponentiel présenté au paragraphe 5.2.2, accompagné d'un prior Gamma pour les activations, i.e.,

$$v_{fn} \sim \text{Exp}([\mathbf{WH}]_{fn}), \tag{6.5}$$
$$h_{kn} \sim \text{G}(\alpha_k, \beta_k). \tag{6.6}$$

Comme nous l'avons vu au paragraphe 5.2.1, lorsque v_{fn} est homogène à un spectrogramme de puissance, le modèle sous-tend un modèle GCM de la TFCT, avec des activations de distribution Gamma.

Dans les deux modèles, on suppose comme toujours les données conditionnellement indépendantes, tel que $p(\mathbf{V}|\mathbf{WH}) = \prod_{fn} p(v_{fn}|[\mathbf{WH}]_{fn})$. Par ailleurs, les deux modèles vérifient $\text{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$. Le paramètre de forme α_k dans le prior pour $\mathbf{H}$ est supposé fixé. Le prior pour $\mathbf{H}$ reste pour le moment conditionnellement dépendant aux paramètres d'échelle $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]^T$, dont le rôle sera traité au paragraphe 6.2.2, ce que nous noterons par $p(\mathbf{H}|\beta)$. Enfin, on suppose les activations indépendantes a priori, de sorte que $p(\mathbf{H}|\boldsymbol{\beta}) = \prod_{kn} p(h_{kn}|\alpha_k, \beta_k)$.

6.2 Comparaison des estimateurs MJLE et MMLE

6.2.1 Définitions

Étant donné le modèle d'observation $p(\mathbf{V}|\mathbf{WH})$ et le prior $p(\mathbf{H}|\boldsymbol{\beta})$, dans lequel $\boldsymbol{\beta}$ est un paramètre d'échelle, l'estimateur MJLE de $\mathbf{W}$, acronyme du vocable anglais *maximum joint likelihood estimation*, est obtenu par optimisation de la log-vraisemblance pénalisée, définie par

$$C_{\text{JL}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) = -\log p(\mathbf{V}, \mathbf{H}|\mathbf{W}, \boldsymbol{\beta}) \tag{6.7}$$
$$= -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{H}|\boldsymbol{\beta}). \tag{6.8}$$

Les initiales JL sont l'abréviation de *joint likelihood*. Dans la mesure où la log-vraisemblance $-\log p(\mathbf{V}, \mathbf{H}|\mathbf{W}, \boldsymbol{\beta})$ sous-tend une divergence, l'optimisation de $C_{\text{JL}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta})$ définit un problème de NMF pénalisée par le terme $-\log p(\mathbf{H}|\boldsymbol{\beta})$. Pour les modèles GaP et GEx considérés, des algorithmes MM peuvent être mis en œuvre en suivant les principes méthodologiques présentés au paragraphe 4.2.

L'estimateur MMLE de $\mathbf{W}$, acronyme du vocable anglais *maximum marginal likelihood estimation*, est obtenu quant à lui par optimisation de la log-vraisemblance marginalisée, définie par

$$C_{\text{ML}}(\mathbf{W}, \boldsymbol{\beta}) = -\log p(\mathbf{V}|\mathbf{W}, \boldsymbol{\beta}) \tag{6.9}$$
$$= -\log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{WH}) p(\mathbf{H}|\boldsymbol{\beta}) d\mathbf{H}. \tag{6.10}$$

Les initiales ML sont l'abréviation de *marginal likelihood*. Des algorithmes de type EM sont présentés au paragraphe 6.2.3 pour l'optimisation de $C_{\mathrm{ML}}(\mathbf{W}, \boldsymbol{\beta})$. À noter que l'estimation MMLE est un pur problème d'apprentissage de dictionnaire dans le sens où la fonction objectif $C_{\mathrm{ML}}(\mathbf{W}, \boldsymbol{\beta})$ ne fait plus apparaître la variable $\mathbf{H}$ dans ses arguments. Une fois l'estimée $\hat{\mathbf{W}}$ obtenue, les coefficients d'activation doivent être reconstruits a posteriori (si nécessaire), par exemple au sens MAP en minimisant $-\log p(\mathbf{V}, \mathbf{H}|\hat{\mathbf{W}})$ par rapport à $\mathbf{H}$.

6.2.2 Comportement par rapport à l'échelle

Une différence significative entre les estimateurs MMLE et MJLE est leur comportement par rapport aux échelles respectives de $\mathbf{W}$ et $\mathbf{H}$. Contrairement à l'estimateur MJLE, l'estimateur MMLE est invariant à l'échelle dans le sens suivant. Soit $\boldsymbol{\Lambda}$ une matrice diagonale non-négative ayant pour coefficients $\{\lambda_k\}$. Par un changement de variable dans l'intégrale apparaissant dans la définition de la vraisemblance marginalisée, voir équation (6.2), on peut montrer la relation suivante

$$C_{\mathrm{ML}}(\mathbf{W}\boldsymbol{\Lambda}^{-1}, \boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{\mathrm{ML}}(\mathbf{W}, \boldsymbol{\beta}). \tag{6.11}$$

Cette relation indique qu'un changement d'échelle coordonné des variables $\mathbf{W}$ et $\mathbf{H}$ laisse la valeur du critère MMLE inchangée. À noter que cette relation n'est pas spécifique au modèles GaP et GEx étudiés mais vraie pour tout modèle $p(\mathbf{V}|\mathbf{WH})$, $p(\mathbf{H}|\boldsymbol{\beta})$ dans lequel $\boldsymbol{\beta}$ est un paramètre d'échelle. La relation (6.11) soulève en corollaire une indétermination d'échelle que l'on pourra lever en fixant par exemple $\beta_k = 1$ et en laissant $\mathbf{W}$ libre. Le critère MJLE vérifie quant à lui la relation suivante (vraie pour tout modèle d'observation $p(\mathbf{V}|\mathbf{WH})$ mais spécifique au choix d'un prior Gamma pour $\mathbf{H}$) :

$$C_{\mathrm{JL}}(\mathbf{W}\boldsymbol{\Lambda}^{-1}, \boldsymbol{\Lambda}\mathbf{H}, \boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{\mathrm{JL}}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) + N \sum_k \log \lambda_k. \tag{6.12}$$

On voit donc qu'un changement d'échelle coordonné des variables $\mathbf{W}$ et $\mathbf{H}$ modifie la valeur du critère MJLE. En plus d'être peu naturelle, cette propriété est problématique dans la mesure où elle conduit à un problème d'optimisation mal conditionné. En effet, le terme $\sum_k \log \lambda_k$ peut être arbitrairement rendu petit, conduisant à une solution dégénérée telle que $\|\mathbf{W}\| \to \infty$, $\|\mathbf{H}\| \to 0$, $\|\boldsymbol{\beta}\| \to 0$. On peut montrer que ce problème peut être résolu en fixant $\boldsymbol{\beta}$ (par exemple, en supposant $\beta_k = 1$) dans le cas où le paramètre d'échelle vérifie $\alpha_k > 1$. Dans l'autre cas, $\alpha_k \leq 1$, l'optimisation du critère MJLE nécessite de contrôler explicitement la norme de $\mathbf{W}$.

6.2.3 Algorithmes EM pour l'optimisation de la vraisemblance marginalisée

Étant donné le modèle probabiliste d'observation $p(\mathbf{V}|\mathbf{WH})$ et le prior $p(\mathbf{H})$ (à la lumière du paragraphe précédent, on suppose désormais $\boldsymbol{\beta}$ fixé), notre objectif est de minimiser la fonction $C_{\mathrm{ML}}(\mathbf{W})$ définie à l'équation (6.2). Le caractère latent de la variable $\mathbf{H}$ dans le formalisme MMLE suggère l'utilisation d'un principe EM, reposant sur l'optimisation itérative de l'espérance conditionnelle des données $\mathbf{V}$ complétées des données latentes $\mathbf{H}$,

$$Q(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V}, \mathbf{H}|\mathbf{W}) p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}}) d\mathbf{H}. \tag{6.13}$$

Malheureusement, dans la plupart des cas et en particulier le cas des modèles GaP et GEx, la fonctionnelle n'est ni évaluable ni optimisable directement. La forme analytique de la distribution a posteriori des variables latentes $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$ n'est en particulier pas disponible. Nous avons eu donc recours à deux types d'alternative, l'algorithme Monte-Carlo EM (MC-EM) [WT90] et l'algorithme EM variationnel (VB-EM, pour *variational Bayes EM*) [BG03].

Algorithme Monte-Carlo EM. Pour notre problème, l'algorithme MC-EM consiste à générer à chaque itération des échantillons $\mathbf{H}^{(i)}$ selon la distribution $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$ et à effectuer une approximation Monte-Carlo de l'intégrale apparaissant dans la définition de $Q(\mathbf{W}|\tilde{\mathbf{W}})$:

$$Q^{\mathrm{MC}}(\mathbf{W}|\tilde{\mathbf{W}}) = -\sum_i \log p(\mathbf{V}, \mathbf{H}^{(i)}|\mathbf{W}). \tag{6.14}$$

L'approximation est asymptotiquement exacte, ce qui offre des garanties fortes à l'algorithme MC-EM. En revanche, en pratique la génération d'un nombre "suffisant" d'échantillons de $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$ est une étape coûteuse qui peut faire de l'algorithme EM une méthode lente. En particulier, pour les modèles GaP et GEx que nous avons considérés, l'échantillonnage stochastique de $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$ est réalisé au moyen d'un échantillonneur de Gibbs basé sur l'introduction de variable latentes tierces (équivalentes aux composantes $\{c_{fkn}\}$ apparaissant dans le modèle GCM). Le paragraphe suivant décrit une solution plus rapide reposant sur une approximation variationnelle de $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$.

Algorithme EM variationnel. Pour notre problème, l'algorithme VB-EM consiste à approcher la distribution a posteriori des données latentes par une distribution permettant le calcul et l'optimisation de la fonctionnelle approchée qui en résulte. Plus précisément, on forme à chaque itération l'approximation $q(\mathbf{H}) \approx p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$, où la distribution dite variationelle $q(\mathbf{H})$ minimise la divergence de KL en distribution entre $q(\mathbf{H})$ et $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$, dans une classe de distribution choisie. Pour le modèle GaP, le choix d'une distribution Gamma s'impose au fil des calculs comme un choix naturel [DF12]. Une distribution inverse-Gamma généralisée (distribution appartenant à la famille exponentielle qui unifie les distributions Gamma et inverse-Gamma) s'impose de la même manière pour le modèle GEx [DF11]. Une fois l'étape d'approximation réalisée, on optimise la fonctionnelle

$$Q^{\mathrm{VB}}(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V}, \mathbf{H}|\mathbf{W}) q(\mathbf{H}) d\mathbf{H} \tag{6.15}$$

au lieu de la fonctionnelle initiale $Q(\mathbf{W}|\tilde{\mathbf{W}})$. La qualité de l'estimateur obtenu pour $\mathbf{W}$ dépend bien évidemment de la qualité de l'approximation faite de $p(\mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$.

6.3 Expériences

Dans une première expérience, nous avons généré des données synthétiques selon le modèle GaP. La matrice dictionnaire "vérité terrain" $\mathbf{W}^\star$, de dimensions $F = 10$ et $K^\star = 5$, est une matrice aléatoire composée de 0 et 1. La matrice activation $\mathbf{H}^\star$ est générée selon une loi exponentielle ($\alpha_k = 1$) de paramètre d'échelle $\beta_k = 1$. Les estimateurs MMLE (obtenus avec les algorithmes VB-EM et MC-EM) et MJLE (obtenu par algorithme MM) sont calculés pour $K = 1, \ldots, 10$. Les valeurs finales des fonctions objectif retournées après convergence sont représentées en fonction de K en figure 6.1. [1]

On constate premièrement que les résultats retournés par VB-EM et MC-EM dans le cas de l'estimation MMLE sont similaires, ce qui montre que l'approximation variationnelle utilisée est une approximation satisfaisante pour le modèle proposé. Ensuite, on constate que la vraisemblance marginale cesse de croître après $K = K^\star = 5$. Une inspection des colonnes de $\hat{\mathbf{W}}$ révèle que cela est dû au fait que $K - K^\star$ colonne(s) sont mises à zéro. Il apparaît ainsi empiriquement que l'estimateur MMLE a la capacité d'auto-régulariser le rang de la décomposition. Cette propriété est confirmée dans les autres expériences qui suivent. A contrario, la vraisemblance conjointe des estimées au sens MJLE continue de croître légèrement pour $K > 5$, ce qui est symptomatique d'un effet de sur-apprentissage.

1. En raison d'un changement de convention entre ce document et les articles dont les figures sont extraites, la figure 6.1 représente des valeurs de log-vraisemblance au lieu de valeurs de log-vraisemblance opposées, d'où le signe '-' apparaissant devant les fonctions C_{ML} et C_{JL} dans les légendes de ces figures.

(a) MMLE par VB-EM (b) MMLE par MC-EM (c) MJLE par MM

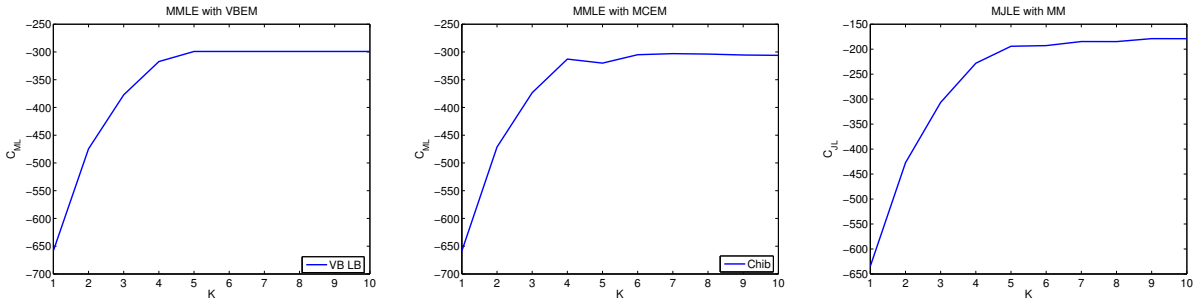


FIGURE 6.1 – Figures (a) & (b) : valeur de la fonction objectif $-C_{\mathrm{ML}}(\hat{\mathbf{W}})$ obtenue pour $K = 1, \ldots, 10$. Dans le cas (a), algorithme VB-EM, la fonction objectif est approchée par l'énergie variationnelle associée à l'approximation $q(\mathbf{H})$. Dans le cas (b), algorithme MC-EM, la fonction objectif est estimée par la méthode de Chibb [Chi95]. Dans les deux cas, voir les détails dans [DF12]. Figure (c) : valeur de la fonction objectif $-C_{\mathrm{JL}}(\hat{\mathbf{W}}, \hat{\mathbf{H}})$. La fonction objectif $-C_{\mathrm{ML}}(\hat{\mathbf{W}})$ devient constante après $K = K^\star = 5$. La fonction $-C_{\mathrm{JL}}(\hat{\mathbf{W}}, \hat{\mathbf{H}})$ continue de croître légèrement, symptôme de sur-apprentissage. (Figures reproduites de [DF12])

Afin de visualiser l'effet d'auto-régularisation du rang produit par l'estimation MMLE, nous avons utilisé le jeu de données *swimmer* présenté à la figure 4.5. Les échantillons originaux ont été dégradés avec un bruit multiplicatif exponentiel et nous avons produit des estimateurs MMLE et MJLE du dictionnaire sous les hypothèse du modèle GEx, avec $K = 16$. Comme précédemment, le prior pour $\mathbf{H}$ est choisi comme exponentiel avec $\alpha_k = \beta_k = 1$. Les résultats, issu de l'article [DF11] reproduit en partie III, sont présentés en figure 6.2. Il apparaît clairement que l'estimateur MMLE retourne les 4 membres du nageur dans leurs 4 positions possibles alors que l'estimateur MJLE sur-apprend les données ; en l'occurrence quelques membres apparaissent comme dupliqués.

Les performances des estimateurs MJLE et MMLE ont été comparées sur des jeux de données réelles dans nos publications [DF11, DF12]. Dans l'article [DF12] nous considérons l'analyse sémantique de données textuelles basée sur le modèle GaP. Les données sont constituées des paroles de 10.000 chansons, sous forme de "sac de mots" (*bag of words*). L'analyse de ces données permet d'extraire des sujets récurrents, sur la base desquelles les chansons peuvent par exemple être classées. Par ailleurs, nous considérons également dans cet article un problème d'interpolation d'images tel que décrit en figure 4.2. Il est montré que le dictionnaire retourné par MMLE possède de meilleures capacités de prédiction que le dictionnaire MJLE. Dans l'article [DF11], nous considérons la décomposition des 40 premières secondes de la chanson *God Only Knows* des Beach Boys au moyen de l'estimateur MMLE et sur la base du modèle GEx. Dans ces trois cas, le même effet d'auto-régulation du rang par l'estimation MMLE est observé.

Nous n'avons pas encore su apporter une explication théorique à ce phénomène surprenant, et en particulier spécifier les conditions précises dans lesquels il apparaît. Ce phénomène apparaît aussi dans un contexte légèrement différent de factorisation bayésienne de matrices réelles, dans un modèle additif Gaussien [NSBT13], mais les résultats théoriques de ce travail ne sont pas facilement applicable aux modèles non-négatifs que nous avons étudiés (en raison notamment de l'absence d'une expression analytique du minimiseur global de $D(\mathbf{V}|\mathbf{WH})$). Dans l'article [DF12] nous proposons néanmoins un début d'explication basé sur une approximation de Laplace de la vraisemblance marginalisée autour de la valeur MAP de $\mathbf{H}$. Cette approximation permet de rapprocher la vraisemblance marginalisée de la vraisemblance conjointe pénalisée par un terme de régularisation de groupe sur les colonnes de $\mathbf{W}$ correspondant bien au phénomène observé.

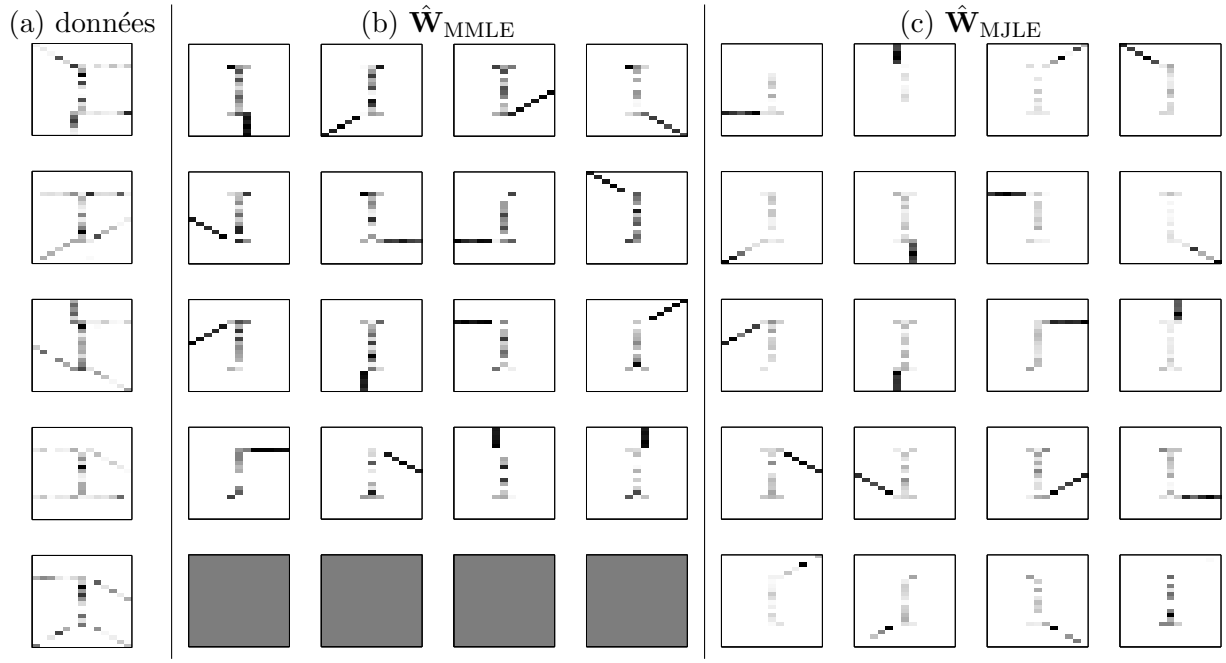


FIGURE 6.2 – Résultats de décomposition des données *swimmer* perturbées par un bruit multiplicatif exponentiel. (a) Échantillons de données. (b) Dictionnaire estimé par MMLE. (c) Dictionnaire estimé par MJLE. Quatre colonnes de $\hat{\mathbf{W}}_{\text{MMLE}}$ ont été mises à zéro alors que l'estimateur MJLE produit un effet de sur-apprentissage, qui se manifeste notamment par des composantes dupliquées. (Figures reproduites de [DF11])

Chapitre 7

Autres travaux (encadrements de thèse)

7.1 Transcription en accords (thèse de Laurent Oudre)

Durant la thèse de Laurent Oudre, co-encadrée avec Yves Grenier à Télécom ParisTech de 2007 à 2010, nous avons proposé des méthodes de transcription en accords basées sur la décomposition non-négative de *chromagrammes*. Un chromagramme est une représentation temps-fréquence particulière, adaptée aux signaux musicaux, dans laquelle l'axe fréquentiel est réduit aux notes de la gamme chromatique (*Do, Ré, Mi, Fa, Sol, La, Si*). Cette représentation est elle-même généralement calculée à l'aide d'une transformée dite à Q-constant. Notre approche générale consiste à identifier le plus proche voisin de chaque trame du chromagramme dans un dictionnaire formé de profils spectraux pré-déterminés d'accords musicaux (par exemple, tous les accords majeurs et mineurs de la gamme chromatique). Cela peut s'interpréter comme un cas particulier de décomposition non-négative, dans laquelle $\mathbf{v}_n$ est une trame de chromagramme, $\mathbf{W}$ est la matrice contenant les profils spectraux et $\mathbf{h}_n$ est contraint d'avoir un seul coefficient non-nul.

Nous avons construit des approches déterministes et probabilistes sur la base de ce principe, impliquant notamment diverses divergences. L'une de nos méthode a obtenu les meilleurs scores parmi 18 autres méthodes sur deux des trois critères d'évaluation pour la tâche "audio chord detection" lors de la campagne d'évaluation internationale MIREX 2009 (Music information retrieval evaluation exchange). [1] Cette méthode a été intégrée au moteur de recherche d'information musicale MuMa, développé par Exalead au sein du projet Quaero. [2]

Les détails de ce travail sont disponibles dans les articles de journaux [OFG11, OGF11].

7.2 Variantes de IS-NMF (thèse d'Augustin Lefèvre)

Durant la thèse d'Augustin Lefèvre, co-encadrée avec Francis Bach (INRIA) de 2009 à 2012, nous nous sommes intéressés à des variantes d'IS-NMF, et en particulier aux deux variantes ci-dessous.

IS-NMF et parcimonie de groupe. La première variante concerne la pénalisation par groupe des coefficients d'activations $\mathbf{h}_n$. Dans un enregistrement musical composé de plusieurs sources, il est attendu que chaque source soit constituée d'une somme de composantes élémentaires (celles dont le spectre est donné par les colonnes de $\mathbf{W}$). Cette structure hiérarchique peut s'interpréter comme une structure de groupe, où chaque groupe correspond à une source. Ainsi, nous nous sommes intéressés à des approches MM pour la résolution de problèmes de la forme

$$\min_{\mathbf{W},\mathbf{H}} D_{\mathrm{IS}}(\mathbf{V}|\mathbf{WH}) + \lambda \sum_n \sum_g S(\|\mathbf{h}_{gn}\|). \tag{7.1}$$

1. `http://www.music-ir.org/mirex/wiki/2009:Audio_Chord_Detection_Results`
2. `http://labs.exalead.com/project/muma`

où $\mathbf{h}_{gn}$ est un sous-vecteur de $\mathbf{h}_n = [\mathbf{h}_{1n}^T, \ldots, \mathbf{h}_{Gn}^T]^T$ et $S(\cdot)$ est un terme induisant de la parcimonie. Les détails de ces travaux, et notamment de leur application à la séparation de sources monocapteur, sont présentés dans l'article [LBF11a].

IS-NMF incrémental. La seconde variante que nous avons abordée concerne le développement d'un algorithme MM incrémental pour IS-NMF. Cette variante peut être utile en contexte de séparation de sources en temps-réel ou encore en contexte "grande masse de données" lorsque les dimensions de $\mathbf{V}$ interdisent l'application de l'algorithme MM en mode "batch". Notre approche s'inspire de travaux en apprentissage de dictionnaire séquentiel réalisés dans l'équipe de Francis Bach, e.g., [MBPS10].

Supposons que nous disposions d'une factorisation $\mathbf{W}^{(N)}\mathbf{H}^{(N)}$ de la matrice $\mathbf{V}^{(N)}$ contenant N échantillons et qu'un nouvel échantillon $\mathbf{v}_{N+1}$ soit présenté. Comment mettre à jour $\mathbf{W}$ et $\mathbf{H}$? En mode batch, on pourrait réaliser la séquence d'opérations suivantes :

$$\mathbf{h}_{N+1} = \arg\min_{\mathbf{h}} D(\mathbf{v}_{N+1}|\mathbf{W}^{(N)}\mathbf{h}) \tag{7.2}$$

$$\mathbf{H}^{(N+1)} = [\mathbf{H}^{(N)}, \mathbf{h}_{N+1}] \tag{7.3}$$

$$\mathbf{W}^{(N+1)} = \arg\min_{\mathbf{W}} D(\mathbf{V}^{(N+1)}|\mathbf{W}\mathbf{H}^{(N+1)}) \tag{7.4}$$

En utilisant les résultats du paragraphe 4.2, on obtiendrait que la mise à jour MM associée à la troisième opération s'écrit

$$w_{fk}^{(N+1)} = \sqrt{\frac{p_{fk}^{(N+1)}}{q_{fk}^{(N+1)}}}, \tag{7.5}$$

avec

$$p_{fk}^{(N+1)} = [w_{fk}^{(N)}]^2 \sum_{n=1}^{N+1} h_{kn}^{(N+1)} \frac{v_{fn}}{[\mathbf{W}^{(N)}\mathbf{H}^{(N+1)}]_{fn}^2}, \tag{7.6}$$

$$q_{fk}^{(N+1)} = \sum_{n=1}^{N+1} h_{kn}^{(N+1)} \frac{1}{[\mathbf{W}^{(N)}\mathbf{H}^{(N+1)}]_{fn}}. \tag{7.7}$$

La mise à jour de $\mathbf{W}^{(N+1)}$ nécessite donc de recalculer les sommes sur n allant de 1 à $N+1$ dans les expressions de p_{fk} et q_{fk} (variables qui peuvent être assimilées à une forme de "statistique exhaustive" dans la mesure où la mise à jour de $\mathbf{W}$ ne dépend que d'elles). C'est cette opération de sommation qui peut apparaître comme rédhibitoire lorsque N est grand. Aussi, notre approche consiste à remplacer les expressions "exactes" de p_{fk} et q_{fk} par une forme approchée basée sur la récursion suivante

$$p_{fk}^{(N+1)} = p_{fk}^{(N)} + [w_{fk}^{(N)}]^2 \left[h_{k(N+1)} \frac{v_{f(N+1)}}{[\mathbf{W}^{(N)}\mathbf{h}_{N+1}]_f^2} \right], \tag{7.8}$$

$$q_{fk}^{(N+1)} = q_{fk}^{(N)} + \left[\frac{h_{k(N+1)}}{[\mathbf{W}^{(N)}\mathbf{h}_{(N+1)}]_f^2} \right]. \tag{7.9}$$

En d'autres termes, seul le dernier terme des sommes sur n dans les équations (7.6) et (7.7), dépendant de la nouvelle observation $\mathbf{v}_{N+1}$, est utilisé dans la mise à jour des statistiques exhaustives. L'approximation vient du fait que $p_{fk}^{(N)}$ et $q_{fk}^{(N)}$ dépendent de valeurs non-actualisées de $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N$ (plus précisément, $\mathbf{h}_n$ a été estimé avec $\mathbf{W} = \mathbf{W}^{(n-1)}$). On peut montrer que la fonction

$$G_N(\mathbf{W}) = \sum_{n=1}^{N} \frac{p_{fk}^{(n)}}{w_{fk}} + q_{fk}^{(n)} w_{fk} \tag{7.10}$$

ainsi construite est une borne supérieure de la fonction objectif $L_N(\mathbf{W}) = \sum_{n=1}^{N} \min_{\mathbf{h}_n} D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$ dont la minimisation est équivalente à la résolution du problème de NMF original défini par l'équation (4.1). Le principal général décrit ici permet (et nécessite parfois) en pratique quelques ajustements, tel que l'utilisation de facteurs d'oubli des anciennes données, la ré-actualisation périodique des coefficients d'activation et le traitement par bloc des nouvelles données. Les détails de ce travail sont disponibles dans l'article [LBF11b].

7.3 Co-factorisation douce de matrices non-négatives (thèse de Nicolas Seichepine)

Dans certains cas, les données sont disponibles dans plusieurs "modalités". On peut penser par exemple aux pistes image et audio d'une vidéo, ou aux images accompagnées de leur légende dans les bases de données photo disponibles sur Internet. Dans ces cas-là, il est souvent souhaitable d'exploiter l'information mutuelle partagée entre les modalités pour un traitement optimal de ces données. À ce titre, et dans le contexte des modèles à facteurs latents, des travaux de l'état de l'art ont proposé de modéliser conjointement les modalités à l'aide d'un facteur latent commun, généralement la matrice d'activation $\mathbf{H}$. En d'autres termes, dans le cas de données bi-modales dont les modalités ont une représentation matricielle $\mathbf{V}_1$ et $\mathbf{V}_2$, un modèle de co-factorisation $\mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H}$, $\mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H}$ peut être approprié. Ce principe est par exemple proposé dans [SG08, YCS11]. Cependant, l'hypothèse que les modalités partagent exactement la même matrice d'activation $\mathbf{H}$ peut être trop forte. Aussi, dans le cadre de la thèse de Nicolas Seichepine, co-encadrée avec Slim Essid et Olivier Cappé à Télécom ParisTech depuis 2012, nous avons proposé une nouvelle forme de co-factorization, dite "douce" et telle que $\mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H}_1$, $\mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H}_2$, $\mathbf{H}_1 \approx \mathbf{H}_2$. La co-factorisation douce est obtenue par résolution du problème d'optimisation suivant

$$\min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}_1,\mathbf{H}_2,\boldsymbol{\lambda}} D_1(\mathbf{V}_1|\mathbf{W}_1\mathbf{H}_1) + \alpha D_2(\mathbf{V}_2|\mathbf{W}_2\mathbf{H}_2) + \beta D_3(\mathbf{H}_1|\boldsymbol{\lambda}\mathbf{H}_2), \tag{7.11}$$

où $D_1(\cdot|\cdot)$, $D_2(\cdot|\cdot)$, $D_3(\cdot|\cdot)$ sont des divergences, et $\boldsymbol{\lambda}$ est une matrice de remise à l'échelle. Nous avons construit un algorithme MM pour ce problème d'optimisation, pour plusieurs cas de divergences.

Cette nouvelle forme de co-factorisation a été appliquée à un problème de regroupement de locuteurs dans des vidéos "talk-show". Le problème consiste simplement à déterminer qui parle quand, et dans notre cas précis concerne directement les archivistes, par exemple de l'INA. Des solutions à ce problème viendraient grandement aider leur travail d'annotation. Le problème de regroupement de locuteur est souvent traité en utilisant la seule piste audio, avec des méthodes de reconnaissance adéquates. Dans le cas des vidéos de talk-show, les pistes son et image présentent une forte corrélation. En effet, la personne qui parle est généralement à l'image, mais pas systématiquement. D'où l'intérêt de la co-factorisation douce : dans ce contexte, les matrices $\mathbf{H}_1$ et $\mathbf{H}_2$, dont les lignes s'avèrent correspondre dans notre contexte à des activations de locuteurs, sont souvent égales mais pas systématiquement. Plus de détail sur ces travaux en cours sont disponibles dans notre première contribution à la co-factorisation douce [SEFC13].

Remarques finales

Le fil conducteur de mes travaux sur la NMF depuis 2007 est la mise en œuvre d'approches probabilistes à ce problème. J'ai cherché à ramener le choix d'une divergence à une hypothèse de modèle. C'est l'usage prédominant et souvent ad-hoc de l'erreur quadratique, alors qu'elle sous-tend un modèle probabiliste mal posé pour les données non-négatives, qui m'a guidé dans cette direction.

Cette démarche appliquée au contexte de la séparation de sources audio, domaine dont j'étais déjà familier, m'a amené à développer des travaux autour du modèle composite Gaussien (GCM) et de la NMF avec la divergence d'Itakura-Saito. Ces travaux ont eu un bon impact dans les communautés concernées comme le révèle le nombre de citations à notre article [FBD09]. L'extension que nous avons faite de ce travail aux signaux multicanaux m'apparaît comme l'une de mes contributions scientifiques les plus fortes. Pour la première fois une méthode générique de séparation de sources musicales, applicable en contexte convolutif sous- ou sur-déterminé, a produit des résultats exploitables dans le cadre d'une application réelle de remastering audio, menée en collaboration avec des ingénieurs du son professionnels.

Ayant établi des liens entre divergences et modèles probabilistes, j'ai souhaité re-considérer l'estimation des facteurs. Il m'a paru intéressant de remettre en cause l'approche classique reposant sur l'optimisation de la log-vraisemblance conjointe. Nous avons proposé un nouvel estimateur, l'estimateur MMLE, mieux posé et offrant une propriété surprenante : la régularisation automatique du nombre de composantes. Nous avons par ailleurs traité cette question de la détermination de l'ordre avec une approche pénalisée explicite, induisant la parcimonie des normes des colonnes ou lignes des facteurs.

Sur le plan algorithmique, le formalisme majoration-minimisation (dont l'algorithme EM est un cas particulier) est celui qui lie la plupart de mes travaux. Ce formalisme que j'ai pu mettre en place grâce à l'aide de Jérôme Idier pour le cas de la NMF avec la β-divergence a joué un rôle important pour la suite de mes travaux, m'offrant une approche d'optimisation générique applicable à de nombreuses variantes de la NMF.

À la lumière des travaux réalisés, je vois actuellement cinq directions de recherche pertinentes au problème la NMF, et qui peuvent concerner les modèles à facteurs latents en général. Elles sont décrites succinctement dans les cinq paragraphes suivants.

Algorithmes à convergence globale. La fonction objectif $D(\mathbf{V}|\mathbf{WH})$ est au mieux bi-convexe si la divergence scalaire $d(x|y)$ sur laquelle elle repose est convexe par rapport à son second argument. Les méthodes d'optimisation alternée considérées dans ce document et dans la grande majorité de la littérature ne peuvent produire qu'un point stationnaire dépendant de l'initialisation. C'est là à mon sens la principale limitation de la NMF dans ses implantations existantes. Quelques travaux récents ont abordé ce problème, soit en tentant de "convexifier" la fonction objectif au moyen d'un changement de variable [Kd12], soit en essayant de minimiser la fonction objectif conjointement par rapport à $\mathbf{W}$ et $\mathbf{H}$ [Rak13]. À ce jour seules les méthodes à base de simulation stochastique peuvent offrir des garanties asymptotiques de convergence globale, dans des problèmes par exemple de caractérisation de densités a posteriori en estimation bayésienne. C'est le mariage de cette littérature et de l'optimisation déterministe qui j'en suis sûr aboutira aux méthodes d'optimisation les plus efficaces. L'article [CP14] offre un exemple de cette nouvelle génération de méthodes.

Algorithmes séquentiels et distribués. À l'heure du *big data*, les algorithmes incrémentaux tels que celui développé lors de la thèse d'Augustin Lefèvre ou dans [MBPS10] sont voués à prendre une importance croissante. Là encore, la non-convexité du problème rend l'étude et la conception de ces algorithmes difficiles, avec cependant de premiers résultats [Sra12, Mai14]. Un problème cousin de l'estimation en ligne est l'estimation distribuée. Imaginons que plusieurs processeurs disposent d'une partie des données et retournent chacun une estimée de $\mathbf{W}$. Comment agréger ces différentes matrices pour produire une unique estimée du dictionnaire ? Ce type d'algorithme, dit de consensus, n'a été que peu étudié dans le contexte de la factorisation de matrices [MJT11].

Théorie de l'estimation par maximum de vraisemblance marginalisée. Une direction de recherche directement liée à l'une de mes contributions personnelles, concerne l'analyse théorique de l'estimateur MMLE basé sur la vraisemblance marginale. Cet estimateur qui a donné des résultats surprenants en pratique mérite une étude plus approfondie. En particulier, il est nécessaire de comprendre d'où vient l'effet observé d'auto-régularisation de l'ordre, et dans quelles conditions précises il peut être attendu. Il est souhaitable d'étudier les propriétés de la fonction objectif, telle que par exemple son éventuelle convexité. Si une telle propriété était vraie, au moins dans certains cas, l'estimation MMLE serait une alternative très intéressante à la NMF classique, qui est toujours un problème au mieux bi-convexe. Enfin, sur le plan algorithmique, il conviendrait de mieux étudier les conditions de validité de l'algorithme VB-EM et de concevoir de nouvelles formes d'algorithmes MC-EM qui épousent plus efficacement les spécificités du problème.

Identifiabilité. Une question liée à l'obtention de solutions globales est celle de l'identifiabilité de la NMF. Ce sujet jusqu'alors peu abordé a connu des développements récents dans le cas non-pénalisé [DS04, Gil12, AGKM12]. Les conditions d'identifiabilité exhibées dans ces articles reposent sur l'existence d'échantillons "purs". En supposant l'existence d'une vérité terrain, les éléments du dictionnaire doivent apparaître de façon non-mélangée dans les données. Cette condition est bien entendu très forte (bien que réaliste dans certains cas, comme par exemple en imagerie hyperspectrale [MBDC$^+$14]). Elle est inhérente au cas non-pénalisé et il est fort à parier que des conditions bien moins restrictives pourraient obtenues dans des cas pénalisés, sous l'hypothèse par exemple de coefficients de $\mathbf{H}$ parcimonieux.

Apprentissage de représentations. Dans de nombreux cas la matrice $\mathbf{V}$ à factoriser est le produit d'un pré-traitement de données brutes. Par exemple, dans le cas des signaux audio, $\mathbf{V}$ est le spectrogramme du signal temporel x. Le choix d'une représentation temps-fréquence sous-tend de nombreux paramètres (tels que la résolution temporelle ou la forme de fenêtre) et de nombreuses variantes sont possibles [Fla98]. Dans le formalisme présenté, ces paramètres sont fixés une fois pour toute. On pourrait envisager de les apprendre en même temps que la factorisation ; par exemple on pourrait imaginer minimiser $D(\mathbf{V}|\mathbf{WH})$ par rapport à $\mathbf{W}$ et $\mathbf{H}$, mais également par rapport à la forme de fenêtre. Autrement dit on pourrait ajuster l'opération temps-fréquence pour quelle transforme les données dans une représentation où l'hypothèse de factorisation est la plus juste possible.

Les méthodes traditionnelles de factorisation de spectrogramme agissent sur des coefficients dits "d'analyse". On pourrait alternativement songer à modéliser les coefficients dits de "synthèse" par une représentation à facteurs latents. Cette approche permettrait notamment de représenter un signal sur une union de dictionnaires avec des résolutions temporelles différentes, à la manière de [DT02] mais en remplaçant les hypothèses usuelles de parcimonie par des hypothèse de "rang faible" de la matrice formée des coefficients de synthèse. De tels travaux sont déjà en cours en collaboration avec Matthieu Kowalski (LSS, Gif-sur-Yvette).

Bibliographie

[AF08] P. Aimé and C. Févotte. La simplification administrative de la gestion des unités de recherche. Rapport de l'Inspection Générale de l'Éducation Nationale et de la Recherche (IGAENR), no 2008-089, Oct. 2008. `http://media.enseignementsup-recherche.gouv.fr/file/Concours_2008/22/3/2008-089simplification_44223.pdf`.

[AGKM12] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization – Provably. In *Proc. Symposium on Theory of Computing (STOC)*, 2012.

[BA98] A. Belouchrani and M. Amin. Blind source separation based on time-frequency signal representations. *IEEE Transactions on Signal Processing*, 46(11) :2888–2897, 1998.

[BBL+07] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1) :155–173, Sep. 2007.

[BG03] M. J. Beal and Z. Ghahramani. The variational Bayesian EM algorithm for incomplete data : with application to scoring graphical model structures. In *Bayesian Statistics 7*, pages 453–464, 2003.

[BHHJ98] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3) :549–559, Sep. 1998.

[BJ06] W. L. Buntine and A. Jakulin. Discrete component analysis. In *Lecture Notes in Computer Science*, volume 3940, pages 1–33. Springer, 2006.

[BTGM04] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. In *Proc. National Academy of Sciences*, pages 4164–4169, Mar. 2004.

[Bur09] C. J. C. Burges. Dimension reduction : A guided tour. *Foundations and Trends in Machine Learning*, 2(4) :275–365, 2009.

[CA10] A. Cichocki and S. Amari. Families of Alpha- Beta- and Gamma- divergences : Flexible and robust measures of similarities. *Entropy*, 12(6) :1532–1568, June 2010.

[Can04] J. F. Canny. GaP : A factor model for discrete data. In *Proc. ACM International Conference on Research and Development of Information Retrieval (SIGIR)*, pages 122–129, 2004.

[CET99] Y. Cao, P. P. B. Eggermont, and S. Terebey. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2) :286–292, Feb. 1999.

[CF02] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *Proc. IEEE Workshop on Multimedia Signal Processing*, 2002.

[CFG07] A. T. Cemgil, C. Févotte, and S. J. Godsill. Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing*, 17(5) :891–913, Sep. 2007. Special issue *Bayesian source separation*, ed. E. E. Kuruoğlu and K. H. Knuth.

[Chi95] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432) :1313–1321, Dec. 1995.

[Com94] P. Comon. Independent component analysis, a new concept ? *Signal processing*, 36(3) :287–314, 1994.

[CP14] P. L. Combettes and J.-C. Pesquet. Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping. Technical report, arXiv, 2014.

[CSDP02] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon. Blind separation of noisy gaussian stationary sources. Application to cosmic microwave background imaging. In *Proc. European Signal Processing Conference (EUSIPCO)*, 2002.

[CZA06] A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. In *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Charleston SC, USA, Mar. 2006.

[CZPA09] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorizations : Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[DCL10] A. Dessein, A. Cont, and G. Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[De 93] A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2) :328–333, 1993.

[DF06] F. Desobry and C. Févotte. Kernel PCA based estimation of the mixing matrix in linear instantaneous mixtures of sparse sources. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[DF11] O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2267–2275, Granada, Spain, Dec. 2011.

[DF12] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10) :5163–5175, Oct. 2012.

[DRdFC07] K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Journal of Mathematical Sciences*, 6(2), June 2007.

[DS04] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts ? In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[DT02] L. Daudet and B. Torrésani. Hybrid representations for audiophonic signal encoding. *Signal Processing*, 82(11) :1595 – 1617, 2002.

[DWM86] M. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorthm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, 5(5) :61 – 66, 1986.

[FBD09] 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, Mar. 2009.

[FC05] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 78–81, Mohonk, NY, USA, Oct. 2005.

[FCC05] D. FitzGerald, M. Cranitch, and E. Coyle. Non-negative tensor factorisation for sound source separation. In *Proc. Irish Signals and Systems Conference*, Dublin, Ireland, Sep. 2005.

[FCC09] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Proc. Irish Signals and Systems Conference*, 2009.

[FD03] C. Févotte and C. Doncarli. A unified presentation of blind source separation methods for convolutive mixtures using block-diagonalization. In *Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA)*, Nara, Japan, Apr. 2003.

[FD04] C. Févotte and C. Doncarli. Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Processing Letters*, 11(3) :386–389, Mar. 2004.

[FDD03] C. Févotte, A. Debiolles, and C. Doncarli. Blind separation of FIR convolutive mixtures : application to speech signals. In *Proc. 1st ISCA Workshop on Non-Linear Speech Processing*, Le Croisic, France, May 2003.

[Fév06] C. Févotte. Bayesian blind separation of audio mixtures with structured priors. In *Proc. 14th European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep. 2006. Special session *Undetermined sparse audio source separation* (invited paper).

[Fév10] C. Févotte. Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition. In W. Wang, editor, *Machine Audition : Principles, Algorithms and Systems*, chapter 11. IGI Global Press, Aug. 2010.

[Fév11] C. Févotte. Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[FFDM04] D. Farina, C. Févotte, C. Doncarli, and R. Merletti. Blind separation of linear instantaneous mixtures of non-stationary surface myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 51(9) :1555 –1567, Sep. 2004.

[FG05] C. Févotte and S. J. Godsill. A Bayesian approach to time-frequency based blind source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, USA, Oct. 2005.

[FG06a] C. Févotte and S. J. Godsill. A Bayesian approach to blind separation of sparse sources. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6) :2174–2188, Nov. 2006.

[FG06b] C. Févotte and S. J. Godsill. Sparse linear regression in unions of bases via Bayesian variable selection. *IEEE Signal Processing Letters*, 13(7) :441–444, Jul. 2006.

[FI11] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9) :2421–2456, Sep. 2011.

[Fla98] P. Flandrin. *Temps-Fréquence*. Hermès, 1998.

[FLH13] C. Févotte, J. Le Roux, and J. R. Hershey. Non-negative dynamical system with application to speech and audio. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[FO10] C. Févotte and A. Ozerov. Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues. In S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, editors, *Proc. 7th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, volume 6684 of *Lecture Notes in Computer Science*, pages 102–115, Málaga, Spain, 2010., 2010. Springer. Long paper.

[FT07] C. Févotte and F. Theis. Pivot selection strategies in Jacobi joint block-diagonalization. In *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 177–187, London, UK, Sep. 2007.

[FTDG08] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill. Sparse linear regression with structured priors and application to denoising of musical audio. *IEEE Transactions on Audio, Speech and Language Processing*, 16(1) :174–185, Jan. 2008.

[Gil12] N. Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research*, 13 :3349 – 3386, 2012.

[HBD10] R. Hennequin, R. Badeau, and B. David. NMF with time-frequency activations to model non stationary audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 445–448, 2010.

[HBFR14] R. Hamon, P. Borgnat, P. Flandrin, and C. Robardet. Nonnegative matrix factorization to find features in temporal networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[HFDZ02] A. Holobar, C. Févotte, C. Doncarli, and D. Zazula. Single autoterms selection for blind source separation in time-frequenccy plane. In *Proc. 11th European Signal Processing Conference (EUSIPCO)*, Toulouse, France, Sep. 2002.

[HFL12] J. Hershey, C. Févotte, and J. Le Roux. Method for transforming non-stationary signals using a dynamic model, Oct. 2012. US Patent 13657077, filed.

[HL04] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58 :30 – 37, 2004.

[Jør87] B. Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 49(2) :127–162, 1987.

[Kd12] V. Krishnamurthy and A. d'Aspremont. Convex algorithms for nonnegative matrix factorization. Technical report, arXiv, 2012.

[KFS12] B. King, C. Févotte, and P. Smaragdis. Optimal cost function and magnitude power for NMF-based speech separation and music interpolation. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sep. 2012.

[LBF11a] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[LBF11b] A. Lefèvre, F. Bach, and C. Févotte. Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, Oct. 2011.

[LC09] L.-H. Lim and P. Comon. Nonnegative approximations of nonnegative tensors. *Journal of Chemometrics*, 23(7-8) :432–441, 2009.

[LDP+13] A. Limem, G. Delmaire, M. Puigt, G. Roussel, and D. Courcot. Non-negative matrix factorization using weighted beta divergence and equality constraints for industrial source apportionment. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sept 2013.

[LFMV02] L. D. Lathauwer, C. Févotte, B. D. Moor, and J. Vandewalle. Jacobi algorithm for joint block diagonalization in blind identification. In *Proc. 23rd Symposium on Information Theory in the Benelux*, pages 155–162, Louvain-la-Neuve, Belgium, May 2002.

[LS99] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401 :788–791, 1999.

[Luc74] L. B. Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal*, 79 :745–754, 1974.

[LYO12] Z. Lu, Z. Yang, and E. Oja. Selecting beta-divergence for nonnegative matrix factorization by score matching. In *Proc. 22nd International Conference on Artificial Neural Networks (ICANN)*, pages 419–426, Lausanne, Switzerland, 2012.

[Mai14] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. Technical report, arXiv, 2014.

[MBDC+14] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi. A signal processing perspective on hyperspectral unmixing : Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1) :67–81, Jan. 2014.

[MBPS10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11 :10–60, 2010.

[MCG97] E. Moulines, J. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 3617–3620, Apr 1997.

[MJT11] L. W. Mackey, M. I. Jordan, and A. Talwalkar. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[NSBT13] S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational bayesian matrix factorization. *The Journal of Machine Learning Research*, 14(1) :1–37, 2013.

[OF10] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3) :550–563, Mar. 2010.

[OFB11] A. Ozerov, C. Févotte, and R. Blouet. Automatic source separation via joint use of segmental information and spatial diversity, Feb. 2011. US Patent 13021692, filed.

[OFBD11] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[OFG11] L. Oudre, C. Févotte, and Y. Grenier. Probabilistic template-based chord recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(8) :2249 – 2259, Nov. 2011.

[OGF11] L. Oudre, Y. Grenier, and C. Févotte. Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7) :2222 – 2233, Sep. 2011.

[OP08] P. D. O'Grady and B. A. Pearlmutter. Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, 72(1-3) :88 – 101, 2008.

[Paa97] P. Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 37(1) :23–25, May 1997.

[PT94] P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5 :111–126, 1994.

[Rak13] A. Rakotomamonjy. Direct optimization of the dictionary learning problem. *IEEE Transactions on Signal Processing*, 61(12) :5495–5506, 2013.

[RDD13] F. Rigaud, B. David, and L. Daudet. A parametric model and estimation techniques for the inharmonicity and tuning of the piano. *Journal of the Acoustical Society of America*, 133(5) :3107–3118, May 2013.

[Ric72] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62 :55–59, 1972.

[SB03] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, Oct. 2003.

[SCY13] U. Simsekli, T. Cemgil, and K. Yilmaz. Learning the beta-divergence in Tweedie compound Poisson matrix factorization models. In *Proc. International Conference on Machine Learning (ICML)*, pages 1409–1417, 2013.

[SEFC13] N. Seichepine, S. Essid, C. Févotte, and O. Cappé. Soft nonnegative matrix co-factorization with application to multimodal speaker diarization. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[SF14] D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[SFM+14] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations : A unified view. *IEEE Signal Processing Magazine*, 31(3) :66–75, May 2014.

[SG08] A. Singh and G. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 358–373. Springer, 2008.

[Sra12] S. Sra. Nonconvex proximal splitting : batch and incremental algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[TF13] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7) :1592 – 1605, July 2013.

[Twe84] M. Tweedie. An index which distinguishes between some important exponential families. In *Proc. Indian Statistical Institute Golden Jubilee International Conference*, 1984.

[VBB10] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Language Processing*, 18 :528 – 537, 2010.

[VGF06] 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, Jul. 2006.

[Vir07] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3) :1066–1074, Mar. 2007.

[WT90] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411) :699–704, 1990.

[Wu07] M. Wu. Collaborative filtering via ensembles of matrix factorizations. In *Proc. KDD Cup and Workshop*, 2007.

[XLG03] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[YCS11] Y. K. Yilmaz, A. T. Cemgil, and U. Simsekli. Generalized Coupled Tensor Factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[YFH06] S. S. Young, P. Fogel, and D. Hawkins. Clustering Scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14(1) :11–13, June 2006.

[ZFVC11] M. Zetlaoui, M. Feinberg, P. Verger, and S. Clémençon. Extraction of food consumption systems by nonnegative matrix factorization (NMF) for the assessment of food choices. *Biometrics*, 67(4) :1647–1658, 2011.

Troisième partie

Articles annexés

(Févotte & Idier, *Neural Computation*, 2011)

LETTER Communicated by Morten Mørup

Algorithms for Nonnegative Matrix Factorization with the β-Divergence

Cédric Févotte
fevotte@telecom-paristech.fr
Laboratoire Traitement et Communication de l'Information (CNRS and Télécom Paris Tech), 75014 Paris, France

Jérôme Idier
Jerome.Idier@irccyn.ec-nantes.fr
Institute de Recherche en Communications et Cybernetique de Nantes (CNRS, Ecole Centrale de Nantes, Ecole des Mines de Nantes, and Université de Nantes), 44000 Nantes, France

This letter describes algorithms for nonnegative matrix factorization (NMF) with the β-divergence (β-NMF). The β-divergence is a family of cost functions parameterized by a single shape parameter β that takes the Euclidean distance, the Kullback-Leibler divergence, and the Itakura-Saito divergence as special cases ($\beta = 2, 1, 0$ respectively). The proposed algorithms are based on a surrogate auxiliary function (a local majorization of the criterion function). We first describe a majorization-minimization algorithm that leads to multiplicative updates, which differ from standard heuristic multiplicative updates by a β-dependent power exponent. The monotonicity of the heuristic algorithm can, however, be proven for $\beta \in (0, 1)$ using the proposed auxiliary function. Then we introduce the concept of the majorization-equalization (ME) algorithm, which produces updates that move along constant level sets of the auxiliary function and lead to larger steps than MM. Simulations on synthetic and real data illustrate the faster convergence of the ME approach. The letter also describes how the proposed algorithms can be adapted to two common variants of NMF: penalized NMF (when a penalty function of the factors is added to the criterion function) and convex NMF (when the dictionary is assumed to belong to a known subspace).

1 Introduction

Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, nonnegative matrix factorization (NMF) is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH}, \tag{1.1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\ K + K\ N \ll F\ N$, hence reducing the data dimension. The factorization is in general only approximate, so that the terms *approximate nonnegative matrix factorization* and *nonnegative matrix approximation* also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, we mention the problems of learning parts of faces and semantic features of text (Lee & Seung, 1999), polyphonic music transcription (Smaragdis & Brown, 2003), object characterization by reflectance spectra analysis (Berry, Browne, Langville, Pauca, & Plemmons, 2007), portfolio diversification (Drakakis, Rickard, de Frein, & Cichocki, 2007), DNA gene expression analysis (Brunet, Tamayo, Golub, & Mesirov, 2004; Gao & Church, 2005), clustering of protein interactions (Greene, Cagney, Krogan, & Cunningham, 2008), and image denoising and inpainting (Mairal, Bach, Ponce, & Sapiro, 2010). The factorization, equation 1.1, is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{1.2}$$

where the notation $\mathbf{A} \geq 0$ expresses nonnegativity of the entries of matrix $\mathbf{A}$ (and not semidefinite positiveness), and where $D(\mathbf{V}|\mathbf{WH})$ is a separable measure of fit such that

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}), \tag{1.3}$$

where $d(x|y)$ is a scalar cost function. What we intend by "cost function" is a positive function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$, with a single minimum for $x = y$.

A popular cost function in NMF is the β-divergence $d_\beta(x|y)$ of Basu, Harris, Hjort, and Jones (1998), Eguchi and Kano (2001), and Cichocki and Amari (2010), defined rigorously in section 2.1. In essence, it is a parameterized cost function with a single parameter β, which takes the Euclidean distance, the generalized Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence as special cases ($\beta = 2, 1$, and 0, respectively). NMF with the β-divergence has been widely used in music signal processing, in particular, for transcription and source separation (O'Grady, 2007; O'Grady & Pearlmutter, 2008; FitzGerald, Cranitch, & Coyle, 2009; Bertin, Févotte, & Badeau, 2009; Févotte, Bertin, & Durrieu, 2009; Vincent, Bertin, & Badeau, 2010; Dessein, Cont, & Lemaitre, 2010; Hennequin, Badeau, & David, 2010). In these works the nonnegative data matrix $\mathbf{V}$ is a spectrogram that is decomposed into elementary spectra with NMF. The parameter β can be tuned so as to optimize transcription or separation accuracy on training data. While popular in music signal processing, NMF with the β-divergence

(shortened as β-NMF in the rest of the letter) can be of interest to any field: the parameter β essentially controls the assumed statistics of the observation noise and can be either fixed or learned from training data or by cross-validation. As Févotte and Cemgil (2009) noted, the values $\beta = 2, 1, 0$, respectively underlie gaussian additive, Poisson, and multiplicative gamma observation noise. The β-divergence offers a continuum of noise statistics that interpolate among these three specific cases (Basu et al., 1998; Eguchi & Kano, 2001; Minami & Eguchi, 2002; Cichocki & Amari, 2010).

The standard β-NMF algorithm used in the above-mentioned papers is presented as a gradient-descent algorithm where the step size is set adaptively and chosen such that the updates are multiplicative, as originally described by Cichocki, Zdunek, and Amari (2006). The same algorithm can be derived from the following heuristic, proposed by Févotte et al. (2009). Let θ be a coefficient of $\mathbf{W}$ or $\mathbf{H}$. As will be seen later, when using the β-divergence, the derivative $\nabla_\theta D(\theta)$ of the criterion $D(\mathbf{V}|\mathbf{WH})$ with respect to (w.r.t) θ can be expressed as the difference of two nonnegative functions such that $\nabla_\theta D(\theta) = \nabla_\theta^+ D(\theta) - \nabla_\theta^- D(\theta)$. Then a heuristic multiplicative algorithm simply writes

$$\theta \leftarrow \theta . \frac{\nabla_\theta^- D(\theta)}{\nabla_\theta^+ D(\theta)}, \tag{1.4}$$

which ensures the nonnegativity of the parameter updates, provided initialization with a nonnegative value. It produces a descent algorithm in the sense that θ is updated toward left (resp. right) when the gradient is positive (resp. negative). A fixed point $\theta^\star$ of the algorithm implies either $\nabla_\theta D(\theta^\star) = 0$ or $\theta^\star = 0$. Monotonicity of this algorithm has been proven by Kompass (2007) for the specific range of values of β for which the divergence $d_\beta(x|y)$ is convex w.r.t y (i.e., $\beta \in [1, 2]$; see section 2.1). The proof is based on a majorization-minimization (MM) procedure: an auxiliary function is built and iteratively minimized for each column of $\mathbf{H}$ (given $\mathbf{W}$) and each row of $\mathbf{W}$ (given $\mathbf{H}$). The auxiliary function is built using Jensen's inequality, thanks to the convexity of the cost for $\beta \in [1, 2]$. However, it was observed in practice that the multiplicative algorithm, 1.4, is still monotone (i.e., decreases the criterion function at each iteration) for values of β out of the "convexity" interval [1, 2], though no proof is available.

This letter studies three descent algorithms for β-NMF, based on an auxiliary function that unifies existing auxiliary functions for the Euclidean distance and KL divergence (De Pierro, 1993; Lee & Seung, 2001), the "generalized divergence" of Kompass (2007), and the IS divergence (Cao, Eggermont, & Terebey, 1999). This auxiliary function was also recently proposed independently by Nakano et al. (2010). The construction of the auxiliary function relies on the decomposition of the criterion function into its convex and concave parts, following the approach of Cao et al.

(1999) for the IS divergence. An auxiliary function to the convex part is constructed using Jensen's inequality, while the concave part is locally majorized by its tangent. It is shown that MM algorithms based on the latter auxiliary function yield multiplicative updates that coincide with the heuristic described by equation 1.4 for $\beta \in [1, 2]$, but differ from a β-dependent power exponent when $\beta \notin [1, 2]$, a result also obtained by Nakano et al. (2010). Additionally, we show that the monotonicity of the heuristic algorithm can be proven for $\beta \in (0, 1)$, using the proposed auxiliary function (it is shown to produce a descent algorithm, though it does not fully minimize the auxiliary function). Then we introduce the concept of the maximization-equalization (ME) algorithm, which produces updates that move along constant level sets of the auxiliary function and leads to larger steps than MM. This is akin to overrelaxation and is shown experimentally to produce faster convergence. Finally, we show how the described MM, ME, and heuristic algorithms can be adapted to two common variants of NMF: penalized NMF (i.e., when a penalty function of $\mathbf{W}$ or $\mathbf{H}$ is added to the criterion function) and convex NMF (when the dictionary is assumed to belong to a known subspace, as proposed by Ding, Li, and Jordan, 2010).

The letter is organized as follows. Section 2 defines and discusses the β-divergence and then exposes in detail the optimization task addressed in this letter. Section 3 recalls the concept of auxiliary function and then introduces a general auxiliary function for the β-NMF problem. Section 4 describes algorithms based on the proposed auxiliary function, namely, MM and ME algorithms, and describes how they relate to the heuristic update, equation 1.4. Section 5 reports simulations and convergence behaviors on synthetic and real data (with audio transcription and face interpolation examples). Section 6 describes extensions of the proposed algorithms to penalized and convex NMF. Section 7 concludes and discusses open questions.

2 Preliminaries

In this section we present the β-divergence and more precisely specify the task that is addressed in this letter. A detailed exposition of the β-divergence can be found in Cichocki & Amari (2010).

2.1 Definition of the β-Divergence. The β-divergence was introduced by Basu et al. (1998) and Eguchi and Kano (2001) and can be defined as

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\log\frac{x}{y} - x + y & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1 & \beta = 0 \end{cases} \tag{2.1}$$

Basu et al. (1998) and Eguchi and Kano (2001) assume $\beta \geq 1$, but the definition domain can be extended to $\beta \in \mathbb{R}$, as suggested by Cichocki et al. (2006), which is the definition domain that is considered in this letter. The β-divergence can be shown continuous in β by using the identity $\lim_{\beta \to 0} (x^\beta - y^\beta)/\beta = \log(x/y)$. The limit cases $\beta = 0$ and $\beta = 1$ correspond to the IS and KL divergences, respectively. The β-divergence coincides up to a factor $1/\beta$ with the "generalized divergence" of Kompass (2007), which, in the context of NMF as well, was separately constructed so as to interpolate between the KL divergence ($\beta = 1$) and the Euclidean distance ($\beta = 2$). The β-divergence is plotted for various values of β on Figure 1. Note that in this letter, we will abusively refer to $d_{\beta=2} = (x - y)^2/2$ as the Euclidean distance, though the latter is formally defined with a square root, and for vectors.

The first and second derivative of $d_\beta(x|y)$ w.r.t y are continuous in β:

$$\begin{aligned} d'_\beta(x|y) &= y^{\beta-2}\,(y - x), \\ d''_\beta(x \mid y) &= y^{\beta-3}\,\left[(\beta - 1)y - (\beta - 2)x\right]. \end{aligned} \tag{2.2}$$

The derivative shows that $d_\beta(x|y)$, as a function of y, has a single minimum in $y = x$ and that it increases with $|y - x|$, justifying its relevance as a measure of fit. The second derivative shows that the β-divergence is convex w.r.t y for $\beta \in [1, 2]$. Outside this interval, the divergence can always be expressed as the sum of a convex, concave, and constant part, such that

$$d_\beta(x|y) = \breve{d}\,(x|y) + \widehat{d}\,(x|y) + \bar{d}(x), \tag{2.3}$$

where $\breve{d}\,(x|y)$ is a convex function of y, $\widehat{d}\,(x|y)$ is a concave function of y, and $\bar{d}(x)$ is a constant of y. The decomposition is not unique, since constant or linear terms (w.r.t y) are both convex and concave or, less trivially, since any convex term can be added to $\breve{d}\,(x|y)$ while subtracted from $\widehat{d}\,(x|y)$. In the following, we will use the natural conventions given in Table 1.

As noted by Févotte et al. (2009), a noteworthy property of the β-divergence is its behavior w.r.t to scale, as the following equation holds for any value of β:

$$d_\beta(\lambda\, x|\lambda\, y) = \lambda^\beta\, d_\beta(x|y). \tag{2.4}$$

It implies that factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest data values, and less precision is to be expected in the estimation of the low-power components, and, conversely, factorizations obtained with $\beta < 0$ will rely more heavily on smallest data values. The IS divergence ($\beta = 0$) is scale invariant, that is, $d_{IS}(\lambda\, x|\lambda\, y) = d_{IS}(x|y)$, and is the only one in the family of β-divergences to possess this property.

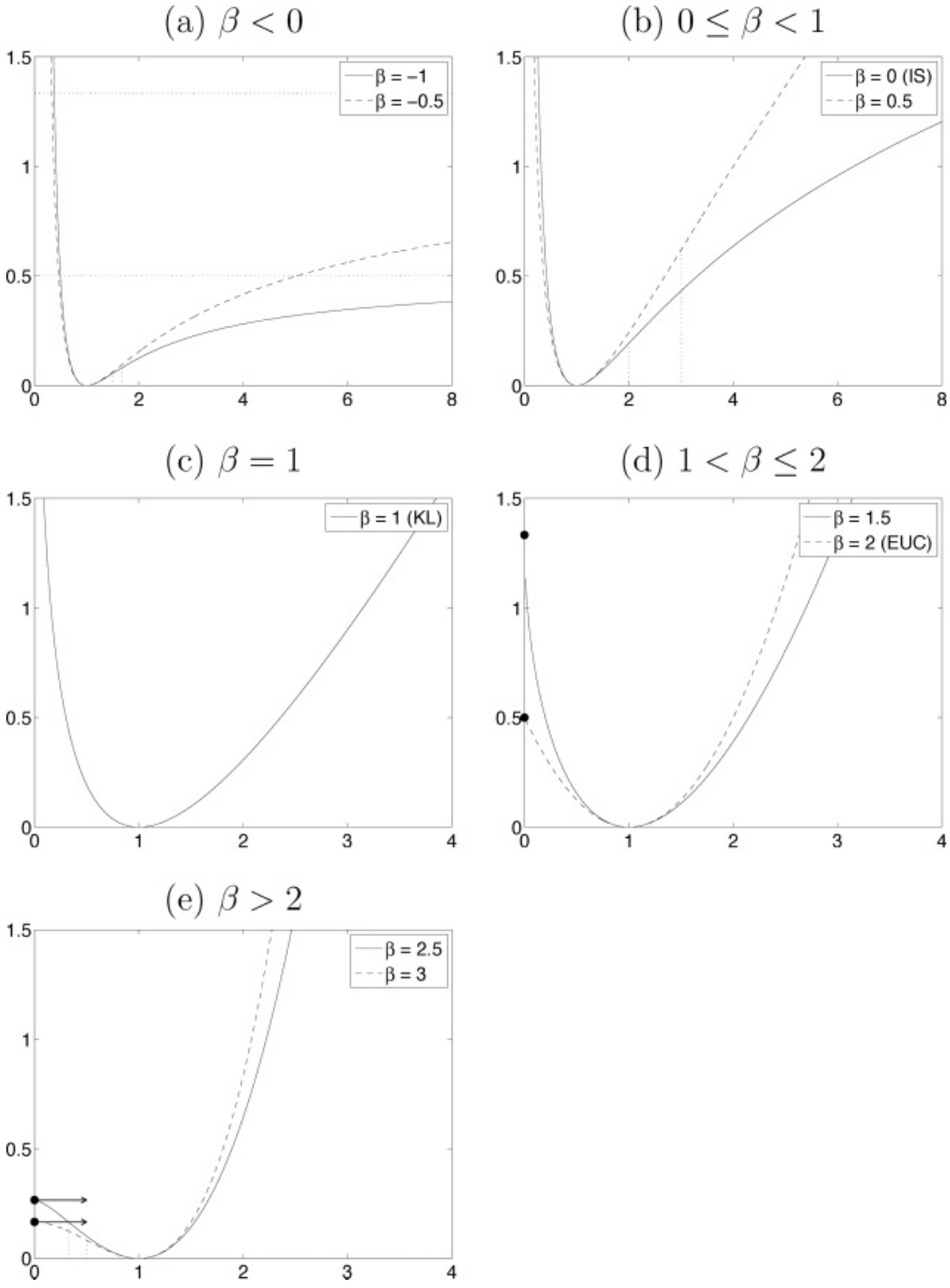


Figure 1: β-divergence $d_\beta(x|y)$ as a function of y (with $x = 1$). Panels a–e illustrate the regimes of the β-divergence for its five characteristic ranges of values of β. The divergence is convex for $1 \leq \beta \leq 2$, as seen in panels c and d. In the other panels, the inflection points are indicated with vertical dotted lines. For $\beta < 0$, the divergence possesses horizontal asymptotes of coordinate $x^\beta/(\beta(\beta-1))$ as $y \to \infty$. For $\beta > 1$, the divergence takes finite value $x^\beta/(\beta(\beta-1))$ at $y = 0$, where the derivative is zero for $\beta > 2$.

Table 1: Example of Differentiable Convex-Concave-Constant Decomposition of the β-Divergence under the Form 2.3.

	$\breve{d}(x\|y)$	$\breve{d}'(x\|y)$	$\hat{d}(x\|y)$	$\hat{d}'(x\|y)$	$\bar{d}(x)$
$\beta < 1$ and $\beta \neq 0$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$
$\beta = 0$	$x\, y^{-1}$	$-x\, y^{-2}$	$\log y$	y^{-1}	$x(\log x - 1)$
$1 \leq \beta \leq 2$	$d_{\beta}(x\|y)$	$d'_{\beta}(x\|y)$	0	0	0
$\beta > 2$	$\frac{1}{\beta} y^{\beta}$	$y^{\beta-1}$	$-\frac{1}{\beta-1} x\, y^{\beta-1}$	$-x\, y^{\beta-2}$	$\frac{1}{\beta(\beta-1)} x^{\beta}$

Factorizations with small, positive values of β are relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks, together with higher-power components such as tonal parts of sustained notes. For example, Févotte et al. (2009) present the results of the decomposition of a piano power spectrogram with IS-NMF and show that components corresponding to very low residual noise and hammer hits on the strings are extracted with great accuracy, while these components are either ignored or severely degraded when using Euclidean or KL divergences. Similarly, the value $\beta = 0.5$ is advocated by FitzGerald et al. (2009) and Hennequin et al. (2010) and has been shown to give optimal results in music transcription based on NMF of the magnitude spectrogram by Vincent et al. (2010).

The β-divergence belongs to the family of Bregman divergences. For $\beta \notin \{0, 1\}$, a suitable Bregman generating function is $\phi(y) = y^{\beta}/(\beta(\beta - 1))$, as noted by Févotte and Cemgil (2009). This function, however, cannot generate the IS and KL divergences by continuity when β tends to 0 or 1. The latter divergences may nonetheless be generated separately, using the functions $\phi(y) = -\log y$ and $\phi(y) = y \log y$, respectively. Cichocki and Amari (2010) give a general Bregman generating function of the β-divergence, continuously defined for all $\beta \in \mathbb{R}$, in the form of $\phi_{\beta \neq 0,1}(y) = (y^{\beta} - \beta y + \beta - 1)/(\beta(\beta - 1))$, $\phi_{\beta=0}(y) = y - \log y - 1$ and $\phi_{\beta=1}(y) = y \log y - y + 1$. NMF with Bregman divergences has been considered by Dhillon and Sra (2005), where the lack of results about the monotonicity of multiplicative algorithms in general has been noted.[1] This letter fills this gap for the specific case of β-divergence.

[1] More precisely, Dhillon and Sra (2005) give proofs of monotonicity for the "reverse" problem of minimizing $D(\mathbf{WH}|\mathbf{V})$ instead of $D(\mathbf{V}|\mathbf{WH})$, while pointing out that monotonicity of multiplicative algorithms based on the heuristic 1.4 for the latter problem is, however, observed in practice.

2.2 Task.

2.2.1 Core Optimization Problem. Like most algorithms in the literature to date, the NMF algorithms we describe in this letter sequentially update $\mathbf{H}$ given $\mathbf{W}$ and then $\mathbf{W}$ given $\mathbf{H}$. These two steps are essentially the same, by symmetry of the factorization ($\mathbf{V} \approx \mathbf{WH}$ is equivalent to $\mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$ and the roles of $\mathbf{W}$ and $\mathbf{H}$ are simply exchanged) and because we are not making any assumption on the relative values of F and N. Hence, we may concentrate on solving the following subproblem,

$$\min_{\mathbf{H}} \; C(\mathbf{H}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH}) \text{ subject to } \mathbf{H} \geq \mathbf{0}, \tag{2.5}$$

with fixed $\mathbf{W}$ and where in the rest of the letter $D(\mathbf{V}|\mathbf{WH})$ is, as of equation 1.3, with $d(x|y) = d_\beta(x|y)$. The criterion function $C(\mathbf{H})$ separates into $\sum_n D(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$, where $\mathbf{v}_n$ and $\mathbf{h}_n$ are the nth column of $\mathbf{V}$ and $\mathbf{H}$, respectively, so that we are essentially left with solving the problem,

$$\min_{\mathbf{h}} \; C(\mathbf{h}) = D(\mathbf{v}|\mathbf{Wh}) \text{ subject to } \mathbf{h} \geq 0, \tag{2.6}$$

where $\mathbf{v} \in \mathbb{R}_+^F$, $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{h} \in \mathbb{R}_+^K$.

2.2.2 KKT Necessary Conditions. An admissible solution $\mathbf{h}^\star$ to problem 2.6 must satisfy the Karush-Kuhn-Tucker (KKT) first-order optimality conditions,

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star).\mathbf{h}^\star = 0, \tag{2.7}$$

$$\nabla_{\mathbf{h}} C(\mathbf{h}^\star) \geq 0, \tag{2.8}$$

$$\mathbf{h}^\star \geq 0, \tag{2.9}$$

where the dot notation . denotes entrywise operations (here term-to-term multiplication) and $\nabla_{\mathbf{h}} C(\mathbf{h})$ denotes the gradient of $C(\mathbf{h})$, given by

$$\nabla_{\mathbf{h}} C(\mathbf{h}) = \mathbf{W}^T \left[d'(v_f|[\mathbf{Wh}]_f)\right]_f \tag{2.10}$$

$$= \mathbf{W}^T[(\mathbf{Wh})^{\cdot(\beta-2)}(\mathbf{Wh} - \mathbf{v})], \tag{2.11}$$

where the notation $[x_f]_f$ refers to the column vector $[x_1, \ldots, x_F]^T$. The KKT conditions 2.7 to 2.9 can be summarized as

$$\min\{\mathbf{h}^\star, \nabla_{\mathbf{h}} C(\mathbf{h}^\star)\} = \mathbf{0}_K, \tag{2.12}$$

where the min operator is entrywise and $\mathbf{0}_K$ is a null vector of dimension K.

2.2.3 Algorithms. In the following, we will say that an algorithm is monotone if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i\geq 0}$, such that $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$ for all $i \geq 0$. An algorithm is said to be convergent if it produces a sequence of iterates $\{\mathbf{h}^{(i)}\}_{i\geq 0}$ that converges to a limit point $\mathbf{h}^\star$ satisfying the KKT conditions 2.7 to 2.9. Monotonicity does not imply convergence in general, nor is it necessary to convergence.

3 An Auxiliary Function for β-NMF

In this section we define the concept of auxiliary function and then exhibit a separable auxiliary function for the β-NMF problem.

3.1 Definition of Auxiliary Function.

Definition 1. *(auxiliary function). The $\mathbb{R}_+^K \times \mathbb{R}_+^K \to \mathbb{R}_+$ mapping $G(\mathbf{h}|\tilde{\mathbf{h}})$ is said to be an auxiliary function to $C(\mathbf{h})$ if and only if*

- $\forall \mathbf{h} \in \mathbb{R}_+^K, C(\mathbf{h}) = G(\mathbf{h}|\mathbf{h})$
- $\forall (\mathbf{h}, \tilde{\mathbf{h}}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K, C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$.

In other words an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ is a majorizing function (or upper bound) of $C(\mathbf{h})$, which is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The optimization of $C(\mathbf{h})$ can be replaced by iterative optimization of $G(\mathbf{h}|\tilde{\mathbf{h}})$. Indeed, any iterate $\mathbf{h}^{(i+1)}$ satisfying

$$G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) \tag{3.1}$$

satisfies $C(\mathbf{h}^{(i+1)}) \leq C(\mathbf{h}^{(i)})$, because we have

$$C(\mathbf{h}^{(i+1)}) \leq G(\mathbf{h}^{(i+1)}|\mathbf{h}^{(i)}) \leq G(\mathbf{h}^{(i)}|\mathbf{h}^{(i)}) = C(\mathbf{h}^{(i)}). \tag{3.2}$$

The iterate $\mathbf{h}^{(i+1)}$ is typically chosen as

$$\mathbf{h}^{(i+1)} = \arg\min_{\mathbf{h}\geq 0} \; G(\mathbf{h}|\mathbf{h}^{(i)}), \tag{3.3}$$

which forms the basis of MM algorithms (see, e.g., Hunter & Lange, 2004, for a tutorial). However, any other iterate $\mathbf{h}^{(i+1)}$ satisfying equation 3.1 produces a monotone algorithm. As such, Figure 2 illustrates the three updates strategies that will be developed in this letter.

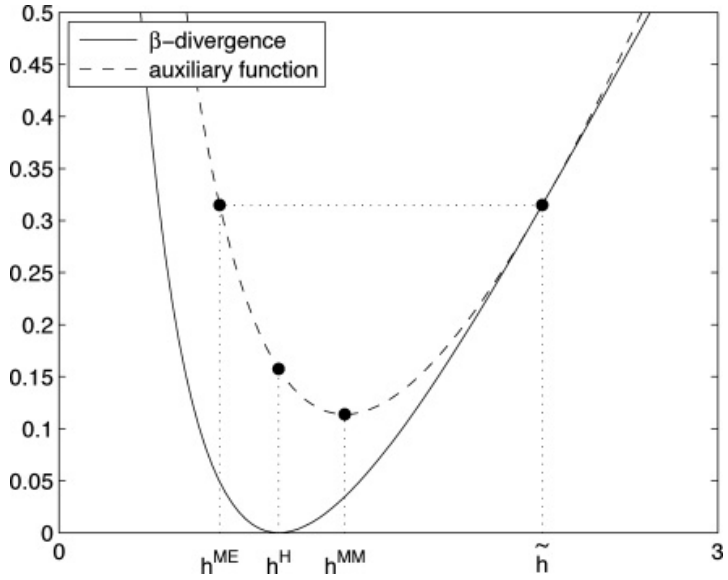


Figure 2: The β-divergence $d_\beta(x|y)$ for $\beta = 0.5$ (with $x = 1$) and its auxiliary function in dimension 1 (with $\tilde{h} = 2.2$). The MM update h^{MM} corresponds to the minimum of the auxiliary function (see section 4.1). The heuristic update h^{H} given by equation 1.4 is discussed in section 4.2 (the heuristic update minimizes the criterion function in the simple one-dimensional case, but this is not true in larger dimensions). The ME update h^{ME} consists of selecting the next update on the other side of the "valley" defined by the auxiliary function, opposite the point defined by the current solution $\tilde{h}$ (see section 4.3).

3.2 Separable Auxiliary Function for β-NMF. In this section, we construct an auxiliary function to $C(\mathbf{h})$ for the specific case of the β-divergence. Our approach follows the one of Cao et al. (1999) for IS divergence and consists of majorizing the convex part of the criterion using Jensen's inequality and majorizing the concave part by its tangent, as detailed in the proof of the following theorem. Here and henceforth, we denote $\mathbf{W}\tilde{\mathbf{h}}$ by $\tilde{\mathbf{v}}$, with entries $[\mathbf{W}\tilde{\mathbf{h}}]_f = \tilde{v}_f$.

Theorem 1 *(auxiliary function for β-NMF). Let $\tilde{\mathbf{h}}$ be such that*

(i) $\forall f, \tilde{v}_f > 0$,
(ii) $\forall k, \tilde{h}_k > 0$.

Then the function $G(\mathbf{h}|\tilde{\mathbf{h}})$ defined by

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left[\sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} \breve{d}\left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \right] + \left[\widehat{d}'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k) + \widehat{d}\,(v_f|\tilde{v}_f) \right] + \bar{d}(v_f) \qquad (3.4)$$

is an auxiliary function to $C(\mathbf{h}) = \sum_f d(v_f|[\mathbf{Wh}]_f)$*, where* $\breve{d}(x|y) + \hat{d}(x|y) + \bar{d}(x)$ *is any differentiable convex-concave-constant decomposition of the* β*-divergence, such as the one defined in Table 1.*

Proof. The condition $G(\mathbf{h}|\mathbf{h}) = C(\mathbf{h})$ is trivially met. The criterion $C(\mathbf{h})$ may be written as

$$C(\mathbf{h}) = \sum_f C_f(\mathbf{h}), \tag{3.5}$$

where $C_f(\mathbf{h}) \stackrel{\text{def}}{=} d(v_f|[\mathbf{Wh}]_f)$. We prove $C(\mathbf{h}) \leq G(\mathbf{h}|\tilde{\mathbf{h}})$ by constructing an auxiliary function to each part $C_f(\mathbf{h})$ of the criterion and, more precisely, by treating the convex and concave part separately. Let us define $\breve{C}_f(\mathbf{h}) \stackrel{\text{def}}{=} \breve{d}(v_f|[\mathbf{Wh}]_f)$ and $\hat{C}_f(\mathbf{h}) \stackrel{\text{def}}{=} \hat{d}(v_f|[\mathbf{Wh}]_f)$, so that we can write

$$C_f(\mathbf{h}) = \breve{C}_f(\mathbf{h}) + \hat{C}_f(\mathbf{h}) + \bar{d}(v_f). \tag{3.6}$$

Convex part: We first prove that

$$\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f}\, \breve{d}\left(v_f|\tilde{v}_f \frac{h_k}{\tilde{h}_k}\right) \tag{3.7}$$

is an auxiliary function to $\breve{C}_f(\mathbf{h})$. The condition $\breve{G}_f(\mathbf{h}|\mathbf{h}) = \breve{C}_f(\mathbf{h})$ is trivially met. The condition $\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \breve{C}_f(\tilde{\mathbf{h}})$ is proven as follows. Let $\mathcal{K}$ be the set of indices k such that $w_{fk} \neq 0$. Define $\forall k \in \mathcal{K}$,

$$\tilde{\lambda}_{fk} = \frac{w_{fk}\tilde{h}_k}{\tilde{v}_f} = \frac{w_{fk}\tilde{h}_k}{\sum_{\ell\in\mathcal{K}} w_{f\ell}\tilde{h}_\ell}. \tag{3.8}$$

We have $\sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk} = 1$ and

$$\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk}\, \breve{d}\left(v_f|\frac{w_{fk}h_k}{\tilde{\lambda}_{fk}}\right) \tag{3.9}$$

$$\geq \breve{d}\left(v_f|\sum_{k\in\mathcal{K}} \tilde{\lambda}_{fk}\frac{w_{fk}h_k}{\tilde{\lambda}_{fk}}\right) \tag{3.10}$$

$$= \breve{d}\left(v_f|\sum_{k=1}^{K} w_{fk}h_k\right) \tag{3.11}$$

$$= \breve{C}_f(\mathbf{h}), \tag{3.12}$$

where we used Jensen's inequality, by convexity of $\breve{d}\,(x|y)$.

Concave part: An auxiliary function $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ to the concave part $\widehat{C}_f(\mathbf{h})$ can be taken as the first-order Taylor approximation to $\widehat{C}_f(\mathbf{h})$ in the vicinity of $\tilde{\mathbf{h}}$:

$$\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \widehat{C}_f(\tilde{\mathbf{h}}) + \nabla^T \widehat{C}_f(\tilde{\mathbf{h}})\,(\mathbf{h} - \tilde{\mathbf{h}}). \tag{3.13}$$

The function satisfies $\widehat{G}_f(\mathbf{h}|\mathbf{h}) = \widehat{C}_f(\mathbf{h})$ by construction and $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) \geq \widehat{C}_f(\mathbf{h})$ by concavity of $\widehat{C}_f(\mathbf{h})$, using the property that the tangent to any point is an upper bound of a concave function.[2] Using

$$\nabla_{h_k} \widehat{C}_f(\mathbf{h}) = w_{fk} \widehat{d}\,'(v_f|[\mathbf{W}\mathbf{h}]_f), \tag{3.14}$$

we give the explicit form for $\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}})$ by

$$\widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) = \widehat{d}\,(v_f|\tilde{v}_f) + \widehat{d}\,'(v_f|\tilde{v}_f) \sum_k w_{fk}(h_k - \tilde{h}_k). \tag{3.15}$$

In the end, a suitable auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to $C(\mathbf{h})$ is obtained by summing up the auxiliary functions constructed for each individual part of the criterion,

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \left(\breve{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \widehat{G}_f(\mathbf{h}|\tilde{\mathbf{h}}) + \bar{d}(v_f) \right), \tag{3.16}$$

which leads to equation 3.4.

3.2.1 Properties of the Auxiliary Function. $G(\mathbf{h}|\tilde{\mathbf{h}})$ is by construction separable in functions of the individual coefficients h_k of $\mathbf{h}$, which allows decoupling the optimization. It is convenient to rewrite the auxiliary function as such in order to derive some of the algorithms of section 4. We may write

$$G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_k G_k(h_k|\tilde{\mathbf{h}}) + cst, \tag{3.17}$$

[2] $\widehat{C}_f(\mathbf{h}) = \widehat{d}\,(v_f|[\mathbf{W}\mathbf{h}]_f)$ is concave as the composition of a concave function and a linear function.

where *cst* is a constant w.r.t $\mathbf{h}$ and

$$\begin{aligned} G_k(h_k|\tilde{\mathbf{h}}) \stackrel{\text{def}}{=} \tilde{h}_k \left[\sum_f \frac{w_{fk}}{\tilde{v}_f} \breve{d}\left(v_f|\tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) \right] \\ + h_k \left[\sum_f w_{fk} \widehat{d}'(v_f|\tilde{v}_f) \right]. \end{aligned} \tag{3.18}$$

The gradient of the auxiliary function is given by

$$\nabla_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\breve{d}'\left(v_f|\tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \widehat{d}'(v_f|\tilde{v}_f) \right]. \tag{3.19}$$

Thanks to the separability of the auxiliary function into its variables, the Hessian matrix is diagonal with

$$\nabla^2_{h_k} G(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f \tilde{v}_f \frac{w_{fk}}{\tilde{h}_k} \breve{d}''\left(v_f|\tilde{v}_f \frac{h_k}{\tilde{h}_k} \right). \tag{3.20}$$

By convexity of $\breve{d}(x|y)$, we have $\breve{d}''(x|y) \geq 0$, which implies positive definiteness of the Hessian matrix and, hence, convexity of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ (convexity more simply derives from the fact that the auxiliary function is built as a sum of convex functions).

3.2.2 Connections with Other Works. The construction of $G(\mathbf{h}|\tilde{\mathbf{h}})$ employs standard mathematical tools (Jensen's inequality, Taylor approximation) that are well known from the MM literature (see, e.g., Hunter & Lange, 2004). For $\beta \in [1, 2]$, $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function built by Kompass (2007), who proposed a generalization of the auxiliary functions proposed by Lee and Seung (2001) for the Euclidean distance ($\beta = 2$) and the generalized KL divergence ($\beta = 1$). For $\beta = 0$ (IS divergence), $G(\mathbf{h}|\tilde{\mathbf{h}})$ coincides with the auxiliary function proposed by Cao et al. (1999). It is worth recalling that in the algorithms proposed by Lee and Seung (2001) the update of $\mathbf{W}$ given $\mathbf{H}$ or $\mathbf{H}$ given $\mathbf{W}$ are instances of well-known algorithms for image restoration (for which $\mathbf{W}$ acts as a fixed, known blurring matrix and $\mathbf{H}$ is a vectorized image to be reconstructed). These algorithms are the iterative space reconstructing algorithm (ISRA) of Daube-Witherspoon and Muehllehner (1986) and the Richardson-Lucy (RL) algorithm of Richardson (1972) and Lucy (1974), which perform nonnegative linear regression with the Euclidean distance and KL divergence, respectively. The ISRA and RL algorithms are shown to be MM algorithms by De Pierro (1993). Similarly, the algorithms proposed by Cao et al. (1999) for nonnegative linear

Table 2: Exponent in the Multiplicative Updates Given by the MM Algorithm.

	$\beta < 1$	$1 \leq \beta \leq 2$	$\beta > 2$
$\gamma(\beta)$	$\frac{1}{2-\beta}$	1	$\frac{1}{\beta-1}$

regression with the IS divergence were designed in the image restoration setting. Finally, let us mention that an auxiliary function based on Jensen's inequality for NMF with the α-divergence (which is always convex w.r.t. to its second argument) is given by Cichocki, Lee, Kim, and Choi (2008).

4 Algorithms for β-NMF

In this section, we describe algorithms for β-NMF based on the auxiliary function constructed in section 3. In the following, $\tilde{\mathbf{h}}$ should be understood as the current iterate $\mathbf{h}^{(i)}$, and we are seeking to obtain $\mathbf{h}^{(i+1)}$ such that equation 3.1 is satisfied.

4.1 Maximization-Minimization Algorithm. An MM algorithm can be derived by minimizing the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ w.r.t. $\mathbf{h}$. Given the convexity and the separability of the auxiliary function, the optimum is obtained by canceling the gradient given by equation 3.19. This is trivially done and leads to the following update,

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right)^{\gamma(\beta)}, \tag{4.1}$$

where $\gamma(\beta)$ is given in Table 2. Note that $\gamma(\beta) \leq 1, \forall \beta$. As suggested in section 1, the gradient of the criterion may be written as the difference of two nonnegative functions such that

$$\nabla_{h_k} C(\tilde{\mathbf{h}}) = \nabla_{h_k}^{+} C(\tilde{\mathbf{h}}) - \nabla_{h_k}^{-} C(\tilde{\mathbf{h}}), \tag{4.2}$$

$$\nabla_{h_k}^{+} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, \tilde{v}_f^{\beta-1}, \tag{4.3}$$

$$\nabla_{h_k}^{-} C(\tilde{\mathbf{h}}) = \sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}, \tag{4.4}$$

so that the update, equation 4.1, can be rewritten in the more interpretable form

$$h_k^{\mathrm{MM}} = \tilde{h}_k \left(\frac{\nabla_{h_k}^- C(\tilde{\mathbf{h}})}{\nabla_{h_k}^+ C(\tilde{\mathbf{h}})} \right)^{\gamma(\beta)}. \tag{4.5}$$

The conclusion is thus that the MM algorithm leads to multiplicative updates, but the latter differ from the "usual ones," obtained by setting $\gamma(\beta) = 1$ for all β and derived heuristically by Cichocki et al. (2006) through gradient descent with adaptative step or by Févotte et al. (2009) by splitting the gradient into two nonnegative functions as discussed above and in section 1. The MM update differs from the heuristic update by the exponent $\gamma(\beta)$, which is not equal to 1 for $\beta \notin [1, 2]$.

4.2 Heuristic Algorithm. This section discusses the properties of the heuristic update proposed by Cichocki et al. (2006) and Févotte et al. (2009) and defined for all β by

$$h_k^{\mathrm{H}} = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1}} \right). \tag{4.6}$$

Very few mathematical results exist for the heuristic update when β falls outside $[1, 2]$, that is, when the β-divergence $d_\beta(x|y)$ is not convex. In such a case, the heuristic update can be erroneously interpreted as an MM algorithm by wrongly applying Jensen's inequality to $C(\mathbf{h})$. Yet in the particular case $\beta = 0$, it holds that each heuristic update produces a decrease of $C(\mathbf{h})$ (Cao et al., 1999). One objective of this section is to extend this result to all values of β between 0 and 1.

Let us first introduce a scalar auxiliary function $g(y|\tilde{y}; x)$:

$$\begin{aligned} \forall y, \tilde{y}, x > 0, \quad g(y|\tilde{y}; x) \;&=\; \breve{d}\,(x|y) + \widehat{d}\,(x|\tilde{y}) \\ &\quad + (y - \tilde{y})\widehat{d}\,'(x|\tilde{y}) + \bar{d}(x), \end{aligned} \tag{4.7}$$

where $\breve{d}\,(x|y)$, $\widehat{d}\,(x|y)$, and $\bar{d}(x|y)$ are defined in Table 1. By immediate application of theorem 1 to the scalar case, $g(y|\tilde{y}; x)$ is an auxiliary function to $d(x|y)$. In particular, $g(\tilde{y}|\tilde{y}; x) = d(x|\tilde{y})$. Then we have the following preliminary result:

Lemma 1. *For all $\beta \in \mathbb{R}$,*

$$G_k(h_k|\tilde{\mathbf{h}}) = \frac{1}{\tilde{h}_k^{\beta-1}} \left(\sum_f w_{fk} \tilde{v}_f^{\beta-1} \right) g(h_k|\tilde{h}_k; h_k^H) + cst. \tag{4.8}$$

Proof. For each of the four possible expressions of $(\widehat{d}\,, \breve{d}\,)$ given in Table 1, the validity of equation 4.8 can be checked straightforwardly by direct verification.

As already mentioned in section 3.1, the MM update, equation 3.3, is not the only way of taking advantage of the auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ to obtain a decrease of $C(\mathbf{h})$: any update satisfying equation 3.1 also ensures that $C(\mathbf{h})$ does not increase. This is a key remark to understand the behavior of the heuristic algorithm for $\beta \in (0, 1)$, given the following property:

Theorem 2. *For all $\beta \in (0, 1)$, and all $\tilde{\mathbf{h}}$ such that conditions (i) and (ii) of theorem 1 hold, the heuristic algorithm produces nonincreasing values of $C(\mathbf{h})$, according to the following inequality:*

$$G(\mathbf{h}^H|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{4.9}$$

Proof. For all $\beta \in (0, 1)$, straightforward calculations yield

$$g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) = \breve{d}\,(x|\tilde{y}) - \breve{d}\,(x|x) - (x - \tilde{y})\widehat{d}\,'(x|\tilde{y}) \tag{4.10}$$

$$= \frac{1}{1-\beta}\,\tilde{y}^{\beta}(1 - \beta + \beta\theta - \theta^{\beta}), \tag{4.11}$$

where $\theta = x/\tilde{y}$. Since $f(\theta) = \theta^{\beta}$ is a concave function of θ, we have $f(\theta) \leq f(1) + (\theta - 1)f'(1)$, which also reads $\theta^{\beta} \leq 1 + (\theta - 1)\beta$. Hence, $g(\tilde{y}|\tilde{y}; x) - g(x|\tilde{y}; x) \geq 0$ for all x, $\tilde{y}$. The latter inequality implies $\forall k$, $g(h_k^{\mathrm{H}}|\tilde{h}_k, h_k^{\mathrm{H}}) \leq g(\tilde{h}_k|\tilde{h}_k, h_k^{\mathrm{H}})$, so that we have $G_k(h_k^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G_k(\tilde{h}_k|\tilde{\mathbf{h}})$ according to equation 4.8, which leads to the result by summation over k.

Cao et al. (1999) show that inequality 4.9 becomes an equality in the case $\beta = 0$, so that each heuristic update yields $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. In this particular case, the heuristic algorithm can be called a majorization-equalization (ME) algorithm, a class of algorithms described in next section. For values of β outside the range $[0, 2]$, inequality 4.9 no longer holds.[3] Of course, this does not mean that the heuristic updates produce increasing values of $C(\mathbf{h})$. On the contrary, numerical simulations tend to indicate that they always produce nonincreasing values of $C(\mathbf{h})$, but proving this is still an open issue for $\beta \notin [0, 2]$. Compared to MM updates, heuristic updates produce larger or equal steps for all β, since it can trivially be shown that

$$\forall k, \quad |h_k^{\mathrm{H}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|. \tag{4.12}$$

[3]Indeed, we can prove that the reversed inequality holds for all $\beta < 0$, while no systematic result is known for $\beta > 2$.

For $\beta \notin [1, 2]$, numerical simulations indicate that the heuristic algorithm is faster than the MM algorithm (and we recall that the two algorithms coincide for $\beta \in [1, 2]$). Given equation 4.12, skipping from the latter to the former has an effect comparable to that of overrelaxation: on average, stretching the steps allows reducing their number to reach convergence. This will be discussed in more detail in section 4.4.

In order to produce even larger steps for $\beta \in [0, 2]$ and yet nonincreasing values of $C(\mathbf{h})$, the following section explores the concept of majorization-equalization.

4.3 Majorization-Equalization Algorithm. Let us introduce the general notion of ME update by the fact that the new iterate $\mathbf{h}^{\mathrm{ME}}$ fulfills

$$G(\mathbf{h}^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}}). \tag{4.13}$$

Equation 4.13 actually defines a level set rather than a single point. Let us concentrate on the following more constrained and manageable condition, given the separability of $G(\mathbf{h}|\tilde{\mathbf{h}})$:

$$\forall k, \quad G_k(h_k^{\mathrm{ME}}|\tilde{\mathbf{h}}) = G_k(\tilde{h}_k|\tilde{\mathbf{h}}).$$

Given equation 4.8, this amounts to solving the following equation for y, for any $\tilde{y}, x > 0$:

$$g(y|\tilde{y}; x) = g(\tilde{y}|\tilde{y}; x). \tag{4.14}$$

Since $g(y|\tilde{y}; x)$ is strictly convex w.r.t. y, equation 4.14 has no more than two solutions, one of them being $\tilde{y}$. By construction, the selection of the other solution (provided that it exists) will provide ME steps that are larger than MM updates—

$$\forall k, \quad |h_k^{\mathrm{ME}} - \tilde{h}_k| \geq |h_k^{\mathrm{MM}} - \tilde{h}_k|, \tag{4.15}$$

as illustrated by Figure 2. To go further on the determination of this solution, a case-by-case analysis must be performed, depending on the range of β.

Case 1. $\beta \in [0, 1)$ In that case, we have

$$g(y|\tilde{y}; x) = \frac{1}{1-\beta} x\, y^{\beta-1} + y\tilde{y}^{\beta-1} + cst. \tag{4.16}$$

Let us remark that

$$\forall\, \tilde{y}, x > 0, \quad \lim_{y\to 0} g(y|\tilde{y}; x) = \lim_{y\to\infty} g(y|\tilde{y}; x) = \infty, \tag{4.17}$$

Table 3: Values of β for Which ME Updates are Closed Form, by Root Extraction of Polynomials of Degree d.

$\beta \leq 0$	$0 \leq \beta \leq 1$	$1 \leq \beta \leq 2$	$\beta \geq 2$	d
0	0	2	2	1
-1	1/2	3/2	3	2
-2	2/3	4/3	4	3
-3	3/4	5/4	5	4

so that equation 4.14 always admits two positive solutions (or one double-positive solution if $\tilde{y} = x$), one of the two being $y = \tilde{y}$. The other one is the solution of interest. However, it is not closed form, except for specific values of β (see Table 3). More precisely, when $\beta = 1 - 1/d$ and d is an integer, the solution can be found by solving the following polynomial equation of degree d, for $z = y^{1/d}$:

$$(1-\beta) \sum_{\ell=1}^{d} \tilde{z}^{d-\ell} z^{\ell} - x = 0, \tag{4.18}$$

where $\tilde{z} = \tilde{y}^{1/d}$. Not surprisingly, the simplest case $\beta = 0$ ($d = 1$) leads us to $y = x$, and thus to $h_k^{\mathrm{ME}} = h_k^{\mathrm{H}}$. The case $\beta = 0.5$ ($d = 2$) is more interesting. The extraction of the positive root of equation 4.18 then provides the following update formula:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4} \left(\sqrt{1 + 8 \frac{h_k^{\mathrm{H}}}{\tilde{h}_k}} - 1 \right)^2. \tag{4.19}$$

This expression does not correspond to a multiplicative update, although it ensures that positivity is maintained.

Case 2. $\beta \in (1, 2]$ In this case, we have

$$g(y|\tilde{y}; x) = \frac{1}{\beta} y^{\beta} - \frac{1}{\beta - 1} x y^{\beta - 1} + cst. \tag{4.20}$$

$g(y|\tilde{y}; x)$ tends toward ∞ for $y \to \infty$, but it remains finite for $y \to 0$. As a consequence, equation 4.14 admits the trivial solution $y = \tilde{y}$ only if $g(\tilde{y}|\tilde{y}; x) > g(0|\tilde{y}; x)$, and also the unwanted solution 0 if $g(\tilde{y}|\tilde{y}; x) = g(0|\tilde{y}; x)$. It is only when $g(\tilde{y}|\tilde{y}; x) < g(0|\tilde{y}; x)$ that a positive, nontrivial solution exists. This solution is closed form for specific values of β given in Table 3. They correspond to $\beta = 1 + 1/d$, where d is an integer. Equation 4.14

then amounts to solving the following polynomial equation of degree d, for $z = y^{1/d}$:

$$\sum_{\ell=0}^{d} \tilde{z}^{d-\ell} z^{\ell} - (d+1)\, x = 0, \tag{4.21}$$

with $\tilde{z} = \tilde{y}^{1/d}$. The simplest case is $\beta = 2$ ($d = 1$), and the solution is then given by $y = 2x - \tilde{y}$ if $\tilde{y} < 2x$, which yields the overrelaxed update

$$h_k^{\mathrm{ME}} = 2h_k^{\mathrm{MM}} - \tilde{h}_k \tag{4.22}$$

provided that $\tilde{h}_k < 2h_k^{\mathrm{MM}}$. Note that this result more simply stems from the fact that when $\beta = 2$, the auxiliary function is parabolic and thus symmetric with respect to h_k^{MM}. In the case $\beta = 1.5$ ($d = 2$), a positive ME update exists if $\tilde{h}_k < 3h_k^{\mathrm{MM}}$, and it takes the following form:

$$h_k^{\mathrm{ME}} = \frac{\tilde{h}_k}{4} \left(\sqrt{12 \frac{h_k^{\mathrm{MM}}}{\tilde{h}_k} - 3} - 1 \right)^2. \tag{4.23}$$

Because we need an update strategy that is defined everywhere, we propose to rely on a linear mixture between the MM update and a prolonged version of ME, defined as

$$h_k^{\theta} = \theta h_k^{\mathrm{pME}} + (1-\theta) h_k^{\mathrm{MM}}, \tag{4.24}$$

where $\theta \in (0, 1)$ and h_k^{pME} prolongs the ME update by 0 when the latter does not exist:

$$h_k^{\mathrm{pME}} = \begin{cases} h_k^{\mathrm{ME}} & \text{if } h_k^{\mathrm{ME}} \text{ is defined} \\ 0 & \text{otherwise.} \end{cases} \tag{4.25}$$

It is mathematically easy to check that h_k^{θ} fulfills equation 3.1 for all $\theta \in [0, 1]$ and that positivity is maintained for all $\theta \in [0, 1)$. In practice, values of θ near 1 may be favored to produce larger steps.

When $\beta < 0$ or $\beta > 2$, similar analyses can be conducted. In particular, there are specific values of β for which a closed-form expression of ME updates is available according to Table 3.

When $\beta < 0$, ME updates always exist since equations 4.16 and 4.17 still hold. Moreover, they provide nonincreasing values of $C(\mathbf{h})$, while the latter monotonicity property is not yet proved for the heuristic algorithm. However, simulations tend to indicate that the heuristic algorithm is faster

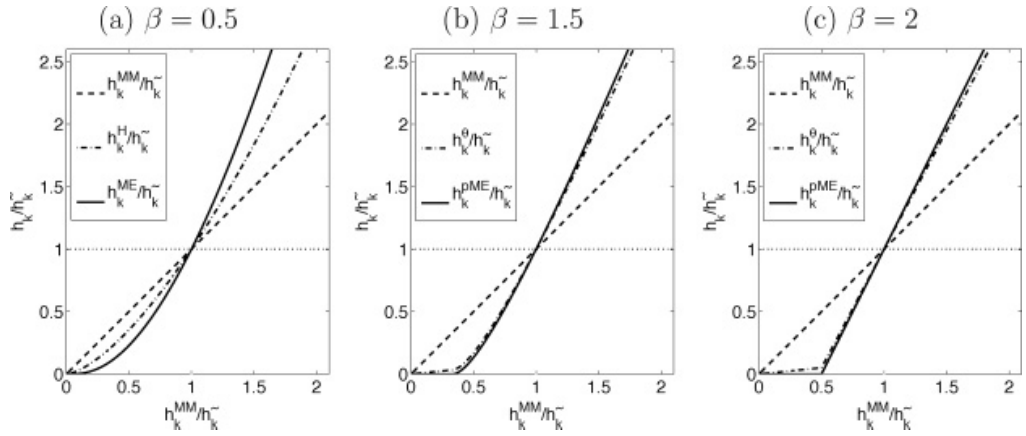


Figure 3: Normalized updates $h_k/\tilde{h}_k$ as functions of $h_k^{\mathrm{MM}}/\tilde{h}_k$ ($\theta = 0.9$). The region between the dotted horizontal line and the solid line corresponds to the steps that fulfill equation 3.1. The larger the departure from the horizontal line, the larger the step is.

than the ME algorithm (which itself is faster than the MM algorithm) in the case $\beta < 0$. This is in conformity with the fact that ME steps can then be proved to be smaller than heuristic steps (on the basis of the reversed inequality mentioned in note 3).

When $\beta > 2$, ME updates do not necessarily exist, akin to the case $\beta \in (1, 2]$. When they exist, they provide nonincreasing values of $C(\mathbf{h})$, while the latter is not yet proved for the heuristic formula. However, since this range of β values is not of practical interest, we will not go further into a detailed analysis here.

4.4 Overrelaxation Properties of the Heuristic and ME Updates. The heuristic and ME updates produce larger steps than the MM update, that is, $|h_k^{\mathrm{H}} - \tilde{h}_k|$ and $|h_k^{\mathrm{ME}} - \tilde{h}_k|$ are larger than $|h_k^{\mathrm{MM}} - \tilde{h}_k|$, for all values of $\beta \in \mathbb{R}$. This is a form of overrelaxation, which will be shown in section 5 to produce faster convergence in practice. The normalized ME (or pME) and heuristic updates studied in previous sections can be written as a function of $h_k^{\mathrm{MM}}/\tilde{h}_k$, such that

$$\frac{h_k}{\tilde{h}_k} = f\left(\frac{h_k^{\mathrm{MM}}}{\tilde{h}_k}\right), \tag{4.26}$$

where h_k is either h_k^{H}, h_k^{ME}, h_k^{pME}, or h_k^{θ}. For the heuristic update, the function is simply given by $f(x) = x^{1/\gamma(\beta)}$. For $\beta \in \{0.5, 1.5, 2\}$, the function f corresponding to the ME/pME update is easily derived from equations 4.19, 4.23, and 4.22. Figure 3 displays the latter functions for the updates studied in this work when $\beta \in \{0.5, 1.5, 2\}$. Overrelaxation appears from the fact that

$h_k < h_k^{\mathrm{MM}}$ whenever $h_k^{\mathrm{MM}} < \tilde{h}_k$ (steps toward left) and $h_k > h_k^{\mathrm{MM}}$ whenever $h_k^{\mathrm{MM}} > \tilde{h}_k$ (steps toward right).

General results about overrelaxation of MM algorithms are given by Salakhutdinov and Roweis (2003), and in particular in the case of NMF. The authors consider the specific case of KL divergence but their study holds for any divergence. They show that in a neighborhood of a stationary point, for any $\eta \in (0, 2)$, relaxed updates h_k^{R} of the form

$$\frac{h_k^{\mathrm{R}}}{\tilde{h}_k} = \left(\frac{h_k^{\mathrm{MM}}}{\tilde{h}_k}\right)^{\eta} \tag{4.27}$$

will converge to the same point as h_k^{MM}, with a different, possibly better, rate of convergence. In particular, the optimal learning rate η, providing the largest rate of convergence, can be computed from the eigenvalues of the Jacobian, at convergence, of the mapping that relates h_k^{MM} at iteration (i) to h_k^{MM} at iteration $(i + 1)$. The optimal learning rate is shown to be always greater than or equal to 1. A similar result was recently obtained by Badeau, Bertin, and Vincent (2010). However, these results do not translate into a practical algorithm, because the latter relaxation property holds only locally, and the computation of the optimal learning rate requires the stationary point to be known. As such, Salakhutdinov and Roweis (2003) propose an adaptive scheme that incrementally proposes values of η greater than 1 at each iteration, and backtrack to $\eta = 1$ when the criterion ceases decreasing.

Our results show that for $\beta \in (0, 1)$, the learning rate $\eta = 1/\gamma(\beta) = 2 - \beta$, corresponding to the heuristic update, ensures descent of the criterion everywhere. The results of Salakhutdinov and Roweis (2003) indicate that the learning rate can be increased to $\eta = 2$ when the algorithm approaches the solution. Note that in the neighborhood of the solution, the Taylor approximation $f(x) \approx f(1) + f'(1)(x - 1)$ applied to $f(x) = x^{\eta}$ implies that

$$(h_k^{\mathrm{R}} - \tilde{h}_k) \approx \eta(h_k^{\mathrm{MM}} - \tilde{h}_k). \tag{4.28}$$

A similar approximation carried out with the ME/pME updates defined by equations 4.19, and 4.23 for $\beta \in \{0.5, 1.5\}$ reveals that in these two cases, $f'(1) = 2$ (and by construction $f(1) = 1$), so that in a neighborhood of the solution, we have

$$(h_k^{\mathrm{ME}} - \tilde{h}_k) \approx 2(h_k^{\mathrm{MM}} - \tilde{h}_k). \tag{4.29}$$

This means that the ME algorithms produce the largest admissible learning rate $\eta = 2$ in the neighborhood of the solution, while avoiding adaptation of the learning rate so as to ensure monotonicity of the criterion. This result holds everywhere for $\beta = 2$, see equation 4.22, by symmetry of the auxiliary function w.r.t to h_k^{MM}. The interested reader may also refer to Lantéri, Roche,

Cuevas, and Aime (2001); and Lantéri, Theys, Richard, and Févotte (2010), for relaxation of multiplicative algorithms using adaptative learning rates computed through line search.

4.5 Implementation and Complexity of the Algorithms. As seen in section 4.4, the update rules of all the studied algorithms can be expressed as functions of the ratio $\nabla^-_{h_k} C(\tilde{\mathbf{h}})/\nabla^+_{h_k} C(\tilde{\mathbf{h}})$, which dominates the algorithmic complexities. Fortunately, the latter ratio takes a simple matrix form that leads to efficient implementations. As such, getting back to the original factorization problem, the heuristic update 4.6 for factors $\mathbf{H}$ and $\mathbf{W}$ can conveniently be expressed in the following matrix form,

$$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]}{\mathbf{W}^T\left[\mathbf{WH}\right]^{\cdot(\beta-1)}}, \tag{4.30}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\frac{\left[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\left[\mathbf{WH}\right]^{\cdot(\beta-1)}\mathbf{H}^T}, \tag{4.31}$$

where the division $\cdot/\cdot$ is here taken entrywise. The MM update simply involves bringing the corrective ratio to the power $\gamma(\beta)$, and the ME update involves applying a function specific to the value of β. Hence, the algorithms have similar complexity $\mathcal{O}(FKN)$, and their implementation takes simple forms. (Matlab implementations of the algorithms discussed in this letter are available online at http://perso.telecom-paristech.fr/~fevotte/Code/code_beta_nmf.zip.)

5 Simulations

In this section, we report performance results of β-NMF algorithms for the specific values $\beta \in \{0.5, 1.5, 2\}$. These values are chosen for their practical interest and because simple ME algorithms exist in their case. This section will evidence the performance improvement brought by the ME approach over the MM or heuristic approaches, with similar computational burden. More precisely, the ME algorithm considered in this section is a mixture of prolonged ME and MM, defined by equation 4.24 and with $\theta = 0.95$, but we will still refer to it as ME for simplicity. The algorithms for all three considered values of β are compared on small-sized synthetic data in section 5.1. The algorithms for $\beta = 0.5$ are analyzed in section 5.2 on the basis of a small music transcription example as this specific value of β has proven efficient for this task (FitzGerald et al., 2009; Vincent et al., 2010; Hennequin et al., 2010).

In the following results we will display the cost values through iterations as well as, following Gonzalez and Zhang (2005), KKT residuals. The residuals allow us to monitor convergence to a stationary point and are here

defined as

$$\mathrm{KKT}(\mathbf{W}) = \| \min \left\{ \mathbf{W}, [(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH} - \mathbf{V})]\mathbf{H}^T \right\} \|_1 / FK, \tag{5.1}$$

$$\mathrm{KKT}(\mathbf{H}) = \| \min \left\{ \mathbf{H}, \mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-2)}.(\mathbf{WH} - \mathbf{V})] \right\} \|_1 / KN. \tag{5.2}$$

They are meant to converge to 0 by equation 2.12. Again, the monotonicity of the heuristic, MM, and ME algorithms does not imply convergence of the iterates to a stationary point. Hence, displaying the KKT residuals allows experimentally checking whether convergence is achieved in practice.

One iteration of each algorithm consists of updating $\mathbf{W}$ given $\mathbf{H}^{(i-1)}$ and $\mathbf{H}$ given $\mathbf{W}^{(i)}$ and then normalizing $\mathbf{W}^{(i)}$ and $\mathbf{H}^{(i)}$ to eliminate trivial scale indeterminacies that leave the cost function unchanged. The normalization step consists of rescaling each column of $\mathbf{W}$ so that $\|\mathbf{w}_k\|_1 = 1$ and rescaling the kth row of $\mathbf{H}$ accordingly. The normalization step is not required but is useful to display and compare the KKT residuals, which are scale sensitive.

5.1 Factorization of Synthetic Data. We consider a synthetic data matrix $\mathbf{V}$ constructed as $\mathbf{V} = \mathbf{W}^*\mathbf{H}^*$ where the ground truth factors are generated as the absolute values of gaussian noise.[4] The matrix can be exactly factorized so that all algorithms should converge to a solution such that $D(\mathbf{V}|\mathbf{WH}) = 0$. The dimensions are $F = 10$, $N = 25$, and $K = 5$. The algorithms (heuristic, MM, and ME for $\beta = 0.5$, MM and ME for $\beta \in \{1.5, 2\}$) are run for 10^5 iterations and initialized with positive random values. Figures 4, 5, and 6 display for each of the three values of β the normalized cost values $D(\mathbf{V}|\mathbf{WH})/FN$, the KKT residuals, and the fit residuals computed as $\|\mathbf{W}^{(i)} - \hat{\mathbf{W}}\|_F / FK$ and $\|\mathbf{H}^{(i)} - \hat{\mathbf{H}}\|_F / KN$, where $\hat{\mathbf{W}}$ and $\hat{\mathbf{H}}$ are the factor estimates at the end of the 10^5 iterations and $\|.\|_F$ is the Frobenius norm. The fit residuals allow measuring the closeness of the current iterates to their end value.

The cost values in all three cases converge to 0 as an exact factorization is reached (oscillations appear in the end iterations as machine precision is reached). Convergence is achieved in all three cases, as shown by both the cost values and KKT residuals. We visually inspected the factorizations returned by the algorithms. For each value of $\beta \in \{0.5, 1.5\}$, the different algorithms appeared to converge to the same solution (and solutions obtained for the two values of β appeared comparable). This was less clear for $\beta = 2$, where ME appeared to reach out a different solution from MM. Still, in this run, ME provides the fastest convergence for every considered value of β. Other runs, obtained from other starting points (obtained randomly), tend to show that when the compared algorithms converge to the same solution, ME converges more quickly. Convergence to a common solution can

[4] For example, in Matlab notation `V = abs(randn(F,K))*abs(randn(K,N))`.

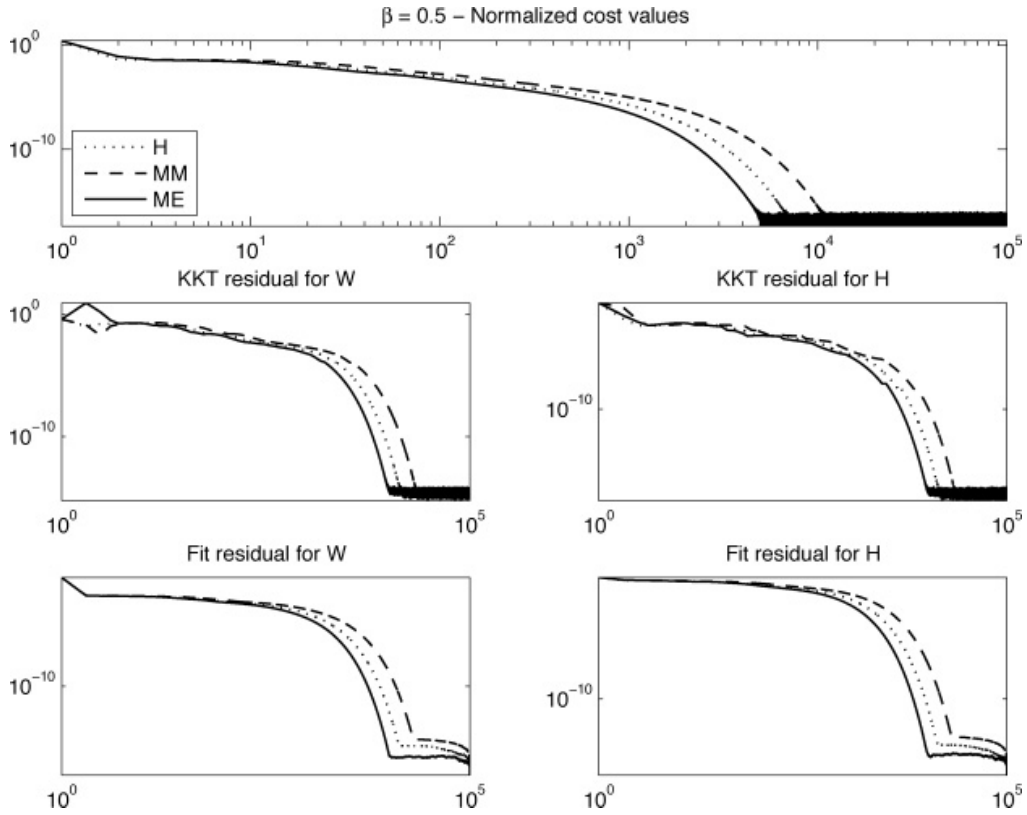


Figure 4: One run of the heuristic (H), ME, and MM algorithms on synthetic data with $\beta = 0.5$. Logarithmic scales for both x- and y-axes.

be controlled in the specific case where $\mathbf{W}$ is fixed and $\beta \in \{1.5, 2\}$, because the objective function is then convex w.r.t. $\mathbf{H}$. In this scenario, ME was found to always converge faster than MM. (These simulations are reported in the companion report available online at http://perso.telecom-paristech.fr/~fevotte/Samples/neco11/beta_nmf_supp.pdf.)

The fit residuals in Figures 4, 5, and 6 show that full convergence will not need to be attained to obtain satisfying solutions for most applications as the fit residual will be considered sufficiently small after a few hundred iterations. Note that the factor iterates do not necessarily converge to the ground-truth values $\mathbf{W}^*$ and $\mathbf{H}^*$ (and this is what we observed) because of the identifiability ambiguities inherent in NMF (Donoho & Stodden, 2004; Laurberg, Christensen, Plumbley, Hansen, and Jensen, 2008).

Using a Matlab implementation run on a Mac 2.6 GHz computer with 2 GB RAM, the CPU time required by each algorithm for the 10^5 iterations is about 60 s for $\beta \in \{0.5, 1.5\}$ and 20 s for $\beta = 2$, including the computation of the cost values and KKT residuals. The ME algorithm is marginally more expensive than MM, itself only slightly more expensive than the heuristic algorithm, for $\beta = 0.5$. The CPU times incurred by the algorithms when $\beta = 2$ is considerably lower thanks to simplifications in equations 4.30

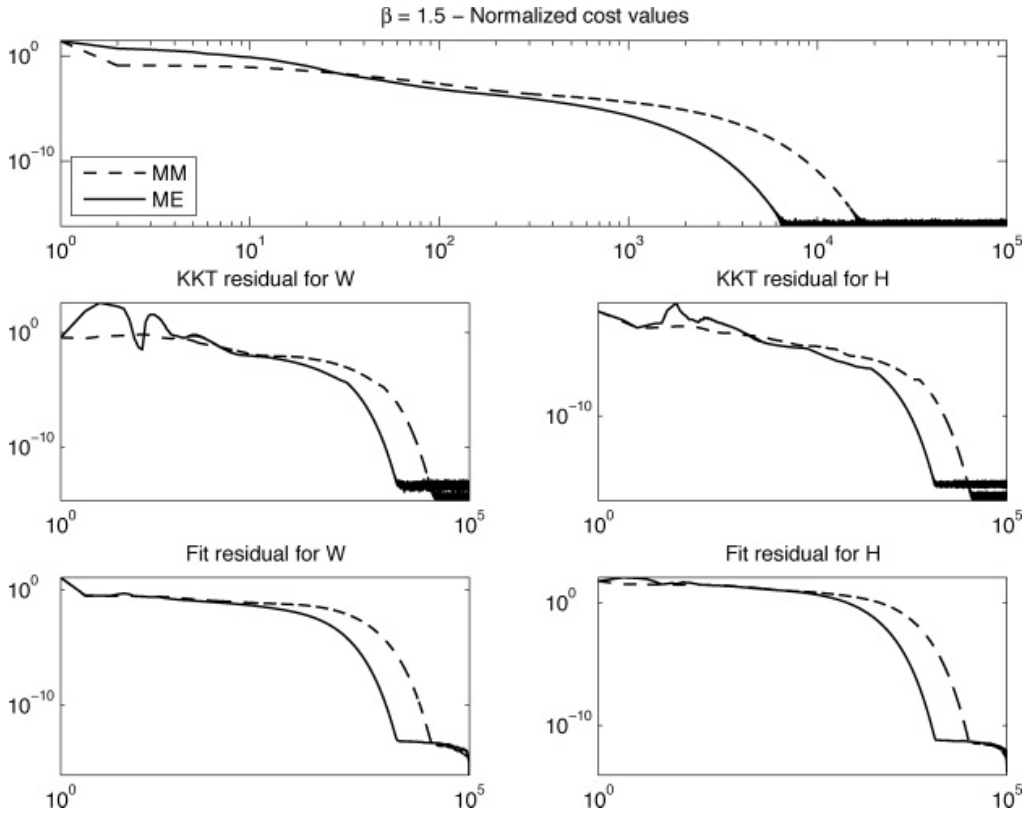


Figure 5: One run of the ME and MM algorithms on synthetic data with $\beta = 1.5$. Logarithmic scales for both x- and y-axes.

and 4.31. Indeed, in the latter case, the term $(\mathbf{WH})\mathbf{H}^T$ appearing at the denominator can more efficiently be computed as $\mathbf{W}(\mathbf{HH}^T)$, which involves a multiplication of matrices with smaller sizes.

5.2 Audio Spectrogram Decomposition. This section addresses the comparison of the heuristic, MM, and ME algorithms for $\beta = 0.5$ applied to an audio spectrogram. We consider the short piano sequence of Févotte et al. (2009), recorded in live conditions, composed of four musical notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. A magnitude spectrogram of the audio signal is computed, leading to nonnegative matrix data $\mathbf{V}$ of size $F = 513$ frequency bins by $N = 674$ time frames. The data are represented in the top left of Figure 7.

As discussed in Févotte et al. (2009), K was set to 6 so as to retrieve in $\mathbf{W}$ the individual spectra of each of the four notes and supplementary spectra corresponding to transients and residual noise. The three algorithms were initialized with common positive random values and run for 10^5 iterations. Figure 7 displays the cost values and KKT residuals along the 10^5 iterations. It was manually checked that the algorithms converged to the

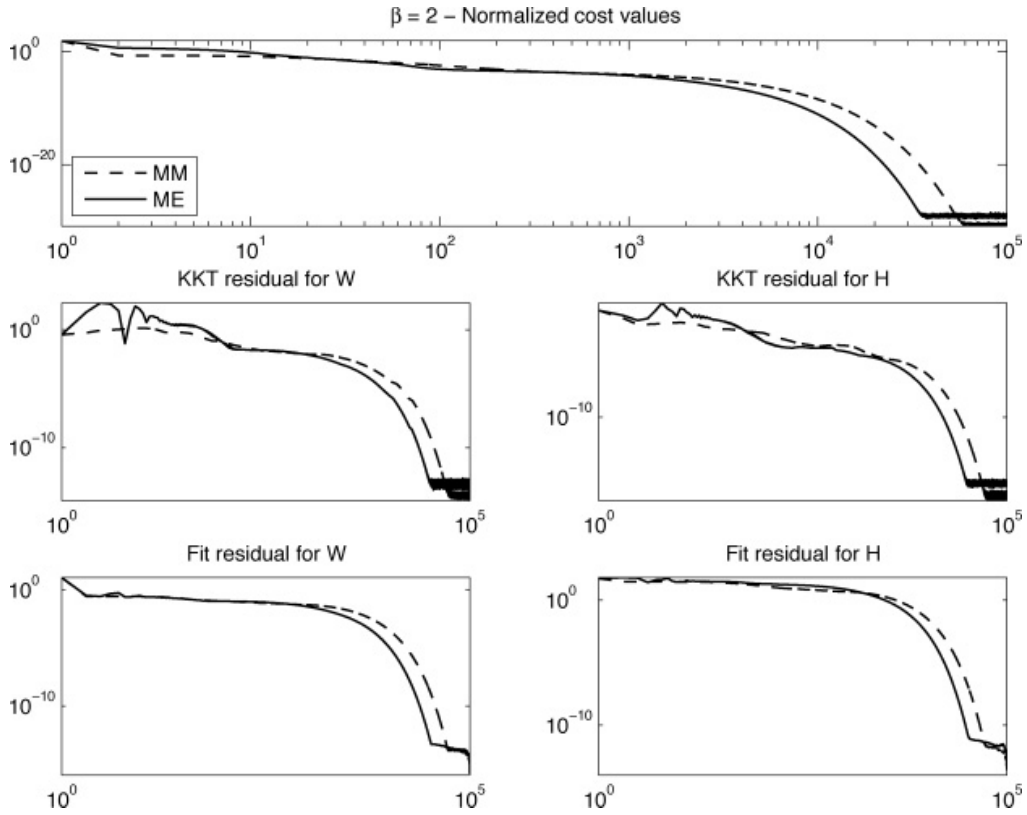


Figure 6: One run of the ME and MM algorithms on synthetic data with $\beta = 2$. Logarithmic scales for both x- and y-axes.

desired ground-truth solution, that is, the notes, transients, and residual noise spectra are correctly unmingled. The three plots show that the ME provides fastest convergence overall, though, judging from the KKT residuals, it appears that convergence is not achieved within the 10^5 iterations. However, the musical pitch values (computed from $\mathbf{W}$ at every iteration) converge to their ground-truth values after only 30, 50, and 580 iterations for ME, heuristic, and MM, respectively. Other initializations yielded two types of results. In a minority of cases, either the heuristic and MM update, on one side, or the ME update, on the other side, converged to a local solution. In the large majority of cases, the three algorithms converge to the same solution, and the results are similar to those of Figure 7: the heuristic algorithm produces the largest decreases of the objective function in the early iterations and is then supplanted by ME. In some runs, the pitch values converged faster with the heuristic algorithm than with ME, and it was found that MM is generally slower than the other two algorithms. These results suggest a mixed update of the form of equation 4.24 where the mixture parameter θ could be made iteration dependent so as to give more weight to the heuristic update in the early iterations and then to ME.

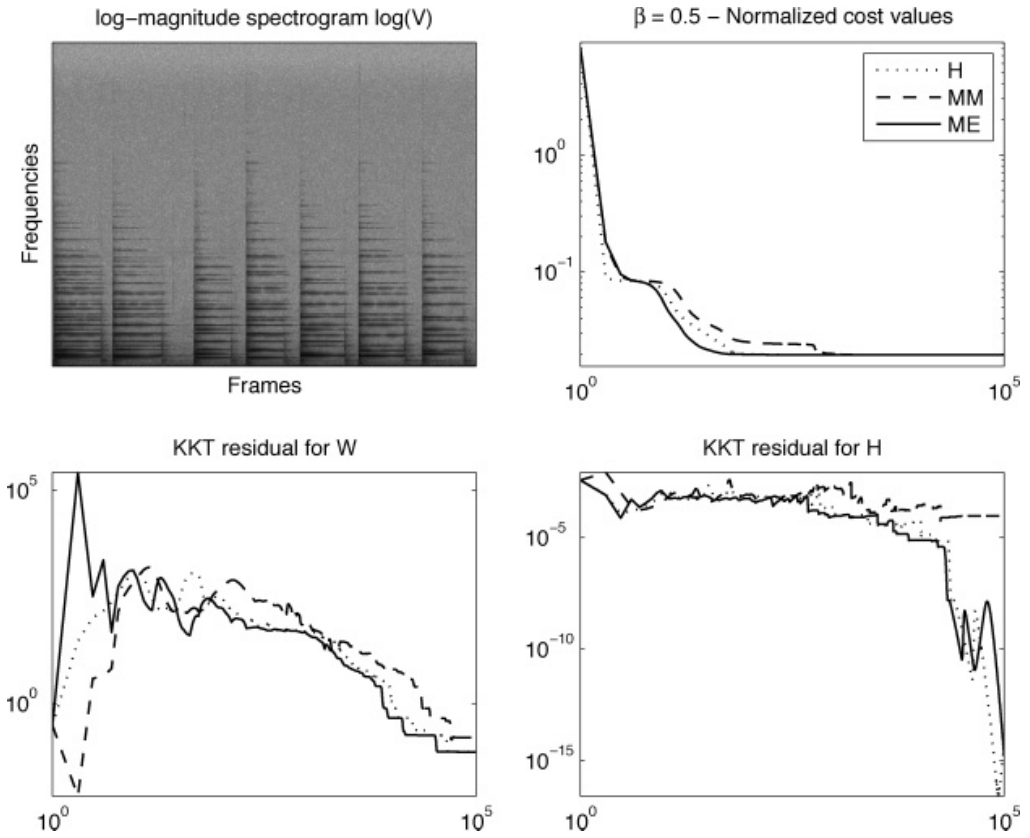


Figure 7: One run of the heuristic (H), MM, and ME algorithms on the piano magnitude spectrogram with $\beta = 0.5$. Logarithmic scales for both x- and y-axes.

5.3 Face Data Decomposition. Finally, in this section we consider decomposition of face data using β-NMF. We use the Olivetti data set, composed of 10 grayscale 8 bits 64×64 face images of 40 people. We retrieved the data in Matlab format from http://cs.nyu.edu/~roweis/data.html. The images are vectorized and form the columns of $\mathbf{V}$, with dimensions $F = 4096$ and $N = 400$. Figure 8 displays the objective functions of one run of the ME and MM algorithms for $\beta \in \{1.5, 2\}$ and illustrates the faster convergence of ME. Other runs led to sensibly similar plots.

As stated in section 1, β-NMF is popular in audio signal processing where the value of β can be controlled so as to improve transcription or separation accuracy. The idea of tuning the value of β so as to optimize performance applies to any NMF-based method for any type of data. As such, to motivate the use of β-NMF in a nonaudio setting, we propose an image interpolation exemple, inspired by Cichocki et al. (2008) and Cemgil (2009), where we show the influence of β on the reconstruction of missing data. We discard 25% of the Olivetti data randomly and produce NMF decompositions using the available data for $\beta \in \{-1, 0, 1, 2, 3\}$ and $K \in \{50, 100, 200\}$. Accounting for the missing data requires minor modifications in the algorithms,

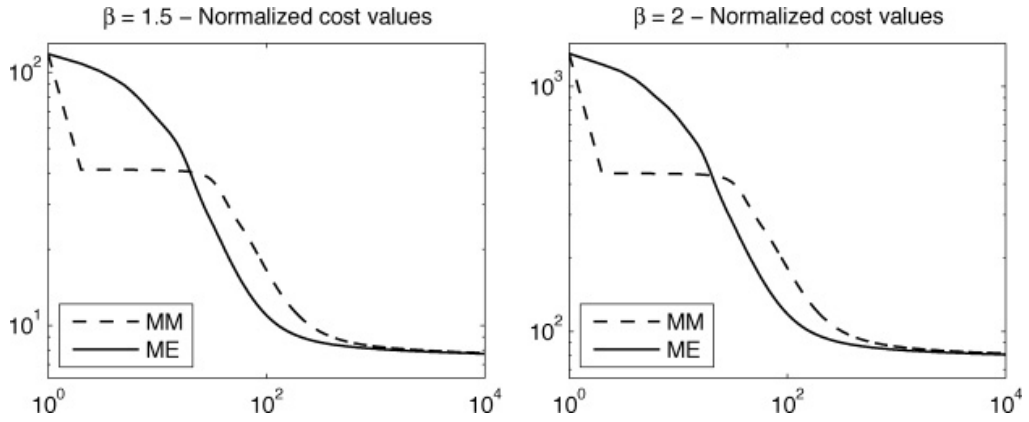
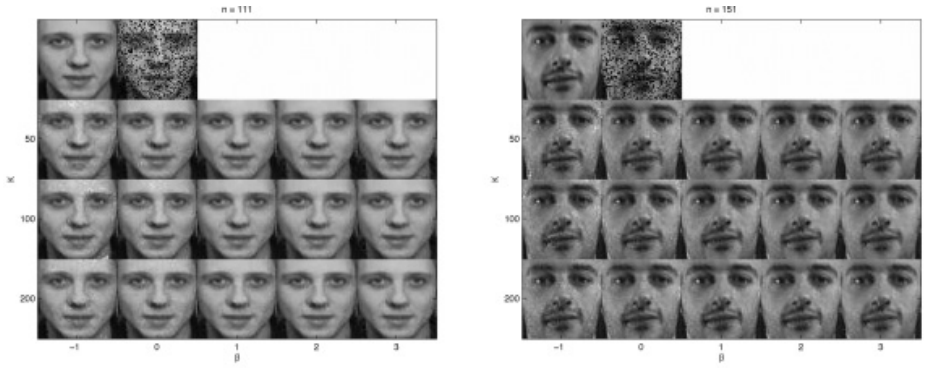


Figure 8: One run of the MM and ME algorithms on the Olivetti data set with $\beta = 1.5$ (left) and $\beta = 2$ (right). Logarithmic scales for both x- and y-axes.



K / β	-1	0	1	2	3
50	28.3	30.7	32.2	31.8	31.2
100	26.6	30.6	32.5	33.2	32.3
200	29.1	31.2	32.4	32.7	32.1

K / β	-1	0	1	2	3
50	26.3	27.9	29.6	29.5	28.8
100	26.2	28.3	29.3	29.8	29.8
200	27.4	28.3	29.0	30.0	30.0

Figure 9: Interpolation results with the Olivetti data set. Original and corrupted data are shown at the top left of each plot. Below are the reconstructions obtained for $K \in \{50, 100, 200\}$ and $\beta \in \{-1, 0, 1, 2, 3\}$. Tables report PSNRs (in dB) of the reconstructions.

basically multiplying $\mathbf{V}$ and its approximate $\mathbf{WH}$ with a binary mask in which zeroes indicate missing pixels (see Ho, 2008; Cichocki et al., 2008; Le Roux, Kameoka, Ono, de Cheveigné, & Sagayama, 2008; Cemgil, 2009; Smaragdis, Raj, & Shashanka, 2009) for similar setups. For simplicity we considered only the MM algorithm, as it is consistently defined for all values of β. It was run from five different initializations for every combination (K, β), and the factorization yielding the lowest end cost value was selected. Figure 9 displays the original image, missing pixels, and reconstructions for

two of the images in the data set. We have also computed the peak signal-to-noise ratio (PSNR) between original and reconstructed images.[5] The maximum mean PSNR value (averaged over all 400 images) is obtained for $\beta = 2$ and $K = 200$. However, since the PSNR value is equivalent to the Euclidean distance between the original and reconstructed image, it is expected that the optimal value of β is biased toward the metric used to assess the quality of reconstruction. Perceptually, we often found the reconstruction obtained with $\beta = 3$ to be more satisfying than with $\beta = 2$.

6 Variants of β-NMF

In this section we briefly discuss how some common variants of NMF, penalized NMF and convex NMF, can be handled under the β-divergence.

6.1 Penalized β-NMF. Supplementary functions of $\mathbf{W}$ or $\mathbf{H}$ (or both) are often added to the cost function, equation 1.3, so as to induce some sort of regularization of the factor estimates or so as to reflect prior belief, for example, in Bayesian maximum a posteriori (MAP) estimation. When such penalty terms are separable in the columns of $\mathbf{H}$ or in the rows of $\mathbf{W}$, penalized NMF essentially amounts to solving the following optimization problem:

$$\min_{\mathbf{h}} C_P(\mathbf{h}) \stackrel{\text{def}}{=} D(\mathbf{v}|\mathbf{W}\mathbf{h}) + L(\mathbf{h}) \text{ subject to } \mathbf{h} \geq 0, \tag{6.1}$$

where $L(\mathbf{h})$ is the penalty term. An auxiliary function to $C_P(\mathbf{h})$ is readily given by

$$G_P(\mathbf{h}|\tilde{\mathbf{h}}) \stackrel{\text{def}}{=} G(\mathbf{h}|\tilde{\mathbf{h}}) + L(\mathbf{h}), \tag{6.2}$$

where $G(\mathbf{h}|\tilde{\mathbf{h}})$ is any auxiliary function to $C(\mathbf{h}) = D(\mathbf{v}|\mathbf{W}\mathbf{h})$. MM or ME algorithms can then be designed on a case-by-case basis. Let us consider a short example for illustration: ℓ_1-norm regularization. In that case, we have

$$L(\mathbf{h}) = \lambda \sum_k h_k, \tag{6.3}$$

where λ is a positive weight parameter. Using the auxiliary function designed in section 3.2 and equation 3.19, the gradient of the penalized

[5] PSNR is a standard evaluation criterion in image reconstruction, defined as $20 \log_{10}(F\,P/\|\mathbf{v} - \hat{\mathbf{v}}\|_2)$, where $\mathbf{v}$ and $\hat{\mathbf{v}}$ denote the vectorized original and reconstructed images and P is the maximum pixel possible value ($P = 255$ in our case).

auxiliary function is

$$\nabla_{h_k} G_L(\mathbf{h}|\tilde{\mathbf{h}}) = \sum_f w_{fk} \left[\breve{d}' \left(v_f | \tilde{v}_f \frac{h_k}{\tilde{h}_k} \right) + \hat{d}'(v_f|\tilde{v}_f) \right] + \lambda.$$

The MM algorithm for ℓ_1-regularized β-NMF takes a very simple form for $\beta \leq 1$, such that

$$h_k = \tilde{h}_k \left(\frac{\sum_f w_{fk}\, v_f\, \tilde{v}_f^{\beta-2}}{\sum_f w_{fk}\, \tilde{v}_f^{\beta-1} + \lambda} \right)^{\gamma(\beta)}. \tag{6.4}$$

This in particular leads to ℓ_1-regularized NMF algorithms for KL-NMF and IS-NMF with proven monotonicity. An update similar to equation 6.4 is obtained for $\beta \geq 2$, but the λ term appears through its sign opposite at the numerator instead of appearing at the denominator. Hence the nonnegativity constraint may become active and must be treated carefully; in that case, our result coincides with similar findings of Pauca, Piper, and Plemmons (2006) and Mørup and Clemmensen (2007) for the specific case of ℓ_1-regularized NMF with the Euclidean distance ($\beta = 2$). In the case $\beta \in (1, 2)$, the MM algorithm does not come up with a simple closed-form update, which supports the fact that in the penalized case, handy algorithms may come only on a case-by-case basis. This is similar to expectation-maximization (EM) procedures for MAP estimation, in which the E-step is essentially unchanged but where the M-step might become intractable because of the penalty term. ME algorithms can also be designed for the ℓ_1-regularized problem, and it can be shown that the results of Table 3 (i.e., the values of β for which a closed-form update exists) still hold in that case.

6.2 Convex β-NMF. In some recent NMF-related works, the dictionary $\mathbf{W}$ is constrained to belong to a known subspace $\mathbf{S} \in \mathbb{R}_+^{F \times M}$ such that

$$\mathbf{W} = \mathbf{SL}, \tag{6.5}$$

where $\mathbf{L} \in \mathbb{R}_+^{M \times K}$. For example, Ding et al. (2010) assume the columns of $\mathbf{W}$ to be linear combinations (with unknown expansion coefficients) of data points (columns of $\mathbf{V}$), so as to enforce the dictionary to be composed of data centroids, while Vincent et al. (2010) assume the dictionary element to be linear combinations of narrow-band spectra so as to enforce harmonicity and smoothness of the dictionary. The term *convex NMF* was introduced by Ding et al. (2010) to express the idea that $\mathbf{W}$ belongs to the convex set of all nonnegative linear combinations of elements of $\mathbf{S}$, but this does not make the optimization problem convex in itself, in the general case.

In this setting, the dictionary update is tantamount to solving

$$\begin{aligned} \min_{\mathbf{L}} \; C_{cv}(\mathbf{L}) &\stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{SLH}) \\ &= \sum_{fn} d\left(v_{fn}|\sum_{mk} s_{fm} l_{mk} h_{kn}\right) \quad \text{subject to } \mathbf{L} \geq 0. \end{aligned} \tag{6.6}$$

In fact, this matricial optimization problem can be turned into vectorial nonnegative linear regression so that the results of section 4 hold. Given some mappings $(f, n) \in \{1, F\} \times \{1, N\} \rightarrow p \in \{1, FN\}$ and $(m, k) \in \{1, M\} \times \{1, K\} \rightarrow q \in \{1, MK\}$, let us introduce the following variables: $\mathbf{T}$ is the matrix of dimension $FN \times MK$ with coefficients $t_{pq} = s_{fm} h_{kn}$, $\underline{\mathbf{v}}$ is the column vector of size FN with coefficients $\underline{v}_p = v_{fn}$, and $\underline{\mathbf{l}}$ is the column vector of size MK with coefficients $\underline{l}_q = l_{mk}$. Then we have

$$D(\mathbf{V}|\mathbf{SLH}) = \sum_{p} d\left(\underline{v}_p|\sum_{q} t_{pq}\, \underline{l}_q\right), \tag{6.7}$$

and thus the estimation of $\mathbf{L}$ amounts to the approximation $\underline{\mathbf{v}} \approx \mathbf{T}\underline{\mathbf{l}}$. As such, any of the algorithms described in section 4 can be employed for this task. As before, the resulting vectorial updates can be turned into matricial updates, leading to simple and efficient implementations. For example, the MM update reads

$$\mathbf{L} \leftarrow \mathbf{L}.\left(\frac{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-2)}.\mathbf{V}\right]\mathbf{H}^T}{\mathbf{S}^T\left[(\mathbf{SLH})^{\cdot(\beta-1)}\right]\mathbf{H}^T}\right)^{\cdot\gamma(\beta)}. \tag{6.8}$$

This result proves the monotonicity of some of the algorithms derived heuristically in Vincent et al. (2010) and also extends the results of Ding et al. (2010) for convex NMF with the Euclidean distance to the more general β-divergence.[6]

7 Conclusion

This letter has addressed NMF with the β-divergence. The problem may be reduced to a mere nonnegative linear regression problem, and our approach is based on the construction of an auxiliary function $G(\mathbf{h}|\tilde{\mathbf{h}})$ that majorizes

[6] More precisely, Ding et al. (2010) consider a "semi"-NMF version where $\mathbf{S} = \mathbf{V}$ and the data are allowed to be real valued, while the nonnegativity constraint is solely imposed on $\mathbf{L}$ and $\mathbf{H}$. Our results do not apply to this more general framework, only to the special case where $\mathbf{V}$ is nonnegative.

the objective function $C(\mathbf{h})$ everywhere and is tight for $\mathbf{h} = \tilde{\mathbf{h}}$. The auxiliary function unifies existing auxiliary functions for the Euclidean distance and the KL divergence (Lee & Seung, 2001), for the generalized divergence of Kompass (2007) (in essence, the β-divergence on its convex part, that is, $\beta \in [1, 2]$) and for the IS divergence (Cao et al., 1999). Various descent algorithms, free of tuning parameters, may then be derived from this auxiliary function. As such, the findings of this letter may be summarized as follows:

- The MM algorithm based on the described auxiliary function is shown to yield multiplicative algorithms for $\beta \in \mathbb{R}$, as described by equation 4.1. For $\beta \in [1, 2]$ (interval of values for which the β-divergence is convex), the MM algorithm coincides with the heuristic algorithm given by equation 4.6, as already known from Kompass (2007).
- In section 4.2, we prove the monotonicity of the heuristic algorithm for $\beta \in (0, 1)$ by proving the inequality $G(\mathbf{h}^{\mathrm{H}}|\tilde{\mathbf{h}}) \leq G(\tilde{\mathbf{h}}|\tilde{\mathbf{h}})$. Hence, with the existing monotonicity results for $\beta = 0$ and $\beta \in [1, 2]$, aggregated, it can now be claimed that the heuristic algorithm is monotone for $\beta \in [0, 2]$, which is the range of values of practical interest that has been considered in the literature.
- In section 4.3, we introduced the concept of maximization-equalization (ME) algorithms. Such algorithms are exhibited for specific values of β, in particular for $\beta \in \{0, 0.5, 1.5, 2\}$, which are values of practical interest. For $\beta = 0$ (IS divergence), the ME algorithm coincides with the heuristic algorithm, whose monotonicity already holds from Cao et al. (1999). For other values of β, the ME algorithms are nonmultiplicative. For $\beta \in \{0.5, 1.5, 2\}$ they amount to solving polynomial equations of order 1 or 2. Simulations have illustrated the faster convergence of the ME approach w.r.t to MM or heuristic, with equivalent complexity.
- In section 6 we considered variants of NMF with the β-divergence. We explained how penalty terms may be handled in the auxiliary function setting; in particular we presented simple multiplicative algorithms for ℓ_1 regularized KL or IS NMF. Then we showed how the algorithms constructed for plain NMF hold for convex NMF, generalizing and proving the monotonicity of existing algorithms.

As for perspectives, this work leaves two important questions unanswered. The first one is the monotonicity of the heuristic algorithm for $\beta \notin [0, 2]$. The monotonicity is observed in practice, but we have not been able to come up with proofs in the presented setting. Either other approaches need to be followed, or a different type of auxiliary function from the one presented here needs to be envisaged. As we suggested in section 2.1, the convex-concave decomposition of the β-divergence is not unique, and decompositions other than the "natural" one employed in this letter may lead

to auxiliary functions that more closely fit the criterion. The second, probably more ambitious, question is the convergence of the sequence of iterates produced by the proposed algorithms to a stationary point. Partial results exist for Euclidean NMF (Lin, 2007a), convergence of multiplicative rules for nonnegative linear regression (i.e., when only one of the two matrices is updated) has been studied in a few cases; see (Titterington, 1987; De Pierro, 1993; Eggermont & LaRiccia, 1998), but general results for NMF with the β-divergence are still lacking. A noteworthy attempt has recently been made by Badeau et al. (2010), which points to difficulties in the convergence study due to the inherent scale ambiguity of factorization models.

Finally, another relevant perspective is the design of new types of β-NMF algorithms. In the Euclidean case, projected gradient methods (Lin, 2007b), second-order active sets methods (Kim, Sra, & Dhillon, 2008), and block-coordinate descent methods (Mairal et al., 2010) have recently been shown to outperform standard multiplicative updates; see also Mørup and Hansen (2009) for a comparison of a selection of algorithms. As such it would be interesting to study how these approaches may extend to the more general β-NMF framework.

Acknowledgments

We thank Francis Bach, Henri Lantéri, Augustin Lefèvre, Cédric Richard, and Céline Theys for inspiring discussions related to optimization in NMF. Many thanks to the reviewers for helpful comments. This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and Applications of Nonnegative Matrix Factorization).

References

Badeau, R., Bertin, N., & Vincent, E. (2010). Stability analysis of multiplicative update algorithms and application to non-negative matrix factorization. *IEEE Transactions on Neural Networks*, *21*(12), 1869–1881.

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, *52*(1), 155–173.

Bertin, N., Févotte, C., & Badeau, R. (2009). A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription. In *Proc. International Conference on Acoustics, Speech and Signal Processing* (pp. 1545–1548). Piscataway, NJ: IEEE.

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of the National Academy of Sciences*, *101*, 4164–4169.

Cao, Y., Eggermont, P. P. B., & Terebey, S. (1999). Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, *8*(2), 286–292.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 785152. doi:10.1155/2009/785152.

Cichocki, A., & Amari, S. (2010). Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, *12*(6), 1532–1568.

Cichocki, A., Lee, H., Kim, Y.-D., & Choi, S. (2008). Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, *29*(9), 1433–1440.

Cichocki, A., Zdunek, R., & Amari, S. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)* (pp. 32–39). Berlin: Springer.

Daube-Witherspoon, M., & Muehllehner, G. (1986). An iterative image space reconstruction algorithm suitable for volume ECT. *IEEE Transactions on Medical Imaging*, *5*(5), 61–66.

De Pierro, A. R. (1993). On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, *12*(2), 328–333.

Dessein, A., Cont, A., & Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proc. 11th International Society for Music Information Retrieval Conference*. Society for Music Information Retrieval.

Dhillon, I. S., & Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems, 19*. Cambridge, MA: MIT Press.

Ding, C. H. Q., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

Donoho, D., & Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems, 16*. Cambridge, MA: MIT Press.

Drakakis, K., Rickard, S., de Frein, R., & Cichocki, A. (2007). Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum*, *3*, 1853–1870.

Eggermont, P. P. B., & LaRiccia, V. N. (1998). *On EM-like algorithms for minimum distance estimation*. http://www.udel.edu/FREC/eggermont/Preprints/emlike.pdf.

Eguchi, S., & Kano, Y. (2001). *Robustifying maximum likelihood estimation*. (Tech. Rep. Research Memo 802). Tokyo: Institute of Statistical Mathematics.

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

Févotte, C., & Cemgil, A. T. (2009). Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)* (pp. 1913–1917). EURASIP.

FitzGerald, D., Cranitch, M., & Coyle, E. (2009). On the use of the beta divergence for musical source separation. In *Proc. Irish Signals and Systems Conference*.

Gao, Y., & Church, G. (2005). Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, *21*, 3970–3975.

Gonzalez, E. F., & Zhang, Y. (2005). *Accelerating the Lee-Seung algorithm for non-negative matrix factorizations* (Tech. Rep.). Houston, TX: Rice University.

Greene, D., Cagney, G., Krogan, N., & Cunningham, P. (2008). Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, *24*(15), 1722–1728.

Hennequin, R., Badeau, R., & David, B. (2010). NMF with time-frequency activations to model non stationary audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)* (pp. 445–448). Piscataway, NJ: IEEE.

Ho, N.-D. (2008). *Nonnegative matrix factorization algorithms and applications*. Unpublished doctoral dissertation, Université Catholique de Louvain.

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, *58*, 30–37.

Kim, D., Sra, S., & Dhillon, I. S. (2008). Fast projection-based methods for the least squares nonnegative matrix approximation problem. *Statistical Analysis and Data Mining*, *1*, 38–51.

Kompass, R. (2007). A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, *19*(3), 780–791.

Lantéri, H., Roche, M., Cuevas, O., & Aime, C. (2001). A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, *81*(5), 945–974.

Lantéri, H., Theys, C., Richard, C., & Févotte, C. (2010). Split gradient method for nonnegative matrix factorization. In *Proc. 18th European Signal Processing Conference (EUSIPCO)*. EURASIP.

Laurberg, H., Christensen, M., Plumbley, M. D., Hansen, L. K., & Jensen, S. H. (2008). Theorems on positive data: On the uniqueness of NMF. *Computational Intelligence and Neuroscience*, article 764206.

Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., & Sagayama, S. (2008). Computational auditory induction by missing-data non-negative matrix factorization. In *Proc. ISCA Workshop on Statistical and Perceptual Audition (SAPA)* (pp. 1–6).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects with nonnegative matrix factorization. *Nature*, *401*, 788–791.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & K.-R. Müller (Eds.), *Advances in neural and information processing systems, 13* (pp. 556–562). Cambridge, MA: MIT Press.

Lin, C.-J. (2007a). On the convergence of multiplicative update algorithms for non-negative matrix factorization. *IEEE Transactions on Neural Networks*, *18*, 1589–1596.

Lin, C.-J. (2007b). Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, *19*, 2756–2779.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astronomical Journal*, *79*, 745–754.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 10–60.

Minami, M., & Eguchi, S. (2002). Robust blind source separation by beta-divergence. *Neural Computation*, *14*, 1859–1886.

Mørup, M., & Clemmensen, L. H. (2007). Multiplicative updates for the LASSO. In *Proc. IEEE Workshop on Machine Learning for Signal Processing*. Piscataway, NJ: IEEE.

Mørup, M., & Hansen, L. K. (2009). Tuning pruning in sparse non-negative matrix factorization. In *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*. EURASIP.

Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N., & Sagayama, S. (2010). Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing*. Piscataway, NJ: IEEE.

O'Grady, P. D. (2007). *Sparse separation of under-determined speech mixtures*. Unpublished doctoral dissertation, National University of Ireland Maynooth.

O'Grady, P. D., & Pearlmutter, B. A. (2008). Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. *Neurocomputing*, *72*(1–3), 88–101.

Pauca, V. P., Piper, J., & Plemmons, R. J. (2006). Nonnegative matrix factorization for spectral data analysis. *Linear Algebra and Its Applications*, *416*, 29–47.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, *62*, 55–59.

Salakhutdinov, R., & Roweis, S. (2003). Adaptive overrelaxed bound optimization methods. In *Proc. International Conference on Machine Learning* (pp. 664–671). CSREA Press.

Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Piscataway, NJ: IEEE Press.

Smaragdis, P., Raj, B., & Shashanka, M. (2009). Missing data imputation for spectral audio signals. In *IEEE International Workshop on Machine Learning for Signal Processing*. Piscataway, NJ: IEEE Press.

Titterington, D. M. (1987). On the iterative image space reconstruction algorithm for ECT. *IEEE Trans. Medical Imaging*, *6*(1), 52–56.

Vincent, E., Bertin, N., & Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech and Language Processing*, *18*, 528–537.

Received October 1, 2010; accepted February 7, 2011.

(Tan & Févotte, *IEEE TPAMI*, 2013)

Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence

Vincent Y.F. Tan, *Member*, *IEEE*, and Cédric Févotte, *Member*, *IEEE*

Abstract—This paper addresses the estimation of the latent dimensionality in nonnegative matrix factorization (NMF) with the β-divergence. The β-divergence is a family of cost functions that includes the squared euclidean distance, Kullback-Leibler (KL) and Itakura-Saito (IS) divergences as special cases. Learning the model order is important as it is necessary to strike the right balance between data fidelity and overfitting. We propose a Bayesian model based on *automatic relevance determination* (ARD) in which the columns of the dictionary matrix and the rows of the activation matrix are tied together through a common scale parameter in their prior. A family of majorization-minimization (MM) algorithms is proposed for maximum a posteriori (MAP) estimation. A subset of scale parameters is driven to a small lower bound in the course of inference, with the effect of pruning the corresponding spurious components. We demonstrate the efficacy and robustness of our algorithms by performing extensive experiments on synthetic data, the `swimmer` dataset, a music decomposition example, and a stock price prediction task.

Index Terms—Nonnegative matrix factorization, model order selection, majorization-minimization, group-sparsity, automatic relevance determination

1 Introduction

Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, nonnegative matrix factorization (NMF) consists in finding a low-rank factorization

$$\mathbf{V} \approx \hat{\mathbf{V}} \triangleq \mathbf{WH}, \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. The common dimension K is usually chosen such that $F\,K + K\,N \ll F\,N$; hence the overall number of parameters to describe the data (i.e., data dimension) is reduced. Early references on NMF include the work of Paatero and Tapper [1] and a seminal contribution by Lee and Seung [2]. Since then, NMF has become a widely used technique for nonsubtractive, parts-based representation of nonnegative data. There are numerous applications of NMF in diverse fields, such as audio signal processing [3], image classification [4], analysis of financial data [5], and bioinformatics [6]. The factorization (1) is usually sought after through the minimization problem:

$$\underset{\mathbf{W},\mathbf{H}}{\text{minimize}}\ D(\mathbf{V}|\mathbf{WH}) \ \text{ subject to } \ \mathbf{W} \geq 0, \mathbf{H} \geq 0, \tag{2}$$

where $\mathbf{A} \geq 0$ means that all entries of the matrix $\mathbf{A}$ are nonnegative (and not positive semidefiniteness). The function $D(\mathbf{V}|\mathbf{WH})$ is a separable measure of fit, i.e.,

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}), \tag{3}$$

where $d(x|y)$ is a scalar cost function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$, and it equals zero when $x = y$. In this paper, we will consider the $d(x|y)$ to be the β-divergence, a family of cost functions parameterized by a single scalar $\beta \in \mathbb{R}$. The squared euclidean (EUC) distance, the generalized Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence are special cases of the β-divergence. NMF with the β-divergence (or, in short, β-NMF) was first considered by Cichocki et al. [7], and more detailed treatments have been proposed in [8], [9], and [10].

1.1 Main Contributions

In most applications, it is crucial that the "right" model order K is selected. If K is too small, the data does not fit the model well. Conversely, if K is too large, overfitting occurs. We seek to find an elegant solution for this dichotomy between data fidelity and overfitting. Traditional model selection techniques such as the Bayesian information criterion (BIC) [11] are not applicable in our setting as the number of parameters is $FK + KN$ and this scales linearly with the number of data points N, whereas BIC assumes that the number of parameters stays constant as the number of data points increases.

To ameliorate this problem, we propose a Bayesian model for β-NMF based on automatic relevance determination (ARD) [12], and in particular, we are inspired by Bayesian principal component analysis (PCA) [13]. We derive *computationally efficient* algorithms with *monotonicity*

- *V.Y.F. Tan is with the Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632, and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. E-mail: vtan@nus.edu.sg.*
- *C. Févotte is with Laboratoire Lagrange (CNRS, Observatoire de la Côte d'Azur & Université de Nice Sophia Antipolis), Parc Valrose, 06000 Nice, France. E-mail: cfevotte@unice.fr.*

Manuscript received 23 Apr. 2012; revised 15 Aug. 2012; accepted 1 Oct. 2012; published online 26 Oct. 2012.
Recommended for acceptance by Y.W. Teh.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2012-04-0312.
Digital Object Identifier no. 10.1109/TPAMI.2012.240.

guarantees to estimate the model order K and to estimate the basis $\mathbf{W}$ and the activation coefficients $\mathbf{H}$. The proposed algorithms are based on the use of auxiliary functions (local majorizations of the objective function). The optimization of these auxiliary functions leads directly to majorization-minimization (MM) algorithms, resulting in efficient multiplicative updates. The monotonicity of the objective function can be proven by leveraging on techniques in [9]. We show via simulations in Section 6 on synthetic data and real datasets (such as a music decomposition example) that the proposed algorithms recover the correct model order and produce better decompositions. We also describe a procedure based on the *method of moments* for adaptive and data-dependent selection of some of the hyperparameters.

1.2 Prior Work

To the best of our knowledge, there is fairly limited literature on model order selection in NMF. References [14] and [15] describe Markov chain Monte Carlo (MCMC) strategies for evaluation of the model evidence in EUC-NMF or KL-NMF. The evidence is calculated for each candidate value of K, and the model with highest evidence is selected. The studies in [16] and [17] describe reversible jump MCMC approaches that allow to sample the model order K, along with any other parameter. These sampling-based methods are computationally intensive. Another class of methods, given in [18], [19], [20], and [21], is closer to the principles that underlie this work; in these works, the number of components K is set to a large value and irrelevant components in $\mathbf{W}$ and $\mathbf{H}$ are driven to zero during inference. A detailed but qualitative comparison between our work and these methods is given in Section 5. In Section 6, we compare the empirical performance of our methods to [18] and [21].

This paper is a significant extension of the authors' conference publication in [22]. First, the cost function in [22] was restricted to be the KL-divergence. In this paper, we consider a continuum of costs parameterized by β, underlying different statistical noise models. We show that this flexibility in the cost function allows for better quality of factorization and model selection on various classes of real-world signals such as audio and images. Second, the algorithms described herein are such that the cost function monotonically decreases to a local minimum whereas the algorithm in [22] is heuristic. Convergence is guaranteed by the MM framework.

1.3 Paper Organization

In Section 2, we state our notation and introduce β-NMF and the MM technique. In Section 3, we present our Bayesian model for β-NMF. Section 4 details ℓ_1- and ℓ_2-ARD for model selection in β-NMF. We then compare the proposed algorithms to other related works in Section 5. In Section 6, we present extensive numerical results to demonstrate the efficacy and robustness of ℓ_1- and ℓ_2-ARD. We conclude the discussion in Section 7.

2 Preliminaries

2.1 Notations

We denote by $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$, the data, dictionary and activation matrices, respectively. These nonnegative matrices are of dimensions $F \times N$, $F \times K$, and $K \times N$, respectively. The entries of these matrices are denoted by v_{fn}, w_{fk}, and h_{kn} respectively. The kth *column* of $\mathbf{W}$ is denoted by $\mathbf{w}_k \in \mathbb{R}_+^F$, and $\underline{h}_k \in \mathbb{R}_+^N$ denotes the kth *row* of $\mathbf{H}$. Thus, $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ and $\mathbf{H} = [\underline{h}_1^T, \ldots, \underline{h}_K^T]^T$.

2.2 NMF with the β-Divergence

This paper considers NMF based on the β-divergence, which we now review. The β-divergence was originally introduced for $\beta \geq 1$ in [23] and [24] and later generalized to $\beta \in \mathbb{R}$ in [7], which is the definition we use here:

$$d_\beta(x|y) \triangleq \begin{cases} \dfrac{x^\beta}{\beta\,(\beta-1)} + \dfrac{y^\beta}{\beta} - \dfrac{x\,y^{\beta-1}}{\beta-1}, & \beta \in \mathbb{R}\backslash\{0,1\}, \\ x\,\log\dfrac{x}{y} - x + y, & \beta = 1, \\ \dfrac{x}{y} - \log\dfrac{x}{y} - 1, & \beta = 0. \end{cases} \quad (4)$$

The limiting cases $\beta = 0$ and $\beta = 1$ correspond to the IS and KL-divergences, respectively. Another case of note is $\beta = 2$, which corresponds to the squared euclidean distance, i.e., $d_{\beta=2}(x|y) = (x-y)^2/2$. The parameter β essentially controls the assumed statistics of the observation noise and can either be fixed or learned from training data by cross-validation. Under certain assumptions, the β-divergence can be mapped to a log-likelihood function for the Tweedie distribution [25], parameterized with respect to its mean. In particular, the values $\beta = 0, 1, 2$ underlie the multiplicative Gamma observation noise, Poisson noise, and Gaussian additive observation noise, respectively. We describe this property in greater detail in Section 3.2. The β-divergence offers a continuum of noise statistics that interpolates between these three specific cases. In the following, we use the notation $D_\beta(\mathbf{V}|\mathbf{WH})$ to denote the separable cost function in (3) with the scalar cost $d = d_\beta$ in (4).

2.3 Majorization-Minimization for β-NMF

We briefly recall some results in [9] on standard β-NMF. In particular, we describe how an MM algorithm [26] that recovers a stationary point of (3) can be derived. The algorithm updates $\mathbf{H}$ given $\mathbf{W}$, and $\mathbf{W}$ given $\mathbf{H}$, and these two steps are essentially the same by the symmetry of $\mathbf{W}$ and $\mathbf{H}$ by transposition ($\mathbf{V} \approx \mathbf{WH}$ is equivalent to $\mathbf{V}^T \approx \mathbf{H}^T\mathbf{W}^T$). Let us thus focus on the optimization of $\mathbf{H}$ given $\mathbf{W}$. The MM framework involves building a (nonnegative) *auxiliary function* $G(\mathbf{H}|\tilde{\mathbf{H}})$ that majorizes the objective $C(\mathbf{H}) = D_\beta(\mathbf{V}|\mathbf{WH})$ everywhere, i.e.,

$$G(\mathbf{H}|\tilde{\mathbf{H}}) \geq C(\mathbf{H}), \quad (5)$$

for all pairs of nonnegative matrices $\mathbf{H}, \tilde{\mathbf{H}} \in \mathbb{R}_+^{K\times N}$. The auxiliary function also matches the cost function whenever its arguments are the same, i.e., for all $\tilde{\mathbf{H}}$,

$$G(\tilde{\mathbf{H}}|\tilde{\mathbf{H}}) = C(\tilde{\mathbf{H}}). \quad (6)$$

If such an auxiliary function exists and the optimization of $G(\mathbf{H}|\tilde{\mathbf{H}})$ over $\mathbf{H}$ for fixed $\tilde{\mathbf{H}}$ is simple, the optimization of $C(\mathbf{H})$ may be replaced by the simpler optimization of $G(\mathbf{H}|\tilde{\mathbf{H}})$ over $\mathbf{H}$. Indeed, any iterate $\mathbf{H}^{(i+1)}$ such that $G(\mathbf{H}^{(i+1)}|\mathbf{H}^{(i)}) \leq G(\mathbf{H}^{(i)}|\mathbf{H}^{(i)})$ reduces the cost since

$$C(\mathbf{H}^{(i+1)}) \leq G(\mathbf{H}^{(i+1)}|\mathbf{H}^{(i)}) \leq G(\mathbf{H}^{(i)}|\mathbf{H}^{(i)}) = C(\mathbf{H}^{(i)}). \quad (7)$$

TABLE 1
The Form of the Auxiliary Function for Various βs [9]

Auxiliary function $G(\mathbf{H}\|\tilde{\mathbf{H}})$	β
$\sum_{kn} q_{kn}h_{kn} - \frac{1}{\beta-1}p_{kn}\tilde{h}_{kn}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + \text{cst}$	$\beta < 1$
$\sum_{kn} q_{kn}h_{kn} - p_{kn}\tilde{h}_{kn}\log\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right) + \text{cst}$	$\beta = 1$
$\sum_{kn} \frac{1}{\beta}q_{kn}\tilde{h}_{kn}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta} - \frac{1}{\beta-1}p_{kn}\tilde{h}_{kn}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + \text{cst}$	$\beta \in (1,2]$
$\sum_{kn} \frac{1}{\beta}q_{kn}\tilde{h}_{kn}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta} - p_{kn}\tilde{h}_{kn} + \text{cst}$	$\beta > 2$

The first inequality follows from (5) and the second from the optimality of $\mathbf{H}^{(i+1)}$. Thus, the MM update is

$$\mathbf{H}^{(i+1)} = \underset{\mathbf{H}\geq 0}{\arg\min}\ G(\mathbf{H}|\mathbf{H}^{(i)}). \tag{8}$$

Note that if $\mathbf{H}^{(i+1)} = \mathbf{H}^{(i)}$, a local minimum is attained since the inequalities in (7) are equalities. The key of the MM approach is thus to build an auxiliary function G which reasonably approximates the original objective at the current iterate $\tilde{\mathbf{H}}$ and such that the function is easy to minimize (over the first variable $\mathbf{H}$). In our setting, the objective function $C(\mathbf{H})$ can be decomposed into the sum of a convex term and a concave term. As such, the construction proposed in [8] and [9] involves majorizing the convex and concave terms separately, using Jensen's inequality and a first-order Taylor approximation, respectively. Denoting $\tilde{v}_{fn} \triangleq [\mathbf{W}\tilde{\mathbf{H}}]_{fn}$ and

$$p_{kn} \triangleq \sum_f w_{fk}v_{fn}\tilde{v}_{fn}^{\beta-2}, \qquad q_{kn} \triangleq \sum_f w_{fk}\tilde{v}_{fn}^{\beta-1}, \tag{9}$$

the resulting auxiliary function can be expressed as in Table 1, where cst denotes constant terms that do not depend on $\mathbf{H}$. In the sequel, the use of the tilde over a parameter will generally denote its *previous* iterate. Minimization of $G(\mathbf{H}|\tilde{\mathbf{H}})$ with respect to (w.r.t) $\mathbf{H}$ thus leads to the following simple update:

$$h_{kn} = \tilde{h}_{kn}\left(\frac{p_{kn}}{q_{kn}}\right)^{\gamma(\beta)}, \tag{10}$$

where the exponent $\gamma(\beta)$ is defined as

$$\gamma(\beta) \triangleq \begin{cases} 1/(2-\beta), & \beta < 1, \\ 1, & 1 \leq \beta \leq 2, \\ 1/(\beta-1), & \beta > 2. \end{cases} \tag{11}$$

3 The Model for Automatic Relevance Determination in β-NMF

In this section, we describe our probabilistic model for NMF. The model involves tying the kth column of $\mathbf{W}$ to the kth row of $\mathbf{H}$ together through a common scale parameter λ_k. If λ_k is driven to zero (or, as we will see, a positive lower bound) during inference, then all entries in the corresponding column of $\mathbf{W}$ and row of $\mathbf{H}$ will also be driven to zero.

3.1 Priors

We are inspired by Bayesian PCA [13], where each element of $\mathbf{W}$ is assigned a Gaussian prior with column-dependent variance-like parameters λ_k. These λ_ks are known as the *relevance weights*. However, our formulation has two main differences vis-à-vis Bayesian PCA. First, there are no nonnegativity constraints in Bayesian PCA. Second, in Bayesian PCA, thanks to the simplicity of the statistical model (multivariate Gaussian observations with Gaussian parameter priors), $\mathbf{H}$ can be easily integrated out of the likelihood, and the optimization can be done over $p(\mathbf{W}, \boldsymbol{\lambda}|\mathbf{V})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}_+^K$ is the vector of relevance weights. We have to maintain the nonnegativity of the elements in $\mathbf{W}$ and $\mathbf{H}$ and also, in our setting, the activation matrix $\mathbf{H}$ cannot be integrated out analytically.

To ameliorate the above-mentioned problems, we propose to tie the columns of $\mathbf{W}$ and the rows of $\mathbf{H}$ together through common scale parameters. This construction is not overconstraining the scales of $\mathbf{W}$ and $\mathbf{H}$ because of the inherent scale indeterminacy between $\mathbf{w}_k$ and $\underline{h}_k$. Moreover, we choose nonnegative priors for $\mathbf{W}$ and $\mathbf{H}$ to ensure that all elements of the basis and activation matrices are nonnegative. We adopt a Bayesian approach and assign $\mathbf{W}$ and $\mathbf{H}$ Half-Normal or Exponential priors. When $\mathbf{W}$ and $\mathbf{H}$ have Half-Normal priors:

$$p(w_{fk}|\lambda_k) = \mathcal{HN}(w_{fk}|\lambda_k), \quad p(h_{kn}|\lambda_k) = \mathcal{HN}(h_{kn}|\lambda_k), \tag{12}$$

where for $x \geq 0$, $\mathcal{HN}(x|\lambda) \triangleq (\frac{2}{\pi\lambda})^{1/2}\exp(-\frac{x^2}{2\lambda})$, and $\mathcal{HN}(x \mid \lambda) = 0$ when $x < 0$. Note that if x is a Gaussian (Normal) random variable, then $|x|$ is a Half-Normal. When $\mathbf{W}$ and $\mathbf{H}$ are assigned Exponential priors:

$$p(w_{fk}|\lambda_k) = \mathcal{E}(w_{fk}|\lambda_k), \quad p(h_{kn}|\lambda_k) = \mathcal{E}(h_{kn}|\lambda_k), \tag{13}$$

where for $x \geq 0$, $\mathcal{E}(x|\lambda) \triangleq \frac{1}{\lambda}\exp(-\frac{x}{\lambda})$, and $\mathcal{E}(x|\lambda) = 0$ otherwise. Note from (12) and (13) that the kth column of $\mathbf{W}$ and the kth row of $\mathbf{H}$ are tied together by a *common* variance-like parameter λ_k, also known as the *relevance weight*. When a particular λ_k is small, that particular column of $\mathbf{W}$ *and* row of $\mathbf{H}$ are not relevant and vice versa. When a row and a column are not relevant, their norms are close to zero and thus can be removed from the factorization without compromising too much on data fidelity. This removal of *common* rows and columns makes the model more parsimonious.

Finally, we impose inverse-Gamma priors on each relevance weight λ_k, i.e.,

$$p(\lambda_k; a, b) = \mathcal{IG}(\lambda_k|a, b) = \frac{b^a}{\Gamma(a)}\lambda_k^{-(a+1)}\exp\left(-\frac{b}{\lambda_k}\right), \tag{14}$$

where a and b are the (nonnegative) shape and scale hyperparameters, respectively. We set a and b to be constant

for all k. We will state how to choose these in a principled manner in Section 4.5. Furthermore, each relevance parameter is independent of every other, i.e., $p(\boldsymbol{\lambda}; a, b) = \prod_{k=1}^{K} p(\lambda_k; a, b)$.

3.2 Likelihood

The β-divergence is related to the family of Tweedie distributions [25]. The relation was noted by Cichocki et al. [27] and detailed in [28]. The Tweedie distribution is a special case of the exponential dispersion model [29], itself a generalization of the more familiar natural exponential family. It is characterized by the simple polynomial relation between its mean and variance:

$$\mathrm{var}[x] = \phi \mu^{2-\beta}, \tag{15}$$

where $\mu = \mathbb{E}[x]$ is the mean, β is the *shape parameter* , and ϕ is referred to as the *dispersion parameter*. The Tweedie distribution is only defined for $\beta \leq 1$ and $\beta \geq 2$. For $\beta \neq 0, 1$, its probability density function (pdf) or probability mass function (pmf) can be written in the following form:

$$\mathcal{T}(x|\mu, \phi, \beta) = h(x, \phi) \exp\left[\frac{1}{\phi}\left(\frac{1}{\beta - 1} x \mu^{\beta-1} - \frac{1}{\beta}\mu^{\beta}\right)\right], \tag{16}$$

where $h(x, \phi)$ is referred to as the *base function*. For $\beta \in \{0, 1\}$, the pdf or pmf takes the appropriate limiting form of (16). The support of $\mathcal{T}(x|\mu, \phi, \beta)$ varies with the value of β, but the set of values that μ can take on is generally $\mathbb{R}^+$, except for $\beta = 2$, for which it is $\mathbb{R}$, and the Tweedie distribution coincides with the Gaussian distribution of mean μ and variance ϕ. For $\beta = 1$ (and $\phi = 1$), the Tweedie distribution coincides with the Poisson distribution. For $\beta = 0$, it coincides with the Gamma distribution with shape parameter $\alpha = 1/\phi$ and scale parameter μ/α.[1] The base function admits a closed form only for $\beta \in \{-1, 0, 1, 2\}$.

Finally, the *deviance* of Tweedie distribution, i.e., the log-likelihood ratio of the saturated ($\mu = x$) and general model, is proportional to the β-divergence. In particular,

$$\log \frac{\mathcal{T}(x|\mu = x, \phi, \beta)}{\mathcal{T}(x|\mu, \phi, \beta)} = \frac{1}{\phi} d_{\beta}(x|\mu), \tag{17}$$

where $d_{\beta}(\,\cdot\,|\,\cdot\,)$ is the scalar cost function defined in (4). As such the β-divergence acts as a minus log-likelihood for the Tweedie distribution whenever the latter is defined. Because the data coefficients $\{v_{fn}\}$ are conditionally independent given $(\mathbf{W}, \mathbf{H})$, the negative log-likelihood function is

$$-\log p(\mathbf{V}|\mathbf{W}, \mathbf{H}) = \frac{1}{\phi} D_{\beta}(\mathbf{V}|\mathbf{WH}) + \mathrm{cst}. \tag{18}$$

3.3 Objective Function

We now form the maximum a posteriori (MAP) objective function for the model described in Sections 3.1 and 3.2. Due to (12), (13), (14), and (18):

$$C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}) \stackrel{\triangle}{=} -\log p(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda}|\mathbf{V}), \tag{19}$$

$$\begin{aligned} &= \frac{1}{\phi} D_{\beta}(\mathbf{V}|\mathbf{WH}) + \sum_{k=1}^{K} \frac{1}{\lambda_k}\left(f(\mathbf{w}_k) + f(\underline{h}_k) + b\right) \\ &\quad + c \log \lambda_k + \mathrm{cst}, \end{aligned} \tag{20}$$

where (20) follows from Bayes' rule and, for the two statistical models,

- Half-Normal model as in (12), $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ and $c = (F + N)/2 + a + 1$;
- Exponential model as in (13), $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $c = F + N + a + 1$.

Observe that for the regularized cost function in (20), the second term is monotonically decreasing in λ_k, while the third term is monotonically increasing in λ_k. Thus, a subset of the λ_ks will be forced to a lower bound, which we specify in Section 4.4, while the others will tend to a larger value. This serves the purpose of pruning irrelevant components out of the model. In fact, the vector of relevance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ can be optimized analytically in (20) leading to an objective function that is a function of $\mathbf{W}$ and $\mathbf{H}$ only, i.e.,

$$\begin{aligned} C(\mathbf{W}, \mathbf{H}) &= \frac{1}{\phi} D_{\beta}(\mathbf{V}|\mathbf{WH}) \\ &\quad + c \sum_{k=1}^{K} \log(f(\mathbf{w}_k) + f(\underline{h}_k) + b) + \mathrm{cst}, \end{aligned} \tag{21}$$

where $\mathrm{cst} = Kc(1 - \log c)$.

In our algorithms, instead of optimizing (21), we keep λ_k as an auxiliary variable for optimizing $C(\mathbf{W}, \mathbf{H}, \boldsymbol{\lambda})$ in (20) to ensure that the columns $\mathbf{H}$ and the rows of $\mathbf{W}$ are decoupled. More precisely, $\mathbf{w}_k$ and $\underline{h}_k$ are conditionally independent given λ_k. In fact, (21) shows that the $\boldsymbol{\lambda}$-optimized objective function $C(\mathbf{W}, \mathbf{H})$ induces sparse regularization among groups, where the groups are pairs of columns and rows, i.e., $\{\mathbf{w}_k, \underline{h}_k\}$. In this sense, our work is related to group LASSO [30] and its variants. See, for example, [31]. The function $x \mapsto \log(x + b)$ in (21) is a sparsity-inducing term and is related to reweighted ℓ_1-minimization [32]. We discuss these connections in greater detail in the supplementary material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.240, [33].

4 Inference Algorithms

In this section, we describe two algorithms for optimizing the objective function (20) for $\mathbf{H}$ given fixed $\mathbf{W}$. The updates for $\mathbf{W}$ are symmetric given $\mathbf{H}$. These algorithms will be based on the MM idea for β-NMF and on the two prior distributions of $\mathbf{W}$ and $\mathbf{H}$. In particular, we use the auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ defined in Table 1 as an upper bound of the data fit term $D_{\beta}(\mathbf{V}|\mathbf{WH})$.

4.1 Algorithm for ℓ_2-ARD β-NMF

We now introduce ℓ_2-ARD β-NMF. In this algorithm, we assume that $\mathbf{W}$ and $\mathbf{H}$ have Half-Normal priors as in (12) and thus the regularizer is

1. We employ the following convention for the Gamma distribution $\mathcal{G}(x; a, b) = x^{a-1} e^{-x/b}/(b^a \Gamma(a))$.

$$R_2(\mathbf{H}) \triangleq \sum_k \frac{1}{\lambda_k} f(\underline{h}_k) = \sum_{kn} \frac{1}{2\lambda_k} h_{kn}^2. \tag{22}$$

The main idea behind the algorithms is as follows: Consider the function $F(\mathbf{H}|\tilde{\mathbf{H}}) \triangleq \phi^{-1}\, G(\mathbf{H}|\tilde{\mathbf{H}}) + R_2(\mathbf{H})$, which is the original auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ times ϕ^{-1} plus the ℓ_2 regularization term. It can, in fact, be easily shown in [9, Section 6] that $F(\mathbf{H}|\tilde{\mathbf{H}})$ is an auxiliary function to the (penalized) objective function in (20). Ideally, we would take the derivative of $F(\mathbf{H}|\tilde{\mathbf{H}})$ w.r.t h_{kn} and set it to zero. Then the updates would proceed in a manner analogous to (10). However, the regularization term $R_2(\mathbf{H})$ does not "fit well" with the form of the auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ in the sense that $\nabla_{\mathbf{H}} F(\mathbf{H}|\tilde{\mathbf{H}}) = 0$ cannot be solved analytically for all $\beta \in \mathbb{R}$. Thus, our idea for ℓ_2-ARD is to consider the cases $\beta \geq 2$ and $\beta < 2$ separately and to find an upper bound of $F(\mathbf{H}|\tilde{\mathbf{H}})$ by some other auxiliary function $J(\mathbf{H}|\tilde{\mathbf{H}})$ so that the resulting equation $\nabla_{\mathbf{H}} J(\mathbf{H}|\tilde{\mathbf{H}}) = 0$ can be solved in closed-form.

To derive our algorithms, we first note the following.

Lemma 1. *For every $\nu > 0$, the function $g_\nu(t) = \frac{1}{t}(\nu^t - 1)$ is monotonically nondecreasing in $t \in \mathbb{R}$. In fact, $g_\nu(t)$ is monotonically increasing unless $\nu = 1$.*

In the above lemma, $g_\nu(0) \triangleq \log \nu$ by L'Hôpital's rule. The proof of this simple result can be found in [34].

We first derive ℓ_2-ARD for $\beta > 2$. The idea is to upper bound the regularizer $R_2(\mathbf{H})$ in (22) elementwise using Lemma 1, and is equivalent to the *moving-term* technique described by Yang and Oja in [34] and [35]. Indeed, we have

$$\frac{1}{2}\left[\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 - 1\right] \leq \frac{1}{\beta}\left[\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^\beta - 1\right], \tag{23}$$

by taking $\nu = h_{kn}/\tilde{h}_{kn}$ in Lemma 1. Thus, for $\beta > 2$,

$$\frac{1}{2\lambda_k} h_{kn}^2 \leq \frac{1}{\lambda_k \beta} \tilde{h}_{kn}^2 \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^\beta + \text{cst}, \tag{24}$$

where cst is a constant w.r.t the optimization variable h_{kn}. We upper bound the regularizer (22) elementwise by (24). The resulting auxiliary function (modified version of $F(\mathbf{H}|\tilde{\mathbf{H}})$) is

$$J(\mathbf{H}|\tilde{\mathbf{H}}) = \frac{1}{\phi}\, G(\mathbf{H}|\tilde{\mathbf{H}}) + \sum_{kn} \frac{1}{\lambda_k \beta} \tilde{h}_{kn}^2 \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^\beta. \tag{25}$$

Note that (24) holds with equality iff $\nu = 1$ or, equivalently, $h_{kn} = \tilde{h}_{kn}$ so (6) holds. Thus, $J(\mathbf{H}|\tilde{\mathbf{H}})$ is indeed an auxiliary function to $F(\mathbf{H}|\tilde{\mathbf{H}})$. Recalling the definition of $G(\mathbf{H}|\tilde{\mathbf{H}})$ for $\beta > 2$ in Table 1, differentiating $J(\mathbf{H}|\tilde{\mathbf{H}})$ w.r.t h_{kn} and setting the result to zero yields the update

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + (\phi/\lambda_k)\tilde{h}_{kn}}\right)^{1/(\beta-1)}. \tag{26}$$

Note that the exponent $1/(\beta-1)$ corresponds to $\gamma(\beta)$ for the $\beta > 2$ case. Also observe that the update is similar to MM for β-NMF (cf. (10)) except that there is an additional term in the denominator.

Algorithm 1. ℓ_2-ARD for β-NMF

Input: Data matrix $\mathbf{V}$, hyperparameter a, tolerance τ
Output: Nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$, nonnegative relevance vector $\boldsymbol{\lambda}$ and model order K_{eff}
Init: Fix K. Initialize $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ to nonnegative values and tolerance parameter $\text{tol} = \infty$
Calculate: $c = (F+N)/2 + a + 1$ and $\xi(\beta)$ as in (31)
Calculate: Hyperparameter b as in (38)
while (tol $< \tau$) **do**

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \left(\frac{\mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}]}{\mathbf{W}^T[(\mathbf{WH})]^{\cdot(\beta-1)} + \phi \mathbf{H}/\text{repmat}(\boldsymbol{\lambda},1,N)}\right)^{\cdot\xi(\beta)}$$

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \left(\frac{[(\mathbf{WH})^{\cdot(\beta-2)}.\mathbf{V}]\mathbf{H}^T}{[(\mathbf{WH})^{\cdot(\beta-1)}]\mathbf{H}^T + \phi \mathbf{W}/\text{repmat}(\boldsymbol{\lambda},F,1)}\right)^{\cdot\xi(\beta)}$$

$\lambda_k \leftarrow [(\frac{1}{2}\sum_f w_{fk}^2 + \frac{1}{2}\sum_n h_{kn}^2) + b]/c$ for all k
$\text{tol} \leftarrow \max_{k=1,\ldots,K} |(\lambda_k - \tilde{\lambda}_k)/\tilde{\lambda}_k|$
end while
Calculate: K_{eff} as in (34)

For the case $\beta \leq 2$, our strategy is not to majorize the regularization term. Rather, we majorize the auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ itself. By applying Lemma 1 with $\nu = h_{kn}/\tilde{h}_{kn}$, we have that for all $\beta \leq 2$:

$$\frac{1}{\beta}\left[\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^\beta - 1\right] \leq \frac{1}{2}\left[\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 - 1\right], \tag{27}$$

which means that

$$\frac{1}{\beta} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^\beta \leq \frac{1}{2} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 + \text{cst}. \tag{28}$$

By replacing the first term of $G(\mathbf{H}|\tilde{\mathbf{H}})$ in Table 1 (for $\beta \leq 2$) with the upper bound above, we have the following new objective function:

$$J(\mathbf{H}|\tilde{\mathbf{H}}) = \sum_{kn} \frac{q_{kn}\tilde{h}_{kn}}{2\phi}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 - \frac{p_{kn}\tilde{h}_{kn}}{\phi(\beta-1)}\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^{\beta-1} + \frac{h_{kn}^2}{2\lambda_k}. \tag{29}$$

Differentiating $J(\mathbf{H}|\tilde{\mathbf{H}})$ w.r.t h_{kn} and setting to zero yields the simple update

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + (\phi/\lambda_k)\tilde{h}_{kn}}\right)^{1/(3-\beta)}. \tag{30}$$

To summarize the algorithm concisely, we define the exponent used in the updates in (26) and (30) as

$$\xi(\beta) \triangleq \begin{cases} 1/(3-\beta) & \beta \leq 2, \\ 1/(\beta-1) & \beta > 2. \end{cases} \tag{31}$$

Finally, we remark that even though the updates in (26) and (30) are easy to implement, we either majorized the regularizer $R_2(\mathbf{H})$ or the auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$. These bounds may be loose and thus may lead to slow convergence in the resulting algorithm. In fact, we can show that for $\beta = 0, 1, 2$, we do not have to resort to upper bounding the original function $F(\mathbf{H}|\tilde{\mathbf{H}}) = \phi^{-1}\, G(\mathbf{H}|\tilde{\mathbf{H}}) + R_2(\mathbf{H})$. Instead, we can choose to solve a polynomial equation to update h_{kn}. The cases $\beta = 0, 1, 2$ correspond to

solving cubic, quadratic, and linear equations in h_{kn}, respectively. It is also true that for all rational β, we can form a polynomial equation in h_{kn}, but the order of the resulting polynomial depends on the exact value of β. See the online supplementary material [33].

4.2 Algorithm for ℓ_1-ARD β-NMF

The derivation of ℓ_1-ARD β-NMF is similar to its ℓ_2 counterpart. We find majorizers for either the likelihood or the regularizer. We omit the derivations and refer the reader to the online supplementary material [33]. In sum:

$$h_{kn} = \tilde{h}_{kn}\left(\frac{p_{kn}}{q_{kn}+\phi/\lambda_k}\right)^{\gamma(\beta)}, \tag{32}$$

where $\gamma(\beta)$ is defined in (11).

Algorithm 2. ℓ_1-ARD for β-NMF

Input: Data matrix $\mathbf{V}$, hyperparameter a, tolerance τ

Output: Nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$, nonnegative relevance vector $\boldsymbol{\lambda}$ and model order K_{eff}

Init: Fix K. Initialize $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ to nonnegative values and tolerance parameter $\text{tol} = \infty$

Calculate: $c = F + N + a + 1$ and $\gamma(\beta)$ as in (11)

Calculate: Hyperparameter b as in (38)

while (tol $< \tau$) **do**

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \left(\frac{\mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-2)}\cdot\mathbf{V}]}{\mathbf{W}^T[(\mathbf{WH})^{\cdot(\beta-1)}]+\phi/\text{repmat}(\boldsymbol{\lambda},1,N)}\right)^{\cdot\gamma(\beta)}$$

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \left(\frac{[(\mathbf{WH})^{\cdot(\beta-2)}\cdot\mathbf{V}]\mathbf{H}^T}{[(\mathbf{WH})^{\cdot(\beta-1)}]\mathbf{H}^T+\phi/\text{repmat}(\boldsymbol{\lambda},F,1)}\right)^{\cdot\gamma(\beta)}$$

$\lambda_k \leftarrow (\sum_f w_{fk} + \sum_n h_{kn} + b)/c$ for all k

$\text{tol} \leftarrow \max_{k=1,\ldots,K}|(\lambda_k - \tilde{\lambda}_k)/\tilde{\lambda}_k|$

end while

Calculate: K_{eff} as in (34)

4.3 Update of λ_k

We have described how to update $\mathbf{H}$ using either ℓ_1-ARD or ℓ_2-ARD. Since $\mathbf{H}$ and $\mathbf{W}$ are related in a symmetric manner, we have also effectively described how to update $\mathbf{W}$. We now describe a simple update rule for the λ_ks. This update is the same for both ℓ_1- and ℓ_2-ARD. We first find the partial derivative of $C(\mathbf{W},\mathbf{H},\boldsymbol{\lambda})$ w.r.t λ_k and set it to zero. This gives the update:

$$\lambda_k = \frac{f(\mathbf{w}_k) + f(\underline{h}_k) + b}{c}, \tag{33}$$

where $f(\cdot)$ and c are defined after (20).

4.4 Stopping Criterion and Determination of K_{eff}

In this section, we describe the stopping criterion and the determination of the effective number of components K_{eff}. Let $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_K)$ and $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1,\ldots,\tilde{\lambda}_K)$ be the vector of relevance weights at the current (updated) and previous iterations, respectively. The algorithm is terminated whenever $\text{tol} \triangleq \max_{k=1,\ldots,K}|(\lambda_k - \tilde{\lambda}_k)/\tilde{\lambda}_k|$ falls below some threshold $\tau > 0$. Note from (33) that iterates of each λ_k are bounded from below as $\lambda_k \geq B \triangleq b/c$ and this bound is attained when $\mathbf{w}_k$ and $\underline{h}_k$ are zero vectors, i.e., the kth column of $\mathbf{W}$ and the kth row of $\mathbf{H}$ are pruned out of the model. After convergence, we set K_{eff} to be the number of components of such that the ratio $(\lambda_k - B)/B$ is strictly larger than τ, i.e.,

$$K_{\text{eff}} \triangleq \left|\left\{k \in \{1,\ldots,K\} : \frac{\lambda_k - B}{B} > \tau\right\}\right|, \tag{34}$$

where $\tau > 0$ is some threshold. We choose this threshold to be the same as that for the tolerance level tol.

The algorithms ℓ_2-ARD and ℓ_1-ARD are detailed in Algorithms 1 and 2, respectively. In the algorithms, we use the notation $\mathbf{A}\cdot\mathbf{B}$ to mean entrywise multiplication of $\mathbf{A}$ and $\mathbf{B}$, $\frac{\mathbf{A}}{\mathbf{B}}$ to mean entrywise division, and $\mathbf{A}^{\cdot\gamma}$ to mean entrywise raising to the γth power. In addition, $\text{repmat}(\boldsymbol{\lambda},1,N)$ denotes the $K\times N$ matrix with each column being the λ vector.

4.5 Choosing the Hyperparameters

4.5.1 Choice of Dispersion Parameter ϕ

The dispersion parameter ϕ represents the tradeoff between the data fidelity and the regularization terms in (20). It needs to be fixed, based on prior knowledge about the noise distribution, or learned from the data using either cross-validation or MAP estimation. In the latter case, ϕ is assigned a prior $p(\phi)$ and the objective $C(\mathbf{W},\mathbf{H},\lambda,\phi)$ can be optimized over ϕ. This is a standard feature in penalized likelihood approaches and has been widely discussed in the literature. In this work, we will not address the estimation of ϕ, but only study the influence of the regularization term on the factorization *given* ϕ. In many cases, it is reasonable to fix ϕ based on prior knowledge. In particular, under the Gaussian noise assumption, $v_{fn} \sim \mathcal{N}(v_{fn}|\hat{v}_{fn},\sigma^2)$, and $\beta = 2$ and $\phi = \sigma^2$. Under the Poisson noise assumption, $v_{fn} \sim \mathcal{P}(v_{fn}|\hat{v}_{fn})$, and $\beta = 1$ and $\phi = 1$. Under multiplicative Gamma noise assumption, $v_{fn} = \hat{v}_{fn}\cdot\epsilon_{fn}$ and ϵ_{fn} is a Gamma noise of mean 1, or equivalently, $v_{fn} \sim \mathcal{G}(v_{kn}|\alpha, \hat{v}_{fn}/\alpha)$, and $\beta = 0$ and $\phi = 1/\alpha$. In audio applications where the power spectrogram is to be factorized, as in Section 6.3, the multiplicative exponential noise model (with $\alpha = 1$) is a generally agreed upon assumption [3] and thus $\phi = 1$.

4.5.2 Choice of Hyperparameters a and b

We now discuss how to choose the hyperparameters a and b in (14) in a data-dependent and principled way. Our method is related to the *method of moments*. We first focus on the selection of b using the sample mean of data, given a. Then the selection of a based on the sample variance of the data is discussed at the end of the section.

Consider the approximation in (1), which can be written element-wise as

$$v_{fn} \approx \hat{v}_{fn} = \sum_k w_{fk}h_{kn}. \tag{35}$$

The statistical models corresponding to shape parameter $\beta \notin (1,2)$ imply that $\mathbb{E}[v_{fn}|\hat{v}_{fn}] = \hat{v}_{fn}$. We extrapolate this property to derive a rule for selecting the hyperparameter b for all $\beta \in \mathbb{R}$ (and for nonnegative real-valued data in general), even though there is no known statistical model governing the noise when $\beta \in (1,2)$. When FN is large, the law of large numbers implies that the sample mean of the elements in $\mathbf{V}$ is close to the population mean (with high probability), i.e.,

$$\hat{\mu}_{\mathbf{V}} \triangleq \frac{1}{FN}\sum_{fn} v_{fn} \approx \mathbb{E}[v_{fn}] = \mathbb{E}[\hat{v}_{fn}] = \sum_k \mathbb{E}[w_{fk}h_{kn}]. \quad (36)$$

We can compute $\mathbb{E}[\hat{v}_{fn}]$ for the Half-Normal and Exponential models using the moments of these distributions and those of the inverse-Gamma for λ_k. These yield

$$\mathbb{E}[\hat{v}_{fn}] = \begin{cases} \dfrac{2Kb}{\pi(a-1)} & \text{Half-Normal}, \\ \dfrac{Kb^2}{(a-1)(a-2)} & \text{Exponential}. \end{cases} \quad (37)$$

By equating these expressions to the empirical mean $\hat{\mu}_{\mathbf{V}}$, we conclude that we can choose b according to

$$\hat{b} = \begin{cases} \dfrac{\pi(a-1)\hat{\mu}_{\mathbf{V}}}{2K} & \ell_2\text{-ARD}, \\ \sqrt{\dfrac{(a-1)(a-2)\hat{\mu}_{\mathbf{V}}}{K}} & \ell_1\text{-ARD}. \end{cases} \quad (38)$$

In summary, $\hat{b} \propto \hat{\mu}_{\mathbf{V}}/K$ and $\hat{b} \propto (\hat{\mu}_{\mathbf{V}}/K)^{1/2}$ for ℓ_2- and ℓ_1-ARD, respectively.

By using the empirical variance of $\mathbf{V}$ and the relation between the mean and variance of the Tweedie distribution in (15), we may also estimate a from the data. The resulting relations are more involved and these calculations are deferred to the online supplementary material [33] for $\beta \in \{0, 1, 2\}$. However, experiments showed that the resulting learning rules for a did not consistently give satisfactory results, especially when FN is not sufficiently large. In particular, the estimates sometimes fall out of the parameter space, which is a known feature of the method of moments. Observe that a appears in the objective function (21) only through $c = (F+N)/2 + a + 1$ (ℓ_2-ARD) or $c = F + N + a + 1$ (ℓ_1-ARD). As such, the influence of a is moderated by $F+N$. Hence, if we want to choose a prior on a that is not too informative, then we should choose a to be small compared to $F+N$. Experiments in Section 6 confirm that smaller values of a (relative to $F+N$) typically produce better results. As discussed in the conclusion, a more robust estimation of a (as well as b and ϕ) would involve a fully Bayesian treatment of our problem, which is left for future work.

5 Connections with Other Works

Our work draws parallels with a few other works on model order selection in NMF. The closest work is [18], which also proposes automatic component pruning via a MAP approach. It was developed during the same period as and independently of our earlier work [22]. An extension to multi-array analysis is also proposed in [19]. In [18], Mørup and Hansen consider NMF with the euclidean and KL costs. They constrained the columns of $\mathbf{W}$ to have unit norm (i.e., $\|\mathbf{w}_k\|_2 = 1$) and assumed that the coefficients of $\mathbf{H}$ are assigned exponential priors $\mathcal{E}(h_{kn}|\lambda_k)$. A noninformative Jeffrey's prior is further assumed on λ_k. Put together, they consider the following optimization over $(\mathbf{W}, \mathbf{H})$:

$$\begin{aligned} \underset{\mathbf{W},\mathbf{H},\boldsymbol{\lambda}}{\text{minimize}} \quad & D(\mathbf{V}|\mathbf{WH}) + \sum_k \frac{1}{\lambda_k}\|\underline{h}_k\|_1 + N\log\lambda_k \\ \text{subject to} \quad & \mathbf{W} \geq 0, \ \ \mathbf{H} \geq 0, \ \ \|\mathbf{w}_k\|_2 = 1, \ \ \forall\, k, \end{aligned} \quad (39)$$

where $D(\cdot|\cdot)$ is either the squared euclidean distance or the KL-divergence. A major difference compared to our objective function in (20) is that this method involves optimizing $\mathbf{W}$ under the constraint $\|\mathbf{w}_k\|_2 = 1$, which is nontrivial. As such, to solve (39), Mørup and Hansen [18] used a change of variables $\mathbf{w}'_k \leftarrow \mathbf{w}_k/\|\mathbf{w}_k\|_2$ and derived a heuristic multiplicative algorithm based on the ratio of negative and positive parts of the new objective function, along the lines of [36]. In contrast, our approach treats $\mathbf{w}_k$ and $\underline{h}_k$ symmetrically and the updates are simple. Furthermore, the pruning approach in [18] only occurs in the rows $\mathbf{H}$ and the corresponding columns of $\mathbf{W}$ may take any nonnegative value (subject to the norm constraint), which makes the estimation of these columns of $\mathbf{W}$ ill-posed (i.e., the parameterization is such that a part of the model is not observable). In contrast, in our approach $\mathbf{w}_k$ and $\underline{h}_k$ are tied together so they converge to zero jointly when λ_k reaches its lower bound.

Our work is also related to the automatic rank determination method in Projective NMF proposed by Yang et al. [20]. Following the principle of PCA, Projective NMF seeks a nonnegative matrix $\mathbf{W}$ such that the projection of $\mathbf{V}$ on the subspace spanned by $\mathbf{W}$ best fits $\mathbf{V}$. In other words, it is assumed that $\mathbf{H} = \mathbf{W}^T\mathbf{V}$. Following ARD in Bayesian PCA as originally described by Bishop [13], Yang et al. consider the additive Gaussian noise model and propose placing half-normal priors with relevance parameters λ_k on the columns of $\mathbf{W}$. They describe how to adapt EM to achieve MAP estimation of $\mathbf{W}$ and its relevance parameters.

Estimation of the model order in the Itakura-Saito NMF (multiplicative exponential noise) was addressed by Hoffman et al. [21]. They employ a nonparametric Bayesian setting in which K is assigned a large value (in principle, infinite), but the model is such that only a finite subset of components is retained. In their model, the coefficients of $\mathbf{W}$ and $\mathbf{H}$ have Gamma priors with fixed hyperparameters and a weight parameter θ_k is placed before each component in the factor model, i.e., $\hat{v}_{fn} = \sum_k \theta_k w_{fk} h_{kn}$. The weight, akin to the relevance parameter in our setting, is assigned a Gamma prior with a sparsity-enforcing shape parameter. A difference with our model is the a priori independence of the factors and the weights. Variational inference is used in [21].

In contrast with the above-mentioned works, the work herein presents a unified framework for model selection in β-NMF. The proposed algorithms have low complexity per iteration and are simple to implement while decreasing the objective function at every iteration. We compare the performance of our algorithms to those in [18] and [21] in Sections 6.3 (music decomposition) and 6.4 (stock price prediction).

6 Experiments

In this section, we present extensive numerical experiments demonstrating the robustness and efficiency of the proposed algorithms for 1) uncovering the correct model order and 2) learning better decompositions for modeling nonnegative data.

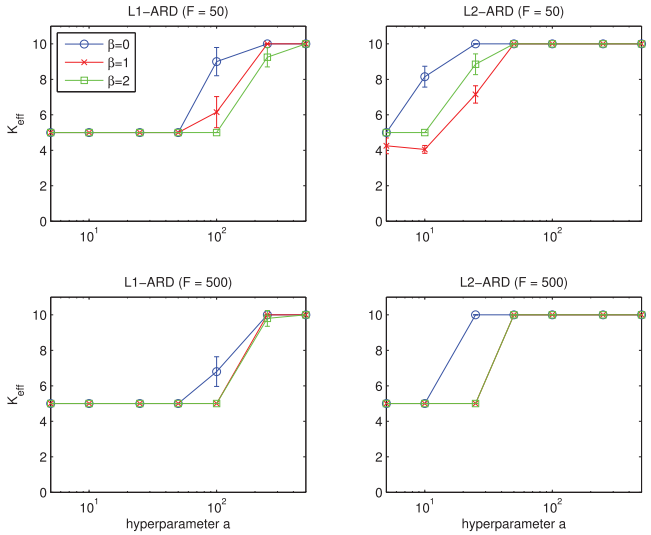


Fig. 1. Estimated number of components as a function of the hyperparameter a (log-linear plot). The true model order is $K_{\text{true}} = 5$. The solid line is the mean across 10 runs and the error bars display $\pm$ the standard deviation.

6.1 Simulations with Synthetic Data

In this section, we describe experiments on synthetic data generated according to the model. In particular, we fixed a pair of hyperparameters $(a_{\text{true}}, b_{\text{true}})$ and sampled $K_{\text{true}} = 5$ relevance weights λ_k according to the inverse-Gamma prior in (14). Conditioned on these relevance weights, we sampled the elements of $\mathbf{W}$ and $\mathbf{H}$ from the Half-Normal or Exponential models depending on whether we chose to use ℓ_2- or ℓ_1-ARD. These models are defined in (12) and (13), respectively. We set $a_{\text{true}} = 50$ and $b_{\text{true}} = 70$ for reasons that will be made clear in the following. We defined the noiseless matrix $\hat{\mathbf{V}}$ as $\mathbf{WH}$. We then generated a noisy matrix $\mathbf{V}$ given $\hat{\mathbf{V}}$ according to the three statistical models $\beta = 0, 1, 2$ corresponding to IS-, KL- and EUC-NMF, respectively. More precisely, the parameters of the noise models are chosen so that the signal-to-noise ratio SNR in dB, defined as $\text{SNR} = 20\log_{10}(\|\hat{\mathbf{V}}\|_F / \|\mathbf{V} - \hat{\mathbf{V}}\|_F)$, is approximately 10 dB for each $\beta \in \{0, 1, 2\}$. For $\beta = 0$, this corresponds to an α, the shape parameter, of approximately 10. For $\beta = 1$, the parameterless Poisson noise model results in an *integer-valued* noisy matrix $\mathbf{V}$. Since there is no noise parameter to select Poisson noise model, we chose b_{true} so that the elements of the data matrix $\mathbf{V}$ are large enough, resulting in an SNR ≈ 10 dB. For the Gaussian observation model ($\beta = 2$), we can analytically solve for the noise variance σ^2 so that the SNR is approximately 10 dB. In addition, we set the number of columns $N = 100$, the initial number of components $K = 2\,K_{\text{true}} = 10$, and chose two different values for F, namely, 50 and 500. The threshold value τ is set to 10^{-7} (refer to Section 4.4). It was observed using this value of the threshold that the iterates of λ_k converged to their limiting values. We ran ℓ_1- and ℓ_2-ARD for $a \in \{5,\ 10,\ 25,\ 50,\ 100,\ 250,\ 500\}$ and using b computed as in Section 4.5.2. The dispersion parameter ϕ is assumed known and set as in the discussion after (18).

To make fair comparisons, the data and the initializations are the same for ℓ_2- and ℓ_1-ARD as well as for every (β, a). We averaged the inferred model order K_{eff} over 10 different runs. The results are displayed in Fig. 1.

First, we observe that ℓ_1-ARD recovers the model order $K_{\text{true}} = 5$ correctly when $a \leq 100$ and $\beta \in \{0, 1, 2\}$. This range includes $a_{\text{true}} = 50$, which is the true hyperparameter we generated the data from. Thus, if we use the correct range of values of a and if the SNR is of the order 10 dB (which is reasonable in most applications), we are able to recover the true model order from the data. On the other hand, from the top right and bottom right plots, we see that ℓ_2-ARD is not as robust in recovering the right latent dimensionality.

Second, note that the quality of estimation is relatively consistent across various βs. The success of the proposed algorithms hinges more on the amount of noise added (i.e., the SNR) compared to which specific β is assumed. However, as discussed in Section 3.2, the shape parameter β should be chosen to reflect our belief in the underlying generative model and the noise statistics.

Third, observe that when more data are available ($F = 500$), the estimation quality improves significantly. This is evidenced by the fact that even ℓ_2-ARD (bottom right plot) performs much better—it selects the right model order for all $a \leq 25$ and $\beta \in \{1, 2\}$. The estimates are also much more consistent across various initializations. Indeed the standard deviations for most sets of experiments is zero, demonstrating that there is little or no variability due to random initializations.

Fig. 2. Sample images of the noisy `swimmer` data. The colormap is adjusted such that black corresponds to the smallest data coefficient value ($v_{fn} = 0$) and white the largest ($v_{fn} = 24$).

6.2 Simulations with the `swimmer` Dataset

In this section, we report experiments on the `swimmer` dataset introduced in [37]. This is a synthetic dataset of $N = 256$ images each of size $F = 32 \times 32 = 1{,}024$. Each image represents a swimmer composed of an invariant torso and four limbs, where each limb can take one of four positions. We set background pixel values to 1 and body pixel values to 10, and generated noisy data with Poisson noise. Sample images of the resulting noisy data are shown in Fig. 2. The "ground truth" number of components for this dataset is $K_{\text{true}} = 16$, which corresponds to all the different limb positions. The torso and background form an invariant component that can be associated with any of the four limbs, or equally split among limbs. The data images are vectorized and arranged in the columns of $\mathbf{V}$.

We applied ℓ_1- and ℓ_2-ARD with $\beta = 1$ (KL-divergence, matching the Poisson noise assumption, and thus $\phi = 1$), $K = 32 = 2\,K_{\text{true}}$ and $\tau = 10^{-6}$. We tried several values for the hyperparameter a, namely, $a \in \{5,\ 10,\ 25,\ 50,\ 75,\ 100,\ 250,\ 500,\ 750,\ 1{,}000\}$, and set b according to (38). For every value of a we ran the algorithms from 10 common positive random initializations. The regularization paths returned by the two algorithms are displayed in Fig. 3. ℓ_1-ARD consistently estimates the correct number of components ($K_{\text{true}} = 16$) up to $a = 500$. Fig. 4 displays the learned basis, objective function, and relevance parameters along iterations

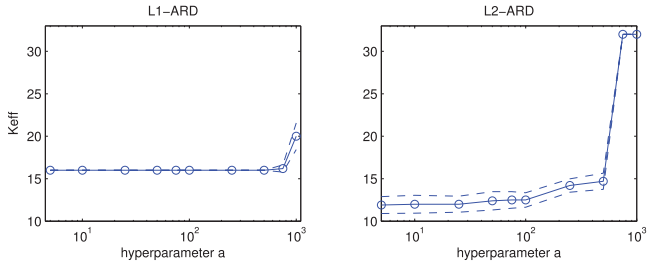


Fig. 3. Estimated number of components K_{eff} as a function of a for ℓ_1- and ℓ_2-ARD. The plain line is the average value of K_{eff} over the 10 runs and dashed lines display $\pm$ the standard deviation.

in one run of ℓ_1-ARD when $a = 100$. It can be seen that the ground truth is perfectly recovered.

In contrast to ℓ_1-ARD, Fig. 3 shows that the value of K_{eff} returned by ℓ_2-ARD is more variable across runs and values of a. Manual inspection reveals that some runs return the correct decomposition when $a = 500$ (and those are the runs with the lowest end value of the objective function, indicating the presence of apparent local minima), but far less consistently than ℓ_1-ARD. Then it might appear that the decomposition strongly overfits the noise for $a \in \{750, 1{,}000\}$. However, visual inspection of learned dictionaries with these values shows that the solutions still make sense. As such, Fig. 5 displays the dictionary learned by ℓ_2-ARD with $a = 1{,}000$. The figure shows that the hierarchy of the decomposition is preserved, despite the fact that the last

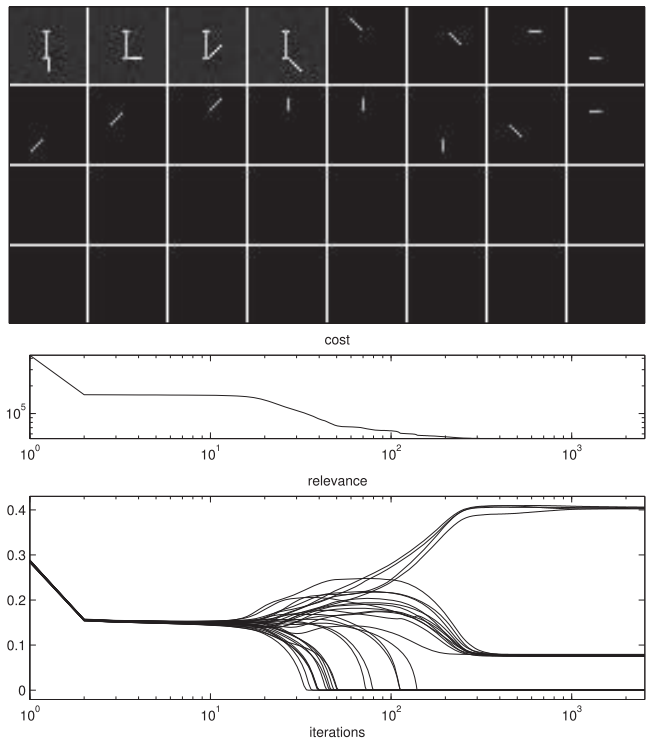


Fig. 4. Top: Dictionary learned in one run of ℓ_1-ARD with $a = 100$. The dictionary elements are presented left to right, top to bottom, by descending order of their relevance λ_k. For improved visualization and fair comparison of the relative importance of the dictionary elements, we display $\mathbf{w}_k$ rescaled by the expectation of h_{kn}, i.e., for ℓ_1-ARD, $\lambda_k \mathbf{w}_k$. The figure colormap is then adjusted to fit the full range of values taken by $\mathbf{W}\,\mathrm{diag}\,\boldsymbol{\lambda}$. Middle: Values of the objective function (21) along iterations (log-log scale). Bottom: Values of $\lambda_k - B$ along iterations (log-linear scale).

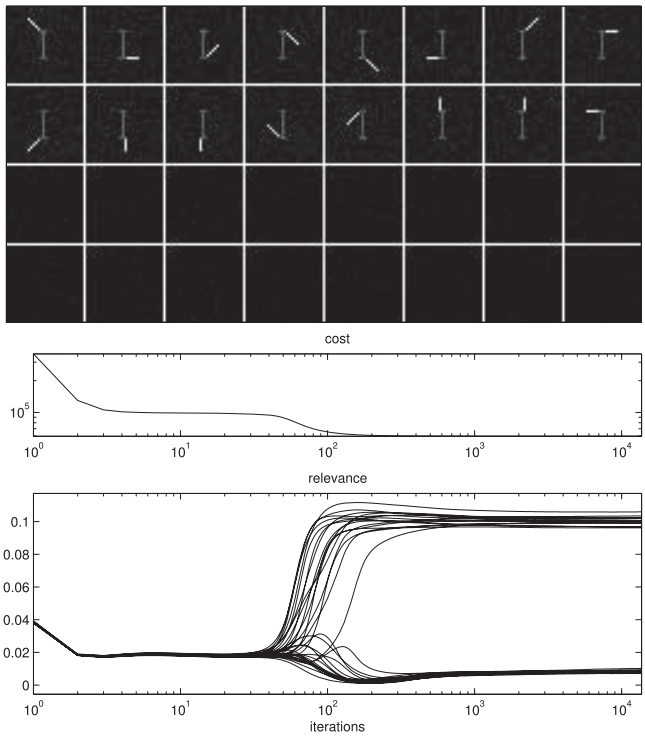


Fig. 5. Top: Dictionary learned by ℓ_2-ARD with $a = 1{,}000$. The dictionary is displayed using the same convention as in Fig. 4, except that the vectors $\mathbf{w}_k$ are now rescaled by the expectation of h_{kn} under the Half-Normal prior, i.e., $(2\lambda_k/\pi)^{1/2}$. Middle: Values of the cost function (21) along iterations (log-log scale). Bottom: Values of $\lambda_k - B$ along iterations (log-linear scale).

16 components capture some residual noise, as a closer inspection would reveal. Thus, despite that fact that pruning is not fully achieved in the 16 extra components, the relevance parameters still give a valid interpretation of what the most significant components are. Fig. 5 shows the evolution of relevance parameters along iterations and it can be seen that the 16 "spurious" components approach the lower bound in the early iterations before they start to fit noise. Note that ℓ_2-ARD returns a solution where the torso is equally shared by the four limbs. This is because the ℓ_2 penalization favors this particular solution over the one returned by ℓ_1-ARD, which favors sparsity of the individual dictionary elements.

With $\tau = 10^{-6}$, the average number of iterations for convergence is approximately $4{,}000 \pm 2{,}000$ for ℓ_1-ARD for all a. The average number of iterations for ℓ_2-ARD is of the same order for $a \leq 500$, and increases to more than 10,000 iterations for $a \geq 750$ because all components are active for these as.

6.3 Music Decomposition

We now consider a music signal decomposition example and illustrate the benefits of ARD in NMF with the IS divergence ($\beta = 0$). Févotte et al. [3] showed that IS-NMF of the power spectrogram underlies a generative statistical model of superimposed Gaussian components, which is relevant to the representation of audio signals. As explained in Sections 3.2 and 4.5, this model is also equivalent to assuming that the power spectrogram is observed in multiplicative exponential noise, i.e., setting $\phi = 1/\alpha = 1$. We investigate the decomposition of the short piano sequence used in [3], a monophonic 15 seconds-long signal x_t recorded in real conditions. The

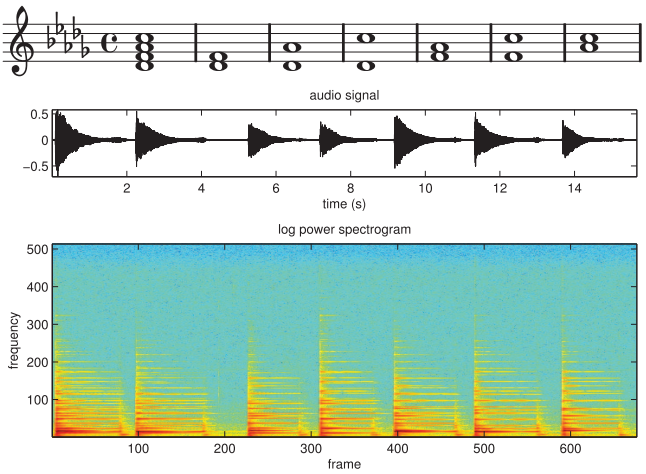


Fig. 6. Three representations of data: Top: original score, middle: time-domain recorded signal, bottom: log-power spectrogram.

sequence is composed of four piano notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. The STFT x_{fn} of the temporal data x_t was computed using a sinebell analysis window of length $L = 1{,}024$ (46 ms) with 50 percent overlap between two adjacent frames, leading to $N = 674$ frames and $F = 513$ frequency bins. The musical score, temporal signal, and log-power spectrogram are shown in Fig. 6. In [3], it was shown that IS-NMF of the power spectrogram $v_{fn} = |x_{fn}|^2$ can correctly separate the spectra of the different notes and other constituents of the signal (sound of hammer on the strings, sound of sustain pedal, etc.).

We set $K = 18$ (three times the "ground truth" number of components) and ran ℓ_2-ARD with $\beta = 0$, $a = 5$, and b computed according to (38). We ran the algorithm from 10 random initializations and selected the solution returned with the lowest final cost. For comparison, we ran standard nonpenalized Itakura-Saito NMF using the multiplicative algorithm described in [3], equivalent to ℓ_2-ARD with $\lambda_k \to \infty$ and $\gamma(\beta) = 1$. We ran IS-NMF 10 times with the same random initializations we used for ARD IS-NMF, and selected the solution with minimum fit. Additionally, we ran the methods by Mørup and Hansen (with KL-divergence) [18] and Hoffman et al. [21]. We used Matlab implementations either publicly available [21] or provided to us by Mørup and Hansen [18]. The best among 10 runs of these methods was selected.

Given an approximate factorization $\mathbf{WH}$ of the data spectrogram $\mathbf{V}$ returned by any of the four algorithms, we proceeded to reconstruct time-domain components by Wiener filtering, following [3]. The STFT estimate $\hat{c}_{k,fn}$ of component k is reconstructed by

$$\hat{c}_{k,fn} = \frac{w_{fk}h_{kn}}{\sum_j w_{fj}h_{jn}} x_{fn}, \tag{40}$$

and the STFT is inverted to produce the temporal component $\hat{c}_{k,t}$.[2] By linearity of the reconstruction and inversion, the decomposition is conservative, i.e., $x_t = \sum_k \hat{c}_{k,t}$.

2. With the approach of Hoffman et al. [21], the columns of $\mathbf{W}$ have to be multiplied by their corresponding weight parameter θ_k prior to reconstruction.

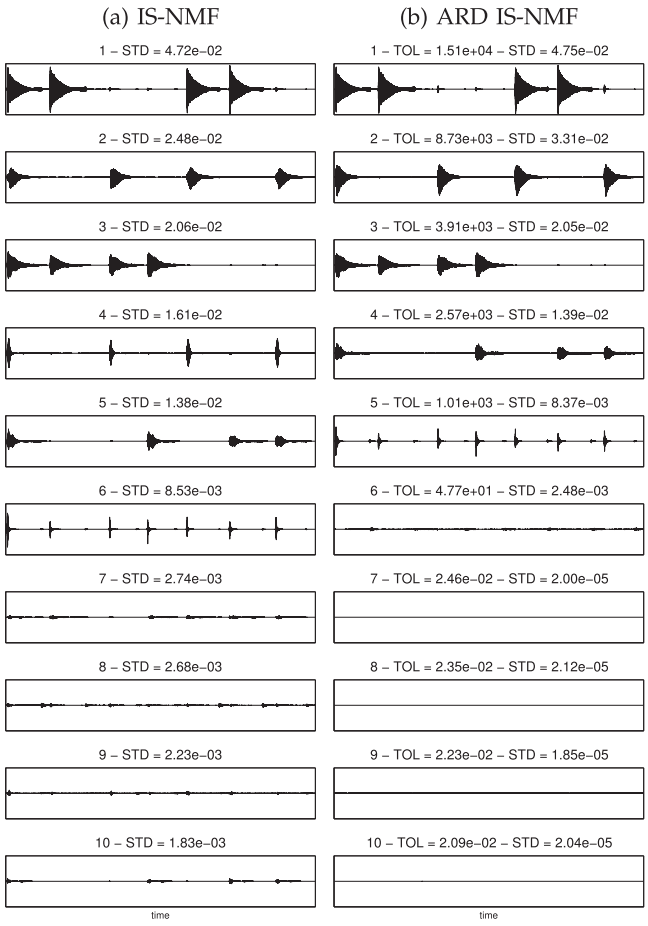


Fig. 7. The first 10 components produced by IS-NMF and ARD IS-NMF. STD denotes the standard deviation of the time samples. TOL is the relevance relative to the bound, i.e., $(\lambda_k - B)/B$. With IS-NMF, the second note of the piece is split into two components ($k = 2$ and $k = 4$).

The components produced by IS-NMF were ordered *by decreasing value of their standard deviations* (computed from the time samples). The components produced by ARD IS-NMF, Mørup and Hansen [18], and Hoffman et al. [21] were ordered *by decreasing value of their relevance weights* ($\{\lambda_k\}$ or $\{\theta_k\}$). Fig. 7 displays the 10 first components produced by IS-NMF and ARD IS-NMF. The y-axes of the two figures are identical so that the component amplitudes are directly comparable. Fig. 8 displays the histograms of the standard deviation values of all 18 components for IS-NMF, ARD IS-NMF, Mørup and Hansen [18], and Hoffman et al. [21].[3]

The histogram in the top right of Fig. 8 indicates that ARD IS-NMF retains six components. This is also confirmed by the value of relative relevance $(\lambda_k - B)/B$ (upon convergence of the relevance weights), displayed with the components in Fig. 7, which drops by a factor of about 2,000 from components 6 to 7. The six components correspond to expected semantic units of the musical sequence: The first four components extract the individual notes and the next two components extract the sound of a hammer hitting the strings and the sound produced by the sustain pedal when it is released. In contrast, IS-NMF has a tendency to overfit; in particular the second note of the piece is split into two

3. The sound files produced by all the approaches are available in the supplementary material, available online. See [33].

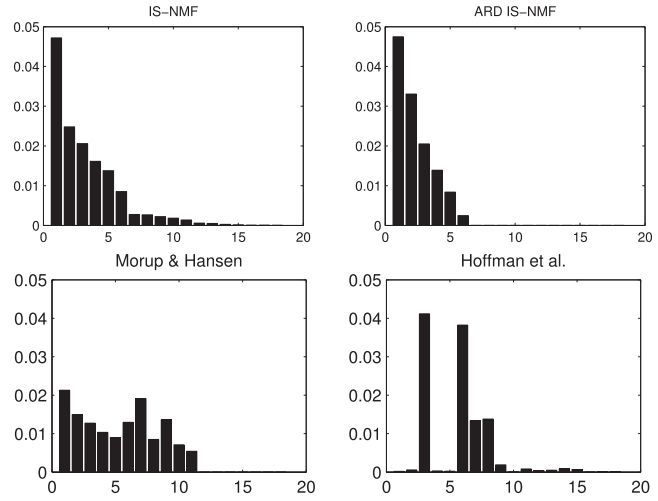


Fig. 8. Histograms of standard deviation values of all 18 components produced by IS-NMF, ARD IS-NMF, Mørup and Hansen [18], and Hoffman et al. [21]. ARD IS-NMF only retains 6 components, which correspond to the expected decomposition, displayed in Fig. 7. On this dataset, the methods proposed in [18] and [21] fail to produce the desired decomposition.

components ($k = 2$ and $k = 4$). The histogram in the bottom left of Fig. 8 shows that the approach of Mørup and Hansen [18] (with the KL-divergence) retains 11 components. Visual inspection of the reconstructed components reveals inaccuracies in the decomposition and significant overfit (some notes are split in subcomponents). The poorness of the results is in part explained by the inadequacy of the KL-divergence (or euclidean distance) for factorization of spectrograms, as discussed in [3]. In contrast, our approach offers flexibility for ARD NMF where the fit-to-data term can be chosen according to the application by setting β to the desired value.

The histogram in the bottom right of Fig. 8 shows that the method by Hoffman et al. [21] retains approximately five components. The decomposition resembles the expected decomposition more closely than [18], except that the hammer attacks are merged with one of the notes. However, it is interesting to note that the distribution of standard deviations does not follow the order of relevance values. This is because the weight parameter θ_k is independent of $\mathbf{W}$ and $\mathbf{H}$ in the prior. As such, the factors are allowed to take very small values while the weight values are not necessarily small.

Finally, we remark that on this data ℓ_1-ARD IS-NMF performed similarly to ℓ_2-ARD IS-NMF and in both cases the retrieved decompositions were fairly robust to the choice of a. We experimented with the same values of a as in previous section and the decompositions and their hierarchies were always found correct. We point out that, as with IS-NMF, initialization is an issue, as other runs did not produce the desired decomposition into notes. However, in our experience the best out of 10 runs always outputs the correct decomposition.

6.4 Prediction of Stock Prices

NMF (with the euclidean and KL costs) has previously been applied on stock data [5] to learn "basis functions" and to cluster companies. In this section, we perform a prediction task on the stock prices of the Dow 30

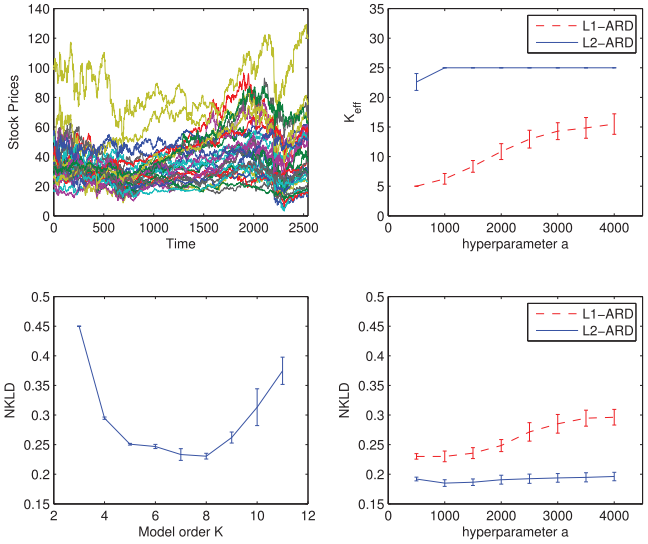


Fig. 9. Top left: The stock data. Top right: Effective model order K_{eff} as a function of a. Bottom: Normalized KL-divergence for KL-NMF (left), ℓ_1- and ℓ_2-ARD KL-NMF (right). Note that the y-axes on both plots are the same. Mørup and Hansen's method [18] yielded an NKLD of 0.37 ± 0.03 (averaged over 10 runs), which is inferior to ℓ_2-ARD, as seen in the bottom right.

companies (comprising the Dow Jones Industrial Average). These are major American companies from various sectors of the economy such as services (e.g., Walmart), consumer goods (e.g., General Motors), and healthcare (e.g., Pfizer). The dataset consists of the stock prices of these $F = 30$ companies from 3 January 2000 to 27 July 2011, a total of $N = 2{,}543$ trading days.[4] The data are displayed in the top left plot of Fig. 9.

In order to test the prediction capabilities of our algorithm, we organized the data into an $F \times N$ matrix $\mathbf{V}$ and removed 50 percent of the entries at random. For the first set of experiments, we performed standard β-NMF with $\beta = 1$, for different values of K, using the observed entries only.[5] We report results for different noninteger values of β in the following. Having performed KL-NMF on the incomplete data, we then estimated the missing entries by multiplying the inferred basis $\mathbf{W}$ and the activation coefficients $\mathbf{H}$ to obtain the estimate $\hat{\mathbf{V}}$. The normalized KL-divergence (NKLD) between the true (missing) stock data and their estimates is then computed as

$$\text{NKLD} \triangleq \frac{1}{|\mathcal{E}|} \sum_{(f,n)\in\mathcal{E}} d_{\text{KL}}(v_{fn}|\hat{v}_{fn}), \tag{41}$$

where $\mathcal{E} \subset \{1,\dots,F\} \times \{1,\dots,N\}$ is the set of missing entries and $d_{\text{KL}}(\cdot\,|\,\cdot)$ is the KL-divergence ($\beta = 1$). The smaller the NKLD, the better the prediction of the missing stock prices and hence the better the decomposition of $\mathbf{V}$ into $\mathbf{W}$ and $\mathbf{H}$. We then did the same for ℓ_1- and ℓ_2-ARD KL-NMF, for different values of a and using $K = 25$. For

4. Stock prices of the Dow 30 companies are provided at the following link: http://www.optiontradingtips.com/resources/historical-data/dow-jones30.html. The raw data consists of four stock prices per company per day. The mean of the four data points is taken to be the representative of the stock price of that company for that day.

5. Accounting for the missing data involves applying a binary mask to $\mathbf{V}$ and $\mathbf{WH}$, where 0 indicates missing entries [38].

KL-NMF, the criterion for termination is chosen so that it mimics that in Section 4.4. Namely, as is commonly done in the NMF literature, we ensured that the columns of $\mathbf{W}$ are normalized to unity. Then, we computed the *NMF relevance weights* $\lambda_k^{\text{NMF}} \triangleq \frac{1}{2}\|\underline{h}_k\|_2^2$. We terminated the algorithm whenever $\text{tol}^{\text{NMF}} \triangleq \max_k |(\lambda_k^{\text{NMF}} - \tilde{\lambda}_k^{\text{NMF}})/\tilde{\lambda}_k^{\text{NMF}}|$ falls below $\tau = 5 \times 10^{-7}$. We averaged the results over 20 random initializations. The NKLDs and the inferred model orders K_{eff} are displayed in Fig. 9.

In the top right plot of Fig. 9, we observe that there is a general increasing trend; as a increases, the inferred model order K_{eff} also increases. In addition, for the same value of a, ℓ_1-ARD prunes more components than ℓ_2-ARD due to its sparsifying effect. This was also observed for synthetic data and the `swimmer` dataset. However, even though ℓ_2-ARD retains almost all the components, the basis and activation coefficients learned model the underlying data better. This is because ℓ_2 penalization methods result in coefficients that are more dense and are known to be better for prediction (rather than sparsification) tasks.

From the bottom left plot of Fig. 9, we observe that when K is too small, the model is not "rich" enough to model the data and hence the NKLD is large. Conversely, when K is too large, the model overfits the data, resulting in a large NKLD. We also observe that ℓ_2-ARD performs spectacularly across a range of values of the hyperparameter a, uniformly better than standard KL-NMF. The NKLD for estimating the missing stock prices hovers around 0.2, whereas KL-NMF results in an NKLD of more than 0.23 for all K. This shows that ℓ_2-ARD produces a decomposition that is more relevant for modeling missing data. Thus, if one does not know the true model order a priori and chooses to use ℓ_2-ARD with some hyperparameter a, the resulting NKLD would be much better than doing KL-NMF even though many components will be retained. In contrast, ℓ_1-ARD does not perform as spectacularly across all values of a but even when a small number of components is retained (at $a = 500$, $K_{\text{eff}} = 5$, NKLD for ℓ_1-ARD ≈ 0.23, NKLD for KL-NMF ≈ 0.25), it performs significantly better than KL-NMF. It is plausible that the stock data fits the assumptions of the Half-Normal model better than the Exponential model and hence ℓ_2-ARD performs better.

For comparison, we also implemented a version of the method by Mørup and Hansen [18] that handles missing data. The mean NKLD value returned over 10 runs is 0.37 ± 0.03, and thus it is clearly inferior to the methods in this paper. The data does not fit the model well.

Finally, in Fig. 10, we demonstrate the effect of varying the shape parameter β and the dispersion parameter ϕ. The distance between the predicted stock prices and the true ones is measured using the NKLD in (41) and the NEUC (the euclidean analogue of the NKLD). We also computed the NIS (the IS analogue of the NKLD), and noted that the results across all three performance metrics are similar so we omit the NIS. We used l_2-ARD, set $a = 1{,}000$, and calculated b using (38). We also chose integer and noninteger values of β to demonstrate the flexibility of l_2-ARD. It is observed that $\beta = 0.5, \phi = 10$ gives the best NKLD and NEUC and that $1 \leq \beta \leq 1.5$ performs well across a wide range of values of ϕ.

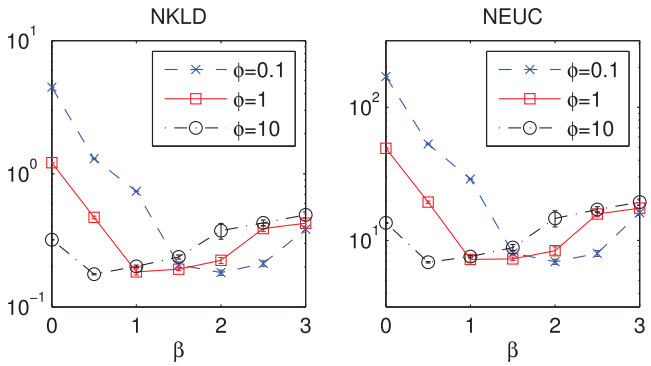


Fig. 10. Effect of varying shape β and dispersion ϕ on prediction performance. Average results over 10 runs.

7 Conclusion

In this paper, we proposed a novel statistical model for β-NMF where the columns of $\mathbf{W}$ and rows $\mathbf{H}$ are tied together through a common scale parameter in their prior, exploiting (and solving) the scale ambiguity between $\mathbf{W}$ and $\mathbf{H}$. MAP estimation reduces to a penalized NMF problem with a group-sparsity inducing regularizing term. A set of MM algorithms accounting for all values of β and either ℓ_1- or ℓ_2-norm group-regularization was presented. They ensure the monotonic decrease of the objective function at each iteration and result in multiplicative update rules of linear complexity in F, K, and N. The updates automatically preserve nonnegativity, given positive initializations, and are easily implemented. The efficiency of our approach was validated on several synthetic and real-world datasets, with competitive performance w.r.t. the state of the art. At the same time, our proposed methods offer improved flexibility over existing approaches (our approach can deal with various types of observation noise and prior structure in a unified framework). Using the method of moments, an effective strategy for the selection of hyperparemeter b given a was proposed and, as a general rule of thumb, we recommend setting a to a small value w.r.t. $F + N$.

There are several avenues for further research: here, we derived a MAP approach that works efficiently, but more sophisticated inference techniques can be envisaged, such as fully Bayesian inference in the model we proposed in Section 3. Following similar treatments in sparse regression [39], [40] or with other forms of matrix factorization [41], one could seek the maximization of $\log p(\mathbf{V}|a, b, \phi)$ using variational or Markov chain Monte-Carlo inference, and in particular handle hyperparameter estimation in a (more) principled way. Other more direct extensions of this work concern the factorization of tensors and online-based methods akin to [42], [43].

Acknowledgments

The authors would like to acknowledge Francis Bach for discussions related to this work, Y. Kenan Yilmaz and A. Taylan Cemgil for discussions on Tweedie distributions, as well as Morten Mørup and Matt Hoffman for sharing their code. They would also like to thank the reviewers whose comments helped to greatly improve the paper. The work of V.Y.F. Tan is supported by A*STAR, Singapore. The

work of C. Févotte is supported by project ANR-09-JCJC-0073-01 TANGERINE (theory and applications of nonnegative matrix factorization).

REFERENCES

[1] P. Paatero and U. Tapper, "Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values," *Environmetrics,* vol. 5, pp. 111-126, 1994.

[2] D.D. Lee and H.S. Seung, "Learning the Parts of Objects with Nonnegative Matrix Factorization," *Nature,* vol. 401, pp. 788-791, 1999.

[3] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation,* vol. 21, pp. 793-830, Mar. 2009.

[4] D. Guillamet, B. Schiele, and J. Vitri, "Analyzing Non-Negative Matrix Factorization for Image Classification," *Proc. Int'l Conf. Pattern Recognition,* 2002.

[5] K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki, "Analysis of Financial Data Using Non-Negative Matrix Factorization," *Int'l J. Math. Sciences,* vol. 6, June 2007.

[6] Y. Gao and G. Church, "Improving Molecular Cancer Class Discovery through Sparse Non-Negative Matrix Factorization," *Bioinformatics,* vol. 21, pp. 3970-3975, 2005.

[7] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms," *Proc. Sixth Int'l Conf. Independent Component Analysis and Blind Signal Separation,* pp. 32-39, Mar. 2006.

[8] 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," *Proc. IEEE Int'l Workshop Machine Learning for Signal Processing,* Sept. 2010.

[9] C. Févotte and J. Idier, "Algorithms for Nonnegative Matrix Factorization with the Beta-Divergence," *Neural Computation,* vol. 23, pp. 2421-2456, Sept. 2011.

[10] A. Cichocki, S. Cruces, and S. Amari, "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization," *Entropy,* vol. 13, pp. 134-170, 2011.

[11] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics,* vol. 6, pp. 461-464, 1978.

[12] D.J.C. Mackay, "Probable Networks and Plausible Predictions—A Review of Practical Bayesian Models for Supervised Neural Networks," *Network: Computation in Neural Systems,* vol. 6, no. 3, pp. 469-505, 1995.

[13] C.M. Bishop, "Bayesian PCA," *Advances in Neural Information Processing Systems,* pp. 382-388, 1999.

[14] A.T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Computational Intelligence and Neuroscience,* vol. 2009, Article ID 785152, p. 17, 2009, doi:10.1155/2009/785152.

[15] M.N. Schmidt, O. Winther, and L.K. Hansen, "Bayesian Non-Negative Matrix Factorization," *Proc. Eighth Int'l Conf. Independent Component Analysis and Signal Separation,* Mar. 2009.

[16] M. Zhong and M. Girolami, "Reversible Jump MCMC for Non-Negative Matrix Factorization," *Proc. Int'l Conf. Artificial Intelligence and Statistics,* p. 8, 2009.

[17] M.N. Schmidt and M. Mørup, "Infinite Non-Negative Matrix Factorizations," *Proc. European Signal Processing Conf.,* 2010.

[18] M. Mørup and L.K. Hansen, "Tuning Pruning in Sparse Non-Negative Matrix Factorization," *Proc. 17th European Signal Processing Conf.,* Aug. 2009.

[19] M. Mørup and L.K. Hansen, "Automatic Relevance Determination for Multiway Models," *J. Chemometrics,* vol. 23, nos. 7/8, pp. 352-363, 2009.

[20] Z. Yang, Z. Zhu, and E. Oja, "Automatic Rank Determination in Projective Nonnegative Matrix Factorization," *Proc. Ninth Int'l Conf. Latent Variable Analysis and Signal Separation,* pp. 514-521, 2010.

[21] M.D. Hoffman, D.M. Blei, and P.R. Cook, "Bayesian Nonparametric Matrix Factorization for Recorded Music," *Proc. Int'l Conf. Machine Learning,* 2010.

[22] V.Y.F. Tan and C. Févotte, "Automatic Relevance Determination in Nonnegative Matrix Factorization," *Proc. Workshop Signal Processing with Adaptative Sparse Structured Representations,* Apr. 2009.

[23] A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones, "Robust and Efficient Estimation by Minimising a Density Power Divergence," *Biometrika,* vol. 85, pp. 549-559, Sept. 1998.

[24] S. Eguchi and Y. Kano, "Robustifying Maximum Likelihood Estimation," technical report, Inst. of Statistical Math., Research Memo. 802, June 2001.

[25] M. Tweedie, "An Index which Distinguishes between Some Important Exponential Families," *Proc. Indian Statistical Inst. of Golden Jubilee Int'l Conf.,* pp. 579-604, 1984.

[26] D.R. Hunter and K. Lange, "A Tutorial on MM Algorithms," *The Am. Statistician,* vol. 58, pp. 30-37, 2004.

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

[28] Y.K. Yilmaz, "Generalized Tensor Factorization," PhD thesis, Boğaziçi Univ., 2012.

[29] B. Jørgensen, "Exponential Dispersion Models," *J. Royal Statistical Soc. Series B (Methodological),* vol. 49, no. 2, p. 127162, 1987.

[30] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *J. Royal Statistical Soc., Series B,* vol. 68, no. 1, pp. 49-67, 2007.

[31] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with Sparsity-Inducing Penalties," *Foundations and Trends in Machine Learning,* vol. 4, no. 1, pp. 1-106, 2012.

[32] E.J. Candès, M.B. Wakin, and S.P. Boyd, "Enhancing Sparsity by Reweighted ℓ_1 Minimization," *J. Fourier Analysis and Applications,* vol. 14, pp. 877-905, Dec. 2008.

[33] V.Y.F. Tan and C. Févotte, "Supplementary Material for 'Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence'," http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.240, 2012.

[34] Z. Yang and E. Oja, "Unified Development of Multiplicative Algorithms for Linear and Quadratic Nonnegative Matrix Factorization," *IEEE Trans. Neural Networks,* vol. 22, no. 12, pp. 1878-1891, Dec. 2011.

[35] Z. Yang and E. Oja, "Linear and Nonlinear Projective Nonnegative Matrix Factorization," *IEEE Trans. Neural Networks,* vol. 21, no. 5, pp. 734-749, May 2010.

[36] J. Eggert and E. Körner, "Sparse Coding and NMF," *Proc. IEEE Int'l Joint Conf. Neural Networks,* pp. 2529-2533, 2004.

[37] D. Donoho and V. Stodden, "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?" *Proc. Advances in Neural Information Processing Systems Conf.,* 2004.

[38] N.-D. Ho, "Nonnegative Matrix Factorization Algorithms and Applications," PhD thesis, Universiteit Katholique de Louvain, 2008.

[39] M.E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Machine Learning Research,* vol. 1, pp. 211-244, 2001.

[40] D.P. Wipf, B.D. Rao, and S. Nagarajan, "Latent Variable Bayesian Models for Promoting Sparsity," *IEEE Trans. Information Theory,* vol. 57, no. 9, pp. 6236-55, Sept. 2011.

[41] R. Salakhutdinov and A. Mnih, "Probabilistic Matrix Factorization," *Proc. Advances in Neural Information Processing Systems Conf.,* vol. 19, 2007.

[42] A. Lefèvre, F. Bach, and C. Févotte, "Online Algorithms for Nonnegative Matrix Factorization with the Itakura-Saito Divergence," *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics,* Oct. 2011.

[43] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding," *J. Machine Learning Research,* vol. 11, pp. 10-60, 2010.

Vincent Y.F. Tan received the BA and MEng degrees in electrical and information sciences tripos (EIST) from the University of Cambridge in 2005 and the PhD degree in electrical engineering and computer science from MIT in 2011, after which he was a postdoctoral researcher at the University of Wisconsin-Madison. He is now a scientist at the Institute for Infocomm Research (I^2R), Singapore, and an adjunct assistant professor in the Department of Electrical and Computer Engineering at the National University of Singapore. During his PhD, he held two summer research internships at Microsoft Research. His research interests include learning and inference in graphical models, statistical signal processing, and network information theory. He received the Charles Lamb prize, a Cambridge University Engineering Department prize awarded to the student who demonstrates the greatest proficiency in the EIST. He also received the MIT EECS Jin-Au Kong outstanding doctoral thesis prize and the A*STAR Philip Yeo prize for outstanding achievements in research. He is a member of the IEEE and of the IEEE "Machine Learning for Signal Processing" technical committee.

Cédric Févotte received the state engineering and PhD degrees in control and computer science from the École Centrale de Nantes, France, in 2000 and 2003, respectively. During his PhD, he was with the Signal Processing Group at the Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN). From 2003 to 2006, he was a research associate with the Signal Processing Laboratory at the University of Cambridge (Engineering Department). He was then a research engineer with the music editing technology start-up company Mist-Technologies (now Audionamix) in Paris. In 2007, he became a CNRS tenured researcher. He was affiliated with LTCI (CNRS & Télécom ParisTech) from 2007 to 2012. Since 2013, he has been with Laboratoire Lagrangre (CNRS, Observatoire de la Côte d'Azur & Université de Nice Sophia Antipolis). His research interests generally concern statistical signal processing and unsupervised machine learning and, in particular, applications to blind source separation and audio signal processing. He is the scientific leader of project TANGERINE (Theory and applications of nonnegative matrix factorization) funded by the French research funding agency ANR and a member of the IEEE and of the IEEE "Machine Learning for Signal Processing" technical committee.

(Févotte, Bertin & Durrieu, *Neural Computation*, 2009)

LETTER Communicated by Andrzej Cichocki

Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis

Cédric Févotte
fevotte@telecom-paristech.fr
Nancy Bertin
nbertin@telecom-paristech.fr
Jean-Louis Durrieu
durrieu@telecom-paristech.fr
CNRS—TELECOM ParisTech, 75014 Paris, France

This letter presents theoretical, algorithmic, and experimental results about nonnegative matrix factorization (NMF) with the Itakura-Saito (IS) divergence. We describe how IS-NMF is underlaid by a well-defined statistical model of superimposed gaussian components and is equivalent to maximum likelihood estimation of variance parameters. This setting can accommodate regularization constraints on the factors through Bayesian priors. In particular, inverse-gamma and gamma Markov chain priors are considered in this work. Estimation can be carried out using a space-alternating generalized expectation-maximization (SAGE) algorithm; this leads to a novel type of NMF algorithm, whose convergence to a stationary point of the IS cost function is guaranteed.

We also discuss the links between the IS divergence and other cost functions used in NMF, in particular, the Euclidean distance and the generalized Kullback-Leibler (KL) divergence. As such, we describe how IS-NMF can also be performed using a gradient multiplicative algorithm (a standard algorithm structure in NMF) whose convergence is observed in practice, though not proven.

Finally, we report a furnished experimental comparative study of Euclidean-NMF, KL-NMF, and IS-NMF algorithms applied to the power spectrogram of a short piano sequence recorded in real conditions, with various initializations and model orders. Then we show how IS-NMF can successfully be employed for denoising and upmix (mono to stereo conversion) of an original piece of early jazz music. These experiments indicate that IS-NMF correctly captures the semantics of audio and is better suited to the representation of music signals than NMF with the usual Euclidean and KL costs.

1 Introduction

Nonnegative matrix factorization (NMF) is a popular dimension-reduction technique, employed for nonsubtractive, part-based representation of nonnegative data. Given a data matrix $\mathbf{V}$ of dimensions $F \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$\mathbf{V} \approx \mathbf{WH}, \tag{1.1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively. K is usually chosen such that $F\,K + K\,N \ll F\,N$, hence reducing the data dimension. Note that the factorization is in general only approximate, so that the terms *approximate nonnegative matrix factorization* and *nonnegative matrix approximation* also appear in the literature. NMF has been used for various problems in diverse fields. To cite a few, we mention the problems of learning parts of faces and semantic features of text (Lee & Seung, 1999), polyphonic music transcription (Smaragdis & Brown, 2003), object characterization by reflectance spectra analysis (Berry, Browne, Langville, Pauca, & Plemmons, 2007), portfolio diversification (Drakakis, Rickard, de Fréin, & Cichocki, 2008), and scotch whiskies clustering (Young, Fogel, & Hawkins, 2006).

In the literature, the factorization, equation 1.1, is usually sought after through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \; D(\mathbf{V} \mid \mathbf{WH}), \tag{1.2}$$

where $D(\mathbf{V} \mid \mathbf{WH})$ is a cost function defined by

$$D(\mathbf{V} \mid \mathbf{WH}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d([\mathbf{V}]_{fn} \mid [\mathbf{WH}]_{fn}), \tag{1.3}$$

and where $d(x \mid y)$ is a scalar cost function. Popular choices are the Euclidean distance, which we here define as

$$d_{EUC}(x \mid y) = \frac{1}{2}(x-y)^2, \tag{1.4}$$

and the (generalized) Kullback-Leibler (KL) divergence, also referred to as I-divergence, defined by

$$d_{KL}(x \mid y) = x\,\log\frac{x}{y} - x + y. \tag{1.5}$$

Both cost functions are positive and take value zero if and only if $x = y$.

Lee and Seung (2001) proposed gradient descent algorithms to solve the minimization problem, equation 1.2, under the latter two cost functions. When a suitable step size is used, the gradient descent update rules are turned into multiplicative rules, under which the cost function is shown to be nonincreasing. The simplicity of the update rules has undoubtedly contributed to the popularity of NMF, and most of the above-mentioned applications are based on Lee and Seung's algorithm for minimization of either the Euclidean distance or the KL divergence.

Nevertheless, some papers have considered NMF under other cost functions and other algorithmic structures. In particular Cichocki and coauthors have devised several types of NMF algorithms for cost functions such as Csiszár divergences (including Amari's α-divergence) and the β-divergence in Cichocki, Zdunek, and Amari (2006), with several other cost functions considered in Cichocki, Amari et al. (2006). Also, Dhillon and Sra (2005) have described multiplicative algorithms for the wide family of Bregman divergences. The choice of the NMF cost function should be driven by the type of data to analyze, and if a good deal of literature is devoted to improving performance of algorithms given a cost function, little literature has been devoted to how to choose a cost function with respect to a particular type of data and application.

In this letter, we are specifically interested in NMF with the Itakura-Saito (IS) divergence, and we demonstrate its relevance to the decomposition of audio spectra. The expression of the IS divergence is given by

$$d_{IS}(x \mid y) = \frac{x}{y} - \log \frac{x}{y} - 1. \tag{1.6}$$

This divergence was obtained by Itakura and Saito (1968) from the maximum likelihood (ML) estimation of short-time speech spectra under autoregressive modeling. It was presented as "a measure of the goodness of fit between two spectra" and became popular in the speech community during the 1970s. It was in particular praised for the good perceptual properties of the reconstructed signals it led to (Gray, Buzo, Gray, & Matsuyama, 1980).

As we shall see, this divergence has other interesting properties. It is in particular scale invariant, meaning that low-energy components of $\mathbf{V}$ bear the same relative importance as high-energy ones. This is relevant to situations in which the coefficients of $\mathbf{V}$ have a large dynamic range, such as in audio short-term spectra. The IS divergence also leads to desirable statistical interpretations of the NMF problem. Indeed, we describe how NMF in this case can be recast as ML estimation of $\mathbf{W}$ and $\mathbf{H}$ in superimposed signals under simple gaussian assumptions. Equivalently, we describe how IS-NMF can be interpreted as ML of $\mathbf{W}$ and $\mathbf{H}$ in multiplicative gamma noise.

The IS divergence belongs to the class of Bregman divergences and is a limit case of the β-divergence. Thus, the gradient descent multiplicative rules given in Dhillon and Sra (2005) and Cichocki, Zdunek et al. (2006),

which coincide in the IS case, can be applied. If convergence of this algorithm is observed in practice, its proof is still an open problem. The statistical framework going along with IS-NMF allows deriving a new type of minimization method, derived from space-alternating expectation-maximization (SAGE), a variant of the standard expectation-maximization (EM) algorithm. This method leads to new update rules, which do not possess a multiplicative structure. The EM setting guarantees convergence of this algorithm to a stationary point of the cost function. Moreover, the statistical framework opens doors to Bayesian approaches for NMF, allowing elaborate priors on $\mathbf{W}$ and $\mathbf{H}$, for which maximum a posteriori (MAP) estimation can again be performed using SAGE. Examples of such priors, yielding regularized estimates of $\mathbf{W}$ and $\mathbf{H}$, are presented in this work.

IS-NMF underlies previous work in the area of automatic music transcription and single-channel audio source separation, but never explicitly so. In particular, our work builds on Benaroya, Gribonval, and Bimbot (2003), Benaroya, Blouet, Févotte, and Cohen (2006), Abdallah and Plumbley (2004), and Plumbley, Abdallah, Blumensath, and Davies (2006), and the connections between IS-NMF and these articles will be discussed.

This letter is organized as follows. Section 2 addresses general properties of IS-NMF. The relation between the IS divergence and other cost functions used in NMF is discussed in section 2.1, section 2.2 addresses scale invariance, and section 2.3 describes the statistical interpretations of IS-NMF. Section 3 presents two IS-NMF algorithms; an existing multiplicative algorithm is described in section 3.1, and section 3.2 introduces a new algorithm derived from SAGE. Section 4 reports an experimental comparative study of Euclidean-NMF, KL-NMF, or IS-NMF algorithms applied to the power spectrogram of a short piano sequence recorded in real conditions, with various initializations and model orders. These experiments show that IS-NMF correctly captures the semantics of the signal and is better suited to the representation of audio than NMF with the usual Euclidean and KL costs. Section 5 presents how IS-NMF can accommodate regularization constraints on $\mathbf{W}$ and $\mathbf{H}$ within a Bayesian framework and how SAGE can be adapted to MAP estimation. In particular, we give update rules for IS-NMF with gamma and inverse-gamma Markov chain priors on the rows of $\mathbf{H}$. In section 6, we present audio restoration results of an original early recording of jazz music; we show how the proposed regularized IS-NMF algorithms can successfully be employed for denoising and upmix (mono to stereo conversion) of the original data. Finally, conclusions and perspectives of this work are given in section 7.

2 Properties of NMF with the Itakura-Saito Divergence

In this section we address the links between the IS divergence and other cost functions used for NMF. Then we discuss its scale invariance property and, finally, describe the statistical interpretations of IS-NMF.

2.1 Relation to Other Divergences Used in NMF.

2.1.1 β-Divergence. As observed by Cichocki, Amari et al. (2006) and Cichocki, Zdunek et al. (2006), the IS divergence is a limit case of the β-divergence introduced by Eguchi and Kano (2001) that we here define as

$$d_\beta(x \mid y) \stackrel{\text{def}}{=} \begin{cases} \dfrac{1}{\beta\,(\beta-1)}\left(x^\beta + (\beta-1)\,y^\beta - \beta\,x\,y^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x\,\log x/y + (y-x) & \beta = 1 \\ \dfrac{x}{y} - \log\dfrac{x}{y} - 1 & \beta = 0. \end{cases} \tag{2.1}$$

Eguchi and Kano (2001) assume $\beta > 1$, but the definition domain can very well be extended to $\beta \in \mathbb{R}$. The β-divergence is shown to be continuous in β by using the identity $\lim_{\beta\to 0}(x^\beta - y^\beta)/\beta = \log(x/y)$. It was considered in NMF by Cichocki, Zdunek et al. (2006) and also coincides up to a factor $1/\beta$ with the generalized divergence of Kompass (2007), which, in the context of NMF as well, was separately constructed so as to interpolate between the KL divergence ($\beta = 1$) and the Euclidean distance ($\beta = 2$). Note that the derivative of $d_\beta(x \mid y)$ with regard to y is also continuous in β and is simply written as

$$\nabla_y\, d_\beta(x \mid y) = y^{\beta-2}\,(y-x). \tag{2.2}$$

The derivative shows that $d_\beta(x|y)$, as a function of y, has a single minimum in $y = x$ and that it increases with $|y-x|$, justifying its relevance as a measure of fit. Figure 1 represents the Euclidean, KL, and IS costs for $x = 1$.

When equation 2.2 is used, the gradients of criterion $D_\beta(\mathbf{V} \mid \mathbf{WH})$ with regard to $\mathbf{W}$ and $\mathbf{H}$ are written as

$$\nabla_{\mathbf{H}}\, D_\beta(\mathbf{V} \mid \mathbf{WH}) = \mathbf{W}^T\left((\mathbf{WH})^{\cdot[\beta-2]} \,.\, (\mathbf{WH} - \mathbf{V})\right) \tag{2.3}$$

$$\nabla_{\mathbf{W}}\, D_\beta(\mathbf{V} \mid \mathbf{WH}) = \left((\mathbf{WH})^{\cdot[\beta-2]} \,.\, (\mathbf{WH} - \mathbf{V})\right)\mathbf{H}^T, \tag{2.4}$$

where . denotes Hadamard entrywise product and $\mathbf{A}^{\cdot[n]}$ denotes the matrix with entries $[\mathbf{A}]_{ij}^n$. The multiplicative gradient descent approach taken in Lee and Seung (2001) and Cichocki, Zdunek et al. (2006) is equivalent to updating each parameter by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion with regard to this parameter, namely, $\theta \leftarrow \theta.[\nabla f(\theta)]_-/[\nabla f(\theta)]_+$, where $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$ and the summands are both nonnegative. This ensures nonnegativity of the parameter updates, provided initialization is with a nonnegative value. A fixed point $\theta^\star$ of the algorithm implies

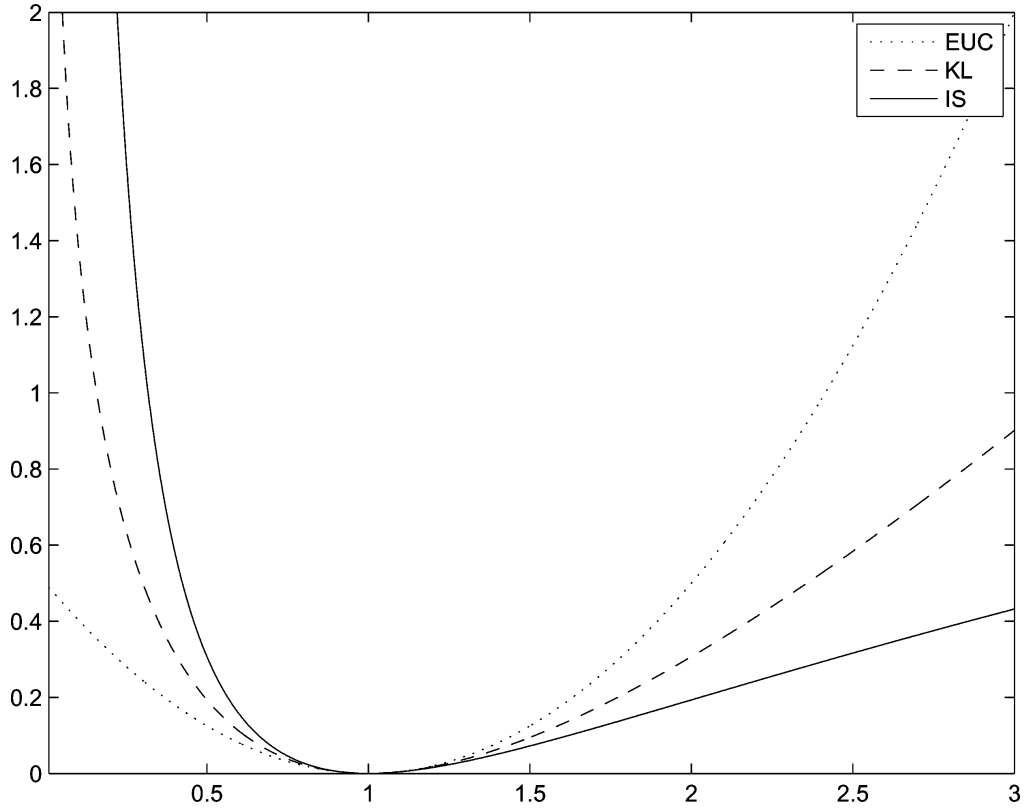


Figure 1: Euclidean, KL, and IS costs $d(x \mid y)$ as a function of y and for $x = 1$. The Euclidean and KL divergences are convex on $(0, \infty)$. The IS divergence is convex on $(0, 2x]$ and concave on $[2x, \infty)$.

either $\nabla f(\theta^\star) = 0$ or $\theta^\star = 0$. This leads to the following updates,

$$\mathbf{H} \leftarrow \mathbf{H}. \frac{\mathbf{W}^T \left((\mathbf{WH})^{\cdot[\beta-2]} . \mathbf{V} \right)}{\mathbf{W}^T \left(\mathbf{WH} \right)^{\cdot[\beta-1]}} \tag{2.5}$$

$$\mathbf{W} \leftarrow \mathbf{W}. \frac{\left((\mathbf{WH})^{\cdot[\beta-2]} . \mathbf{V} \right) \mathbf{H}^T}{\left(\mathbf{WH} \right)^{\cdot[\beta-1]} \mathbf{H}^T}, \tag{2.6}$$

where $\frac{\mathbf{A}}{\mathbf{B}}$ denotes the matrix $\mathbf{A}.\mathbf{B}^{\cdot[-1]}$. Lee and Seung (1999) showed that $D_\beta(\mathbf{V} \mid \mathbf{W}\,\mathbf{H})$ is nonincreasing under the latter updates for $\beta = 2$ (Euclidean distance) and $\beta = 1$ (KL divergence). Kompass (2007) generalizes the proof to the case $1 \leq \beta \leq 2$. In practice, we observe that the criterion is still nonincreasing under updates 2.5 and 2.6 for $\beta < 1$ and $\beta > 2$ (and in particular for $\beta = 0$, corresponding to the IS divergence), but no proof is available. Indeed, the proof Kompass gives makes use of the convexity of $d_\beta(x|y)$ as a function of y, which is true only for $1 \leq \beta \leq 2$. In the rest of the letter, *EUC-NMF* will be used as shorthand for *Euclidean-NMF*.

2.1.2 Bregman Divergences. The IS divergence belongs to the class of Bregman divergences, defined as $d_\phi(x|y) = \phi(x) - \phi(y) - \nabla\phi(y)\,(x - y)$, where ϕ is a strictly convex function of $\mathbb{R}$ that has a continuous derivative $\nabla\phi$. The IS divergence is obtained with $\phi(y) = -\log(y)$. Using the same approach as in the previous paragraph, Dhillon and Sra (2005) derive the following update rules for minimization of $D_\phi(\mathbf{V} \mid \mathbf{WH})$:

$$\mathbf{H} \leftarrow \mathbf{H}.\ \frac{\mathbf{W}^T\ (\nabla^2\phi(\mathbf{WH}).\mathbf{V})}{\mathbf{W}^T\ (\nabla^2\phi(\mathbf{WH}).\mathbf{WH})} \tag{2.7}$$

$$\mathbf{W} \leftarrow \mathbf{W}.\ \frac{(\nabla^2\phi(\mathbf{WH}).\mathbf{V})\,\mathbf{H}^T}{(\nabla^2\phi(\mathbf{WH}).\mathbf{WH})\,\mathbf{H}^T}. \tag{2.8}$$

Again, the authors observed in practice continual descent of $D_\phi(\mathbf{V} \mid \mathbf{WH})$ under these rules, but a proof of convergence is yet to be found. Note that equations 2.5 and 2.6 coincide with equations 2.7 and 2.8 for the IS divergence.

2.2 Scale Invariance. The following property holds for any value of β:

$$d_\beta(\gamma\ x \mid \gamma\ y) = \gamma^\beta\ d_\beta(x \mid y). \tag{2.9}$$

It implies that the IS divergence is scale invariant (i.e., $d_{IS}(\gamma x \mid \gamma y) = d_{IS}(x \mid y)$) and is the only one of the β-divergence family to possess this property. *Scale invariance* means that same relative weight is given to small and large coefficients of $\mathbf{V}$ in cost function (see equation 1.3) in the sense that a bad fit of the factorization for a low-power coefficient $[\mathbf{V}]_{fn}$ will cost as much as a bad fit for a higher-power coefficient $[\mathbf{V}]_{f'n'}$. In contrast, factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest coefficients, and less precision is to be expected in the estimation of the low-power components.

The scale invariance of the IS divergence is relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along frequency f and also usually comprise low-power transient components such as note attacks, together with higher-power components such as tonal parts of sustained notes. The results of the decomposition of a piano spectrogram presented in section 4 confirm these expectations by showing that IS-NMF extracts components corresponding to very low residual noise and hammer hits on the strings with great accuracy. These components are either ignored or severely degraded when using Euclidean or KL divergences.

2.3 Statistical Interpretations. We now turn to statistical interpretations of IS-NMF, which lead to the new EM-based algorithm described in section 3.

2.3.1 Notations. The entries of matrices $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$ are denoted v_{fn}, w_{fk}, and h_{kn}, respectively. Lowercase bold letters in general denote columns, such that $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, while lowercase plain letters with a single index denote rows, such that $\mathbf{H} = [h_1^T, \ldots, h_K^T]^T$. We also define the matrix $\hat{\mathbf{V}} = \mathbf{WH}$, whose entries are denoted $\hat{v}_{fn}$. Where these conventions clash, the intended meaning should be clear from the context.

2.3.2 Sum of Gaussian Components.

Theorem 1 *(IS-NMF as ML estimation in sum of gaussian components). Consider the generative model defined by,* $\forall n = 1, \ldots, N$,

$$\boldsymbol{x}_n = \sum_{k=1}^{K} \boldsymbol{c}_{k,n}, \tag{2.10}$$

where $\boldsymbol{x}_n$ *and* $\boldsymbol{c}_{k,n}$ *belong to* $\mathbb{C}^{F\times 1}$ *and*

$$\boldsymbol{c}_{k,n} \sim \mathcal{N}_c(0, h_{kn}\, diag\,(\boldsymbol{w}_k)), \tag{2.11}$$

where $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *denotes the proper multivariate complex gaussian distribution and where the components* $\boldsymbol{c}_{1,n}, \ldots, \boldsymbol{c}_{K,n}$ *are mutually independent and individually independently distributed. Define* $\boldsymbol{V}$ *as the matrix with entries* $v_{fn} = |x_{fn}|^2$*. Then, maximum likelihood estimation of* $\boldsymbol{W}$ *and* $\boldsymbol{H}$ *from* $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ *is equivalent to NMF of* $\boldsymbol{V}$ *into* $\boldsymbol{V} \approx \boldsymbol{WH}$*, where the Itakura-Saito divergence is used.*

Proof. Under the assumptions of theorem 1 and using the expression of $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given in appendix A, the minus log-likelihood function $C_{ML,1}(\mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{X} \mid \mathbf{W}, \mathbf{H})$ simply factorizes as

$$C_{ML,1}(\mathbf{W}, \mathbf{H}) = -\sum_{n=1}^{N}\sum_{f=1}^{F} \log \mathcal{N}_c\left(x_{fn} \mid 0, \sum_k w_{fk}\, h_{kn}\right) \tag{2.12}$$

$$= NF\ \log\pi + \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(\sum_k w_{fk}\, h_{kn}\right) + \frac{|x_{fn}|^2}{\left(\sum_k w_{fk}\, h_{kn}\right)} \tag{2.13}$$

$$\stackrel{\text{c}}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\left(|x_{fn}|^2 \mid \sum_k w_{fk}\, h_{kn}\right), \tag{2.14}$$

where $\stackrel{\text{c}}{=}$ denotes equality up to constant terms. The minimization of $C_{ML,1}(\mathbf{W}, \mathbf{H})$ with regard to $\mathbf{W}$ and $\mathbf{H}$ thus amounts to the NMF $\mathbf{V} \approx \mathbf{WH}$ with the IS divergence. Note that theorem 1 holds also for real-valued

gaussian components. In that case, $C_{ML,1}(\mathbf{W}, \mathbf{H})$ equals $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ up to a constant and a factor 1/2.

The generative model, equation 2.10, was introduced by Benaroya et al. (2003, 2006) for single-channel audio source separation. In that context, $\mathbf{x}_n = [x_{1n}, \ldots, x_{fn}, \ldots, x_{Fn}]^T$ is the short-time Fourier transform (STFT) of an audio signal x, where $n = 1, \ldots, N$ is a frame index and $f = 1, \ldots, F$ is a frequency index. The signal x is assumed to be the sum of two sources, $x = s_1 + s_2$, and the STFTs of the sources are modeled as $\mathbf{s}_{1,n} = \sum_{k=1}^{K_1} \mathbf{c}_{k,n}$ and $\mathbf{s}_{2,n} = \sum_{k=K_1+1}^{K_1+K_2} \mathbf{c}_{k,n}$, with $K_1 + K_2 = K$. This means that each source STFT is modeled as a sum of elementary components, each characterized by a power spectral density (PSD) $\mathbf{w}_k$ modulated in time by frame-dependent activation coefficients h_{kn}. The PSDs characterizing each source are learned on training data, before the mixture spectrogram $|\mathbf{X}|^{\cdot[2]}$ is decomposed onto the known dictionary $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{K_1}, \mathbf{w}_{K_1+1}, \ldots, \mathbf{w}_{K_1+K_2}]$. However, in these articles, the PSDs and the activation coefficients are estimated separately using somewhat ad hoc strategies (the PSDs are learned with vector quantization) and the equivalence between ML estimation and IS-NMF is not fully exploited.

Complex gaussian modeling of STFT frames of audio signals has been widely used in signal processing and has proven to be a satisfying model for many applications, in particular for audio denoising (see, e.g., Cohen & Gannot, 2007, for a review). But while denoising settings typically assume one observation frame $\mathbf{x}_n$ to be the sum of a source frame and a noise frame, IS-NMF in essence extends this modeling by assuming that one observation frame is the sum of several gaussian frames with different covariances.

The generative model, equation 2.10, may also be viewed as a generalization of well-known models of composite signals. For example, inference in superimposed components with gaussian structure can be tracked back to Feder and Weinstein (1988). In the latter article, however, the components are assumed stationary and solely modeled by their PSD $\mathbf{w}_k$, which in turn is parameterized by a set of parameters of interest $\boldsymbol{\theta}_k$, to be estimated. One extension brought in equation 2.10 is the addition of the amplitude parameters $\mathbf{H}$. This, however, has the inconvenience of making the total number of parameters $F\ K + K\ N$ dependent on N, with the consequence of losing the asymptotical optimality properties of ML estimation. But note that it is precisely the addition of the amplitude parameters in the model that allows $\mathbf{W}$ to be treated as a set of possibly identifiable parameters. Indeed, if h_{kn} is set to 1 for all k and n, the variance of $\mathbf{x}_n$ becomes $\sum_k \mathbf{w}_k$ for all n (i.e., is equal to the sum of the parameters). This would obviously make each PSD $\mathbf{w}_k$ not uniquely identifiable.

Interestingly, the equivalence between IS-NMF and ML inference in the sum of gaussian components provides means of reconstructing the components $\mathbf{c}_{k,n}$ with a sense of statistical optimality, which contrasts with NMF using other costs where methods of reconstructing components from the

factorization **WH** are somewhat ad hoc (see below). Indeed, given **W** and **H**, minimum mean square error (MMSE) estimates can be obtained through Wiener filtering, such that

$$\hat{c}_{k,fn} = \frac{w_{fk}\, h_{kn}}{\sum_{l=1}^{K} w_{fl}\, h_{ln}} x_{fn}. \tag{2.15}$$

Because the Wiener gains sum up to 1 for a fixed entry (f, n), the decomposition is conservative:

$$\mathbf{x}_n = \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}. \tag{2.16}$$

Note that a consequence of Wiener reconstruction is that the phase of all components $\hat{c}_{k,fn}$ is equal to the phase of x_{fn}.

Most works in audio have considered the NMF of magnitude spectra $|\mathbf{X}|$ instead of power spectra $|\mathbf{X}|^{\cdot[2]}$ (see, e.g., Smaragdis & Brown, 2003; Smaragdis, 2007; Virtanen, 2007; Bertin, Badeau, & Richard, 2007). In that case, it can be noted (see, e.g., Virtanen, Cemgil, & Godsill, 2008) that KL-NMF is related to the ML problem of estimating **W** and **H** in the model structure

$$|\mathbf{x}_n| = \sum_{k=1}^{K} |\mathbf{c}_{k,n}| \tag{2.17}$$

under Poissonian assumptions, that is, $|c_{k,fn}| \sim \mathcal{P}(w_{fk}h_{kn})$, where $\mathcal{P}(\lambda)$ is the Poisson distribution, defined in appendix A. Indeed, the sum of Poisson random variables being Poissonian itself (with the shape parameters summing up as well), one obtains $|x_{fn}| \sim \mathcal{P}(\sum_{k=1}^{K} w_{fk}h_{kn})$. Then it can easily be seen that the likelihood $-\log p(\mathbf{X} \mid \mathbf{W}, \mathbf{H})$ is equal up to a constant to $D_{KL}(\,|\mathbf{X}| \,|\, \mathbf{W}\,\mathbf{H})$. Here, **W** is homogeneous to a magnitude spectrum and not to a power spectrum. After factorization, component estimates are typically formed using the phase of the observations (Virtanen, 2007) such that

$$\hat{c}_{k,fn} = w_{fk}\, h_{kn}\, \arg(x_{fn}), \tag{2.18}$$

where $\arg(x)$ denotes the phase of complex scalar x. This approach is worth a few comments. First, the Poisson distribution is formerly defined only for integers, which impairs the statistical interpretation of KL-NMF on uncountable data such as audio spectra (but one could assume an appropriate data scaling and a very fine quantization to work around this).[1]

[1] Actually, KL-NMF has interesting parallels with inference in probabilistic latent variable models of histogram data; see Shashanka, Raj, and Smaragdis (2008a).

Second, this approach enforces nonnegativity in a somehow arbitrary way by taking the absolute value of data $\mathbf{X}$. In contrast, with gaussian modeling, nonnegativity arises naturally through the variance fitting problem equivalence. Similarly, the reconstruction method enforces the components to have same phase as observation coefficients, while this is a consequence of Wiener filtering only in the gaussian modeling framework. Last, the component reconstruction method is not statistically grounded and is not conservative: $\mathbf{x}_n \approx \sum_{k=1}^{K} \hat{\mathbf{c}}_{k,n}$. Note that Wiener reconstruction is used with KL-NMF of the magnitude spectrum $|\mathbf{X}|$ by Smaragdis (2007), where it is presented as spectral filtering, and its conservativity is pointed out.

2.3.3 Multiplicative Noise.

Theorem 2 *(IS-NMF as ML estimation in gamma multiplicative noise). Consider the generative model*

$$\boldsymbol{V} = (\boldsymbol{W}\boldsymbol{H}) \,.\, \boldsymbol{E}, \tag{2.19}$$

where $\boldsymbol{E}$ is multiplicative independent and identically distributed (i.i.d.) gamma noise with mean 1. Then, maximum likelihood estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$ is equivalent to NMF of $\boldsymbol{V}$ into $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$, where the Itakura-Saito divergence is used.

Proof. Let us note $\{e_{fn}\}$, the entries of $\mathbf{E}$. We have $v_{fn} = \hat{v}_{fn} e_{fn}$, with $p(e_{fn}) = \mathcal{G}(e_{fn} \mid \alpha, \beta)$, and where $\mathcal{G}(x \mid \alpha, \beta)$ is the gamma probability density function (PDF) defined in appendix A. Under the iid noise assumption, the minus log likelihood $C_{ML,2}(\mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{V} \mid \mathbf{W}, \mathbf{H})$ is

$$C_{ML,2}(\mathbf{W}, \mathbf{H}) = -\sum_{f,n} \log p(v_{fn} \mid \hat{v}_{fn}) \tag{2.20}$$

$$= -\sum_{f,n} \log \mathcal{G}\left(v_{fn}/\hat{v}_{fn} \mid \alpha, \beta\right)/\hat{v}_{fn} \tag{2.21}$$

$$\stackrel{c}{=} \beta \sum_{f,n} \frac{v_{fn}}{\hat{v}_{fn}} - \frac{\alpha}{\beta} \log \frac{v_{fn}}{\hat{v}_{fn}} - 1. \tag{2.22}$$

The ratio α/β is simply the mean of the gamma distribution. When it is equal to 1, we obtain that $C_{ML,2}(\boldsymbol{\theta})$ is equal to $D_{IS}(\mathbf{V} \mid \hat{\mathbf{V}}) = D_{IS}(\mathbf{V} \mid \mathbf{WH})$ up to a positive factor and a constant.

The multiplicative noise equivalence explains the scale invariance of the IS divergence because the noise acts as a scale factor on $\hat{v}_{fn}$. In contrasts EUC-NMF is equivalent to the ML likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in additive iid gaussian noise. The influence of additive noise is greater on coefficients of $\hat{\mathbf{V}}$ with small amplitude (i.e., low SNR) than on the largest

ones. As to KL-NMF, it corresponds to neither multiplicative nor additive noise but to ML estimation in Poisson noise.[2] To summarize, we have

$$\text{EUC-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \mathcal{N}(v_{fn}|\hat{v}_{fn}, \sigma^2), \tag{2.23}$$

$$\text{KL-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \mathcal{P}(v_{fn}|\hat{v}_{fn}), \tag{2.24}$$

$$\text{IS-NMF: } p(v_{fn} \mid \hat{v}_{fn}) = \frac{1}{\hat{v}_{fn}} \mathcal{G}\left(\frac{v_{fn}}{\hat{v}_{fn}}\middle|\alpha, \alpha\right), \tag{2.25}$$

and in all cases, $\mathrm{E}\{v_{fn} \mid \hat{v}_{fn}\} = \hat{v}_{fn}$.

Theorem 2 reports in essence how Abdallah and Plumbley (2004) derive a "statistically motivated error measure," which happens to be the IS divergence, in the very similar context of nonnegative sparse coding (see also developments in Plumbley et al., 2006). Pointing out the scale invariance of this measure, this work leads Virtanen (2007) to consider the IS divergence (but again without referring to it as such) for NMF in the context of single-channel source separation, but the algorithm is applied to the magnitude spectra instead of the power spectra, losing statistical coherence, and the sources are reconstructed through equation 2.18 instead of Wiener filtering.

3 Algorithms for NMF with the Itakura-Saito Divergence

In this section, we describe two algorithms for IS-NMF. The first one has a multiplicative structure and is only a special case of the derivations of section 2.1. The second one is a novel type, EM based, and is derived from the statistical presentation of IS-NMF given in theorem 1.

3.1 Multiplicative Gradient Descent Algorithm.

A multiplicative gradient descent IS-NMF algorithm is obtained by setting either $\beta = 0$ in equations 2.5 and 2.6 or $\phi(y) = -\log(y)$ in equations 2.7 and 2.8. The resulting update rules coincide and lead to algorithm 1:

Algorithm 1: IS-NMF/MU
Input: nonnegative matrix $\mathbf{V}$
Output: nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \mathbf{WH}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
for $i = 1$: n_{iter} **do**
$\mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\left((\mathbf{WH})^{\cdot[-2]}.\mathbf{V}\right)}{\mathbf{W}^T(\mathbf{WH})^{\cdot[-1]}}$
$\mathbf{W} \leftarrow \mathbf{W}.\frac{\left((\mathbf{WH})^{\cdot[-2]}.\mathbf{V}\right)\mathbf{H}^T}{(\mathbf{WH})^{\cdot[-1]}\mathbf{H}^T}$
Normalize $\mathbf{W}$ and $\mathbf{H}$
end for

[2] KL-NMF is wrongly presented as ML estimation in additive Poisson noise in numerous publications.

These update rules were also obtained by Abdallah and Plumbley (2004), prior to Dhillon and Sra (2005) and Cichocki, Zdunek et al. (2006). In the following, we refer to this algorithm as IS-NMF/MU. This algorithm includes a normalization step at every iteration, which eliminates trivial scale indeterminacies, leaving the cost function unchanged. We impose $\|\mathbf{w}_k\|_2 = 1$ and scale h_k accordingly. Again, we emphasize that continual descent of the cost function is observed in practice with this algorithm but that a proof of convergence is yet to be found.

3.2 SAGE Algorithm. We now describe an EM-based algorithm for estimating the parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$, derived from the statistical formalism introduced in theorem 1. The additive structure of the generative model, equation 2.10, allows updating the parameters describing each component $\mathbf{C}_k \stackrel{\text{def}}{=} [\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}]$ separately, using SAGE (Fessler & Hero, 1994). SAGE is an extension of EM for data models with particular structures, including data generated by superimposed components. It is known to converge faster in iterations than standard EM, though one iteration of SAGE is usually more computationally demanding than EM as it usually requires updating the sufficient statistics "more often." Let us consider a partition of the parameter space $\boldsymbol{\theta} = \bigcup_{k=1}^{K} \boldsymbol{\theta}_k$ with

$$\boldsymbol{\theta}_k = \{\mathbf{w}_k, h_k\}, \tag{3.1}$$

where we recall that $\mathbf{w}_k$ is the kth column of $\mathbf{W}$ and h_k is the kth row of $\mathbf{H}$. The SAGE algorithm involves choosing for each subset of parameters $\boldsymbol{\theta}_k$ a hidden-data space that is complete for this particular subset. Here, the hidden-data space for $\boldsymbol{\theta}_k$ is simply chosen to be $\mathbf{C}_k \stackrel{\text{def}}{=} [\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}]$. An EM-like functional is then built for each subset $\boldsymbol{\theta}_k$ as the conditional expectation of the minus log likelihood of $\mathbf{C}_k$:

$$Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\mathbf{C}_k \mid \boldsymbol{\theta}_k)\, p(\mathbf{C}_k \mid \mathbf{X}, \boldsymbol{\theta}')\, d\mathbf{C}_k. \tag{3.2}$$

One iteration i of the SAGE algorithm then consists of computing (E-step) and minimizing (M-step) $Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')$ for $k = 1, \ldots, K$. Note that $\boldsymbol{\theta}'$ always contains the most up-to-date parameter values, and not only the values at iteration $i - 1$ as in standard EM. This leads to the increase in computational burden, which is mild in our case.

The derivations of the SAGE algorithm for IS-NMF are detailed in appendix B. However, for a fixed k, the E-step merely consists of computing the posterior power $\mathbf{V}_k$ of component $\mathbf{C}_k$, defined by $[\mathbf{V}_k]_{fn} = v_{k,fn} = |\mu_{k,fn}^{post}|^2 + \lambda_{k,fn}^{post}$, where $\mu_{k,fn}^{post}$ and $\lambda_{k,fn}^{post}$ are the posterior mean and variance

of $c_{k,fn}$, given by

$$\mu_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}}\, x_{fn}, \tag{3.3}$$

$$\lambda_{k,fn}^{post} = \frac{w_{fk}\, h_{kn}}{\sum_l w_{fl}\, h_{ln}} \sum_{l\neq k} w_{fl}\, h_{ln}. \tag{3.4}$$

The M-step is then shown to amount to the following one-component NMF problem,

$$\min_{\mathbf{w}_k,\, h_k \geq \mathbf{0}} D_{IS}(\mathbf{V}'_k \mid \mathbf{w}_k\, h_k), \tag{3.5}$$

where $\mathbf{V}'_k$ denotes $\mathbf{V}_k$ as computed from $\boldsymbol{\theta}'$. Interestingly, in the one-component case, the gradients simplify to

$$\nabla_{h_{kn}}\, Q_k^{ML}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') = \frac{F}{h_{kn}} - \frac{1}{h_{kn}^2} \sum_{f=1}^{F} \frac{v'_{k,fn}}{w_{fk}}, \tag{3.6}$$

$$\nabla_{w_{fk}}\, Q_k^{ML}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') = \frac{N}{w_{fk}} - \frac{1}{w_{fk}^2} \sum_{n=1}^{N} \frac{v'_{k,fn}}{h_{kn}}. \tag{3.7}$$

The gradients are easily zeroed, leading to the following updates,

$$h_{kn}^{(i+1)} = \frac{1}{F} \sum_f \frac{v'_{k,fn}}{w_{fk}^{(i)}}, \tag{3.8}$$

$$w_{fk}^{(i+1)} = \frac{1}{N} \sum_n \frac{v'_{k,fn}}{h_{kn}^{(i+1)}}, \tag{3.9}$$

which guarantees $Q_k^{ML}(\mathbf{w}_k^{(i+1)}, h_k^{(i+1)} \mid \boldsymbol{\theta}') \leq Q_k^{ML}(\mathbf{w}_k^{(i)}, h_k^{(i)} \mid \boldsymbol{\theta}')$. This can also be written in matrix form, as shown in algorithm 2, which summarizes the SAGE algorithm for IS-NMF:

Algorithm 2: IS-NMF/EM
Input: nonnegative matrix $\mathbf{V}$
Output: nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{V} \approx \mathbf{WH}$
Initialize $\mathbf{W}$ and $\mathbf{H}$ with nonnegative values
for $i = 1$: n_{iter} **do**
  for $k = 1$: K **do**
    Compute $\mathbf{G}_k = \frac{\mathbf{w}_k h_k}{\mathbf{WH}}$ `% Wiener gain`
    Compute $\mathbf{V}_k = \mathbf{G}_k^{.[2]}.\mathbf{V} + (1-\mathbf{G}_k).(\mathbf{w}_k h_k)$ `% Posterior power of` $\mathbf{C}_k$

$h_k \leftarrow \frac{1}{F}(\mathbf{w}_k^{[-1]})^T \mathbf{V}_k$ % Update row k of $\mathbf{H}$
$\mathbf{w}_k \leftarrow \frac{1}{N}\mathbf{V}_k (h_k^{[-1]})^T$ % Update column k of $\mathbf{W}$
Normalize $\mathbf{w}_k$ and h_k
end for
end for

% Note that $\mathbf{WH}$ needs to be computed only once, at initialization, and be subsequently updated as $\mathbf{WH} - \mathbf{w}_k^{old} h_k^{old} + \mathbf{w}_k^{new} h_k^{new}$.

In the following, we refer to this algorithm as IS-NMF/EM.

IS-NMF/EM and IS-NMF/MU have the same complexity $\mathcal{O}(12\,F\,K\,N)$ per iteration, but can lead to different run times, as shown in the results below. Indeed, in our Matlab implementation, the operations in IS-NMF/MU can be efficiently vectorized using matrix entrywise multiplication, while IS-NMF/EM requires looping over the components, which is more time-consuming.

The convergence of IS-NMF/EM to a stationary point of $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ is granted by property of SAGE. However, it can converge only to a point in the interior domain of the parameter space; $\mathbf{W}$ and $\mathbf{H}$ cannot take entries equal to zero. This is seen in equation 3.5: if w_{fk} or h_{kn} is zero, then the cost $d_{IS}(v'_{k,fn} \mid w_{fk}h_{kn})$ becomes infinite. This is not a feature shared by IS-NMF/MU, which does not a priori exclude zero coefficients in $\mathbf{W}$ and $\mathbf{H}$ (but excludes $\hat{v}_{fn} = 0$, which would lead to a division by zero). However, because zero coefficients are invariant under multiplicative updates (see section 2.1), if IS-NMF/MU attains a fixed-point solution with zero entries, then it cannot be determined if the limit point is a stationary point. Yet if the limit point does not take zero entries (i.e., it belongs to the interior of the parameter space), then it is a stationary point, which may or may not be a local minimum. This is stressed by Berry et al. (2007) for EUC-NMF but holds for IS-NMF/MU as well.

Note that SAGE has been used in the context of single-channel source separation by Ozerov, Philippe, Bimbot, and Gribonval (2007) for inference on a model somehow related to the IS-NMF model, equation 2.10. Indeed, these authors address voice and music separation using a generative model of the form $\mathbf{x}_n = \mathbf{c}_{V,n} + \mathbf{c}_{M,n}$ where the first component represents voice and the second one represents music. Then each component is given a gaussian mixture model (GMM). The GMM parameters for voice are learned from training data, while the music parameters are adapted to data. Though related, the GMM and NMF models are quite different in essence. The first one expresses the signal as a sum of two components that each can take different states. The second one expresses the signal as a sum of K components, each representative of one object. It cannot be claimed that one model is better than the other; rather, they address different characteristics. It is anticipated that the two models can be used jointly within the SAGE framework, for example, by modeling voice $\mathbf{c}_{V,n}$ with a GMM (i.e., a specific

component with many states) and music $\mathbf{c}_{M,n}$ with an NMF model (i.e., a composite signal with many components).

4 Analysis of a Short Piano Excerpt

In this section, we report an experimental comparative study of the NMF algorithms applied to the spectrogram of a short monophonic piano sequence. In the first step, we compare the results of multiplicative Euclidean, KL, and IS NMF algorithms for several values of K before we more specifically compare the multiplicative and EM-based algorithms for IS-NMF in the second step.

4.1 Experimental Setup. A piano sequence played from the score given in Figure 2 on a Yamaha DisKlavier MX100A upright piano was recorded in a small-size room by a Schoeps omnidirectional microphone, placed about 15 cm (6 inches) above the opened body of the piano. The sequence is composed of four notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures. The 15.6-seconds-long recorded signal was downsampled to $\nu_s = 22{,}050$ Hz, yielding $T = 339{,}501$ samples. A STFT $\mathbf{X}$ of x was computed using a sinebell analysis window of length $L = 1024$ (46 ms) with 50% overlap between two frames, leading to $N = 674$ frames and $F = 513$ frequency bins. The time-domain signal x and its log-power spectrogram are represented in Figure 2.

IS-NMF/MU, IS-NMF/EM, and the multiplicative gradient descent NMF algorithms with Euclidean and KL costs were implemented in Matlab and run on data $\mathbf{V} = |\mathbf{X}|^{\cdot[2]}$. Note that in the following, the terms *EUC-NMF* and *KL-NMF* will implicitly refer to the multiplicative implementation of these NMF techniques. All algorithms were run for several values of the number of components, more specifically, for $K = 1, \ldots, 10$. For each value of K, 10 runs of each algorithm were produced from 10 random initializations of $\mathbf{W}$ and $\mathbf{H}$, chosen, in Matlab notation, as `W = abs(randn(F,K)) + ones(F,K)` and `H = abs(randn(K,N)) + ones(K,N)`. The algorithms were run for $n_{iter} = 5000$ iterations.

4.2 Pitch Estimation. In the following results, it will be observed that some of the basis elements (columns of $\mathbf{W}$) have a pitched structure, characteristic of individual musical notes. If pitch estimation is not the objective per se of the following study, it is informative to check if correct pitch values can be inferred from the factorization. As such, a fundamental frequency (or pitch) estimator is applied using the method described in Vincent, Bertin, and Badeau (2007). It consists of computing dot products of $\mathbf{w}_k$ with a set of J frequency combs and retaining the pitch number corresponding to the largest dot product. Each comb is a cosine function with period f_j, scaled and shifted to the amplitude interval [0 1], which takes its maximum value

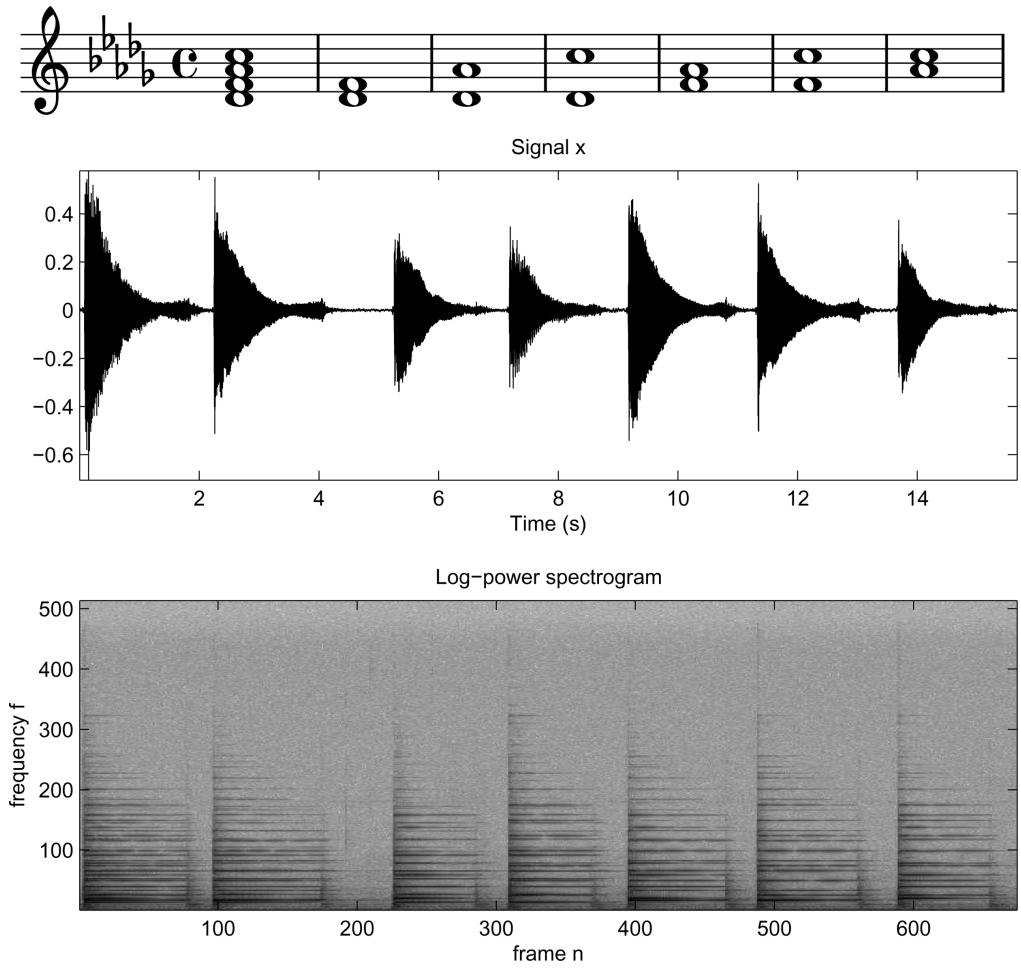


Figure 2: Three representations of data: (Top) Original score. (Middle) Time-domain recorded signal x. (Bottom) Log-power spectrogram $\log |\mathbf{X}|^{\cdot[2]}$. The four notes read $D_4^\flat$ (pitch 61), F_4 (pitch 65), $A_4^\flat$ (pitch 68), and C_5 (pitch 72). Together they form a $D^\flat$ major seventh chord. In the recorded interpretation, the third chord is slightly out of tempo.

1 at bins multiple of f_j. The set of fundamental frequency bins $f_j = \frac{\nu_j}{\nu_s} L$ is indexed on the MIDI logarithmic scale, such that

$$\nu_j = 440 \times 2^{\frac{p_j - 69}{12}}. \tag{4.1}$$

The piano note range usually goes from $p_{\min} = 21$, that is, note A_0 with fundamental frequency $f_{\min} = 27.5$ Hz, to $p_{\max} = 108$, that is, note C_8 with frequency $f_{\max} = 4186$ Hz. Two adjacent keys are separated by a semitone ($\Delta p = 1$). The MIDI pitch numbers of the notes pictured in Figure 2 are 61 ($D_4^\flat$), 65 (F_4), 68 ($A_4^\flat$), and 72 (C_5) and were chosen arbitrarily. In our implementation of the pitch estimator, the MIDI range was sampled from 20.6 to 108.4 with step 0.2. In the following, an arbitrary pitch value of 0 will be given to unpitched basis elements. The classification of pitched and

Table 1: Run Times in Seconds of 1000 Iterations of the NMF Algorithms Applied to the Piano Data.

K	1	2	3	4	5	10	$\mathcal{O}(.)$
EUC-NMF	17	18	20	24	27	37	$4FKN + 2K^2(F+N)$
KL-NMF	90	90	92	100	107	117	$8FKN$
IS-NMF/MU	127	127	129	135	138	149	$12FKN$
IS-NMF/EM	81	110	142	171	204	376	$12FKN$

Notes: This was implemented in Matlab on a 2.16 GHz Intel Core 2 Duo iMac with 2 GB RAM. The run times include the computation of the cost function at each iteration (for possible convergence monitoring). The last column shows the algorithm complexities per iteration, expressed in number of flops (addition, subtraction, multiplication, division). The complexity of EUC-NMF assumes $K < F, N$.

unpitched elements was done manually by looking at the basis elements and listening to the component reconstructions.

4.3 Results and Discussion

4.3.1 Convergence Behavior and Algorithm Complexities. Run times of 1000 iterations of each of the four algorithms are shown in Table 1, together with the algorithm complexities. Figure 3 shows for each algorithm and for every value of K the final cost values of the 10 runs, after the 5000 algorithm iterations. A first observation is that the minimum and maximum cost values differ: $K > 4$ in the Euclidean case, $K > 3$ in the KL case, and $K > 2$ in the IS case. This means either that the algorithms have failed to converge after 5000 iterations in some cases or the presence of local minima. Figure 4 displays for all four algorithms the evolution of the cost functions along the 5000 iterations for all 10 runs in the case $K = 6$.

4.3.2 Evolution of the Factorizations with Order K. In this paragraph, we examine in detail the underlying semantics of the factorizations obtained with all three cost functions. We address only the comparison of factorizations obtained from the three multiplicative algorithms. IS-NMF/EM and IS-NMF/MU will be more specifically compared in the next paragraph. Otherwise stated, the factorizations studied are those obtained from the run yielding the minimum cost value among the 10 runs. Figures 5 to 8 display the columns of $\mathbf{W}$ and corresponding rows of $\mathbf{H}$. The columns of $\mathbf{W}$ are represented against frequency bin f on the left (in $\log_{10}$ amplitude scale), and the rows of $\mathbf{H}$ are represented against frame index n on the right (in linear amplitude scale). Pitched components are displayed first (top to bottom, in ascending order of estimated pitch value), followed by the unpitched components. We reproduce only part of the results in this letter, but the factorizations obtained with all four algorithms for $K = 4, 5, 6$ are available online at http://www.tsi.enst.fr/∼fevotte/Samples/is-nmf, together with

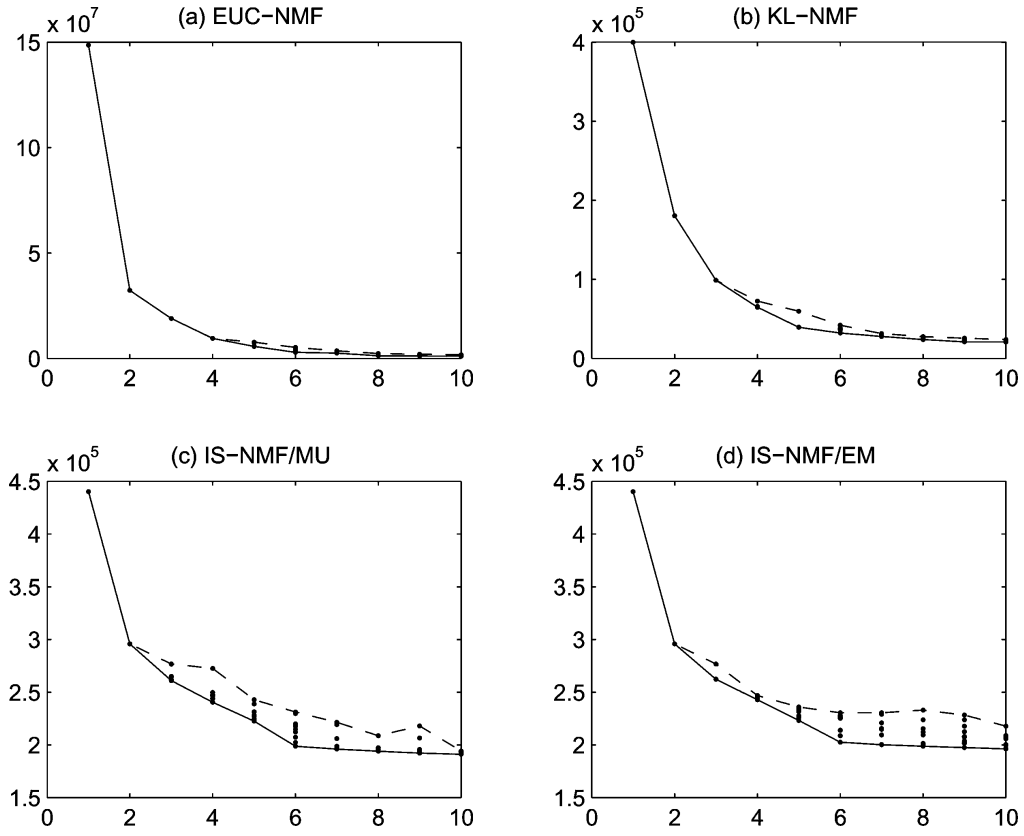


Figure 3: Cost values after 5000 iterations, obtained from 10 random initializations. (a) Euclidean distance. (b) KL divergence. (c) IS divergence (using IS-NMF/MU). (d) IS divergence (using IS-NMF/EM). On each plot, the solid line connects all minimum cost values, and the dashed line connects all maximum cost values.

sound reconstructions of the individual components. Component STFTs $\hat{\mathbf{C}}_k$ were computed by applying the Wiener filter, equation 2.15, to $\mathbf{X}$ using the factors $\mathbf{W}$ and $\mathbf{H}$ obtained with all three cost functions. Time-domain components c_k were then reconstructed by inverting the STFTs using an adequate overlap-add procedure with dual synthesis window. By conservativity of Wiener reconstruction and linearity of the inverse STFT, the time domain decomposition is also conservative, such that

$$x = \sum_{k=1}^{K} c_k. \tag{4.2}$$

Common sense suggests that choosing as many components as notes forms a sensible guess for the value of K so as to obtain a meaningful factorization of $|\mathbf{X}|^{\cdot[2]}$, where each component would be expected to represent one and only one note. The factorizations obtained with all three costs for $K = 4$ prove that this is not the case. Euclidean and KL-NMF rather

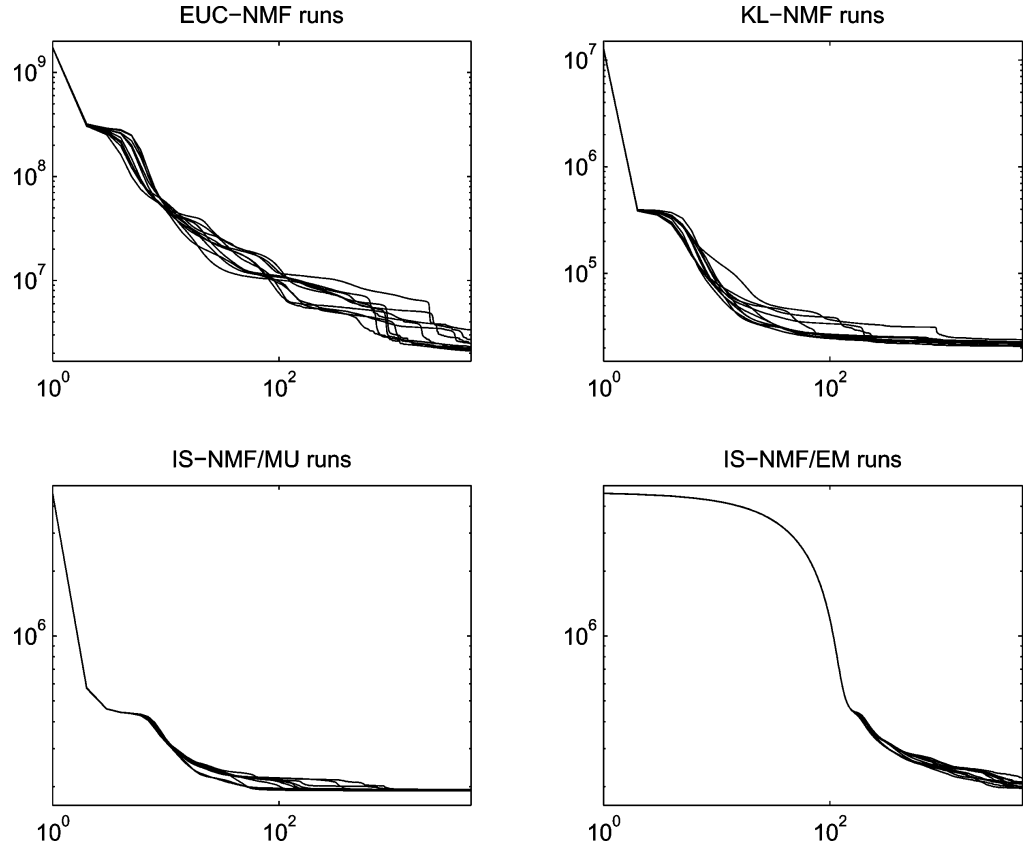


Figure 4: Evolution in log-log scale of the cost functions along the 5000 iterations of all 10 runs of the four algorithms in the specific case of $K = 6$.

successfully extract notes 65 and 68 into separate components (second and third), but notes 61 and 72 are melted into the first component, while a fourth component seems to capture transient events corresponding to the note attacks (the sound of the hammer hitting the string) and the sound produced by the release of the sustain pedal. The first two components obtained with IS-NMF have a similar interpretation to those given by EUC-NMF and KL-NMF. However, the two other components differ in nature: the third component comprises note 68 and transients, while the fourth component is akin to residual noise. It is interesting to notice how this last component, though of much lower energy than the other components (on the order of 1 compared to 10^4 for the others) bears equal importance in the decomposition. This is undoubtedly a consequence of the scale invariance property of the IS divergence discussed in section 2.2.

A fully separated factorization (at least as intended) is obtained for $K = 5$ with KL-NMF, as displayed in Figure 5. This results in four components, each made up of a single note, and a fifth component containing sound events corresponding to note attacks and pedal releases. However, these latter events are not well localized in time and suffer from an unnatural tremolo effect (oscillating variations in amplitudes), as can be heard from

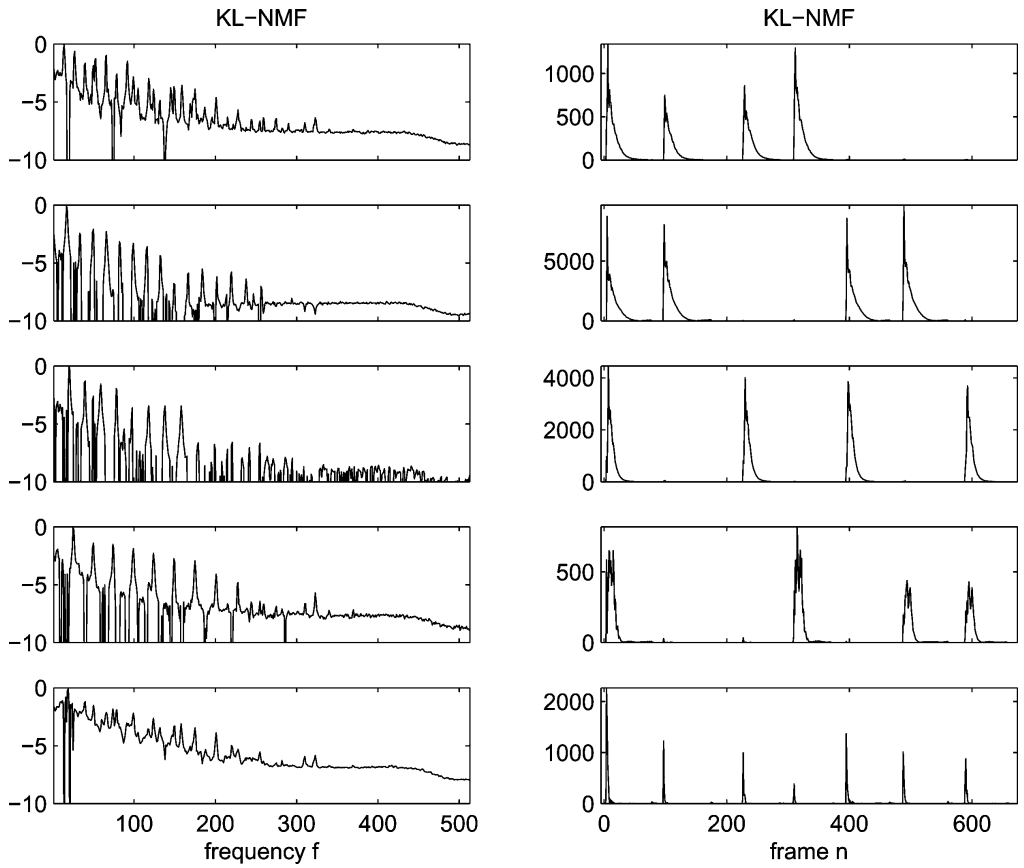


Figure 5: KL-NMF with $K = 5$. Pitch estimates: [61 65 68 72.2 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

the reconstructed sound files. Surprisingly, the decomposition obtained with EUC-NMF by setting $K = 5$ results in splitting the second component of the $K = 4$ decomposition into two components with estimated pitches 65 and 65.4 instead of actually demixing the third component, which comprises notes 61 and 72. As for IS-NMF, the first component now groups notes 61 and 68; the second and third components, respectively, capture notes 65 and 72; the fourth component is still akin to residual noise; and the fifth component perfectly renders the attacks and releases.

Full separation of the individual notes is finally obtained with Euclidean and IS costs for $K = 6$, as shown in Figures 6 and 7. KL-NMF produces an extra component (with pitch estimate 81) that is not clearly interpretable and is in particular not akin to residual noise as could have been hoped for. The decomposition obtained with the IS cost describes as follows. The four first components correspond to individual notes whose pitch estimate matches exactly the pitches of the notes played. The visual aspect of the PSDs is much better than the basis elements learned from EUC-NMF and KL-NMF. The fifth component captures the hammer hits and pedal releases with great accuracy, and the sixth component is akin to residual noise.

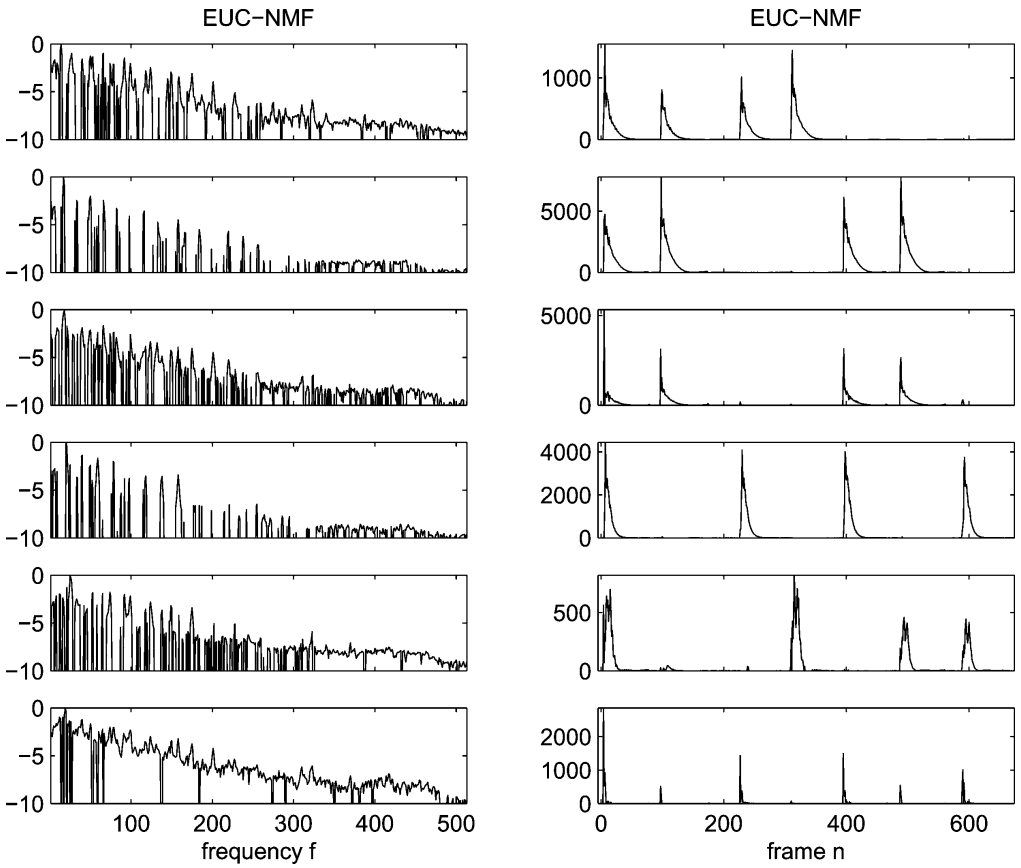


Figure 6: EUC-NMF with $K = 6$. Pitch estimates: [61 65 65.4 68 72 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

When the decomposition is carried beyond $K = 6$, EUC-NMF and KL-NMF split existing components into several subcomponents (such as components capturing sustained and decaying parts of one note) with pitch in the neighborhood of the note fundamental frequency. In contrast, IS-NMF/MU spends the extra components in fine-tuning the representation of the low-energy components—residual noise and transient events (as such, the hammer hits and pedal releases eventually get split in two distinct components). For $K = 10$, the pitch estimates read EUC-NMF: [61 64.8 64.8 65 65 65.8 68 68.4 72.2 0], KL-NMF: [61 61 65 65 66 68 72 80.2 0 0], IS-NMF/MU: [61 61 65 68 72 0 0 0 0 0]. If note 61 is indeed split into two components with IS-NMF/MU, one of the two components is actually inaudible.

The message of this experimental study is that the nature of the decomposition obtained with IS-NMF, and its progression as K increases, is in accord with an object-based representation of music, close to our own comprehension of sound. Entities with well-defined semantics emerge from the decomposition (individual notes, hammer hits, pedal releases, residual noise), while the decompositions obtained from the Euclidean and KL costs are less interpretable from this perspective. These conclusions do not

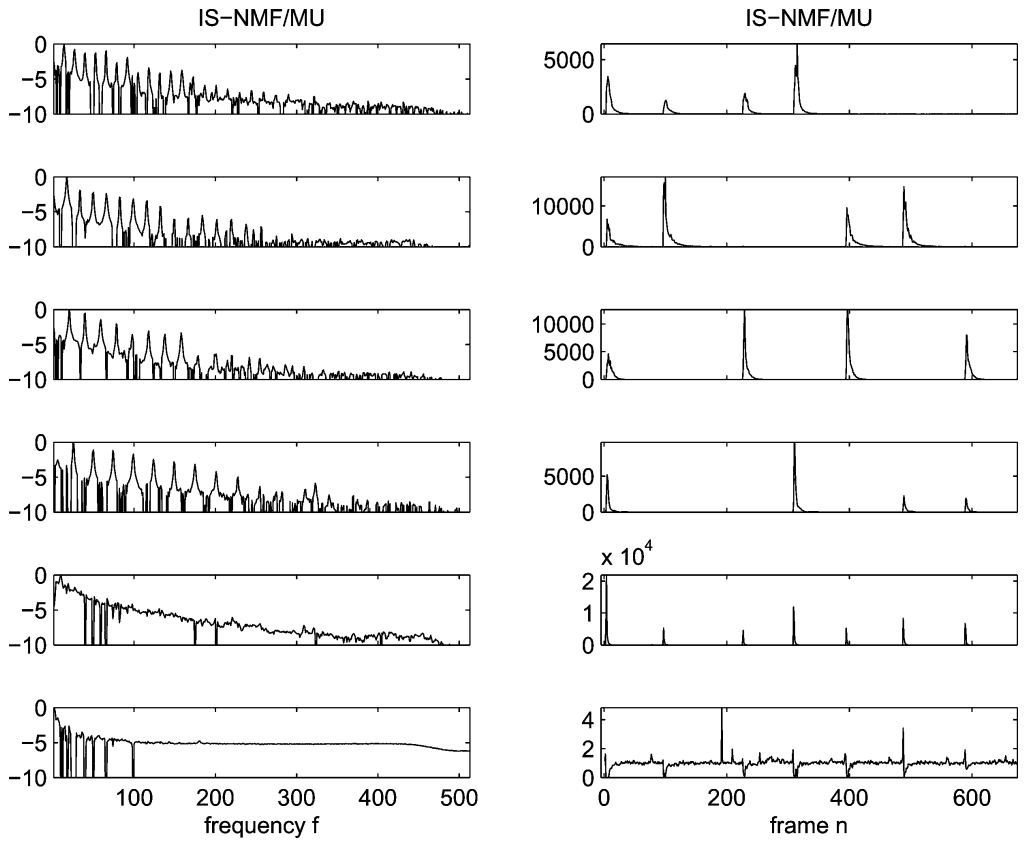


Figure 7: IS-NMF/MU with $K = 6$. Pitch estimates: [61 65 68 72 0 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

always hold when the factorization is not the one yielding the lowest-cost values from the 10 runs. As such, we also examined the factorizations with highest-cost values (with all three cost functions), and we found out that they did not reveal the same semantics, which was not always easily interpretable. The upside, however, is that the lowest IS cost values correspond to the most desirable factorizations, so that IS-NMF "makes sense."

4.3.3 Comparison of Multiplicative and EM-Based IS-NMF. Algorithms IS-NMF/MU and IS-NMF/EM are designed to address the same task of minimizing the cost $D_{IS}(\mathbf{V} \mid \mathbf{WH})$, so that the achieved factorization should be identical in nature, provided they complete this task. As such, the progression of the factorization provided by IS-NMF/EM is similar to the one observed for IS-NMF/MU, described in the previous paragraph. However, the resulting factorizations are not exactly equivalent, because IS-NMF/EM does not inherently allow zeros in the factors (see section 3.2). This feature can be desirable for $\mathbf{W}$, as the presence of sharp notches in the spectrum may not be physically realistic for audio, but can be considered a drawback as far as $\mathbf{H}$ is concerned. Indeed, the rows of $\mathbf{H}$ being akin to activation

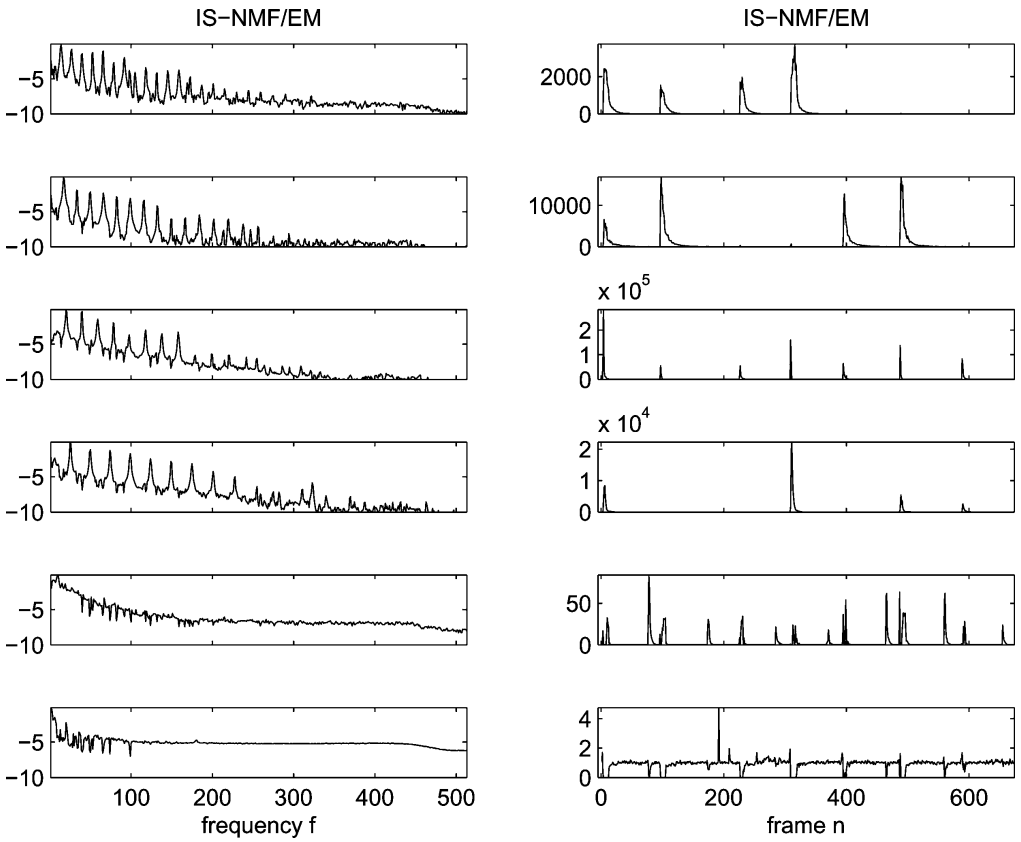


Figure 8: IS-NMF/EM with $K = 6$. Pitch estimates: [61 65 68 72 0 0]. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Rows of $\mathbf{H}$.

coefficients, when a sound object k is not present in frame n, then h_{kn} should be strictly zero. These remarks probably explain the factorization obtained from IS-NMF/EM with $K = 6$, displayed in Figure 8. The notches present in the PSDs learned with IS-NMF/MU, as seen in Figure 7, have disappeared from the PSDs on Figure 8, which exhibit better regularity. Unfortunately, IS-NMF/EM does not fully separate out the note attacks in the fifth component as IS-NMF/MU does. Indeed, some parts of the attacks appear in the second component, and the rest appear in the fifth component, which also contains the pedal releases. This is possibly explained by the a priori high sparsity of a transients component, which can be handled by IS-NMF/MU but not IS-NMF/EM (because it does not allow zero values in $\mathbf{H}$). Increasing the number of components K or the number of algorithm iterations n_{iter} does not solve this specific issue.

Regarding the compared convergence of the algorithms, IS-NMF/MU decreases the cost function much faster in the initial iterations and, with this data set, attains lower final cost values than IS-NMF/EM, as shown in Figure 3 or 4 for $K = 6$. Although the two algorithms have the same complexity, the run time per iteration of IS-NMF/MU is smaller than IS-NMF/EM for $K > 3$ (see Table 1).

5 Regularized IS-NMF

We now describe how the statistical setting of IS-NMF can be exploited to incorporate regularization constraints and prior information in the factors estimates.

5.1 Bayesian Setting. We consider a Bayesian setting where $\mathbf{W}$ and $\mathbf{H}$ are given (independent) prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$. We are looking for a joint MAP estimate of $\mathbf{W}$ and $\mathbf{H}$ through minimization of criterion

$$C_{MAP}(\mathbf{W}, \mathbf{H}) \stackrel{\text{def}}{=} -\log p(\mathbf{W}, \mathbf{H} \mid \mathbf{X}) \tag{5.1}$$

$$\stackrel{\text{c}}{=} D_{IS}(\mathbf{V} \mid \mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}). \tag{5.2}$$

When independent priors of the form $p(\mathbf{W}) = \prod_k p(\mathbf{w}_k)$ and $p(\mathbf{H}) = \prod_k p(h_k)$ are used, then the SAGE algorithm presented in section 3.2 can be used again for MAP estimation. In that case, the functionals to be minimized for each component k are

$$Q_k^{MAP}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}') \stackrel{\text{def}}{=} -\int_{\mathbf{C}_k} \log p(\boldsymbol{\theta}_k \mid \mathbf{C}_k)\, p(\mathbf{C}_k \mid \mathbf{X}, \boldsymbol{\theta}')\, d\mathbf{C}_k \tag{5.3}$$

$$\stackrel{\text{c}}{=} Q_k^{ML}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') - \log p(\mathbf{w}_k) - \log p(h_k). \tag{5.4}$$

Thus, the E-step still amounts to computing $Q_k^{ML}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}')$, as done in section 3.2, and only the M-step is changed by the regularization constraints $-\log p(\mathbf{w}_k)$ and $-\log p(h_k)$, which now need to be taken into account.

Next, we more specifically consider Markov chain priors favoring smoothness over the rows of $\mathbf{H}$. In the following results, no prior structure will be assumed for $\mathbf{W}$ (i.e., $\mathbf{W}$ is estimated through ML). However, we stress that the methodology presented for the rows of $\mathbf{H}$ can be transposed to the columns of $\mathbf{W}$, that prior structures can be imposed on both $\mathbf{W}$ and $\mathbf{H}$, and that these structures need not belong to the same class of models. Note also that since the components are treated separately, each can be given a different type of model (e.g., some components could be assigned a GMM, as discussed at the end of section 3.2).

We assume the following prior structure for h_k,

$$p(h_k) = \prod_{n=2}^{N} p\big(h_{kn} \mid h_{k(n-1)}\big)\; p(h_{k1}), \tag{5.5}$$

where $p(h_{kn} \mid h_{k(n-1)})$ is a PDF with mode $h_{k(n-1)}$. The motivation behind this prior is to constrain h_{kn} not to differ significantly from its value at entry $n-1$, hence favoring smoothness of the estimate. Possible PDF choices are, for $n = 2, \ldots, N$,

$$p\big(h_{kn} \mid h_{k(n-1)}\big) = \mathcal{IG}\big(h_{kn} \mid \alpha, (\alpha+1)\, h_{k(n-1)}\big) \tag{5.6}$$

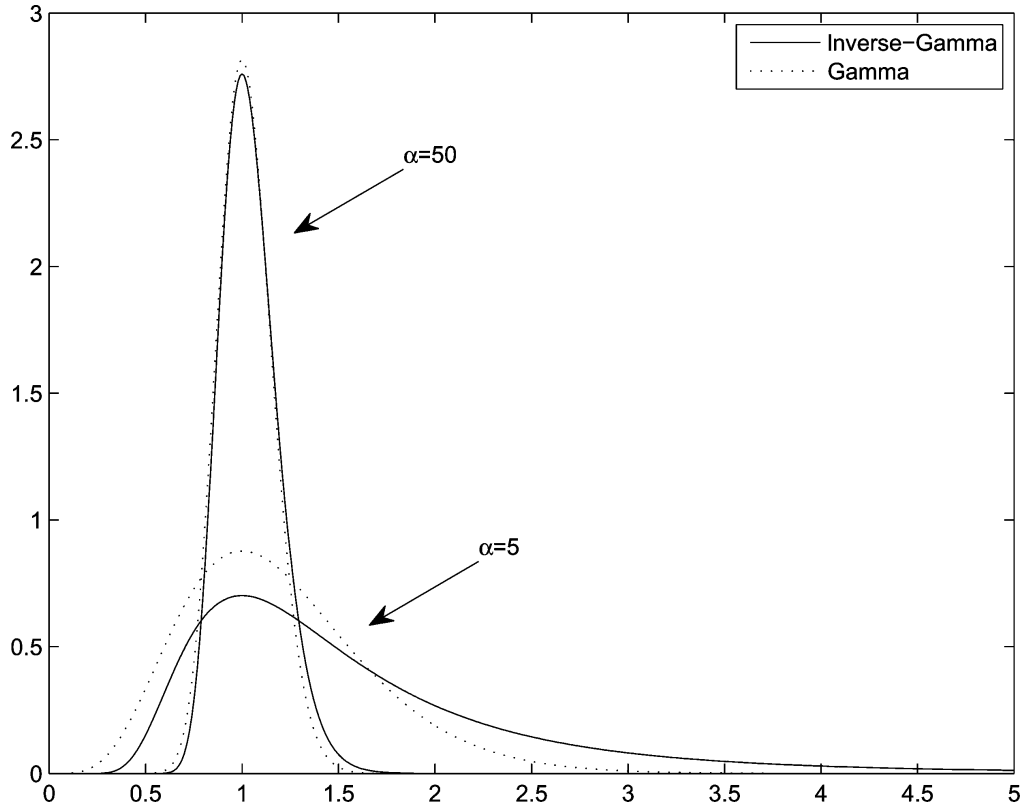


Figure 9: Prior PDFs $\mathcal{IG}(h_{kn} \mid \alpha - 1, \alpha\, h_{k(n-1)})$ (solid line) and $\mathcal{G}(h_{kn} \mid \alpha + 1, \alpha / h_{k(n-1)})$ (dashed line) for $h_{k(n-1)} = 1$ and for $\alpha = \{5, 50\}$.

and

$$p\big(h_{kn} \mid h_{k(n-1)}\big) = \mathcal{G}\big(h_{kn} \mid \alpha, (\alpha - 1)/h_{k(n-1)}\big), \tag{5.7}$$

where $\mathcal{G}(x|\alpha, \beta)$ is the previously introduced gamma PDF, with mode $(\alpha - 1)/\beta$ (for $\alpha \geq 1$) and $\mathcal{IG}(x|\alpha, \beta)$ is the inverse-gamma PDF (see appendix A), with mode $\beta/(\alpha + 1)$. Both priors are constructed so that their mode is obtained for $h_{kn} = h_{k(n-1)}$. α is a shape parameter that controls the sharpness of the prior around its mode. A high value of α will increase sharpness and thus accentuate the smoothness of h_k, while a low value of α will render the prior more diffuse and thus less constraining. The two priors become very similar for large values of α (see Figure 9). In the following, h_{k1} is assigned the scale-invariant Jeffreys noninformative prior $p(h_{k1}) \propto 1/h_{k1}$.

5.2 New Updates. Under prior structure 5.5, the derivative of $Q_k^{MAP}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}')$ with regard to h_{kn} writes $\forall n = 2, \ldots, N-1$,

$$\begin{aligned}\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') = \nabla_{h_{kn}} Q_k^{ML}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') - \nabla_{h_{kn}} \log p\big(h_{k(n+1)} \mid h_{kn}\big) \\ - \nabla_{h_{kn}} \log p\big(h_{kn} \mid h_{k(n-1)}\big).\end{aligned} \tag{5.8}$$

Table 2: Coefficients of the Order 2 Polynomial to Solve in Order to Update h_{kn} in Bayesian IS-NMF with a Markov Chain Prior.

	p_2	p_1	p_0
Inverse-Gamma Markov chain			
h_{k1}	$(\alpha+1)/h_{k2}$	$F-\alpha+1$	$-F\ \hat{h}_{k1}^{ML}$
h_{kn}	$(\alpha+1)/h_{k(n+1)}$	$F+1$	$-F\ \hat{h}_{kn}^{ML}-(\alpha+1)\,h_{k(n-1)}$
h_{kN}	0	$F+\alpha+1$	$-F\ \hat{h}_{kN}^{ML}-(\alpha+1)\,h_{k(N-1)}$
Gamma Markov chain			
h_{k1}	0	$F+\alpha+1$	$-F\ \hat{h}_{k1}^{ML}-(\alpha-1)\,h_{k2}$
h_{kn}	$(\alpha-1)/h_{k(n-1)}$	$F+1$	$-F\ \hat{h}_{kn}^{ML}-(\alpha-1)\,h_{k(n+1)}$
h_{kN}	$(\alpha-1)/h_{k(N-1)}$	$F-\alpha+1$	$-F\ \hat{h}_{kN}^{ML}$

Note: $\hat{h}_{kn}^{ML}$ denotes the ML update, given by equation 3.8.

This is shown to be equal to

$$\nabla_{h_{kn}} Q_k^{MAP}(\mathbf{w}_k, h_k \mid \boldsymbol{\theta}') = \frac{1}{h_{kn}^2}\big(p_2\, h_{kn}^2 + p_1\, h_{kn} + p_0\big), \tag{5.9}$$

where the values of p_0, p_1, and p_2 are specific to the type of prior employed (gamma or inverse-gamma chains), as given in Table 2. Updating h_{kn} then simply amounts to solving an order 2 polynomial. The polynomial has only one nonnegative root, given by

$$h_{kn} = \frac{\sqrt{p_1^2 - 4\,p_2\,p_0} - p_1}{2\,p_2}. \tag{5.10}$$

The coefficients h_{k1} and h_{kN} at the borders of the Markov chain require specific updates, but they also require solving polynomials of order 2 or 1, with coefficients given in Table 2 as well.

Note that the difference between the updates with the gamma and inverse-gamma chains prior mainly amounts to interchanging the positions of $h_{k(n-1)}$ and $h_{k(n+1)}$ in p_0 and p_2. Interestingly, using a backward gamma chain prior $p(h_k) = \prod_{n=1}^{N-1} p(h_{kn} \mid h_{k(n+1)})\ p(h_{kN})$ with shape parameter α is actually equivalent (in terms of MAP updates) to using a forward inverse-gamma chain prior as in equation 5.5 with shape parameter $\alpha - 2$. Respectively, using a backward inverse-gamma chain prior with shape parameter α is equivalent to using a forward-gamma chain prior with shape parameter $\alpha + 2$.

Note that Virtanen et al. (2008) recently considered gamma chains for regularization of KL-NMF. The modeling proposed in their work is, however, different from ours. Their gamma chain prior is constructed

in a hierarchical setting, by introducing extra auxiliary variables, so as to ensure conjugacy of the priors with the Poisson observation model. Estimation of the factors is then carried out with the standard gradient descent multiplicative approach, and single-channel source separation results are presented from the factorization of the magnitude spectrogram $|\mathbf{X}|$ with component reconstruction 2.18. Regularized NMF algorithms for the Euclidean and KL costs with norm-2 constraints on $h_{kn} - h_{k(n-1)}$ have also been considered by Chen, Cichocki, and Rutkowski (2006) and Virtanen (2007). Finally, we also wish to mention that Shashanka, Raj, and Smaragdis (2008b) have recently derived a regularized version of KL-NMF with sparsity constraints in a Bayesian setting.

6 Learning the Semantics of Music with IS-NMF

The aim of the experimental study proposed in section 4 was to analyze the results of several NMF algorithms on a short, simple, and well-defined musical sequence with respect to the cost function, initialization, and model order. We now present the results of NMF on a long polyphonic recording. Our goal is to examine how much of the semantics NMF can learn from the signal, with a fixed number of components and a fixed random initialization. This is not easily assessed numerically in the most general context, but quantitative evaluations could be performed on specific tasks in simulation settings. Such tasks could include music transcription, as in Abdallah and Plumbley (2004), single-channel source separation, as in Benaroya et al. (2006, 2003), or content-based music retrieval based on NMF features.

Rather than choosing and addressing one of these specific tasks, we use NMF in an actual audio restoration scenario, where the purpose is to denoise and upmix original monophonic material (one channel) to stereo (two channels). This task is very close to single-channel source separation, with the difference that we are not aiming at perfectly separating each of the sources, but rather isolating subsets of coherent components that can be given different directions of arrival in the stereo remaster so as to render a sensation of spatial diversity. We will show in particular that the addition of smoothness constraints on the rows of $\mathbf{H}$ leads to more pleasing component reconstructions and brings out the pitched structure of some of the learned PSDs better.

6.1 Experimental Setup. We address the decomposition of a 108-second-long music excerpt from "My Heart (Will Always Lead Me Back to You)" recorded by Louis Armstrong and His Hot Five in the 1920s. The band features (to our best hearing) a trumpet, a clarinet, a trombone, a piano, and a double bass. The data are original unprocessed mono material containing substantial noise. The signal was downsampled to $\nu_s =$ 11,025 kHz, yielding $T =$ 1,191,735 samples. The STFT $\mathbf{X}$ of x was computed using a sinebell analysis window of length $L = 256$ (23 ms) with 50%

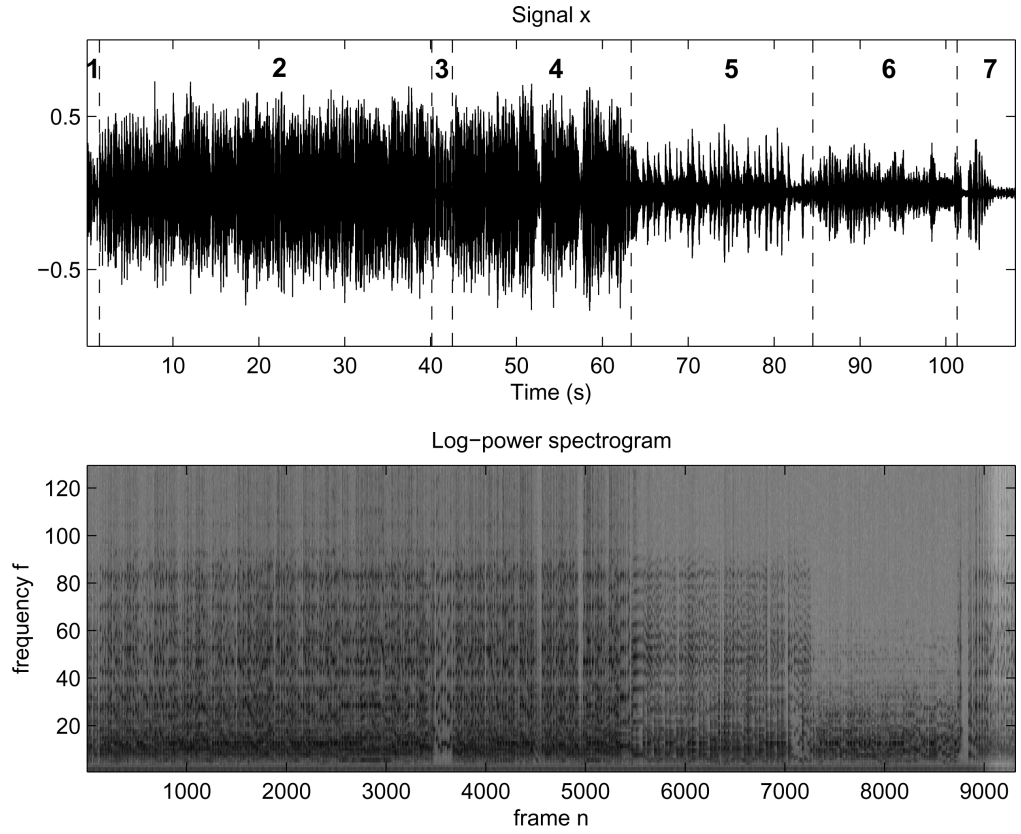


Figure 10: Original Louis Armstrong data. (Top) Time-domain recorded signal x. (Bottom) Log-power spectrogram. The vertical dashed lines on the top plot identify successive phases in the music piece, which we annotated manually: (2,4,7) all instruments, (1) clarinet only, (3) trumpet solo, (5) clarinet and piano, (6) piano solo.

overlap between two frames, leading to $N = 9312$ frames and $F = 129$ frequency bins. The time domain signal x and its log-power spectrogram are represented in Figure 10.

We applied EUC-NMF, KL-NMF, IS-NMF/MU, and IS-NMF/EM to $\mathbf{V} = |\mathbf{X}|^{\cdot[2]}$, as well as a regularized version of IS-NMF, as described in section 5. We used the inverse-gamma Markov chain prior (see equation 5.6) with α arbitrarily set to 10. We refer to this algorithm as IS-NMF/IG. Among many trials, this value of α provided a good trade-off between the smoothness of the component reconstructions and adequacy to data. Experiments with the gamma Markov chain prior, equation 5.6, did not lead to significant differences in the results and are not reported here.

The number of components K was arbitrarily set to 10. All five algorithms were run for $n_{iter} = 5000$ iterations and were initialized with the same random values. For comparison, we also applied KL-NMF to the magnitude spectrogram $|\mathbf{X}|$ with component reconstruction described by equation 2.18, as this can be considered state-of-the-art methodology for NMF-based single-channel audio source separation (Virtanen, 2007).

6.2 Results and Discussion. For conciseness, we here display only the decomposition obtained with IS-NMF/IG (see Figure 11) because it leads to the best results as far as our audio restoration task is concerned. (All decompositions and component reconstructions obtained from all NMF algorithms are available online at http://www.tsi.enst.fr/∼fevotte/Samples/is-nmf.) Figure 11 displays the estimated basis functions $\mathbf{W}$ in log-scale on the left and represents on the right the time-domain signal components reconstructed from Wiener filtering.

Figure 12 displays the evolution of the IS cost along the 5000 iterations with IS-NMF/MU, IS-NMF/EM, and IS-NMF/IG. In this case, IS-NMF/EM achieves a lower cost than IS-NMF/MU. The run times of 1000 iterations of the algorithms were, respectively: EUC-NMF, 1.9 min; KL-NMF, 6.8 min; IS-NMF/MU, 8.7 min; IS-NMF/EM, 23.2 min; and IS-NMF/IG, 32.2 min.

The comparison of the decompositions obtained with the three cost functions (Euclidean, KL, and IS), through visual inspection of $\mathbf{W}$ and listening to the components c_k, shows again that the IS divergence leads to the most interpretable results. In particular, some of the columns of matrix $\mathbf{W}$ produced by all three IS-NMF algorithms have a clear pitched structure, which indicates that some notes have been extracted. Furthermore, one of the components captures the hiss noise from the recording. Discarding this component from the reconstruction of x yields satisfying denoising (this is particularly noticeable during the piano solo, where the input SNR is low). Surprisingly, most of the rhythmic accompaniment (piano and double bass) is isolated in a single component (component 1 of IS-NMF/MU, component 2 of IS-NMF/EM and IS-NMF/IG), though its spectral content is clearly evolving in time. A similar effect happens with IS-NMF/IG and the trombone, which is mostly contained by component 7.

While we do not have a definite explanation for this, we believe that this is a consequence of Wiener reconstruction. Indeed, the Wiener component reconstruction is seen only as a set of K masking filters applied to x_{fn}, so that it does not constrain the spectrum of component k to be exactly $\mathbf{w}_k$ (up to amplitude h_{kn}), as the reconstruction method described by equation 2.18 does. So if one assumes that the IS-NMF model, equation 2.10, adequately captures some of the sound entities present in the mix (in our case, that would be the preponderant notes or chords and the noise), then the other entities are bound to be relegated in remaining components by conservativity of the decomposition $x = \sum_{k=1}^{K} c_k$.

As anticipated, the addition of frame-persistency constraints with IS-NMF/IG has an impact on the learned basis $\mathbf{W}$. In particular, some of the components exhibit a more pronounced pitched structure. But more important, the regularization yields more pleasing sound reconstructions, which is particularly noticeable when listening to the accompaniment component obtained from IS-NMF/MU (component 1) or IS-NMF/EM (component 2) on the one side and from IS-NMF/IG (component 2) on the other side. Note also that in every case, the sound quality of Wiener

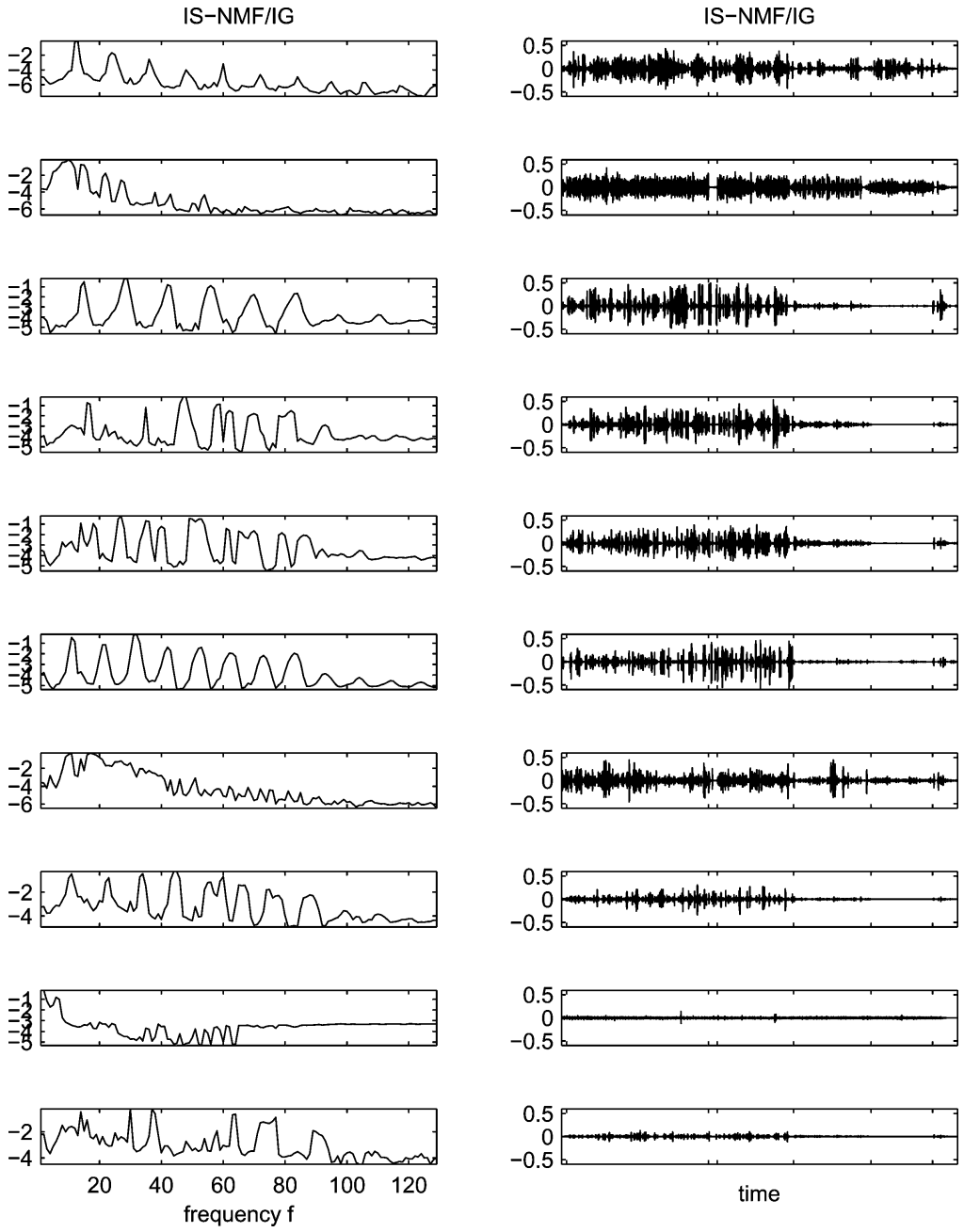


Figure 11: Decomposition of Louis Armstrong music data with IS-NMF/IG. (Left) Columns of $\mathbf{W}$ ($\log_{10}$ scale). (Right) Reconstructed components c_k. The x-axis ticks correspond to the temporal segmentation border lines displayed with signal x on Figure 10. Component 2 captures most of the acompaniment, component 7 most of the trombone, and component 9 most of the hiss noise. Summing up the other components leads to extracting the trumpet and clarinet, together with some piano notes.

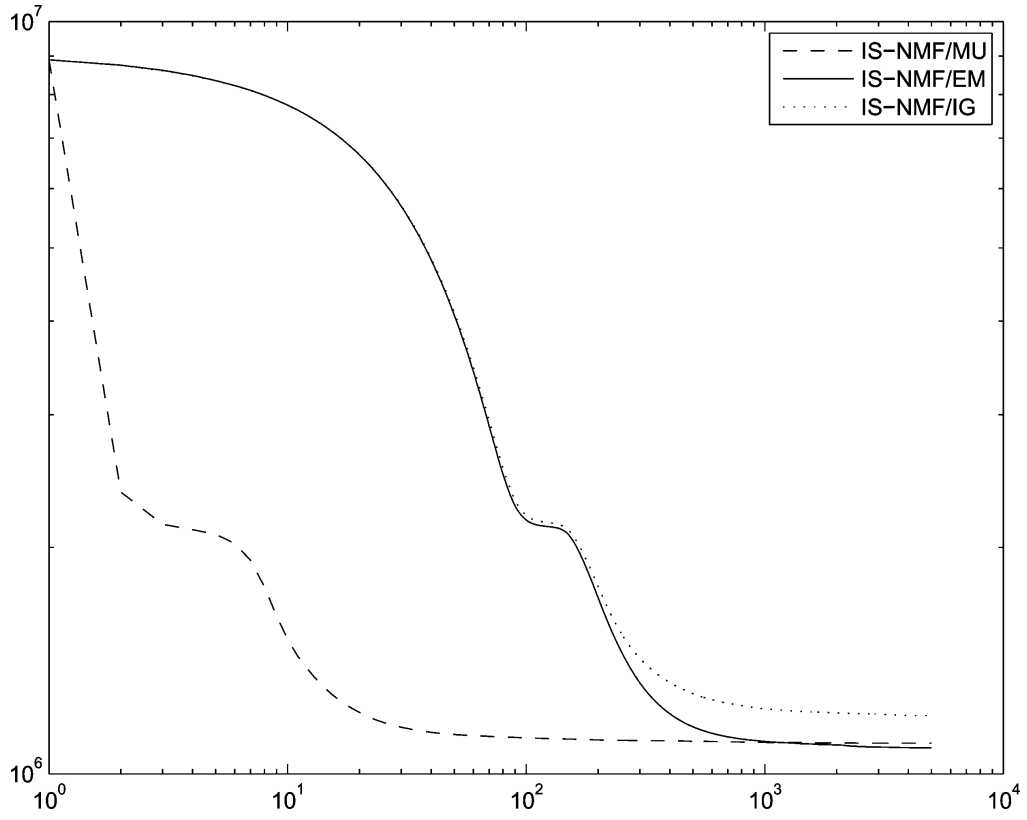


Figure 12: Evolution in log-log scale of the IS cost function along the 5000 iterations of IS-NMF/MU, IS-NMF/EM, and IS-NMF/IG, initialized with the same random values, with $K = 10$.

reconstructions is far better than state-of-the-art KL-NMF of $|\mathbf{X}|$ and ad hoc reconstruction described by equation 2.18.

To conclude this study, we provide online a restored version of the original recording, produced from the IS-NMF/IG decomposition. This is, to our best knowledge, the first use of NMF in an actual audio restoration scenario. The restoration includes denoising (by discarding component 9, which is regarded as noise) and upmixing. A stereo mix is produced by dispatching parts of each component to the left and right channels, hence simulating directions of arrival. As such, we manually created a mix where the components are arranged from 54 degrees left to 54 degrees right, such that the wind instruments (trumpet, clarinet, trombone) are placed left and the stringed instruments (piano, double bass) are placed right. While this stereo mix does render a sensation of spatialization, we emphasize that its quality could undoubtedly be improved with appropriate sound engineering skills.

The originality of our restoration approach lies in the joint noise removal and upmix (as opposed to a suboptimal sequential approach) and the genuine content-based remastering, as opposed to standard techniques based, for example, on phase delays or equalization.

7 Conclusions

We have presented modeling and algorithmic aspects of NMF with the Itakura-Saito divergence. On the modeling side, the following three features of IS-NMF have been demonstrated in this letter:

- IS-NMF is underlaid by a statistical model of superimposed gaussian components.
- This model is relevant to the representation of audio signals.
- This model can accommodate regularization constraints through Bayesian approaches.

On the algorithmic side, we have proposed a novel type of NMF algorithm, IS-NMF/EM, derived from SAGE, a variant of the EM algorithm. The convergence of this algorithm to a stationary point of the cost function $D_{IS}(\mathbf{V} \mid \mathbf{WH})$ is guaranteed by EM. This new algorithm was compared to an existing algorithm, IS-NMF/MU, whose convergence has not been proved, though it has been observed in practice. This letter also reports an experimental comparative study of the standard EUC-NMF and KL-NMF algorithms, together with the two described IS-NMF algorithms, applied to a given data set (a short piano sequence), with various random initializations and model orders. Such a furnished experimental study was, to our best knowledge, not yet available. This letter also reports a proof of concept of the use of IS-NMF for audio restoration, with an actual example. Finally, we believe we have shed light on the statistical implications of NMF with all of three cost functions.

We have shown how smoothness constraints on $\mathbf{W}$ and $\mathbf{H}$ can easily be handled in a Bayesian setting with IS-NMF. As such, we have shown how Markov chains' prior structures can improve both the auditory quality of the component reconstructions and the interpretability of the basis elements. The Bayesian setting opens doors to even more elaborate prior structures that can better fit the specificities of data. For music signals, we believe that two promising lines of research lay in (1) the use of switching state models for the rows of $\mathbf{H}$ that explicitly model the possibility for h_{kn} to be strictly zero with a certain prior probability (and time persistency could be favored by modeling the state sequence with a discrete Markov chain) and (2) the use of models that explicitly take into account the pitched structure of some of the columns of $\mathbf{W}$ and where the fundamental frequency could act as a model parameter. These models fit into the problem of object-based representation of sound, an active area of research in the music information retrieval and auditory scene analysis communities.

In section 4 we compared the factorization results of a short piano power spectrogram, obtained from three cost functions, given a common algorithmic structure: standard multiplicative updates. The experiments illustrate the slow convergence of this type of algorithm, which has has already been pointed out in other work (Cichocki, Amari et al., 2006; Berry et al., 2007;

Lin, 2007). If the proposed IS-NMF/EM does not improve on this issue, its strength is, however, to offer enough flexibility to accommodate Bayesian approaches. We believe we have made our point that the IS cost is well suited to the factorization of audio power spectrograms (i.e., independent of the type of algorithm used); future work will address the development of faster IS-NMF algorithms. Following developments for other cost functions, we intend to investigate projected gradient techniques (Lin, 2007), exponentiated gradient descent and generalizations (Cichocki, Amari et al., 2006), quasi-Newton second-order methods (Zdunek & Cichocki, 2007), and multilayered approaches (Cichocki & Zdunek , 2006).

Key issues that still need to be resolved in NMF concern identifiability and order selection. A related issue is the investigation into the presence of local minima in cost functions and ways to avoid them. In that matter, Markov chain Monte Carlo (MCMC) sampling techniques could be used as a diagnostic tool to better understand the topography of the criteria to minimize. While it is not clear whether these techniques can be applied to EUC-NMF or KL-NMF, they can readily be applied to IS-NMF, using its underlying gaussian composite structure the same way that IS-NMF/EM does. As to the avoidance of local minima, techniques inherited from simulated annealing could be applied with IS-NMF in either MCMC or EM inference.

Regarding order selection, usual criteria such as the Bayesian information criterion or Akaike's criterion (see, e.g., Stoica & Selén, 2004) cannot be directly applied to IS-NMF because the number of parameters ($F\ K + K\ N$) is not constant with regard to the number of observations N. This feature breaks the validity of the assumptions in which these criteria have been designed. As such, a final promising line of research concerns the design of methods characterizing $p(\mathbf{V} \mid \mathbf{W})$ instead of $p(\mathbf{V} \mid \mathbf{W}, \mathbf{H})$, treating $\mathbf{H}$ as a latent variable, as in independent component analysis (MacKay, 1996; Lewicki & Sejnowski, 2000). Besides allowing for model order selection, such approaches should lead to more reliable estimation of the basis $\mathbf{W}$.

Appendix A: Standard Distributions

Proper complex gaussian	$\mathcal{N}_c(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \|\pi\ \boldsymbol{\Sigma}\|^{-1}\ \exp -(\mathbf{x}-\boldsymbol{\mu})^H\ \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$
Poisson	$\mathcal{P}(x\|\lambda) = \exp(-\lambda)\ \frac{\lambda^x}{x!}$
Gamma	$\mathcal{G}(u \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ u^{\alpha-1}\ \exp(-\beta\, u),\ u \geq 0$
Inverse-gamma	$\mathcal{IG}(u \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ u^{-(\alpha+1)}\ \exp(-\frac{\beta}{u}),\ u \geq 0$

The inverse-gamma distribution is the distribution of $1/X$ when X is gamma distributed.

Appendix B: Derivations of the SAGE Algorithm

In this appendix, we detail the derivations leading to algorithm 2. The functions involved in the definition of $Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')$, given by equation 3.2,

can be derived as follows. For the hidden data minus log likelihood,

$$-\log p(\mathbf{C}_k \mid \boldsymbol{\theta}_k) = -\sum_{n=1}^{N}\sum_{f=1}^{F} \log \mathcal{N}_c(c_{k,fn} \mid 0, h_{kn}\, w_{fk}) \tag{B.1}$$

$$\stackrel{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\, h_{kn}\right) + \frac{|c_{k,fn}|^2}{w_{fk}\, h_{kn}}. \tag{B.2}$$

Then the hidden-data posterior is obtained through Wiener filtering, yielding

$$p(\mathbf{C}_k \mid \mathbf{X}, \boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{f=1}^{F} \mathcal{N}_c\left(c_{k,fn} \mid \mu_{k,fn}^{post}, \lambda_{k,fn}^{post}\right), \tag{B.3}$$

with $\mu_{k,fn}^{post}$ and $\lambda_{k,fn}^{post}$ given by equations 3.3 and 3.4. The E-step is performed by taking the expectation of equation B.2 with regard to the hidden-data posterior, leading to

$$Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}') \stackrel{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(w_{fk}\, h_{kn}\right) + \frac{\left|\mu_{k,fn}^{post\,\prime}\right|^2 + \lambda_{k,fn}^{post\,\prime}}{w_{fk}\, h_{kn}} \tag{B.4}$$

$$\stackrel{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} d_{IS}\left(\left|\mu_{k,fn}^{post\,\prime}\right|^2 + \lambda_{k,fn}^{post\,\prime} \mid w_{fk}\, h_{kn}\right). \tag{B.5}$$

The M-step thus amounts to minimizing $D_{IS}(\mathbf{V}'_k \mid \mathbf{w}_k\, h_k)$ with regard to $\mathbf{w}_k \geq 0$ and $h_k \geq 0$, as stated in section 3.2.

Acknowledgments

We acknowledge Roland Badeau, Olivier Cappé, Jean-François Cardoso, Maurice Charbit, and Alexey Ozerov for discussions related to this work. Many thanks to Gaël Richard for comments and suggestions about this manuscript and to Simon Godsill for helping us with the Louis Armstrong original data.

References

Abdallah, S. A., & Plumbley, M. D. (2004). Polyphonic transcription by nonnegative sparse coding of power spectra. In *5th International Symposium of Music Information Retrieval (ISMIR'04)* (pp. 318–325). Vienna: Austrian Computer Society.

Benaroya, L., Blouet, R., Févotte, C., & Cohen, I. (2006). Single sensor source separation using multiple-window STFT representation. In *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*. Paris: Télécom Paris.

Benaroya, L., Gribonval, R., & Bimbot, F. (2003). Non negative sparse representation for Wiener based source separation with a single sensor. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'03)* (pp. 613–616). Los Alamitos, CA: IEEE.

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis, 52*(1), 155–173.

Bertin, N., Badeau, R., & Richard, G. (2007). Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*. Los Alamitos, CA: IEEE.

Chen, Z., Cichocki, A., & Rutkowski, T. M. (2006). Constrained non-negative matrix factorization method for EEG analysis in early detection of Alzheimer's disease. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*. Los Alamitos, CA: IEEE.

Cichocki, A., Amari, S.-I., Zdunek, R., Kompass, R., Hori, G., & He, Z. (2006). Extended SMART algorithms for non-negative matrix factorization. In *Proc. of the International Conference on Artificial Intelligence and Soft Computing (ICAISC'06)* (pp. 548–562). Calgary, Canada: ACTA Press.

Cichocki, A., & Zdunek, R. (2006). Multilayer nonnegative matrix factorization. *Electronics Letters, 42*(16), 947–948.

Cichocki, A., Zdunek, R., & Amari, S. (2006). Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)* (pp. 32–39). Berlin: Springer.

Cohen, I., & Gannot, S. (2007). Spectral enhancement methods. In M. M. Sondhi, J. Benesty, & Y. Huang (Eds.), *Springer handbook of speech processing*. New York: Springer.

Dhillon, I. S., & Sra, S. (2005). Generalized nonnegative matrix approximations with Bregman divergences. In M. I. Jordan, Y. Le Cun, & S. A. Solla (Eds.), *Advances in neural information processing systems, 19*. Cambridge, MA: MIT Press.

Drakakis, K., Rickard, S., de Fréin, R., & Cichocki, A. (2008). Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum, 3*, 1853–1870.

Eguchi, S., & Kano, Y. (2001). *Robustifying maximum likelihood estimation*. (Research Memo 802). Tokyo: Institute of Statistical Mathematics. Available online at http://www.ism.ac.jp/~eguchi/pdf/Robustify_MLE.pdf.

Feder, M., & Weinstein, E. (1988). Parameter estimation of superimposed signals using the EM algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 36*, 477–489.

Fessler, J. A., & Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing, 42*, 2664–2677.

Gray, R. M., Buzo, A., Gray, A. H., & Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*, 367–376.

Itakura, F., & Saito, S. (1968). Analysis synthesis telephony based on the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics* (pp. C–17–C–20). Los Alamitos, CA: IEEE.

Kompass, R. (2007). A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation, 19*, 780–791.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects with nonnegative matrix factorization. *Nature, 401*, 788–791.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural and information processing systems, 13* (pp. 556–562). Cambridge, MA: MIT Press.

Lewicki, M. S., & Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation, 12*, 337–365.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. *Neural Computation, 19*, 2756–2779.

MacKay, D. (1996). *Maximum likelihood and covariant algorithms for independent component analysis*. Unpublished manuscript. Available online at http://www.inference.phy.cam.ac.uk/mackay/ica.pdf.

Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing, 15*, 1564–1578.

Plumbley, M. D., Abdallah, S. A., Blumensath, T., & Davies, M. E. (2006). Sparse representations of polyphonic music. *Signal Processing, 86*(3), 417–431.

Shashanka, M., Raj, B., & Smaragdis, P. (2008a). Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience, 2008*.

Shashanka, M., Raj, B., & Smaragdis, P. (2008b). Sparse overcomplete latent variable decomposition of counts data. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems, 20* (pp. 1313–1320). Cambridge, MA: MIT Press.

Smaragdis, P. (2007). Convolutive speech bases and their application to speech separation. *IEEE Transactions on Audio, Speech, and Language Processing, 15*, 1–12.

Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Los Alamitos, CA: IEEE.

Stoica, P., & Selén, Y. (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine, 21*, 36–47.

Vincent, E., Bertin, N., & Badeau, R. (2007). Two nonnegative matrix factorization methods for polyphonic pitch transcription. In *Proc. of the Music Information Retrieval Evaluation eXchange (MIREX)*. Available online at http://www.music-ir.org.

Virtanen, T. (2007). Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing, 15*, 1066–1074.

Virtanen, T., Cemgil, A. T., & Godsill, S. (2008). Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)* (pp. 1825–1828). Los Alamitos, CA: IEEE.

Young, S. S., Fogel, P., & Hawkins, D. (2006). Clustering scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association, 14*, 11–13.

Zdunek, R., & Cichocki, A. (2007). Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing, 87*, 1904–1916.

Received April 29, 2008; accepted July 3, 2008.

(Ozerov & Févotte, *IEEE TASLP*, 2010)

Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation

Alexey Ozerov, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***Abstract*—We consider inference in a general data-driven object-based model of multichannel audio data, assumed generated as a possibly underdetermined convolutive mixture of source signals. We work in the short-time Fourier transform (STFT) domain, where convolution is routinely approximated as linear instantaneous mixing in each frequency band. Each source STFT is given a model inspired from nonnegative matrix factorization (NMF) with the Itakura–Saito divergence, which underlies a statistical model of superimposed Gaussian components. We address estimation of the mixing and source parameters using two methods. The first one consists of maximizing the exact joint likelihood of the multichannel data using an expectation-maximization (EM) algorithm. The second method consists of maximizing the sum of individual likelihoods of all channels using a multiplicative update algorithm inspired from NMF methodology. Our decomposition algorithms are applied to stereo audio source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings. Our EM method produces competitive results with respect to state-of-the-art as illustrated on two tasks from the international Signal Separation Evaluation Campaign (SiSEC 2008).**

***Index Terms*—Expectation-maximization (EM) algorithm, multichannel audio, nonnegative matrix factorization (NMF), nonnegative tensor factorization (NTF), underdetermined convolutive blind source separation (BSS).**

I. Introduction

NONNEGATIVE matrix factorization (NMF) is an unsupervised data decomposition technique with effervescent popularity in the fields of machine learning and signal/image processing [1]. Much research about this topic has been driven by applications in audio, where the data matrix is taken as the magnitude or power spectrogram of a sound signal. NMF was for example applied with success to automatic music transcription [2], [3] and audio source separation [4], [5]. The factorization amounts to decomposing the spectrogram data into a sum of rank-1 spectrograms, each of which being the expression of an elementary spectral pattern amplitude-modulated in time. However, while most music recordings are available in multichannel format (typically, stereo), NMF in its standard setting is only suited to single-channel data. Extensions to multichannel data have been considered, either by stacking up the spectrograms of each channel into a single matrix [6] or by considering nonnegative tensor factorization (NTF) under a parallel factor analysis (PARAFAC) structure, where the channel spectrograms form the slices of a 3-valence tensor [7]. These approaches inherently assume that the original sources have been mixed instantaneously, which in modern music mixing is not realistic, and they require a posterior binding step so as to group the elementary components into instrumental sources. Furthermore they do not exploit the redundancy between the channels in an optimal way, as will be shown later.

The aim of this work is to remedy these drawbacks. We formulate a multichannel NMF model that accounts for convolutive mixing. The source spectrograms are modeled through NMF and the mixing filters serve to identify the elementary components pertaining to each source. We consider more precisely I sampled signals $\tilde{x}_i(t)$ $(i = 1, \ldots, I,\ t = 1, \ldots, T)$ generated as convolutive noisy mixtures of J point source signals $\tilde{s}_j(t)$ $(i = 1, \ldots, J)$ such that

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau) \tilde{s}_j(t-\tau) + \tilde{b}_i(t) \tag{1}$$

where $\tilde{a}_{ij}(\tau)$ is the finite-impulse response of some (causal) filter and $\tilde{b}_i(t)$ is some additive noise. The time-domain mixing given by (1) can be approximated in the short-time Fourier transform (STFT) domain as

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \tag{2}$$

where $x_{i,fn}$, $s_{j,fn}$ and $b_{i,fn}$ are the complex-valued STFTs of the corresponding time signals, $a_{ij,f}$ is the complex-valued discrete Fourier transform of filter $\tilde{a}_{ij}(\tau)$, $f = 1, \ldots, F$ is a frequency bin index, and $n = 1, \ldots, N$ is a time frame index. Equation (2) holds when the filter length L is assumed "significantly" shorter than the STFT window size $(2F - 2)$ [8]. Equation (2) can be rewritten in matrix form, such that

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn} \tag{3}$$

where $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$, $\mathbf{s}_{fn} = [s_{1,fn}, \ldots, s_{J,fn}]^T$, $\mathbf{b}_{fn} = [b_{1,fn}, \ldots, b_{I,fn}]^T$, and $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$.

Manuscript received December 24, 2008; revised August 17, 2009. Current version published February 10, 2010. This work was supported in part by the French ANR project SARAH (StAndardisation du Remastering Audio Haute-Définition). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

A. Ozerov was with the Institut Telecom, Telecom ParisTech, CNRS LTCI, 75014 Paris, France. He is now with the METISS Team of IRISA/INRIA, 35042 Rennes Cedex, France (e-mail: alexey.ozerov@irisa.fr).

C. Févotte is with CNRS LTCI, Telecom ParisTech, 75014 Paris, France (e-mail: cedric.fevotte@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2031510

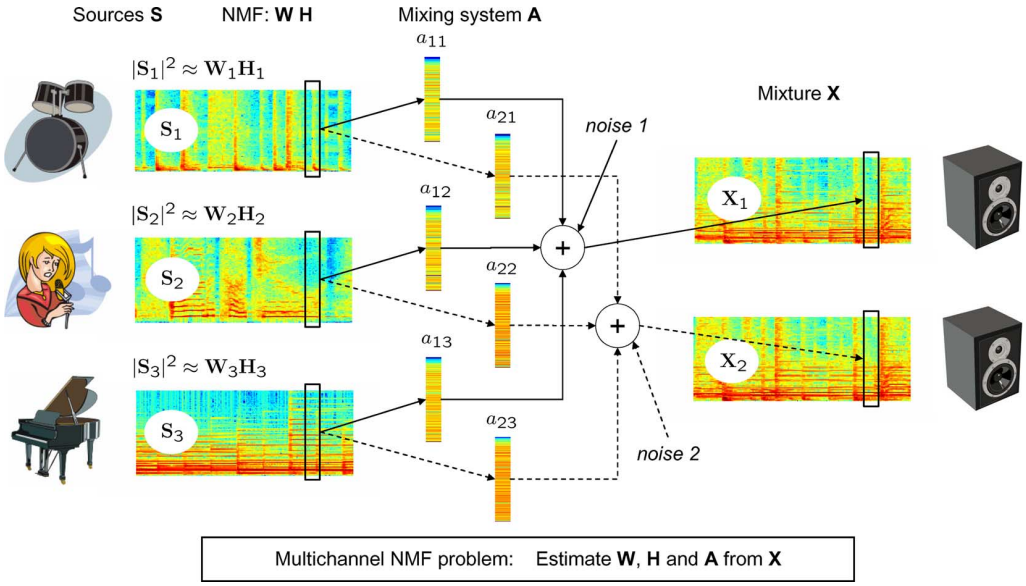


Fig. 1. Representation of convolutive mixing system and formulation of Multichannel NMF problem.

A key ingredient of this work is to model the $F \times N$ power spectrogram $|\mathbf{S}_j|^2 = [|s_{j,fn}|^2]_{fn}$ of source j as a product of two nonnegative matrices $\mathbf{W}_j$ and $\mathbf{H}_j$, such that

$$|\mathbf{S}_j|^2 \approx \mathbf{W}_j\mathbf{H}_j. \tag{4}$$

Given the observed mixture STFTs $\mathbf{X} = \{x_{i,fn}\}_{ifn}$, we are interested in joint estimating the source spectrogram factors $\{\mathbf{W}_j, \mathbf{H}_j\}_j$ and the mixing system $\{\mathbf{A}_f\}_f$, as illustrated in Fig. 1. Our problem splits into two subtasks: 1) defining suitable estimation criteria, and 2) designing algorithms optimizing these criteria.

We adopt a statistical setting in which each source STFT is modeled as a sum of latent Gaussian components, a model introduced by Benaroya *et al.* [9] in a supervised single-channel audio source separation context. A connection between full maximum-likelihood (ML) estimation of the variance parameters in this model and NMF using the Itakura–Saito (IS) divergence was pointed out in [10]. Given this source model, hereafter referred to as *NMF model*, we introduce two estimation criteria together with corresponding inference methods.

- The first method consists of maximizing the exact joint log-likelihood of the multichannel data using an expectation-maximization (EM) algorithm [11]. This method fully exploits the redundancy between the channels, in a statistically optimal way. It draws parallels with several model-based multichannel source separation methods [12]–[18], as described throughout the paper.
- The second method consists of maximizing the sum of individual log-likelihoods of all channels using a multiplicative update (MU) algorithm inspired from NMF methodology. This approach relates to the above-mentioned NTF techniques [6], [7]. However, in contrast to standard NTF which inherently assumes instantaneous mixing, our approach addresses a more general convolutive structure and does not require the posterior binding of the elementary components into J sources.

The general multichannel NMF framework we describe yields a data-driven object-based representation of multichannel data that may benefit many tasks in audio, such as transcription or object-based coding. In this article we will more specifically focus on the convolutive blind source separation (BSS) problem, and as such we also address means of reconstructing source signal estimates from the set of estimated parameters. Our decompositions are conservative in the sense that the spatial source estimates sum up to the original mix. The mixing parameters may also be changed without degrading audio quality, so that music remastering is one potential application of our work. Remixes of well-known songs retrieved from commercial CD recordings are proposed in the results section.

Many convolutive BSS methods have been designed under model (3). Typically, an instantaneous independent component analysis (ICA) algorithm is applied to data $\{\mathbf{x}_{fn}\}_{n=1,\ldots,N}$ in each frequency subband f, yielding a set of J source subband estimates per frequency bin. This approach is usually referred to as frequency-domain ICA (FD-ICA) [19]. The source labels remain however unknown because of the ICA standard permutation indeterminacy, leading to the well-known FD-ICA permutation alignment problem, which cannot be solved without using additional *a priori* knowledge about the sources and/or about the mixing filters. For example in [20] the sources in different frequency bins are grouped *a posteriori* relying on their temporal correlation, thus using prior knowledge about the sources, and in [21], [22] the sources and the filters are estimated assuming a particular structure of convolutive filters, i.e., using prior knowledge about the filters. The permutation ambiguity arises from the individual processing of each subband, which implicitly assumes mutual independence of one source's subbands. This is not the case in our work where our source model implies a coupling of the frequency bands, and joint estimation of the source

parameters and mixing coefficients frees us from the permutation alignment problem.

Our EM-based method is related to some multichannel source separation techniques employing Gaussian mixture models (GMMs) as source models. Univariate independent and identically distributed (i.i.d.) GMMs have been used to model source samples in the time domain for separation of instantaneous [12], [13] and convolutive [12] mixtures. However, such time-domain GMMs are not of the most relevance for audio as they do not model temporal correlations in the signal. In [14], Attias proposes to model the sources in the STFT domain using multivariate GMMs, hence taking into account temporal correlations in the audio signal, assumed stationary in each window frame. The author develops a source separation method for convolutive mixtures, supervised in the sense that the source models are pre-trained in advance. A similar approach with log-spectral domain GMMs is developed by Weiss *et al.* in [15]. Arberet *et al.* [16] propose a multivariate GMM-based separation method for instantaneous mixing that involves a computationally efficient strategy for learning the source GMMs separately, using intermediate source estimates obtained by some BSS method. As compared to these works, we use a different source model (the NMF model), which might be considered more suitable than the GMM for musical signals. Indeed, the NMF is well suited to polyphony as it basically takes the source to be a sum of elementary components with characteristic spectral signatures. In contrast, the GMM takes the source as a single component with many states, each representative of a characteristic spectral signature, but not mixed *per se*. To put it in an other way, in the NMF model a summation occurs in the STFT domain (or equivalently, in the time domain), while in the GMM the summation occurs on the distribution of the frames. Moreover, as discussed later, the computational complexity of inference in our model grows linearly with the number of components while the complexity of standard inference in GMMs grows combinatorially.

The remaining of this paper is organized as follows. NMF source model and noise model are introduced in Section II. Section III is devoted to the definition of our two estimation criteria, with corresponding optimization algorithms. Section IV presents results of our methods to stereo source separation in various settings, including blind and supervised separation of music and speech sources in synthetic instantaneous and convolutive mixtures, as well as in professionally produced music recordings. Conclusions are drawn in Section V. Preliminary aspects of this work are presented in [23]. We here considerably extend on the simulations part as well as on the theoretical developments related to our algorithms.

II. Models

A. Sources

Let $K \geq J$ and $\{\mathcal{K}_j\}_{j=1}^{J}$ be a nontrivial partition of $\mathcal{K} = \{1, \ldots, K\}$. Following [9], [10], we assume the complex random variable $s_{j,fn}$ to be a sum of $\#\mathcal{K}_j$ latent components, such that

$$s_{j,fn} = \sum_{k \in \mathcal{K}_j} c_{k,fn} \quad \text{with} \quad c_{k,fn} \sim \mathcal{N}_c(0, w_{fk} h_{kn}) \tag{5}$$

where $w_{fk}, h_{kn} \in \mathbb{R}^+$ and $\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the *proper* complex Gaussian distribution [24] with probability density function (pdf)

$$N_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{|\pi \boldsymbol{\Sigma}|} \exp\left[-(\mathbf{x} - \boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]. \tag{6}$$

In the rest of the paper, the quantities $s_{j,fn}$ and $c_{k,fn}$ are, respectively, referred to as "source" and "component". The components are assumed *mutually* independent and *individually* independent across frequency f and frame n. It follows that

$$s_{j,fn} \sim \mathcal{N}_c\left(0, \sum_{k \in \mathcal{K}_j} w_{fk} h_{kn}\right). \tag{7}$$

Denoting $\mathbf{S}_j$ the $F \times N$ STFT matrix $[s_{j,fn}]_{fn}$ of source j and introducing the matrices $\mathbf{W}_j = [w_{fk}]_{f, k \in \mathcal{K}_j}$ and $\mathbf{H}_j = [h_{kn}]_{k \in \mathcal{K}_j, n}$, respectively, of dimensions $F \times \#\mathcal{K}_j$ and $\#\mathcal{K}_j \times N$, it is easily shown [10] that the minus log-likelihood of the parameters describing source j writes

$$-\log p(\mathbf{S}_j | \mathbf{W}_j, \mathbf{H}_j) \stackrel{c}{=} \sum_{fn} d_{IS}\left(|s_{j,fn}|^2 | [\mathbf{W}_j \mathbf{H}_j]_{fn}\right)$$

where "$\stackrel{c}{=}$" denotes equality up to a constant and

$$d_{IS}(x|y) = \frac{x}{y} - \log \frac{x}{y} - 1 \tag{8}$$

is the IS divergence. In other words, ML estimation of $\mathbf{W}_j$ and $\mathbf{H}_j$ given source STFT $\mathbf{S}_j$ is equivalent to NMF of the power spectrogram $|\mathbf{S}_j|^2$ into $\mathbf{W}_j \mathbf{H}_j$, where the IS divergence is used. MU and EM algorithms for IS-NMF are, respectively, described in [25], [26] and in [10]; in essence, this paper describes a generalization of these algorithms to a multichannel multisource scenario. In the following, we will use the notation $\mathbf{P}_j = \mathbf{W}_j \mathbf{H}_j$, i.e., $p_{j,fn} = \mathbb{E}\{|s_{j,fn}|^2\}$.

Our source model is related to the GMM used for example in [14], [16] in the same source separation context, with the difference that one source frame is here modeled as a sum of $\#\mathcal{K}_j$ elementary components while in the GMM one source frame is modeled as a process which can take one of many states, each characterized by a covariance matrix. The computational complexity of inference in our model with our algorithms described next grows linearly with the total number of components while the derivation of the equivalent EM algorithm for GMM leads to an algorithm that has combinatorial complexity with the number of states [12], [13], [15]. It is possible to achieve linear complexity in the GMM case also, but at the price of approximate inference [14], [16]. Note that all considered algorithms, either for the NMF model or GMM, only ensure convergence to a stationary point of the objective function, and, as a consequence, the final result depends strongly on the parameters initialization. We wish to emphasize that we here take a fully data-driven approach in the sense that no parameter is pre-trained.

B. Noise

In the most general case, we may assume noisy data and the following algorithms can easily accommodate estimation of noise statistics under Gaussian independent assumptions and given covariance structures such as $\boldsymbol{\Sigma}_{\mathbf{b},fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}$ or $\boldsymbol{\Sigma}_{\mathbf{b},n}$. In

this paper, we consider for simplicity stationary and spatially uncorrelated noise such that

$$b_{i,fn} \sim \mathcal{N}_c\left(0, \sigma^2_{i,f}\right) \tag{9}$$

and $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma^2_{i,f}]_i)$. The musical data we consider in Section IV-A is not noisy in the usual sense, but the noise component can account for model discrepancy and/or quantization noise. Moreover, this noise component is required in the EM algorithm to prevent from potential numerical instabilities (see Section III-A1 below) and slow convergence (see Section III-A6 below). In Section IV-D, we will consider several scenarios: when the variances are equal and fixed to a small value $\tilde{\sigma}^2$, when the variances are estimated from data, and most importantly when annealing is performed via the noise variance, so as to speed up convergence as well as favor global solutions.

C. Convolutive Mixing Model Revisited

With (5), the mixing model (3) can be recast as

$$\mathbf{x}_{fn} = \mathring{\mathbf{A}}_f \mathbf{c}_{fn} + \mathbf{b}_{fn} \tag{10}$$

where $\mathbf{c}_{fn} = [c_{1,fn}, \ldots, c_{K,fn}]^T \in \mathbb{C}^{K\times 1}$ and $\mathring{\mathbf{A}}_f$ is the so called "augmented mixing matrix" of dimension $I \times K$, with elements defined by $\mathring{a}_{ik,f} = a_{ij,f}$ if and only if $k \in \mathcal{K}_j$. Thus, for every frequency bin f, our model is basically a linear mixing model with I channels and K elementary Gaussian sources $c_{k,fn}$, with structured mixing coefficients (i.e., subsets of elementary sources are mixed identically). Subsequently, we will note $\mathbf{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}([w_{fk}h_{kn}]_k)$ the covariance of $\mathbf{c}_{fn}$.

III. Methods

A. Maximization of Exact Likelihood With EM

1) Criterion: Let $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}, \mathbf{\Sigma}_\mathbf{b}\}$ be the set of all parameters, where $\mathbf{A}$ is the $I \times J \times F$ tensor with entries $a_{ij,f}$, $\mathbf{W}$ is the $F \times K$ matrix with entries w_{fk}, $\mathbf{H}$ is the $K \times N$ matrix with entries h_{kn}, and $\mathbf{\Sigma}_\mathbf{b}$ are the noise covariance parameters. Under previous assumptions, data vector $\mathbf{x}_{fn}$ has a zero-mean proper Gaussian distribution with covariance

$$\mathbf{\Sigma}_{\mathbf{x},fn}(\boldsymbol{\theta}) = \mathbf{A}_f \mathbf{\Sigma}_{\mathbf{s},fn} \mathbf{A}_f^H + \mathbf{\Sigma}_{\mathbf{b},f} \tag{11}$$

where $\mathbf{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}([p_{j,fn}]_j)$ is the covariance of $\mathbf{s}_{fn}$. ML estimation is consequently shown to amount to minimization of

$$C_1(\boldsymbol{\theta}) = \sum_{fn} \mathrm{trace}\left(\mathbf{x}_{fn}\mathbf{x}_{fn}^H \mathbf{\Sigma}^{-1}_{\mathbf{x},fn}\right) + \log\det \mathbf{\Sigma}_{\mathbf{x},fn}. \tag{12}$$

The noise covariance term $\mathbf{\Sigma}_{\mathbf{b},f}$ appears necessary so as to prevent from ill-conditioned inverses that occur if 1) $\mathrm{rank}(\mathbf{A}_f) < I$, and in particular if $I > J$, i.e., in the overdetermined case, or if 2) $\mathbf{\Sigma}_{\mathbf{s},fn}$ has more than $(J - I)$ null diagonal coefficients in the underdetermined case $(I < J)$. Case 2) might happen in regions of the time–frequency plane where sources are inactive.

For fixed f and n, the BSS problem described by (3) and (12), and the following EM algorithm, is reminiscent of works by Cardoso *et al.*, see, e.g., [27] for the square noise-free case, [17] for other cases and [18] for use in an audio setting. In these papers, a grid of the representation domain is chosen, in each cell of which the source statistics are assumed constant. This is not required in our case where we instead solve F parallel linear instantaneous mixtures tied across frequency by the source model.[1]

2) Indeterminacies: Criterion (12) suffers from obvious scale, phase and permutation indeterminacies.[2] Regarding scale and phase, let $\hat{\boldsymbol{\theta}} = \{\{\mathbf{A}_f\}_f, \{\mathbf{W}_j, \mathbf{H}_j\}_j\}$ be a minimizer of (12) and let $\{\mathbf{D}_f\}_f$ and $\{\mathbf{\Lambda}_j\}_j$ be sets of respectively *complex* and *nonnegative* diagonal matrices. Then, the set

$$\tilde{\boldsymbol{\theta}} = \left\{\left\{\mathbf{A}_f\mathbf{D}_f^{-1}\right\}_f, \left\{\mathrm{diag}\left(\left[|d_{jj,f}|^2\right]_f\right)\mathbf{W}_j\mathbf{\Lambda}_j^{-1}\right\}_j, \{\mathbf{\Lambda}_j\mathbf{H}_j\}_j\right\}$$

leads to $\mathbf{\Sigma}_{\mathbf{x},fn}(\hat{\boldsymbol{\theta}}) = \mathbf{\Sigma}_{\mathbf{x},fn}(\tilde{\boldsymbol{\theta}})$, hence same likelihood value. Similarly, permuted diagonal matrices would also leave the criterion unchanged. In practice, we remove the scale and phase ambiguity by imposing $\sum_i |a_{ij,f}|^2 = 1$ and $a_{1j,f} \in \mathbb{R}^+$ (and scaling the rows of $\mathbf{W}_j$ accordingly) and then by imposing $\sum_f w_{fk} = 1$ (and scaling the rows of $\mathbf{H}_j$ accordingly). With these conventions, the columns of $\mathbf{A}_f$ convey normalized mixing proportions between the channels, the columns of $\mathbf{W}$ convey normalized frequency shapes and all time-dependent amplitude information is relegated into $\mathbf{H}$.

3) Algorithm: We derive an EM algorithm based on *complete data* $\{\mathbf{X}, \mathbf{C}\}$, where $\mathbf{C}$ is the $K \times F \times N$ STFT tensor with coefficients $c_{k,fn}$. The complete data pdfs $\{p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11] or [29, Appendix]) and the set $\{\mathbf{R}_{\mathbf{xx},f}, \mathbf{R}_{\mathbf{xs},f}, \mathbf{R}_{\mathbf{ss},f}, \{u_{k,fn}\}_{kn}\}_f$ defined by

$$\mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \quad \mathbf{R}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{s}_{fn}^H \tag{13}$$

$$\mathbf{R}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \mathbf{s}_{fn}\mathbf{s}_{fn}^H, \quad u_{k,fn} = |c_{k,fn}|^2 \tag{14}$$

is shown to be a *natural (sufficient) statistics* [29] for this family. Thus, one iteration of EM consists of computing the expectation of the natural statistics conditionally on the current parameter estimates (E step) and of reestimating the parameters using the updated natural statistics, which amounts to maximizing the conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \int[\log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})]p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}')d\mathbf{C}$ (M step). The resulting updates are given in Algorithm 1, with more details given in Appendix A.

Algorithm 1 EM algorithm (one iteration)

- **E step**. Conditional expectations of natural statistics:

$$\widehat{\mathbf{R}}_{\mathbf{xx},f} = \mathbf{R}_{\mathbf{xx},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\mathbf{x}_{fn}^H, \tag{15}$$

$$\widehat{\mathbf{R}}_{\mathbf{xs},f} = \frac{1}{N}\sum_n \mathbf{x}_{fn}\hat{\mathbf{s}}_{fn}^H, \tag{16}$$

$$\widehat{\mathbf{R}}_{\mathbf{ss},f} = \frac{1}{N}\sum_n \hat{\mathbf{s}}_{fn}\hat{\mathbf{s}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{s},fn} - \mathbf{G}_{\mathbf{s},fn}\mathbf{A}_f\mathbf{\Sigma}_{\mathbf{s},fn} \tag{17}$$

$$\hat{u}_{k,fn} = \left[\hat{\mathbf{c}}_{fn}\hat{\mathbf{c}}_{fn}^H + \mathbf{\Sigma}_{\mathbf{c},fn} - \mathbf{G}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f\mathbf{\Sigma}_{\mathbf{c},fn}\right]_{k,k} \tag{18}$$

[1] In [17] and [27], the ML criterion can be recast as a measure of fit between observed and parameterized covariances, where the measure of deviation writes $D(\mathbf{\Sigma}_1|\mathbf{\Sigma}_2) = \mathrm{trace}(\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}) - \log\det\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1} - I$ and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are positive definite matrices of size $I \times I$ (note that the IS divergence is obtained in the special case $I = 1$). The measure is simply the KL divergence between the pdfs of two zero-mean Gaussians with covariances $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Such a formulation cannot be used in our case because $\mathbf{\Sigma}_1 = \mathbf{x}_{fn}\mathbf{x}_{fn}^H$ is not invertible for $I > 1$.

[2] There might also be other less obvious indeterminacies, such as those inherent to NMF (see, e.g., [28]), but this study is here left aside.

where
$$\hat{\mathbf{s}}_{fn} = \mathbf{G}_{\mathbf{s},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{s},fn} = \boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1} \tag{19}$$
$$\hat{\mathbf{c}}_{fn} = \mathbf{G}_{\mathbf{c},fn}\mathbf{x}_{fn}, \quad \mathbf{G}_{\mathbf{c},fn} = \boldsymbol{\Sigma}_{\mathbf{c},fn}\mathring{\mathbf{A}}_f^H\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1} \tag{20}$$
$$\boldsymbol{\Sigma}_{\mathbf{x},fn} = \mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},fn}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f} \tag{21}$$
$$\boldsymbol{\Sigma}_{\mathbf{s},fn} = \mathrm{diag}\left(\left[\sum_{k\in\mathcal{K}_j} w_{fk}h_{kn}\right]_j\right) \tag{22}$$
$$\boldsymbol{\Sigma}_{\mathbf{c},fn} = \mathrm{diag}\left([w_{fk}h_{kn}]_k\right) \tag{23}$$

and $\mathring{\mathbf{A}}_f$ is defined in Section II-C.

- **M step**. Update the parameters:

$$\mathbf{A}_f = \widehat{\mathbf{R}}_{\mathbf{xs},f}\widehat{\mathbf{R}}_{\mathbf{ss},f}^{-1}, \tag{24}$$
$$\boldsymbol{\Sigma}_{\mathbf{b},f} = \mathrm{diag}\left(\widehat{\mathbf{R}}_{\mathbf{xx},f} - \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{xs},f}^H - \widehat{\mathbf{R}}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\widehat{\mathbf{R}}_{\mathbf{ss},f}\mathbf{A}_f^H\right) \tag{25}$$
$$w_{fk} = \frac{1}{N}\sum_n \frac{\hat{u}_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f \frac{\hat{u}_{k,fn}}{w_{fk}}. \tag{26}$$

- Normalize $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-A2.

4) Implementation Issues: The computation of the source Wiener gain $\mathbf{G}_{\mathbf{s},fn}$ given by (19) requires the inversion of the $I \times I$ matrix $\boldsymbol{\Sigma}_{\mathbf{x},fn}$ at every time–frequency (TF) point. When $I > J$ (overdetermined case) it may be preferable for sake of computational efficiency to use the following alternative formulation of $\mathbf{G}_{\mathbf{s},fn}$, obtained using Woodbury matrix identity [30]

$$\mathbf{G}_{\mathbf{s},fn} = \boldsymbol{\Xi}_{\mathbf{s},fn}^{-1}\mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1} \tag{27}$$

with

$$\boldsymbol{\Xi}_{\mathbf{s},fn} = \mathbf{A}_f^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f + \boldsymbol{\Sigma}_{\mathbf{s},fn}^{-1}. \tag{28}$$

This second formulation requires the inversion of the $J \times J$ matrix $\boldsymbol{\Xi}_{\mathbf{s},fn}$ instead of the inversion of the $I \times I$ matrix $\boldsymbol{\Sigma}_{\mathbf{x},fn}$. The same idea applies to the computation of $\mathbf{G}_{\mathbf{c},fn}$, (20), if $I > K$. Thus, this second formulation may become interesting in practice only if $I > J$ and $I > K$, i.e., if $I > K$ (recall that $K \geq J$). As we only consider undetermined mixtures in the experimental part of this article ($I < J$), we turn to the original formulation given by (19). As we more precisely consider stereo mixtures, we only need inverting 2×2 matrices per TF point and our MATLAB code was efficiently vectorized so as to manipulate time–frequency matrices directly, thanks to Cramer's explicit matrix inversion formula. Note also that we only need to compute the diagonal elements of the $K \times K$ matrix in (18). Hence, the computational complexity of one EM algorithm iteration grows linearly (and not quadratically) with the number of components.

5) Linear Instantaneous Case: Linear instantaneous mixing is a special case of interest, that concerns for example "pan pot" mixing. Here, the mixing matrix is real-valued and shared between all the frequency subbands, i.e., $\mathbf{A}_f = \mathbf{A}_{\mathrm{inst}} \in \mathbb{R}^{I\times J}$. In that case, (24) needs only be replaced by

$$\mathbf{A}_{\mathrm{inst}} = \Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{xs},f}\right\}\left[\Re\left\{\sum_f \widehat{\mathbf{R}}_{\mathbf{ss},f}\right\}\right]^{-1}. \tag{29}$$

6) Simulated Annealing: If one computes $\mathbf{A}_f$ through (24), (16), (17), (19), and (21), assuming $\boldsymbol{\Sigma}_{\mathbf{b},f} = 0$, one has $\mathbf{A}_f = \mathbf{A}_f$ as result. Thus, by continuity, when the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{b},f}$ tends to zero, the resulting update rule for $\mathbf{A}_f$ tends to $\mathbf{A}_f \leftarrow \mathbf{A}_f$. Hence, the convergence of $\mathbf{A}_f$ becomes very slow for small values of $\sigma_{i,f}^2$. To overcome this difficulty and also favor global convergence, we have tested in the experimental section several simulated annealing strategies. In our framework, simulated annealing consists in setting the noise variances $\sigma_{i,f}^2$ to a common iteration-dependent value $\sigma_{i,f}^2(iter)$, initialized with an arbitrary large value $\hat{\sigma}_{i,f}^2$ and gradually decreased through iterations to a small value $\check{\sigma}_{i,f}^2$. Besides improving convergence speed, this scheme should also favor convergence to global solutions, as typical of annealing algorithms: the cost function is rendered flatter in the first iterations due to the (assumed) presence of high noise, smoothing out local minima, and is gradually brought back to its exact shape in the subsequent iterations.

7) Reconstruction of the Sources: Minimum mean square error (MMSE) estimates $\hat{\mathbf{s}}_{fn} = \mathbb{E}[\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}]$ of the source STFTs are directly retrieved using Wiener filter of (19). Time-domain sources may then be obtained through inverse STFT using an adequate overlap-add procedure with dual synthesis window (see e.g., [31]).

By conservativity of Wiener reconstruction the spatial images of the estimated sources and of the estimated noise sum up to the original mix in STFT domain, i.e., $\hat{\mathbf{A}}_f$, $\hat{\mathbf{s}}_{fn}$, and $\hat{\mathbf{b}}_{fn} = \boldsymbol{\Sigma}_{\mathbf{b},f}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}\mathbf{x}_{fn}$ satisfy (3). Thanks to linearity of the inverse-STFT, the reconstruction is conservative in the time domain as well.

B. Maximization of Individual Likelihoods With MU Rules

1) Criterion: We now consider a different approach consisting of maximizing the sum of individual channel log-likelihoods $\sum_i \log p(\mathbf{X}_i|\boldsymbol{\theta})$, hence discarding mutual information between the channels. This is equivalent to setting the off-diagonal terms of $\mathbf{x}_{fn}\mathbf{x}_{fn}^H$ and $\boldsymbol{\Sigma}_{\mathbf{x},fn}$ to zero in criterion (12), leading to minimization of cost

$$C_2(\boldsymbol{\theta}) = \sum_{ifn} d_{IS}\left(|x_{i,fn}|^2|\hat{v}_{i,fn}\right) \tag{30}$$

where $\hat{v}_{i,fn}$ is the structure defined by

$$\hat{v}_{i,fn} = \sum_j q_{ij,f}\underbrace{\sum_{k\in\mathcal{K}_j} w_{fk}h_{kn}}_{p_{j,fn}} \quad \left(+\sigma_{i,f}^2\right) \tag{31}$$

and $q_{ij,f} = |a_{ij,f}|^2$. For a fixed channel i, $\hat{v}_{i,fn}$ is basically the sum of the source variances modulated by the mixing weights. A noise variance term $\sigma_{i,f}^2$ might be considered, either fixed or to be estimated, but we will simply set it to zero as we will not here encounter the issues described in Section III-A6 about convergence of EM in noise-free observations.

Criterion (30) may also be read as the ML criterion corresponding to the model where the contributions of each component (and thus, of each source) to each channel would be different and independent realizations of the same Gaussian process, as opposed to the same realization. In other words, this

assumption amounts to changing our observation and source models given by (2) and (5) to

$$x_{i,fn} = \sum_{j=1} a_{ij,f} s^{(i)}_{j,fn} + b_{i,fn} \quad (32)$$

$$s^{(i)}_{j,fn} = \sum_{k \in \mathcal{K}_j} c^{(i)}_{k,fn} \quad \text{with} \quad c^{(i)}_{k,fn} \sim \mathcal{N}_c(0, w_{fk}h_{kn}) \quad (33)$$

and thus changing (7) to

$$s^{(i)}_{j,fn} \sim \mathcal{N}_c\left(0, \sum_{k \in \mathcal{K}_j} w_{fk}h_{kn}\right) \quad (34)$$

where $c^{(i)}_{k,fn}$ (resp. $s^{(i)}_{j,fn}$) denotes the contribution of component k (resp. source j) to channel i, and these contributions are assumed independent over channels (i.e., over i).

Our approach differs from the NTF approach of [6], [7] where the following PARAFAC structure [32] is considered

$$\hat{v}^{NTF}_{i,fn} = \sum_k q^{NTF}_{ik} w_{fk}h_{kn}. \quad (35)$$

It is only a sum of $I \times F \times N$ rank-1 tensors and amounts to assuming that $\hat{\mathbf{V}}^{NTF}_i = [\hat{v}^{NTF}_{i,fn}]_{fn}$ is a linear combination of $F \times N$ time–frequency patterns $\mathbf{w}_k h_k$, where $\mathbf{w}_k$ is column k of $\mathbf{W}$ and h_k is row k of $\mathbf{H}$. It intrinsically implies a linear instantaneous mixture and requires a postprocessing binding step in order to group the K elementary patterns into J sources, based on clustering of the ratios $\{q^{NTF}_{1k}/q^{NTF}_{2k}\}_k$ (in the stereo case). To ease comparison, our model can be rewritten as

$$\hat{v}_{i,fn} = \sum_k \overset{\circ}{q}_{ik,f} w_{fk}h_{kn} \quad (36)$$

subject to the constraint $\overset{\circ}{q}_{ik,f} = q_{ij,f}$ if and only if $k \in \mathcal{K}_j$ (with the notation introduced in Section II-C, we have also $\overset{\circ}{q}_{ik,f} = |\overset{\circ}{a}_{ik,f}|^2$). Hence, our model has the following merits with respect to (w.r.t.) the PARAFAC-NTF model: 1) it accounts for convolutive mixing by considering frequency-dependent mixing proportions ($\overset{\circ}{q}_{ik,f}$ instead of q^{NTF}_{ik}) and 2) the constraint that the K mixing proportions $\{\overset{\circ}{q}_{ik,f}\}_k$ can only take J possible values implies that the clustering of the components is taken care of within the decomposition as opposed to after the decomposition.

We have here chosen to use the IS divergence as a measure of fit in (30) because it connects with the optimal inference setting of Section III-A and because it was shown a relevant cost for factorization of audio power spectrograms [10], but other costs could be considered, such as the standard Euclidean distance and the generalized Kullback–Leibler (KL) divergence, which are the costs considered in [6] and [7].

2) Indeterminacies: Criterion (30) suffers from same scale, phase and permutations ambiguities as criterion (12), with the exception that ambiguity on the phase of $a_{ij,f}$ is now total as this parameter only appears through its squared-modulus. In the following, the scales are fixed as in Section III-A2.

3) Algorithm: We describe for the minimization of $C_2(\boldsymbol{\theta})$ an iterative MU algorithm inspired from NMF methodology [1], [33], [34]. Continual descent of the criterion under this algorithm was observed in practice. The algorithm simply consists of updating each scalar parameter θ_l by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t. this parameter, namely

$$\theta_l \leftarrow \theta_l \frac{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-}{[\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+} \quad (37)$$

where $\nabla_{\theta_l} C_2(\boldsymbol{\theta}) = [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_+ - [\nabla_{\theta_l} C_2(\boldsymbol{\theta})]_-$ and the summands are both nonnegative [10]. Not any cost function gradient may be separated in two such summands, but this is the case for the Euclidean, KL and IS costs, and more generally the β-divergence of which they are specific cases [10], [26]. This scheme automatically ensures the non-negativity of the parameter updates, provided initialization with a nonnegative value.

The resulting parameter updates are described in Algorithm 2, where "." indicates element-wise matrix operations, $\mathbf{1}_{N\times 1}$ is a N-vector of ones, $\mathbf{q}_{ij}$ is the $F \times 1$ vector $[q_{ij,f}]_f$ and $\mathbf{V}_i$ (resp. $\hat{\mathbf{V}}_i$) is the $F \times N$ matrix $[|x_{i,fn}|^2]_{fn}$ (resp. $[\hat{v}_{i,fn}]_{fn}$). Some details about the derivation of the algorithm are given in Appendix B.

Algorithm 2 MU rules (one iteration)

- Update $\mathbf{Q}$

$$\mathbf{q}_{ij} \leftarrow \mathbf{q}_{ij} . \frac{\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N\times 1}}{\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right] \mathbf{1}_{N\times 1}}. \quad (38)$$

- Update $\mathbf{W}$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j . \frac{\sum_{i=1}^I \mathrm{diag}(\mathbf{q}_{ij}) \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right) \mathbf{H}_j^T}{\sum_{i=1}^I \mathrm{diag}(\mathbf{q}_{ij}) \hat{\mathbf{V}}_i^{-1} \mathbf{H}_j^T}. \quad (39)$$

- Update $\mathbf{H}$

$$\mathbf{H}_j \leftarrow \mathbf{H}_j . \frac{\sum_{i=1}^I \left(\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j\right)^T \left(\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i\right)}{\sum_{i=1}^I \left(\mathrm{diag}(\mathbf{q}_{ij}) \mathbf{W}_j\right)^T \hat{\mathbf{V}}_i^{-1}}. \quad (40)$$

- Normalize $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ according to Section III-B2.

4) Linear Instantaneous Case: In the linear instantaneous case, when $q_{ij,f} = q_{ij}$, we obtain the following update rule for the mixing matrix coefficients:

$$q_{ij} \leftarrow q_{ij} . \frac{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-2} . \mathbf{V}_i . (\mathbf{W}_j \mathbf{H}_j)\right]}{\mathrm{sum}\left[\hat{\mathbf{V}}_i^{-1} . (\mathbf{W}_j \mathbf{H}_j)\right]} \quad (41)$$

where $\mathrm{sum}[\mathbf{M}]$ is the sum of all coefficients in $\mathbf{M}$. Then, $\mathrm{diag}(\mathbf{q}_{ij})$ needs only be replaced by q_{ij} in (39) and (40). The overall algorithm yields a specific case of PARAFAC-NTF which directly assigns the elementary components to J directions of arrival (DOA). This scheme however requires to fix in advance the partition $\{\mathcal{K}_j\}_{j=1}^J$ of $\mathcal{K} = \{1, \ldots, K\}$, i.e., assign a given number of components per DOA. In the specific linear instantaneous case, multiplicative updates for the whole matrices $\mathbf{Q}$, $\mathbf{W}$, $\mathbf{H}$ can be exhibited (instead of individual updates for q_{ij}, $\mathbf{W}_j$, $\mathbf{H}_j$), but are not given here for conciseness. They are similar in form to [33], [34] and lead to a faster MATLAB implementation.

5) Reconstruction of the Source Images: Criterion (30) being equivalent to the ML criterion under the model defined by (32) and (33), the MMSE estimate $\hat{s}^{(i)\mathrm{im}}_{j,fn} = \mathbb{E}[s^{(i)\mathrm{im}}_{j,fn} | \mathbf{x}_{fn}; \boldsymbol{\theta}]$ of the

image $s_{j,fn}^{(i)\mathrm{im}} \stackrel{\mathrm{def}}{=} a_{ij,f} s_{j,fn}^{(i)}$ of source j in channel i is computed through

$$\hat{s}_{j,fn}^{(i)\mathrm{im}} = \frac{q_{ij,f} p_{i,fn}}{\hat{v}_{i,fn}} x_{i,fn} \tag{42}$$

i.e., by Wiener filtering of each channel. A noise component (if any) can similarly be reconstructed as $\hat{b}_{i,fn} = (\sigma_{i,f}^2 / \hat{v}_{i,fn}) x_{i,fn}$. Overall the decomposition is conservative, i.e., $\sum_j \hat{s}_{j,fn}^{(i)\mathrm{im}} + \hat{b}_{i,fn} = x_{i,fn}$.

IV. Experiments

In this section, we first describe the test data and evaluation criteria, and then proceed with experiments. All the audio datasets and separation results are available from our demo web page [35]. MATLAB implementations of the proposed algorithms are also available from the authors' web pages.

A. Datasets

Four audio datasets have been considered and are described below.

- **Dataset A** consists of two synthetic stereo mixtures, one instantaneous the other convolutive, of $J = 3$ musical sources (drums, lead vocals and piano) created using 17-s excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Commons License at [36] and downsampled to 16 kHz. The mixing parameters (instantaneous mixing matrix and the convolutive filters) were taken from the 2008 Signal Separation Evaluation Campaign (SiSEC'08) "under-determined speech and music mixtures" task development datasets [37], and are described below.
- **Dataset B** consists of synthetic (instantaneous and convolutive) and live-recorded (convolutive) stereo mixtures of speech and music sources, corresponding to the test data for the 2007 Stereo Audio Source Separation Evaluation Campaign (SASSEC'07) [38]. It also coincides with development dataset `dev2` of SiSEC'08 "under-determined speech and music mixtures" task. All the mixtures are 10 s long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. The synthetic convolutive filters were generated with the Roomsim toolbox [39]. They simulate a pair of omnidirectional microphones placed 1 m apart in a room of dimensions $4.45 \times 3.55 \times 2.5$ m with reverberation time 130 ms, which correspond to the setting employed for the live-recorded mixtures. The distances between the sources and the center of the microphone pair vary between 80 cm and 1.20 m. For all mixtures the source directions of arrival vary between $-60°$ and $+60°$ with a minimal spacing of 15° (for more details see [37]).
- **Dataset C** consists of SiSEC'08 test and development datasets for task "professionally produced music recordings". The test dataset consists of two excerpts (of about 22 s long) from two different professionally produced stereo songs, namely "Que pena tanto faz" by Tamy and "Roads" by Bearlin. The development dataset consists of two other excerpts (of about 12 s long) from the same songs, with all original stereo tracks provided separately. All recordings are sampled at 44 kHz (CD quality).
- **Dataset D** consists of three excerpts of length between 25 and 50 s taken from three professionally produced stereo recordings of well-known pop and reggae songs, and downsampled to 22 kHz.

TABLE I
STFT Window Lengths Used in Different Experiments

experiment section	dataset	window length		sampling freq. (Hz)
		samples	milliseconds	
IV-D, IV-E	A	1024	64	16000
IV-F	B - inst.	1024	64	16000
	B - conv.	2048	128	16000
IV-G	C	2048	46	44100
IV-H	D	2048	93	22050

B. Source Separation Evaluation Criteria

In order to evaluate our multichannel NMF algorithms in terms of audio source separation we use the signal-to-distortion ratio (SDR) numerical criterion defined in [38], which essentially compares the reconstructed source images with the original ones. The quality of the mixing system estimates was assessed with the mixing error ratio (MER) described at [37], which is an SNR-like criterion expressed in decibels. MATLAB routines for computing these criteria were obtained from the SiSEC'08 web page [37]. These evaluation criteria can only be computed when the original source spatial images (and mixing systems) are available. When not (i.e., for datasets C and D), separation performance is assessed perceptually and informally by listening to the separated source images, available online at [35].

C. Algorithm Parameters

1) STFT Parameters: In all the experiments below we used STFTs with half-overlapping sine windows, using the STFT computation tools for MATLAB available from [37]. The choice of the STFT window size is rather important, and is a matter of compromise between 1) good frequency resolution and validity of the convolutive mixing approximation of (2) and 2) validity of the assumption of source local stationarity. We have tried various window sizes (powers of 2) for every experiment, and the most satisfactory window sizes are reported in Table I.

2) Model Order: In our case the model order parameters consist of the total number of components K and the allocation of the components among the J sources, i.e., the partition $\{\mathcal{K}_1, \ldots, \mathcal{K}_J\}$. The value of J may be set by hand to the number of instrumental sources in the recording, although, as we shall discussed later, the existence of non-point sources or the existence of sources mixed similarly might render the choice of J trickier. The choice of the number of components per source may raise more questions. As a first guess one may choose a high value, so that the model can account for all of the diversity of the source; basically, one may think of one component per note or elementary sound object. This leads to increased flexibility in the model, but, at the same time, can lead to data overfitting (in case of few data), and favors the existence of local minima,

thus rendering optimization more difficult, as well as more intensive. Interestingly, it has been noted in [10] that, given a limited number of components, IS-NMF is also able to learn higher level structures in the musical signal. One or a few components can capture a large part of one source or a subset of sources, so that a coherent sound decomposition can be achieved to some extent. A similar behavior was logically observed in our multichannel scenario, with even more success as the spatial information helps to discriminate between the sources. Hence, satisfying source separation results could be obtained with small values of K.

In the experiments of Sections IV-D and IV-E we set $\#\mathcal{K}_j = 4$; however, this has minor importance there as the aim of these experiments is merely to investigate the algorithms behavior, and not to obtain optimal source separation performance. In the experiments of Sections IV-F and IV-G, $\#\mathcal{K}_j$ is chosen by hand through trials so as to obtain most satisfying results. In the experiment of Section IV-H the total number of components is arbitrary set to either $K = 15$ or 20, depending on the recording, and the numbers of components per source $\#\mathcal{K}_j$ are chosen automatically by the initialization procedure, see below.

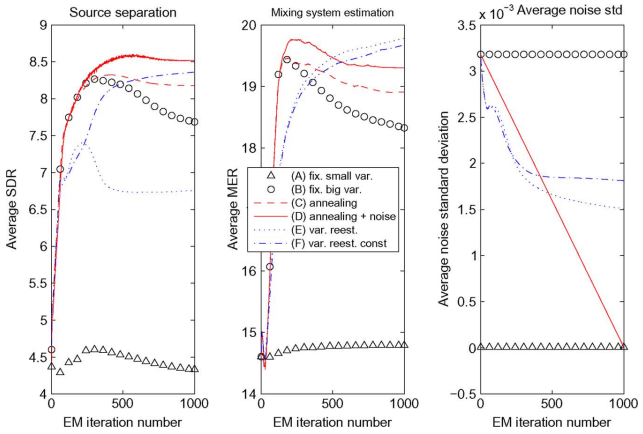


Fig. 2. EM algorithm results on convolutive mixture of dataset A, using various noise variance update schemes. (Left) Average source separation SDR. (Middle) average mixing system identification MER. (Right) average noise standard deviation. (A) Triangles: small fixed noise variance. (B) Circles: larger fixed noise variance. (C) Dashed line: annealing. (D)Solid line: annealing with noise injection. (E) Dotted line: diagonal noise covariance reestimation. (F) Dash-dotted line: isotropic noise variance reestimation.

D. Dealing With the Noise Part in the EM Algorithm

In this section, we experiment strategies for updating the noise parameters in the EM algorithm. We here arbitrarily use the convolutive mixture of dataset A and set the total number of components to $K = 12$, equally distributed between $J = 3$ sources. Our EM algorithm being sensitive to parameters initialization, we used the following *perturbed oracle* initializations so as to ensure "good" initialization: factors $\mathbf{W}$ and $\mathbf{H}$ as computed from the original sources using IS-NMF [10] and original mixing system $\mathbf{A}$, all perturbed with high level additive noise. We have tested the following noise update schemes.

- (**A**): $\mathbf{\Sigma}_{\mathbf{b},f} = \tilde{\sigma}^2\mathbf{I}_I$, with fixed $\tilde{\sigma}^2$ set to 16-bit PCM quantization noise variance.
- (**B**): $\mathbf{\Sigma}_{\mathbf{b},f} = \hat{\sigma}_f^2\mathbf{I}_I$, with fixed $\hat{\sigma}_f^2$ set to the average channel empirical variance in every frequency band divided by 100, i.e., $100\hat{\sigma}_f^2 = \sum_{in} |x_{i,fn}|^2/IN$.
- (**C**): $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ with standard deviation σ_f decreasing linearly through iterations from $\hat{\sigma}_f$ to $\tilde{\sigma}$. This is what we refer to as simulated annealing.
- (**D**): Same strategy as (C), but with adding a random noise with covariance $\mathbf{\Sigma}_{\mathbf{b},f}$ to $\mathbf{X}$ at every EM iteration. We refer to this as annealing with noise injection.
- (**E**): $\mathbf{\Sigma}_{\mathbf{b},f} = \mathrm{diag}([\sigma_{i,f}^2]_i)$ is reestimated with update (25).
- (**F**): Noise covariance is reestimated like in scheme E, but under the more constrained structure $\mathbf{\Sigma}_{\mathbf{b},f} = \sigma_f^2\mathbf{I}_I$ (isotropic noise in each subband). In that case, operator $\mathrm{diag}(\cdot)$ in (25) needs to be replaced with $\mathrm{trace}(\cdot)\mathbf{I}_I/I$.

The algorithm was run for 1000 iterations in each case and the results are presented in Fig. 2, which displays the average SDR and MER along iterations, as well as the noise standard deviations $\sigma_{i,f}$, averaged over all channels i and frequencies f. As explained in Section III-A6, we observe that with a small fixed noise variance (scheme A), the mixing parameters stagnate. With a fixed larger noise variance (scheme B) convergence starts well but then performance drops due to artificially high noise variance. Simulated annealing (scheme C) overcomes this problem, and artificial noise injection (scheme D) even improves the results (both in terms of source separation and mixing system estimation). Noise variance reestimation allows to obtain performances almost similar to annealing, but only in the case when the variance is constrained to be the same in both channels (scheme F). However, we observed that faster convergence is obtained in general using annealing with noise injection (scheme D) for similar results.

Finally, it should be noted that for the schemes with annealing (C and D) both the average SDR and MER start decreasing from about 400 iterations (for SDR) and 200 iterations (for MER). We believe this is because the final noise variance $\tilde{\sigma}^2$ (set to 16-bit PCM quantization noise variance) might be too small to account for discrepancy in the convolutive mixing equation STFT-approximation (2). Indeed, with scheme F (constrained reestimated variance) the average noise standard deviation seem to be converging to a value in the range of 0.002 (see right plot of Fig. 2), which is much larger than $\tilde{\sigma}$. Thus, if computation time is not an issue, scheme F can be considered the most advantageous because this is the only scheme to systematically increase both the average SDR and MER at every iteration and it allows to adjust a suitable noise level adaptively. However, as we want to keep the number of iterations low (e.g., 300–500) for sake of short computation time, we will resort to scheme D in the following experiments.

E. Convergence and Separation Performance

In this experiment we wish to check consistency of optimization of the proposed criteria with respect to source separation performance improvement, in the least as measured by the SDR. We used both mixtures of dataset A (instantaneous and convolutive) and ran 1000 iterations of both algorithms (EM and MU) from ten different perturbed oracle initializations, obtained as in previous section. Again we used $K = 12$ components, equally split into $J = 3$ sources. Figs. 3 and 4 report results for the instantaneous and convolutive mixtures, respectively. Plots on

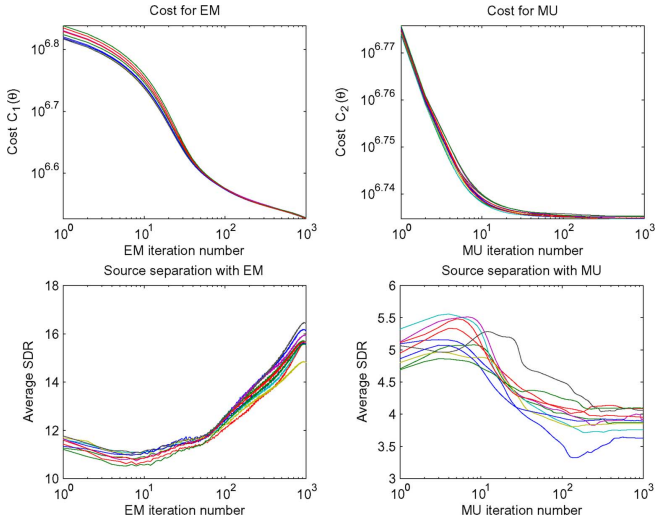


Fig. 3. Ten runs of EM and MU from ten perturbed oracle initializations using instantaneous mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

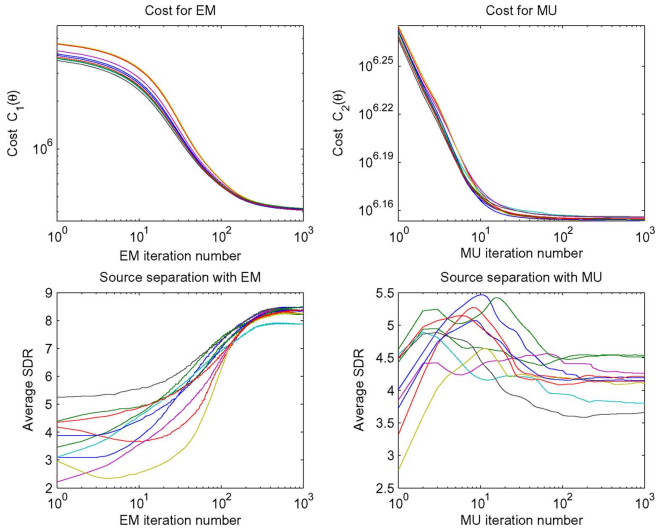


Fig. 4. Ten runs of EM and MU from ten perturbed oracle initializations using convolutive mixture of dataset A. (Top) cost functions. (Bottom) average SDRs.

top row display in log-scale the cost functions $C_1(\boldsymbol{\theta})$ and $C_2(\boldsymbol{\theta})$ w.r.t. iterations for all ten runs. Note that cost $C_1(\boldsymbol{\theta})$ is not positive in general, see (12), so that we have added a common large constant value to all curves so as to ensure positivity, and to be able plotting cost value in the logarithmic scale. Plots on bottom row display the average SDRs.

The results show that maximization of the joint likelihood with the EM algorithm leads to consistent improvement of source separation performance in term of SDR, in the sense that final average SDR values are higher than values at initialization. This is not the case with MU, which results in nearly every case in worsening the SDR values obtained from oracle initialization. This is undoubtedly a consequence of discarding mutual information between the channels.

As for computational loads, our MATLAB implementation of EM (resp. MU) algorithm takes about 80 min (resp. 20 min) per 1000 iterations, for this particular experiment with 17-s stereo mixture (sampled at 16 kHz), $J = 3$ sources, and $K = 12$ components.

F. Blind Separation of Under-Determined Speech and Music Mixtures

In this section, we compare our algorithms with the methods that achieved competitive results at the SASSEC'07 evaluation campaign for the tasks of underdetermined mixtures of respectively speech and music signals, in both instantaneous and convolutive cases. We used the same data and evaluation criteria as in the campaign. More precisely, our algorithms are compared in the instantaneous case to the method of Vincent [40], based on source STFT reconstruction using a minimum l_0 norm constraint given a mixing matrix estimate obtained with the method of Arberet *et al.* [41]. In the convolutive case, our algorithms are compared to the method of Sawada, based on frequency-dependent complex-valued mixing matrices estimation [42], and *a posteriori* grouping relying on temporal correlations between sources in different frequency bins [20]. We used the outputs of these methods to initialize our own algorithms. In the linear instantaneous case, we were given MATLAB implementations of [40] and [41]. In the convolutive case, we simply downloaded the source image estimates from the SASSEC'07 web page [43]. In both cases we built initializations of $\mathbf{W}$ and $\mathbf{H}$ based on NMF of the source spectrogram estimates.[3]

We have found satisfactory separation results through trials using $\#\mathcal{K}_j = 4$ components for musical sources and $\#\mathcal{K}_j = 10$ components for speech sources. More components seem to be needed for speech so as to account for its higher variability (e.g., vibrato). The EM and MU algorithms were run for 500 iterations, final source separation SDR results together with reference methods results are displayed in Table II.[4] The EM method yields a significant separation improvement for all linear instantaneous mixtures. Improvement is also obtained in the convolutive case for most source estimates, but is less significant in terms of SDRs. However, and maybe most importantly, we believe our source estimates to be generally more pleasant to listen to. Indeed, one drawback of sparsity-based, nonlinear source reconstruction is musical noise, originating from unnatural, isolated time-frequency atoms scattered over the time–frequency plane. In contrast, our Wiener source estimates, obtained as a linear combination of data in each TF cell, appear to be less prone to such artifacts as can be listened to at demo web page [35]. We have entered our EM algorithm to the "under-determined speech and music mixtures" task of SiSEC'08 for instantaneous mixtures, and our results can be compared to other

[3]However, in that case we used KL-NMF instead of IS-NMF, not to fit the lower-energy residual artifacts and interferences, to which IS-NMF might be overly sensitive as a consequence of its scale-invariance. This seemed to lead to better initializations indeed.

[4]The reference algorithms performances in Table II do not always coincide with those given on the SASSEC'07 web page [43]. In the instantaneous case, this is because we have not used the exact same implementation of the l_0 minimization algorithm [40] that was used for SASSEC. In the convolutive case, this is because we have removed the dc component from all speech signals (including reference, source image estimates, and mixtures) using high-pass filtering, in order to avoid numerical instabilities.

TABLE II
SOURCE SEPARATION RESULTS FOR SASSEC DATA IN TERMS OF SDR (dB)

Linear instantaneous mixtures

	female4				male4				nodrums			wdrums			average
	s1	s2	s3	s4	s1	s2	s3	s4	s1	s2	s3	s1	s2	s3	
l_0 min.	12.6	6.1	4.7	7.3	15.6	2.7	5.3	6.9	21.2	1.7	15.8	-0.5	3.1	28.4	9.6
EM	14.2	7.8	5.9	8.6	16.8	3.5	8.2	9.6	27.1	7.6	21.4	0.9	4.6	29.8	12.3
MU	3.9	0.9	0.1	2.2	8.6	-0.7	2.8	2.9	8.8	-6.4	3.3	10.0	2.9	19.3	4.4

Synthetic convolutive mixtures (1m)

Sawada	5.2	5.3	3.2	2.6	4.5	0.6	4.9	2.3	3.0	1.0	-1.6	4.4	-12.7	0.6	1.3
EM	7.7	6.4	4.1	3.2	6.2	0.4	5.5	2.7	4.1	1.0	-1.8	3.9	-12.4	1.3	1.9
MU	5.2	3.3	2.7	1.4	3.4	-0.9	3.0	1.7	2.8	1.0	-2.0	5.9	-10.9	1.9	1.1

Live-recorded convolutive mixtures (1m)

Sawada	4.1	3.8	6.0	3.3	3.0	1.6	4.8	2.4	4.1	5.1	-3.8	4.1	4.5	6.0	3.5
EM	5.3	3.6	7.2	4.3	3.5	2.1	5.6	3.1	4.5	7.3	-4.5	4.9	5.5	8.0	4.3
MU	1.6	-0.2	4.3	1.8	1.1	0.0	2.8	2.1	3.9	3.6	-4.9	4.1	4.5	7.5	2.4

methods in [44], and online at [45]. Note that among the ten algorithms participating in this task our algorithm outperformed all the other competing methods by at least 1 dB for all separation measures (SDR, ISR, SIR, and SAR), see [44, Table 2].

G. Supervised Separation of Professionally Produced Music Recordings

We here apply our algorithms to the separation of the professionally produced music recordings of dataset B. This is a supervised setting in the sense that training data is available to learn the source spectral patterns $\mathbf{W}$ and filters. The following procedure is used.

- Learn mixing parameters $\{a_{ij,f}^{tr}\}_{if}$, spectral patterns $\mathbf{W}_j^{tr}$, and activation coefficients $\mathbf{H}_j^{tr}$ from available training signal images of source j (using 200 iterations of EM/MU); discard $\mathbf{H}_j^{tr}$.
- Clamp $\mathbf{A}$ and $\mathbf{W}$ to their trained values $\mathbf{A}^{tr}$ and $\mathbf{W}^{tr}$ and reestimate activation coefficients $\mathbf{H}$ from test data $\mathbf{X}$ (using 200 iterations of EM/MU).
- Reconstruct source image estimates from $\mathbf{A}^{tr}$, $\mathbf{W}^{tr}$ and $\mathbf{H}$.

Except for the training of mixing coefficient, the procedure is similar in spirit to supervised single-channel separation schemes proposed, e.g., in [9] and [46].

One important issue with professionally produced modern music mixtures is that they do not always comply with the mixing assumptions of (3). This might be due to nonlinear sound effects (e.g., dynamic range compression), to reverberation times longer than the analysis window length, and maybe most importantly to when the *point source* assumption does not hold anymore, i.e., when the channels of a stereo instrumental track cannot be represented as a convolution of the *same* source signal. The latter situation might happen when a sufficiently voluminous musical instrument (e.g., piano, drums, acoustic guitar) is recorded with several microphones placed close to the instrument. As such, the guitar track of the "Que pena tanto faz" song from dataset C is a non-point source image. Such tracks may be modeled as a sum of several point sources, with different mixing filters.

For the "Que pena tanto faz" song, the vocal part is modeled as an instantaneously mixed point source image with $\#\mathcal{K}_1 = 8$ components while the guitar part is modeled as a sum of three convolutively mixed point source images, each modeled with $\#\mathcal{K}_2 = \#\mathcal{K}_3 = \#\mathcal{K}_4 = 3$ components. For the "Roads" song, the bass and vocals parts are each modeled as instantaneously mixed point source images with six components, the piano part is modeled as a convolutive point source image with six components and finally, the residual background music (sum of remaining tracks) is modeled as a sum of three convolutive point source images with four components. The audio results, available at [35], tend to show better performance of the EM approach, especially on the "Roads" song. Our results can be compared to those of the other methods that entered the "professionally produced music recordings" task of SiSEC'08 in [44], and online at [47].

H. Blind Separation of Professionally Produced Music Recordings

In the last experiment, we have tested the EM and MU algorithms for the separation of professionally produced music recordings (commercial CD excerpts) in a fully unsupervised (blind) setting. We used the following parameter initialization procedure, inspired from [48], which yielded satisfactory results.

- Stack left and right mixture STFTs so as to create a $2F \times N$ complex-valued matrix $\mathbf{X}_{2\text{ch}} = [\mathbf{X}_L^T \; \mathbf{X}_R^T]^T$.
- Produce a K-components IS-NMF decomposition of $|\mathbf{X}_{2\text{ch}}|^2 \approx \mathbf{W}_{2\text{ch}}\mathbf{H}_{2\text{ch}}$.
- Initialize $\mathbf{W}$ as the average of $\mathbf{W}_L$ and $\mathbf{W}_R$, where $\mathbf{W}_{2\text{ch}} = [\mathbf{W}_L^T \; \mathbf{W}_R^T]^T$. Initialize $\mathbf{H} = \mathbf{H}_{2\text{ch}}$.
- Reconstruct K components $\widehat{\mathbf{C}}_{2\text{ch},k} = [\widehat{\mathbf{C}}_{L,k}^T \; \widehat{\mathbf{C}}_{R,k}^T]^T$ from $\mathbf{X}_{2\text{ch}}$, $\mathbf{W}_{2\text{ch}}$, and $\mathbf{H}_{2\text{ch}}$, using single-channel Wiener filtering (see, e.g., [10]). Produce K ad-hoc left and right component-dependent mixing filters estimates by averaging $\widehat{\mathbf{C}}_{L,k}/\Phi$ and $\widehat{\mathbf{C}}_{R,k}/\Phi$ over frames, with $\Phi = \arg(\widehat{\mathbf{C}}_{L,k})$, and normalizing according to Section III-A2. Cluster the resulting filter estimates with the K-means algorithm, whose output can be used to

define the partition $\{\mathcal{K}_j\}_{j=1}^{J}$ (using cluster indices) and a mixing system estimate $\hat{\mathbf{A}}$ (using cluster centroids).

Depending on the recording we set the number of sources J to 3 or 4 and used a total of $K = 15$ to 20 components. The EM and MU algorithms were run for 300 iterations in every case. On these specific examples the superiority of the EM method w.r.t. the MU method is not as clear as with previous datasets. A likely reason is the existence of nonpoint sources breaking the validity of mixing assumptions (2). In such precise cases, choosing not to exploit inter-channel dependencies might be better, because our model of these dependencies is now wrong. Looking for suitable probabilistic models of nonpoint sources is a new and interesting research direction.

In some cases the source image estimates contain several musical instruments and some musical instruments are spread over several source images. Besides poor initialization, this can be explained by 1) sources mixed similarly (e.g, same directions of arrival), and thus impossible to separate in our fully blind setting, 2) nonpoint sources, not well represented by our model and thus split into different source image estimates.

One way to possibly refine separation results is to reconstruct individual stereo component images (i.e., obtained via Wiener filtering (20) in case of EM method, or via (42) by replacing $p_{i,fn}$ with $w_{fk}h_{kn}$ in case of MU method), and manually group them through listening, either to separate sources mixed similarly, or to reconstruct multidirectional sound sources that better match our understanding/perception of a single source.

Finally, to show the potential of our source separation approach for music remixing, we have created some remixes using the blindly separated source images and/or the manually regrouped ones. The remixes were created in Audacity [49] by simply re-panning the source image estimates between left and right channels and by changing their gains. The audio results can be listened to at [35].

V. Conclusion

We have presented a general probabilistic framework for the representation of multichannel audio, under possibly underdetermined and noisy convolutive mixing assumptions. We have introduced two inference methods: an EM algorithm for the maximization of the channels joint log-likelihood and a MU algorithm for the maximization of the sum of individual channel log-likelihoods. The complexity of these algorithms grows linearly with the number of model components, and make them thus suitable to real-world audio mixtures with any number of sources. The corresponding CPU computational loads are in the order of a few hours for a song, which may be considered reasonable for applications such as remixing, where real-time is not an issue.

We have applied our decomposition algorithms to stereo source separation in various settings, covering blind and supervised separation, music and speech sources, synthetic instantaneous and convolutive mixtures, as well as professionally produced music recordings.

The EM algorithm was shown to outperform state-of-the-art methods, given appropriate initializations. Both our methods have indeed been found sensitive to parameter initialization, but we have come up with two satisfying initialization schemes. The first one, described in Section IV-F, consists in using the output of a different separation algorithm. We show that our EM algorithm improves the separation results in almost all cases. The second scheme, described in Section IV-H, consists in a single-channel NMF decomposition followed by K-means filters clustering. Our experiments tend to show that the NMF model is more suitable to music than speech: music sources can be represented by a small number of components to attain good separation performance, and informal listening indicates better separation of music signals.

Given that the mixed signals follow the mixing and point source assumptions inherent to (2), the EM method gives better separation results than the MU method, because between-channel dependencies are optimally exploited. However, the performance of the EM method may significantly drop when these assumptions are not verified. In contrast, we have observed that the MU method, which relies on a weaker model of between-channel dependencies, yields more even results overall and higher robustness to model discrepancies (that may for example occur in professionally produced recordings).

Let us now mention some further research directions. Algorithms faster than EM (both in terms of convergence rate and CPU time per iteration) would be desirable for optimization of the joint likelihood (12). As such, we envisage turning to Newton gradient optimization, as inspired from [50]. Mixed strategies could also be considered, consisting of employing EM in the first few iterations to get a sharp decrease of the likelihood before switching to faster gradient search once in the neighborhood of a solution.

Bayesian extensions of our algorithm are readily available, using for example priors favoring sparse activation coefficients h_k, or even sparse filters $q_{ij,f}$ like in [51]. Minor changes are required in the MU rules so as to yield algorithms for maximum *a posteriori* (MAP) estimation. More complex priors structure can also be envisaged within the EM method, such as Markov chains favoring smoothness of the activation coefficients $\mathbf{H}$ [10].

An important perspective is automatic order selection. In our case, that concerns the total number of components K, the number of sources J and the partition $\{\mathcal{K}_j\}_j$. Regarding the total number of components K, ideas from *automatic relevance determination* can be explored, see [52] in a NMF setting. Then the problem of partitioning can be viewed as a clustering problem with unknown number of clusters J, which is a typical machine learning problem.

While we have assessed the validity of our model in terms of source separation, our decompositions more generally provide a data-driven object-based representation of multichannel audio that could be relevant to other problems such as audio transcription, indexing and object-based coding. As such, it will be interesting to investigate the semantics revealed by the learnt spectral patterns $\mathbf{W}$ and activation coefficients $\mathbf{H}$.

Finally, as discussed in Section IV-H, new models should be considered for professionally produced music recordings, dealing with nonpoint sources, nonlinear sound effects, such as dynamic range compression, and long reverberation times.

APPENDIX A

APPENDIX A
EM ALGORITHM DERIVATION OUTLINE

The complete data minus log-likelihood can be written as

$$\begin{aligned}
&-\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) \\
&\quad = -\log p(\mathbf{X}|\mathbf{C},\boldsymbol{\theta}) - \log p(\mathbf{C}|\boldsymbol{\theta}) \\
&\quad \stackrel{\mathrm{c}}{=} \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + (\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})^H\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}(\mathbf{x}_{fn}-\mathbf{A}_f\mathbf{s}_{fn})\right] \\
&\qquad + \sum_k\sum_{fn}\left[\log(h_{k,n}w_{k,f}) + \frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\quad = \sum_{fn}\left[\log|\boldsymbol{\Sigma}_{\mathbf{b},f}| + \sum_k\log(h_{k,n}w_{k,f}) + \sum_k\frac{|c_{k,fn}|^2}{h_{k,n}w_{k,f}}\right] \\
&\qquad + N\sum_f \mathrm{trace}\Big[\boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xx},f} - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&\qquad\qquad - \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}^{-1}\mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\Big]
\end{aligned} \tag{43}$$

with $\mathbf{R}_{\mathbf{xx},f}$, $\mathbf{R}_{\mathbf{xs},f}$, $\mathbf{R}_{\mathbf{ss},f}$, and $u_{k,fn}$ defined by (13) and (14). Thus, we have shown that the complete data log-likelihood can be represented in the following form:

$$\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta}) = \langle\eta(\boldsymbol{\theta}),\mathbf{T}(\mathbf{X},\mathbf{C})\rangle + \nu(\boldsymbol{\theta}) \tag{44}$$

where $\mathbf{T}(\mathbf{X},\mathbf{C})$ is a vector of all scalar elements of $\mathbf{t}(\mathbf{X},\mathbf{C}) \triangleq \{\mathbf{R}_{\mathbf{xx},f},\mathbf{R}_{\mathbf{xs},f},\mathbf{R}_{\mathbf{ss},f},\{u_{k,fn}\}_{kn}\}_f$, and $\eta(\boldsymbol{\theta})$ and $\nu(\boldsymbol{\theta})$ are some vector and scalar functions of parameters. That means that the complete data pdfs $\{p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ form an *exponential family* (see, e.g., [11], [29]) and complete data statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$ is a *natural (sufficient) statistics* [11], [29] for this family. To derive an EM algorithm in this special case one needs to 1) solve complete data ML criterion (thanks to (44) this solution can be always expressed as a function of natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$), and 2) replace in this solution $\mathbf{t}(\mathbf{X},\mathbf{C})$ by its conditional expectation $\hat{\mathbf{t}}(\mathbf{X},\boldsymbol{\theta}') \triangleq \int\mathbf{t}(\mathbf{X},\mathbf{C})p(\mathbf{C}|\mathbf{X},\boldsymbol{\theta}')d\mathbf{C}$ using model $\boldsymbol{\theta}'$ estimated at the previous step of EM.

To solve the complete data ML criterion, we first compute the derivatives of $\log p(\mathbf{X},\mathbf{C}|\boldsymbol{\theta})$ (43) w.r.t. model parameters $\boldsymbol{\theta}$ (see [53] for issues regarding derivation w.r.t. complex-valued parameters), set them to zero and solve the corresponding equations (subject to the constraint that $\boldsymbol{\Sigma}_{\mathbf{b},f}$ is diagonal), and we have:[5]

$$\mathbf{A}_f = \mathbf{R}_{\mathbf{xs},f}\mathbf{R}_{\mathbf{ss},f}^{-1} \tag{45}$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{b},f} = \mathrm{diag}\big(&\mathbf{R}_{\mathbf{xx},f} - \mathbf{A}_f\mathbf{R}_{\mathbf{xs},f}^H \\
&- \mathbf{R}_{\mathbf{xs},f}\mathbf{A}_f^H + \mathbf{A}_f\mathbf{R}_{\mathbf{ss},f}\mathbf{A}_f^H\big)
\end{aligned} \tag{46}$$

$$w_{fk} = \frac{1}{N}\sum_n\frac{u_{k,fn}}{h_{kn}}, \quad h_{kn} = \frac{1}{F}\sum_f\frac{u_{k,fn}}{w_{fk}}. \tag{47}$$

[5]Bayesian MAP estimation can be carried out instead of ML by simply adding a prior term $-\log p(\boldsymbol{\theta})$ to the right part of (43) and solving the corresponding complete data MAP criterion.

Our EM algorithm is strictly speaking only a *Generalized* EM algorithm [54] because it only ensures $Q(\boldsymbol{\theta}^{m+1}|\boldsymbol{\theta}^m) \geq Q(\boldsymbol{\theta}^m|\boldsymbol{\theta}^m)$. Indeed, in (47) $\mathbf{W}$ is still a function of $\mathbf{H}$, and reversely, $\mathbf{H}$ is a function of $\mathbf{W}$.

To finish derivation of our EM algorithm we need to compute conditional expectation of the natural statistics $\mathbf{t}(\mathbf{X},\mathbf{C})$. It can be shown that given $\mathbf{x}_{fn}$ the source vector $\mathbf{s}_{fn}$ is a proper Gaussian random vector, i.e.,

$$p(\mathbf{s}_{fn}|\mathbf{x}_{fn};\boldsymbol{\theta}) = N_c\left(\mathbf{s}_{fn};\hat{\mathbf{s}}_{fn},\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\mathrm{post}}\right) \tag{48}$$

with mean vector $\hat{\mathbf{s}}_{fn}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\mathrm{post}}$ as follows:

$$\begin{aligned}
\hat{\mathbf{s}}_{fn} &= \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{x}_{fn}, \\
\boldsymbol{\Sigma}_{\mathbf{s},fn}^{\mathrm{post}} &= \boldsymbol{\Sigma}_{\mathbf{s},f} - \boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H\left(\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}\mathbf{A}_f^H + \boldsymbol{\Sigma}_{\mathbf{b},f}\right)^{-1}\mathbf{A}_f\boldsymbol{\Sigma}_{\mathbf{s},f}.
\end{aligned}$$

Computing conditional expectations of $\mathbf{R}_{\mathbf{xs},f}$ and $\mathbf{R}_{\mathbf{ss},f}$ using (48) leads to (16) and (17) of EM Algorithm 1. Very similar derivations can be done to compute the conditional expectations of $u_{k,fn}$. To that matter, one only needs to compute the posterior distribution of $\mathbf{c}_{fn}$ instead of $\mathbf{s}_{fn}$, using mixing equation (10) instead of mixing equation (3).

APPENDIX B
MU ALGORITHM DERIVATION OUTLINE

Let θ be a scalar parameter of the set $\{\mathbf{Q},\mathbf{W},\mathbf{H}\}$. The derivative of cost $C_2(\boldsymbol{\theta})$, given by (30), w.r.t. θ simply writes

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{ifn}(\nabla_\theta\hat{v}_{i,fn})d'_{IS}(v_{i,fn}|\hat{v}_{i,fn}) \tag{49}$$

where $d'_{IS}(x|y)$ is the derivative of $d_{IS}(x|y)$ w.r.t. y given by

$$d'_{IS}(x|y) = \frac{1}{y} - \frac{x}{y^2}. \tag{50}$$

Using (49), we obtain the following derivatives:

$$\begin{aligned}
\nabla_{q_{ij,f}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{n=1}^N p_{j,fn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{w_{jfk}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{n=1}^N q_{ij,f}h_{j,kn}d'(v_{i,fn}|\hat{v}_{i,fn}) \\
\nabla_{h_{jkn}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\sum_{f=1}^F q_{ij,f}w_{j,fk}d'(v_{i,fn}|\hat{v}_{i,fn})
\end{aligned}$$

which can be written in the following matrix forms:

$$\begin{aligned}
\nabla_{\mathbf{q}_{ij}}D(\mathbf{V}|\hat{\mathbf{V}}) &= \left(\hat{\mathbf{V}}_i^{\cdot-1}\mathbf{P}_j - \hat{\mathbf{V}}_i^{\cdot-2}.\mathbf{V}_i.\mathbf{P}_j\right)\mathbf{1}_{N\times1} \\
\nabla_{\mathbf{W}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I\mathrm{diag}(\mathbf{q}_{ij})\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right)\mathbf{H}_j^T \\
\nabla_{\mathbf{H}_j}D(\mathbf{V}|\hat{\mathbf{V}}) &= \sum_{i=1}^I(\mathrm{diag}(\mathbf{q}_{ij})\mathbf{W}_j)^T\left(\hat{\mathbf{V}}_i^{\cdot-2}.(\hat{\mathbf{V}}_i - \mathbf{V}_i)\right).
\end{aligned}$$

Hence, the update rules given in Algorithm 2, following the multiplicative update strategy described in Section III-B3.

ACKNOWLEDGMENT

The authors would like to thank S. Arberet for kindly sharing his implementation of DEMIX algorithm [41], all the organizers of SiSEC'08 for well-prepared evaluation campaign, as well as the anonymous reviewers for their valuable comments.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2003, pp. 177–180.

[3] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, 2007, pp. 65–68.

[4] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] P. Smaragdis, "Convolutive speech bases and their application to speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[6] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, Mar. 2006, pp. 666–673.

[7] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. Irish Signals Syst. Conf.*, Dublin, Ireland, Sep. 2005, pp. 8–12.

[8] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[9] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, Hong Kong, 2003, pp. 613–616.

[10] 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, Mar. 2009.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[12] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'97)*, Apr. 1997, pp. 3617–3620.

[13] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, pp. 803–851, 1999.

[14] H. Attias, "New EM algorithms for source separation and deconvolution," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, pp. 297–300.

[15] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," in *Proc. Interspeech'08*, 2008, pp. 419–422.

[16] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'09)*, 2009, pp. 751–758.

[17] J.-F. Cardoso, H. Snoussi, J. Delabrouille, and G. Patanchon, "Blind separation of noisy Gaussian stationary sources. Application to cosmic microwave background imaging," in *Proc. 11th Eur. Signal Process. Conf. (EUSIPCO'02)*, 2002, pp. 561–564.

[18] C. Févotte and J.-F. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, Mohonk, NY, Oct. 2005, pp. 78–81.

[19] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'97)*, New Paltz, NY, Oct. 1997, 4 pp..

[20] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE Int. Symp. Circuits Syst. (ISCAS'07)*, May 27–30, 2007, pp. 3247–3250.

[21] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. Neural Inf. Process. Syst. (NIPS 19)*, 2007, pp. 953–960.

[22] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, Oct. 2007, pp. 147–150.

[23] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, Apr. 2009, pp. 3137–3140.

[24] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1293–1302, Jul. 1993.

[25] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Symp. Music Inf. Retrieval (ISMIR'04)*, Oct. 2004, pp. 318–325.

[26] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Proc. 6th Int. Conf. Ind. Compon. Anal. Blind Signal Separation (ICA'06)*, Charleston, SC, 2006, pp. 32–39.

[27] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[28] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: On the uniqueness of NMF," *Comput. Intell. Neurosci.*, vol. 2008, pp. 1–9, 2008.

[29] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[31] M. Goodwin, "The STFT, sinusoidal models, and speech modification," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York: Springer, 2008, ch. 12, pp. 229–258.

[32] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics Intell. Lab. Syst.*, vol. 38, no. 2, pp. 149–171, Oct. 1997.

[33] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recognition Lett.*, vol. 22, no. 12, pp. 1255–1261, 2001.

[34] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd Int. Conf. Mach. Learn.*, Bonn, Germany, 2005, pp. 792–799, ACM.

[35] Example Web Page [Online]. Available: http://www.irisa.fr/metiss/ozerov/demos.html#ieee_taslp09

[36] S. Hurley, Call for Remixes: Shannon Hurley [Online]. Available: http://www.ccmixter.org/shannon-hurley

[37] in *Signal Separation Evaluation Campaign (SiSEC 2008)*, 2008 [Online]. Available: http://www.sisec.wiki.irisa.fr

[38] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 552–559, Springer.

[39] D. Campbell, Roomsim Toolbox [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/5184

[40] E. Vincent, "Complex nonconvex lp norm minimization for underdetermined source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07)*, 2007, pp. 430–437.

[41] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'06)*, 2006, pp. 536–543.

[42] P. D. O'Grady and P. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA)*, 2004, pp. 428–435.

[43] in *Stereo Audio Source Separation Evaluation Campaign (SASSEC 2007)*, 2007 [Online]. Available: http://www.sassec.gforge.inria.fr/

[44] E. Vincent, S. Araki, and P. Bofilld, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'09)*, 2009, pp. 734–741 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_ICA09.pdf

[45] in *SiSEC 2008 Under-Determined Speech and Music Mixtures Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_underdetermined/

[46] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 414–421.

[47] in *SiSEC 2008 Professionally Produced Music Recordings Task Results*, 2008 [Online]. Available: http://www.sassec.gforge.inria.fr/SiSEC_professional/

[48] S. Winter, H. Sawada, S. Araki, and S. Makino, "Hierarchical clustering applied to overcomplete BSS for convolutive mixtures," in *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Process. (SAPA 2004)*, Oct. 2004, pp. 652–660.

[49] "Audacity: The Free, Cross-Platform Sound Editor," [Online]. Available: http://www.audacity.sourceforge.net/

[50] J.-F. Cardoso and M. Martin, "A flexible component model for precision ICA," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'07)*, London, U.K., Sep. 2007, pp. 1–8.

[51] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 839–847, Mar. 2006.

[52] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop Signal Process. Adaptative Sparse Structured Representations (SPARS'05)*, Saint-Malo, France, Apr. 2009.

[53] A. van den Bos, "Complex gradient and Hessian," *IEE Proc. Vision, Image, Signal Process.*, vol. 141, pp. 380–382, 1994.

[54] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.

Alexey Ozerov received the M.Sc. degree in mathematics from the Saint-Petersburg State University, Saint-Petersburg, Russia, in 1999, the M.Sc. degree in applied mathematics from the University of Bordeaux 1, Bordeaux, France, in 2003, and the Ph.D. degree in signal processing from the University of Rennes 1, Rennes, France, in 2006.

He worked towards the Ph.D. degree from 2003 to 2006 in the labs of France Telecom R&D and in collaboration with the IRISA institute. Earlier, From 1999 to 2002, he worked at Terayon Communication Systems as a R&D Software Engineer, first in Saint-Petersburg and then in Prague, Czech Republic. He was for one year (2007) in the Sound and Image Processing Lab at KTH (Royal Institute of Technology), Stockholm, Sweden, and for one year and half (2008-2009) with TELECOM ParisTech / CNRS LTCI—Signal and Image Processing (TSI) Department. Currently, he is with the METISS team of IRISA/INRIA—Rennes as a Postdoctoral Researcher. His research interests include audio source separation, source coding, and automatic speech recognition.

Cédric Févotte received the State Engineering degree and the M.Sc. degree in control and computer science from École Centrale de Nantes, Nantes, France, in 2000, and the Ph.D. degree in 2003 from the University of Nantes.

From 2003 to 2006, he was a Research Associate with the Signal Processing Laboratory at the University of Cambridge, Cambridge, U.K., working on Bayesian approaches to audio signal processing tasks such as audio source separation, denoising, and feature extraction. From May 2006 to February 2007, he was a Research Engineer with the start-up company Mist-Technologies (Paris), working on mono/stereo to 5.1 surround sound upmix solutions. In March 2007, he joined CNRS LTCI/Telecom ParisTech, first as a Research Associate and then as a CNRS tenured Research Scientist in November 2007. His research interests generally concern statistical signal processing and unsupervised machine learning with audio applications.

(Dikmen & Févotte, *IEEE TSP*, 2012)

Maximum Marginal Likelihood Estimation for Nonnegative Dictionary Learning in the Gamma-Poisson Model

Onur Dikmen, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***Abstract*—In this paper we describe an alternative to standard nonnegative matrix factorization (NMF) for nonnegative dictionary learning, i.e., the task of learning a dictionary with nonnegative values from nonnegative data, under the assumption of nonnegative expansion coefficients. A popular cost function used for NMF is the Kullback-Leibler divergence, which underlies a Poisson observation model. NMF can thus be considered as maximization of the *joint likelihood* of the dictionary and the expansion coefficients. This approach lacks optimality because the number of parameters (which include the expansion coefficients) grows with the number of observations. In this paper we describe variational Bayes and Monte-Carlo EM algorithms for optimization of the *marginal likelihood*, i.e., the likelihood of the dictionary where the expansion coefficients have been integrated out (given a Gamma prior). We compare the output of both maximum joint likelihood estimation (i.e., standard NMF) and maximum marginal likelihood estimation (MMLE) on real and synthetical datasets. In particular we present face reconstruction results on CBCL dataset and text retrieval results over the musiXmatch dataset, a collection of word counts in song lyrics. The MMLE approach is shown to prevent overfitting by automatically pruning out irrelevant dictionary columns, i.e., embedding automatic model order selection.**

***Index Terms*—Automatic relevance determination, model order selection, Monte Carlo EM, nonnegative matrix factorization, sparse coding, variational EM.**

I. Introduction

NONNEGATIVE matrix factorization (NMF) [1] is a popular method for nonnegative dictionary learning based on matrix decomposition. The goal is to approximate a $F \times N$ nonnegative matrix $\boldsymbol{V}$ as the product of two nonnegative matrices, $\boldsymbol{W}$ (dictionary) and $\boldsymbol{H}$ (expansion coefficients), of sizes $F \times K$ and $K \times N$, respectively. These two matrices can be estimated via minimizing a measure of fit between $\boldsymbol{V}$ and $\boldsymbol{WH}$. One such popular measure is the (generalized) Kullback-Leibler (KL) divergence

$$D_{KL}(\boldsymbol{A}|\boldsymbol{B}) = \sum_{f=1}^{F}\sum_{n=1}^{N}\left(a_{fn}\log\frac{a_{fn}}{b_{fn}} - a_{fn} + b_{fn}\right)$$

which is always nonnegative, convex with respect to (w.r.t) each factor (but not w.r.t both factors jointly) and is equal to zero if and only if $\boldsymbol{A} = \boldsymbol{B}$. Minimization of the fit w.r.t the factors can be carried out with a fast, iterative algorithm based on multiplicative updates as described in [1] and based on the Richardson-Lucy algorithm [2], [3]. This approach also coincides with maximum joint likelihood estimation of $\boldsymbol{W}$ and $\boldsymbol{H}$ when $\boldsymbol{V}$ is assumed generated by a Poisson observation model, as will be later recalled.

A criticism of NMF for nonnegative dictionary learning is that little can be said about the asymptotical optimality of the learnt dictionary $\boldsymbol{W}$. This is because the total number of parameters $FK+KN$ considered for maximum likelihood estimation grows with the number of observations N. As such, in this paper we seek to optimize the marginal likelihood of $\boldsymbol{W}$ given by

$$p(\boldsymbol{V}|\boldsymbol{W}) = \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})p(\boldsymbol{H})\,d\boldsymbol{H}, \tag{1}$$

where $p(\boldsymbol{H})$ is an assumed prior distribution for the expansion coefficients. Our approach is similar in spirit to independent component analysis (ICA), where the likelihood of the "mixing matrix" is obtained through marginalization of the latent independent components, see, e.g., [4].

In order to compute the marginal likelihood in (1), we define a prior distribution for $\boldsymbol{H}$. We choose a Gamma distribution, which is conjugate to the Poisson likelihood, mostly for algorithmic convenience as will be apparent further in the paper. The Gamma distribution takes the sparsity-inducing exponential distribution as a special case. We leave the dictionary $\boldsymbol{W}$ to be a deterministic variable. As such, our model coincides with the Gamma-Poisson (GaP) model of Canny [5], which constitutes the base for many elaborate models for text clustering and image interpolation (e.g., [6]–[9]). In Discrete Component Analysis (DCA) of Buntine & Jakulin [8] a Dirichlet prior is assumed for $\boldsymbol{W}$, whereas in Cemgil's work [9] the prior is a Gamma distribution. In these works, maximum a posteriori (MAP) estimate or full posterior distribution of $\boldsymbol{W}$ (thus taken as a random varianble) have been sought after. In this work we do not wish to make any prior assumption on $\boldsymbol{W}$ and rather look for the maximum likelihood estimator of $\boldsymbol{W}$ in the basic GaP model. Our work bears methodological resemblance to [8] and [9], but we here pursue a different objective.

The main motivation in this paper is to learn the dictionary parameters from the marginal model in order to prevent the overfitting in standard KL-NMF, which stems from the growing number of parameters with the number of observations. At this

Manuscript received July 22, 2011; revised December 04, 2011, March 21, 2012, and June 10, 2012; accepted June 12, 2012. Date of publication July 05, 2012; date of current version September 11, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Samson Lasaulce. This work was supported by Project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization).

O. Dikmen was with the CNRS LTCI; Télécom ParisTech, Paris 75014, France. He is now with the Department of Information and Computer Science, Aalto University, Espoo 02150, Finland (e-mail: onur.dikmen@aalto.fi).

C. Févotte is with the CNRS LTCI; Télécom ParisTech, Paris 75014, France (e-mail: fevotte@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2012.2207117

stage a parallel can be made between our approach and Latent Dirichlet Allocation (LDA) [10], which was proposed as a remedy to the overfitting problem of Latent Semantic Indexing (pLSI) [11]. LDA and pLSI are based on the same model (a discrete observation model in which word f in document n is generated with probability $\sum_k w_{fk}h_{kn}$ under the constraint that $\sum_f w_{fk} = 1$ and $\boldsymbol{h}_n = [h_{1n}, \ldots, h_{Kn}]^T$ has a Dirichlet prior), but LDA seeks marginal likelihood estimation whereas pLSI performs joint likelihood estimation (i.e., the dictionary and the expansion matrix are updated together).

This paper describes two approximate learning algorithms for the maximization of the marginal likelihood. The integration in (1) is not tractable, so the expansion variables $\boldsymbol{H}$ cannot be analytically integrated out of the model. In addition, an exact expectation-maximization (EM) algorithm cannot be pursued, because the exact form of the posterior distribution of $\boldsymbol{H}$ is not available either. Our EM algorithms, variational Bayes (VBEM) and Monte Carlo EM (MCEM), are based on inferring the posterior distribution of $\boldsymbol{H}$ in the E-step and maximizing $\boldsymbol{W}$ in the M-step. VBEM maximizes a lower bound of the marginal likelihood $p(\boldsymbol{V}|\boldsymbol{W})$, whereas the functional that MCEM maximizes at each step is asymptotically exact. In the experiments section, we verify that these two methods perform similarly and the lower bound that VBEM optimizes is a tight approximation of criterion (1), as numerically confirmed using the more computationally intensive but asymptotically optimal Chib's method [12]. Computation of the criterion itself is for example needed for classification tasks based on likelihood ratios. In this paper we also describe novel algorithms for maximum joint likelihood estimation (MJLE), equivalent to standard KL-NMF with a penalty term on $\boldsymbol{H}$ that stems from the assumed prior, that extend previous work by Canny [5].

We compare MJLE and MMLE on synthetical and real data and indeed show that MMLE avoids the problem of overfitting by assigning "unnecessary" columns of the dictionary to zero, i.e., performing automatic model order selection. With MMLE, the estimated dictionaries are more column-sparse (in the sense that many columns become negligible when K is overestimated) than those of MJLE and this makes the dictionaries more interpretable. We will in particular demonstrate this feature on the musiXmatch dataset, a large collection of word counts from song lyrics, which has recently been made available [13].

The rest of this paper is organized as follows. Section II describes the generative model and presents the two dictionary estimators considered in this paper. Sections III and IV describe algorithms proposed for these estimators. Section V reports results on real and synthetical data and in particular illustrates a very desirable feature of the marginal likelihood approach: automatic order selection. Section VI concludes the paper. The VBEM algorithm of Section IV was introduced in [14] and was mainly applied to audio data. Here, we also introduce the asymptotically exact MCEM algorithm and verify the results obtained by VBEM. We describe Chib's method [12] for this model to discuss the tightness of the variational lower bound. We also present novel algorithms for MJLE for a large range of shape parameters. The two estimators, MJLE and MMLE, are compared in face reconstruction and text retrieval applications.

II. Model and Estimators

A. GaP Model

The generative model assumed for the observations $v_{fn} = [\boldsymbol{V}]_{fn}$ is

$$v_{fn} \sim \mathcal{P}\left(v_{fn}|\sum_k w_{fk}h_{kn}\right) \tag{2}$$

where $\mathcal{P}$ denotes the Poisson distribution, defined by $\mathcal{P}(x|\lambda) = \exp(-\lambda)\,\frac{\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$ The data is assumed independently distributed given $\boldsymbol{W}$ and $\boldsymbol{H}$. Using the superposition property of the Poisson distribution, the generative model can equivalently be written as a *composite model* such that

$$v_{fn} = \sum_{k=1}^{K} c_{k,fn}, \quad c_{k,fn} \sim \mathcal{P}(c_{k,fn}|w_{fk}h_{kn}) \tag{3}$$

where the components $c_{k,fn}$ act as *latent variables*. In the remainder of the text, $\boldsymbol{C}$ will denote the $K \times F \times N$ matrix consisting of $\{c_{k,fn}\}$ and $\boldsymbol{c}_{fn}$ will represent the $K \times 1$ vector $[c_{1,fn}, c_{2,fn}, \ldots, c_{K,fn}]^T$. The posterior distribution of $\boldsymbol{c}_{fn}$ is analytically available and is a multinomial distribution. Similarly, posterior distribution of each $c_{k,fn}$ is binomial. As can be seen from (3), the introduction of the components allows us to break the coupling $\sum_k w_{fk}h_{kn}$ in the probability density function of the observation model. This fact will be used in the data augmentation algorithms described in Sections IV-A and IV-B.

We further take the expansion coefficients h_{kn} to be random variables with Gamma prior, such that $h_{kn} \sim \mathcal{G}(h_{kn}|\alpha_k, \beta_k)$, where $\mathcal{G}(x|\alpha, \beta) = [\beta^\alpha \Gamma(\alpha)]^{-1}\, x^{\alpha-1}\, \exp\left(\frac{-x}{\beta}\right)$, $x \geq 0$, $\alpha > 0$, $\beta > 0$. The Gamma distribution is a prior of choice for its conjugacy with the Poisson distribution, and will facilitate some algorithm derivations to be presented next. Under these assumptions our model coincides with the Gamma-Poisson (GaP) model of [5], [8] which has been used in text analysis. In the rest of the paper, the scale parameters β_k will be fixed, so as to remedy the scale ambivalence between k^{th} column of $\boldsymbol{W}$ (denoted $\boldsymbol{w}_k$ in the following) and k^{th} row of $\boldsymbol{H}$, as more thoroughly discussed in Section II-D. We will denote $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_K]^T$. The shape parameters α_k are also fixed. The dictionary $\boldsymbol{W}$ is taken as a free deterministic parameter.

B. Maximum Joint Likelihood Estimation (MJLE)

The MJLE estimator of $\boldsymbol{W}$ is obtained by maximization (under nonnegativity of all the parameters) of the joint penalized log-likelihood likelihood of $\boldsymbol{W}$, $\boldsymbol{H}$ and $\boldsymbol{\beta}$, defined by

$$\begin{aligned} C_{JL}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\beta}) &\stackrel{\text{def}}{=} \log p(\boldsymbol{V}|\boldsymbol{W}, \boldsymbol{H}) + \log p(\boldsymbol{H}|\boldsymbol{\beta}) \\ &= -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{H}) + cst \end{aligned}$$

where cst denotes terms constant w.r.t $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{\beta}$ and where

$$L_{\boldsymbol{\beta}}(\boldsymbol{H}) \stackrel{\text{def}}{=} \sum_{kn} \frac{h_{kn}}{\beta_k} - (\alpha_k - 1)\log h_{kn} + \alpha_k \log \beta_k.$$

As it appears, MJLE in the GaP model is equivalent to penalized KL-NMF [9], [15], with penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{H})$. In Section III, we will present minorization-maximization (MM) algorithms

for MJLE for different α_k values. For a comprehensive study about MM algorithms for NMF, see [16].

C. Maximum Marginal Likelihood Estimation (MMLE)

The MMLE estimator of $\boldsymbol{W}$ is obtained by maximization of the marginal log-likelihood of $\boldsymbol{W}$, defined by

$$C_{ML}(\boldsymbol{W},\boldsymbol{\beta}) \stackrel{\text{def}}{=} \log p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{\beta}) = \log \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}) p(\boldsymbol{H}|\boldsymbol{\beta})\, d\boldsymbol{H}.$$

This integral is intractable, i.e., it is not possible to obtain the marginal model analytically. Note that in Bayesian estimation the term *marginal likelihood* is sometimes used as a synonym for the *model evidence*, which would be the likelihood of data given the model, i.e., where all random parameters (including $\boldsymbol{W}$ as well) have been marginalized. This full Bayesian approach has been considered for example in [9] and [17] for the Poisson and Gaussian additive noise models, respectively. In [8], $\boldsymbol{W}$ again has a prior distribution and is estimated with a maximum a posteriori approach. Let us emphasize again that in our setting $\boldsymbol{W}$ is taken as a deterministic parameter and that the term "marginal likelihood" here refers to the likelihood of $\boldsymbol{W}$ where $\boldsymbol{H}$ has been integrated out. In Section IV, we will describe approximate EM algorithms for evaluating and maximizing C_{ML}.

D. Scales

The MJLE and MMLE estimators of $\boldsymbol{W}$ have different behaviors w.r.t scale $\boldsymbol{\beta}$. The MMLE objective function $C_{ML}(\boldsymbol{W},\boldsymbol{\beta})$ is scale-invariant, in the following sense. Let $\boldsymbol{\Lambda}$ be a nonnegative diagonal matrix with coefficients λ_k. Then we have the following property:

$$C_{ML}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{ML}(\boldsymbol{W},\boldsymbol{\beta}).$$

This is shown using the change of variable $h_{kn} = \lambda_k \tilde{h}_{kn}$ and the property that if $X \sim \mathcal{G}(\alpha,\beta)$ then $\lambda X \sim \mathcal{G}(\alpha,\lambda\beta)$. More precisely, we have

$$\begin{aligned} C_{ML}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{\beta}) &= \int_{\boldsymbol{H}} p(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{H}) \prod_{kn} \mathcal{G}(h_{kn}|\alpha_k,\lambda_k\beta_k) d\boldsymbol{H} \\ &= \int_{\tilde{\boldsymbol{H}}} p(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\tilde{\boldsymbol{H}}) \prod_{kn} \lambda_k \mathcal{G}(\lambda_k\tilde{h}_{kn}|\alpha_k,\lambda_k\beta_k) d\tilde{\boldsymbol{H}} \\ &= \int_{\tilde{\boldsymbol{H}}} p(\boldsymbol{V}|\boldsymbol{W},\tilde{\boldsymbol{H}}) \prod_{kn} \mathcal{G}(\tilde{h}_{kn}|\alpha_k,\beta_k) d\tilde{\boldsymbol{H}} \\ &= C_{ML}(\boldsymbol{W},\boldsymbol{\beta}). \end{aligned}$$

As such, we may fix β_k to any arbitrary value, because of this simple linear mapping that exists between dictionary solutions obtained for different scale parameters.

The MJLE objective function $C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta})$ behaves differently w.r.t scale, and as it appears, in a potentially problematic way. Indeed, the following expression holds:

$$C_{JL}(\boldsymbol{W}\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{H},\boldsymbol{\Lambda}\boldsymbol{\beta}) = C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta}) - N\sum_k \log\lambda_k.$$

This implies that maximization of $C_{JL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\beta})$ under mere nonnegativity constraints can only lead to a degenerate solution $(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$ such that $\boldsymbol{W}\to\infty$, $\boldsymbol{H}\to 0$ and $\boldsymbol{\beta}\to 0$. This can be shown by contradiction: assume that $\boldsymbol{W}^\star$ is finite, then for any nonnegative diagonal matrix $\boldsymbol{\Lambda}$ such that $\lambda_k < 1$ we obtain that $C_{JL}(\boldsymbol{W}^\star\boldsymbol{\Lambda}^{-1},\boldsymbol{\Lambda}\boldsymbol{H}^\star,\boldsymbol{\Lambda}\boldsymbol{\beta}^\star) > C_{JL}(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$, which contradicts the fact that $(\boldsymbol{W}^\star,\boldsymbol{H}^\star,\boldsymbol{\beta}^\star)$ is the optimum. As such, joint estimation of $\boldsymbol{W}$, $\boldsymbol{H}$, $\boldsymbol{\beta}$ requires to control the norm of $\boldsymbol{W}$ to prevent from degeneracy. In this paper will not try to estimate $\boldsymbol{\beta}$ (either by maximization of C_{JL} or by any other mean such as cross-validation) but we will rather concentrate on the properties of the two estimators (MJLE and MMLE) based on the same generative model, i.e., with both α_k and β_k fixed to arbitrary values. As such, we will remove from now on the dependency of C_{JL} and C_{ML} on $\boldsymbol{\beta}$. Fixing $\boldsymbol{\beta}$ removes the degeneracy problem only when $\alpha_k > 1$. Indeed, it is easy to check that when $\alpha \leq 1$ the penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ can still be made arbitrarily small as λ_k goes to zero, thus encouraging the solution $\boldsymbol{W}\to\infty$. As such, controlling the norm of $\boldsymbol{W}$ is still needed in that case. A set of algorithms for MJLE, depending on the value of α_k and the required constraint on $\boldsymbol{W}$ are presented in the next section. A conclusion of this section is that, besides the question of its asymptotical optimality, MJLE is not an as well-posed problem as MMLE.

III. Algorithms For MJLE

A. Algorithms For $\alpha_k > 1$

1) Canny's Algorithm: Given the discussion of Section II-D, when $\boldsymbol{\beta}$ is fixed and $\alpha_k > 1$, the norm of $\boldsymbol{W}$ does not need to be necessarily controlled. In that setting, an iterative algorithm that guarantees to increase $C_{JL}(\boldsymbol{W},\boldsymbol{H})$ at every iteration is described in [5]. Though not clearly stated as such, it is essentially an EM algorithm based on the set of components $c_{k,fn}$ introduced in Section II-A, that updates $\boldsymbol{H}$ given $\boldsymbol{W}$ and then $\boldsymbol{W}$ given $\boldsymbol{H}$. It can also be seen as a MM algorithm where each update is based on the maximization of a surrogate auxiliary function, i.e., a lower bound of the original objective function which is tight at the current parameter estimate [14]. This strategy ensures the increase of the joint likelihood after every update of $\boldsymbol{W}$ and $\boldsymbol{H}$. Using its concavity w.r.t $\boldsymbol{H}$ given $\boldsymbol{W}$ or $\boldsymbol{H}$ given $\boldsymbol{W}$, the fit to data term $-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H})$ can easily be minorized using a Jensen's type inequality, such that, $\forall\tilde{\boldsymbol{H}}$

$$-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) \geq -\sum_{fkn} \frac{w_{fk}\tilde{h}_{kn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} d_{KL}\left(v_{fn}|[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}\frac{h_{kn}}{\tilde{h}_{kn}}\right) \quad (4)$$

with equality when $\boldsymbol{H} = \tilde{\boldsymbol{H}}$, or similarly, $\forall\tilde{\boldsymbol{W}}$

$$-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) \geq -\sum_{fkn} \frac{\tilde{w}_{fk}h_{kn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}} d_{KL}\left(v_{fn}|[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}\frac{w_{fk}}{\tilde{w}_{fk}}\right) \quad (5)$$

with equality when $\boldsymbol{W} = \tilde{\boldsymbol{W}}$. Iterative maximization of the lower bound (5) w.r.t $\boldsymbol{W}$ leads to the well known multiplicative algorithm described by [1]–[3]

$$w_{fk} = \tilde{w}_{fk}\frac{\sum_n \frac{h_{kn}v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\sum_n h_{kn}}.$$

The penalty term $-L_{\boldsymbol{\beta}}(\boldsymbol{H})$ needs solely to be added to the right side of (4) to obtain a suitable auxiliary function for $\boldsymbol{H}$, which, when maximized w.r.t $\boldsymbol{H}$, leads to

$$h_{kn} = \frac{\tilde{h}_{kn} \sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} + (\alpha_k - 1)}{\frac{1}{\beta_k} + \sum_f w_{fk}}.$$

2) Algorithm for Norm-Constrained $\boldsymbol{W}$: It is also possible to derive an algorithm for maximizing $C_{JL}(\boldsymbol{W}, \boldsymbol{H})$ under the constraint that $\|\boldsymbol{w}_k\| = 1$, where we will take $\|\cdot\|$ as the ℓ_1 norm. As discussed in Section II-D, though not necessary when β_k is fixed (and when $\alpha_k > 1$), this is needed when β_k has to be estimated as well. In that case we want to solve

$$\max_{\boldsymbol{W},\boldsymbol{H}} C_{JL}(\boldsymbol{W}, \boldsymbol{H}) = -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{H}) + cst \quad s.t \quad \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0,\ \|\boldsymbol{w}\|_k = 1. \tag{6}$$

Following [18], [19], the latter problem is equivalent to the following surrogate optimization problem, that involves a scale-invariant objective function:

$$\max_{\boldsymbol{W},\boldsymbol{H}} \bar{C}_{JL}(\boldsymbol{W}, \boldsymbol{H}) \stackrel{\text{def}}{=} -D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}) - L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) + cst \quad s.t \quad \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0 \tag{7}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\|w_1\|, \ldots, \|w_K\|\}$. The equivalence between (6) and(7) is explained as follows. Let $(\boldsymbol{W}, \boldsymbol{H})$ be a pair of nonnegative matrices and let $(\boldsymbol{W}^\bullet = \boldsymbol{W}\boldsymbol{\Lambda}^{-1}, \boldsymbol{H}^\bullet = \boldsymbol{\Lambda}\boldsymbol{H})$ be their rescaled equivalents. Then, we have $\bar{C}_{JL}(\boldsymbol{W}, \boldsymbol{H}) = C_{JL}(\boldsymbol{W}^\bullet, \boldsymbol{H}^\bullet)$, and $\boldsymbol{W}^\bullet$ satisfies the constraint $\|\boldsymbol{w}_k^\bullet\| = 1$ by construction. As such, one may solve (7), free of scale constraint, and then rescale its solution to obtain a solution to (6). Using same recipe as in previous section, with $L_{\boldsymbol{\beta}}(\boldsymbol{H})$ changed into $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$, we can obtain the following update for $\boldsymbol{H}$:

$$h_{kn} = \frac{\tilde{h}_{kn} \sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}} + (\alpha_k - 1)}{\left(1 + \frac{1}{\beta_k}\right) \sum_f w_{fk}}. \tag{8}$$

The penalty term $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ also depends on $\boldsymbol{W}$ through λ_k. As such, the dictionary update is also changed. Unfortunately simply adding $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$ to the right side of (5) does not lead to an auxiliary function for $\boldsymbol{W}$ that can be maximized in close form. The penalty term $-L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H})$, and more precisely its log part needs to be minorized as well. We may write

$$-L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) = -\sum_{fkn} \frac{w_{fk} h_{kn}}{\beta_k} + N \sum_k (\alpha_k - 1) \log \|\boldsymbol{w}_k\| + cst$$

where cst represents terms constants w.r.t $\boldsymbol{W}$.[1] By concavity of $\log x$ and Jensen's inequality we may write

$$(\alpha_k - 1) \log \|\boldsymbol{w}_k\| \geq (\alpha_k - 1) \sum_f \frac{\tilde{w}_{fk}}{\|\tilde{\boldsymbol{w}}_k\|} \log \left(\|\tilde{\boldsymbol{w}}_k\| \frac{w_{fk}}{\tilde{w}_{fk}} \right)$$

[1]Except when otherwise specified, in the following we will abusively denote by cst any irrelevant term constant w.r.t the variable of the function in which it appears.

with equality when $\boldsymbol{w}_k = \tilde{\boldsymbol{w}}_k$. This minorization of the log part of the penalty term leads to close form maximization of the resulting auxiliary function, which writes

$$w_{fk} = \tilde{w}_{fk} \frac{\frac{(\alpha_k - 1)N}{\|\tilde{\boldsymbol{w}}_k\|} + \sum_n \frac{h_{kn} v_{fn}}{[\tilde{\boldsymbol{W}}\tilde{\boldsymbol{H}}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}. \tag{9}$$

B. Algorithm for $\alpha_k \leq 1$

As discussed in Section II-D, when $\alpha_k \leq 1$, the norm of the dictionary needs to be controlled to prevent from degeneracy, β_k being treated either as a fixed or free parameter.

1) Algorithm for $\alpha_k = 1$: Canny's algorithm is not applicable anymore for $\alpha_k = 1$ but the derivations of Section III-A-II still hold. They simplify to the following update rules:

$$h_{kn} = \tilde{h}_{kn} \frac{\sum_f \frac{w_{fk} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_f w_{fk}} \tag{10}$$

$$w_{fk} = \tilde{w}_{fk} \frac{\sum_n \frac{h_{kn} v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}]_{fn}}}{\left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}. \tag{11}$$

Note the symmetry of the update rules (the roles of $\boldsymbol{W}$ and $\boldsymbol{H}$ are simply exchanged), due to the symmetry of the penalty term, that reduces to $L_{\boldsymbol{\beta}}(\boldsymbol{\Lambda}\boldsymbol{H}) = \sum_{fkn} \frac{w_{fk} h_{kn}}{\beta_k} + cst$.

2) Algorithm for $\alpha_k < 1$: When $\alpha_k < 1$, then $(\alpha_k - 1)$ becomes negative and the minorization of the log part of the penalty term used for the update of $\boldsymbol{W}$ in Section III-A-II does not hold anymore. It is possible to use instead a first order Taylor approximation of $\log \|\boldsymbol{w}_k\|$ (using the property that a concave function is majorized by its tangent), that leads to

$$w_{fk} \leftarrow \tilde{w}_{fk} \frac{\sum_n \frac{h_{kn} v_{fn}}{[\boldsymbol{W}\tilde{\boldsymbol{H}}]_{fn}}}{\frac{(1-\alpha_k)N}{\|\tilde{\boldsymbol{w}}_k\|} + \left(1 + \frac{1}{\beta_k}\right) \sum_n h_{kn}}$$

i.e., the term $\frac{(1-\alpha_k)N}{\|\tilde{\boldsymbol{w}}_k\|}$ that appeared in the numerator of (9) is moved to the denominator (after changing its sign). The main source of difficulty when $\alpha_k < 1$ lies in the update of $\boldsymbol{H}$. The update given by (8) still maximizes the auxiliary function, but not under the nonnegativity constraint. The update may fail to satisfy the nonnegative constraint because of the $(\alpha_k - 1)$ term at the numerator which is now negative. Truncating h_{kn} to zero when it fails to satisfy the nonnegative constraint does provide a valid ascent algorithm, but any coefficient that hits zero will remain zero. Other schemes could be envisaged for $\boldsymbol{H}$ (such as projected gradient descent) but we will not pursue this issue as it is out of main scope of this paper, given that Gamma shape parameters less than one are rarely used in practice.

IV. Algorithms for MMLE

We investigate a set of Expectation-Maximization (EM) [20] algorithms for MMLE. The EM algorithm converges to a stationary point of the likelihood function by iteratively evaluating (E-step) and maximizing (M-step) the expected log-likelihood of the data completed by some latent data. For example, for the

observation model $p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H})$ in (2) with the prior distribution $p(\boldsymbol{H})$, an EM algorithm can be built on the complete data set $(\boldsymbol{V},\boldsymbol{H})$ by iteratively evaluating and maximizing the functional defined by

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) = \int_{\boldsymbol{H}} \log p(\boldsymbol{V},\boldsymbol{H}|\boldsymbol{W}) p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})\, d\boldsymbol{H}. \quad (12)$$

As it appears the posterior distribution $p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is not analytically available and the functional can neither be evaluated nor maximized. An other EM algorithm can be built on the larger complete data set $(\boldsymbol{C},\boldsymbol{H})$, exploiting the composite model representation of (3), leading to[2]

$$Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) = \int_{\boldsymbol{C},\boldsymbol{H}} \log p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{W}) p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})\, d\boldsymbol{C}\, d\boldsymbol{H}. \quad (13)$$

While this functional is still intractable, the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ in the composite model representation is easier to infer using variational Bayes or Markov chain Monte Carlo methods. In Sections IV-A and IV-B, we will describe two EM algorithms where the E-steps are based on these computational methods.

A. Variational Bayes EM (VBEM)

A variational EM algorithm [21] for the maximization of $C_{ML}(\boldsymbol{W})$ based on the functional $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ can be constructed as follows. As explained above, we resort to a variational approximation of the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ that renders all derivations tractable, though at the cost of approximate inference. The two steps of the variational EM are described next.

1) E-Step: A variational approximation $q(\boldsymbol{C},\boldsymbol{H})$ of the exact posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is computed at every iteration of the EM algorithm and plugged in $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$. As fundamental to variational approximations, the computation of $q(\boldsymbol{C},\boldsymbol{H})$ relies on the minimization of the KL divergence (in *distribution*) between $q(\boldsymbol{C},\boldsymbol{H})$ and $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$, given a parametric form of $q(\boldsymbol{C},\boldsymbol{H})$. The variational objective function may be decomposed as

$$\begin{aligned}\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})] \\ = \log p(\boldsymbol{V}|\tilde{\boldsymbol{W}}) + \mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})]. \quad (14)\end{aligned}$$

Because the marginal likelihood $\log p(\boldsymbol{V}|\tilde{\boldsymbol{W}})$ is independent of $q(\boldsymbol{C},\boldsymbol{H})$, the minimization of the variational objective may be replaced by the (simpler) maximization of

$$\mathrm{L}[q(\boldsymbol{C},\boldsymbol{H})] = -\mathrm{KL}[q(\boldsymbol{C},\boldsymbol{H})|p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})]$$

which forms a lower bound of the marginal likelihood $\log p(\boldsymbol{V}|\tilde{\boldsymbol{W}})$ (thanks to nonnegativity of the KL divergence). It can be shown that, given the expression of $p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})$, the following form of variational distribution appears as a natural choice (in particular for tractability):

$$q(\boldsymbol{C},\boldsymbol{H}) = \prod_{f=1}^{F}\prod_{n=1}^{N} q(\boldsymbol{c}_{fn}) \prod_{k=1}^{K}\prod_{n=1}^{N} q(h_{kn})$$

where $\boldsymbol{c}_{fn}$ denotes the vector $[c_{1,fn}, c_{2,fn}, \ldots, c_{K,fn}]^T$ as in Section II-A, $q(\boldsymbol{c}_{fn})$ is multinomial with probabilities $p_{k,fn}$ and $q(h_{kn})$ is a Gamma distribution with shape and scale parameters $\bar{\alpha}_{kn}$ and $\bar{\beta}_{kn}$. The factors $q(h_{kn})$ and $q(\boldsymbol{c}_{fn})$ can be shown to satisfy the following fixed point equations [21]:

$$\log q(h_{kn}) = \langle \log p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})\rangle_{q(\boldsymbol{H}_{-kn})q(\boldsymbol{C})} + cst \quad (15)$$

$$\log q(\boldsymbol{c}_{fn}) = \langle \log p(\boldsymbol{C},\boldsymbol{H}|\tilde{\boldsymbol{W}})\rangle_{q(\boldsymbol{H})q(\boldsymbol{C}_{-fn})} + cst, \quad (16)$$

where $\langle . \rangle_\pi$ denotes expectation under probability distribution π and $\boldsymbol{A}_{-ij}$ refer to the set of coefficients of $\boldsymbol{A}$ excluding a_{ij}. In particular, the optimal variational distribution $q(\boldsymbol{c}_{fn})$ satisfies

$$\begin{aligned}\log q(\boldsymbol{c}_{fn}) = \sum_k c_{k,fn} \log \tilde{w}_{fk} \\ + c_{k,fn}\langle \log h_{kn}\rangle - \log\Gamma(c_{k,fn}+1) + cst,\end{aligned}$$

which lead to the following fixed point update for its probability parameters

$$p_{k,fn} = \frac{\tilde{w}_{fk}\exp(\langle \log h_{kn}\rangle)}{\sum_l \tilde{w}_{fl}\exp(\langle \log h_{ln}\rangle)}$$

where the expectation is w.r.t the variational distribution $q(h_{kn})$ and $\langle \log h_{kn}\rangle = \psi(\bar{\alpha}_{kn}) + \log\bar{\beta}_k$.[3] Similarly, the optimal variational distribution $q(h_{kn})$ satisfies

$$\begin{aligned}\log q(h_{kn}) = -\left(\sum_f \tilde{w}_{fk} + \frac{1}{\beta_k}\right) h_{kn} \\ + \left(\sum_f \langle c_{k,fn}\rangle + \alpha_k - 1\right)\log h_{kn} + cst\end{aligned}$$

from which the updates are found as

$$\begin{aligned}\bar{\alpha}_{kn} &= \alpha_k + \sum_f \langle c_{k,fn}\rangle \\ \frac{1}{\bar{\beta}_{kn}} &= \frac{1}{\bar{\beta}_k} = \frac{1}{\beta_k} + \sum_f \tilde{w}_{fk}\end{aligned}$$

where the expectation is w.r.t $q(c_{k,fn})$ and has the analytical form $\langle c_{k,fn}\rangle = p_{k,fn} v_{fn}$.

2) M-Step: Given the variational distribution $q(\boldsymbol{C},\boldsymbol{H})$ obtained in the E-step, the EM functional can be approximated as

$$\begin{aligned}Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx \sum_{fn}\sum_k -w_{fk}\langle h_{kn}\rangle \\ + \langle c_{k,fn}\rangle(\log w_{fk} + \langle \log h_{kn}\rangle) \\ - \langle \log\Gamma(c_{k,fn}+1)\rangle \\ + \sum_{kn}(\alpha_k - 1)\langle h_{kn}\rangle - \frac{\langle h_{kn}\rangle}{\beta_k} \\ - \alpha_k \log\beta_k - \log\Gamma(\alpha_k)\end{aligned}$$

where $\langle h_{kn}\rangle = \bar{\alpha}_{kn}\bar{\beta}_{kn}$. Maximization of $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ leads to the update rule

$$w_{fk}^{\mathrm{VBEM}} = \frac{\sum_n \langle c_{k,fn}\rangle}{\sum_n \langle h_{kn}\rangle}.$$

[2]Note that $\boldsymbol{V}$ is encompassed in $\boldsymbol{C}$ because of the relation $v_{fn} = \sum_k c_{k,fn}$ and as such needs not to appear in the complete set, see [20].

[3]ψ is the digamma function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$.

If the value of $\langle c_{k,fn}\rangle$ is plugged in, it is easy to see that this is a multiplicative update rule

$$w_{fk}^{\mathrm{VBEM}} = \tilde{w}_{fk}\frac{\sum_n \frac{\exp(\langle \log h_{kn}\rangle)v_{fn}}{[\tilde{\boldsymbol{W}}\exp(\langle \log \boldsymbol{H}\rangle)]_{fn}}}{\sum_n \langle h_{kn}\rangle} \tag{17}$$

which remains nonnegative provided that the initial value of $\tilde{\boldsymbol{W}}$ in nonnegative.

Note that one could contemplate plugging the variational distribution $q(\boldsymbol{H})$ into $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ so as to produce an alternative EM algorithm, but the integration incurred in $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ is still intractable.

B. Monte Carlo EM (MCEM)

In this section, we describe how the dictionary parameters can be optimized using Monte Carlo EM (MCEM) [22]. The algorithm consists of an E-step where the posterior distribution $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ is inferred by a Gibbs sampler [23]. This is a Markov chain Monte Carlo (MCMC) method, in which the chain is constructed by drawing samples from full conditional distributions of (blocks of) variables. Then, the EM functional, either Q_1 or Q_2, is approximated using Monte Carlo integration and maximized w.r.t. $\boldsymbol{W}$. Contrary to VBEM, the Monte Carlo approximation is asymptotically exact.

1) E-Step: In our model, the full conditional distributions of component variables, $\boldsymbol{C}$, are multinomial and those of expansion coefficients, $\boldsymbol{H}$, are Gamma. So, it is highly convenient to use the Gibbs sampler that samples $\boldsymbol{H}$ given $\boldsymbol{C}$ and $\boldsymbol{C}$ given $\boldsymbol{H}$. At iteration $i+1$, a new sample $h_{kn}^{(i+1)}$ is drawn from a Gamma distribution $\mathcal{G}(\bar{\alpha}_{kn},\bar{\beta}_k)$ with parameters given by:

$$\bar{\alpha}_{kn} = \alpha_{kn} + \sum_f c_{k,fn}^{(i)}$$
$$\frac{1}{\bar{\beta}_{kn}} = \frac{1}{\bar{\beta}_k} = \frac{1}{\beta_k} + \sum_f \tilde{w}_{fk}$$

and a new sample $\boldsymbol{c}_{fn}^{(i+1)}$ is subsequently drawn from a multinomial distribution with probabilities

$$p_{k,fn} = \frac{\tilde{w}_{fk}h_{kn}^{(i+1)}}{\sum_l \tilde{w}_{fl}h_{ln}^{(i+1)}}.$$

2) M-Step: Once a set of samples from $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ have been obtained (after a burn in period), the functional $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ and $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ may be approximated as

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\log p(\boldsymbol{V}|\boldsymbol{W},\boldsymbol{H}^{(i)}) \tag{18}$$

$$Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\log p(\boldsymbol{C}^{(i)},\boldsymbol{H}^{(i)}|\boldsymbol{W}). \tag{19}$$

The maximization of $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, approximated as in (19), is available in close form and leads to

$$w_{fk}^{\mathrm{MCEM2}} = \frac{\sum_{in} c_{k,fn}^{(i)}}{\sum_{in} h_{kn}^{(i)}}. \tag{20}$$

The maximization of $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, approximated as in (18), is not available in close form and requires an optimization procedure. Equation (18) reduces to

$$Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}}) \approx -\frac{1}{N_s}\sum_i^{N_s} D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}^{(i)}) + cst.$$

By minorizing the individual terms $-D_{KL}(\boldsymbol{V}|\boldsymbol{W}\boldsymbol{H}^{(i)})$ as in (5) and summing up the auxiliary functions corresponding to all samples, one gets a general auxiliary function whose iterative maximization leads to following fixed point equation:

$$w_{fk} \leftarrow w_{fk}\frac{\sum_{in}\frac{h_{kn}^{(i)}v_{fn}}{[\boldsymbol{W}\boldsymbol{H}^{(i)}]_{fn}}}{\sum_{in} h_{kn}^{(i)}}.$$

In practice we perform only one iteration of the fixed point equation (which only increases $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ instead of fully maximizing it), starting from the current EM update $\tilde{\boldsymbol{W}}$, so that

$$w_{fk}^{\mathrm{MCEM1}} = \tilde{w}_{fk}\frac{\sum_{in}\frac{h_{kn}^{(i)}v_{fn}}{[\tilde{\boldsymbol{W}}\boldsymbol{H}^{(i)}]_{fn}}}{\sum_{in} h_{kn}^{(i)}}. \tag{21}$$

As it appears the update (21) based on $Q_1(\boldsymbol{W}|\tilde{\boldsymbol{W}})$ is the Rao-Blackwellized version of the update (20) based on $Q_2(\boldsymbol{W}|\tilde{\boldsymbol{W}})$, i.e.,

$$w_{fk}^{\mathrm{MCEM1}} = \frac{\sum_{in}\mathrm{E}[c_{k,fn}|\boldsymbol{V},\tilde{\boldsymbol{W}},\boldsymbol{H}^{(i)}]}{\sum_{in} h_{kn}^{(i)}}.$$

Rao-Blackwellization is known to produce updates with lesser variance [24] and as such, we will use update(21) in practice.

Marginal Likelihood From Gibbs Samples (Chib's Method): While in VBEM an estimation of the criterion $C_{ML}(\tilde{\boldsymbol{W}})$ comes as a by-product in the form of the lower bound $\mathrm{L}[q(\boldsymbol{C},\boldsymbol{H})]$, the MCEM approach does not come with such an estimation. It is however possible to "recycle" the samples of $\boldsymbol{C}$ obtained during E-step to form an estimation of the objective, using Chib's method [12]. Chib's approach is based on the following relation (Bayes' theorem)

$$C_{ML}(\tilde{\boldsymbol{W}}) = \log p(\boldsymbol{V}|\tilde{\boldsymbol{W}},\boldsymbol{H}^*) + \log p(\boldsymbol{H}^*) - \log p(\boldsymbol{H}^*|\boldsymbol{V},\tilde{\boldsymbol{W}}) \tag{22}$$

which holds for any $\boldsymbol{H}^*$, but recommendation is to use the posterior mode $\arg\max p(\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ in practice. The first two terms in the right side of (22) are easy to compute, and are given by the generative model. The last term is not readily available, however it can be estimated by the following Monte Carlo integration

$$p(\boldsymbol{H}^*|\boldsymbol{V},\tilde{\boldsymbol{W}}) = \int p(\boldsymbol{H}^*|\boldsymbol{C},\tilde{\boldsymbol{W}})p(\boldsymbol{C}|\boldsymbol{V},\tilde{\boldsymbol{W}})\,d\boldsymbol{C} \approx \frac{1}{N_s}\sum_{i=1}^{N_s} p(\boldsymbol{H}^*|\boldsymbol{C}^{(i)}\tilde{\boldsymbol{W}}),$$

where $\boldsymbol{C}^{(i)}$ are the samples drawn from the posterior $p(\boldsymbol{C},\boldsymbol{H}|\boldsymbol{V},\tilde{\boldsymbol{W}})$ during Gibbs sampling.

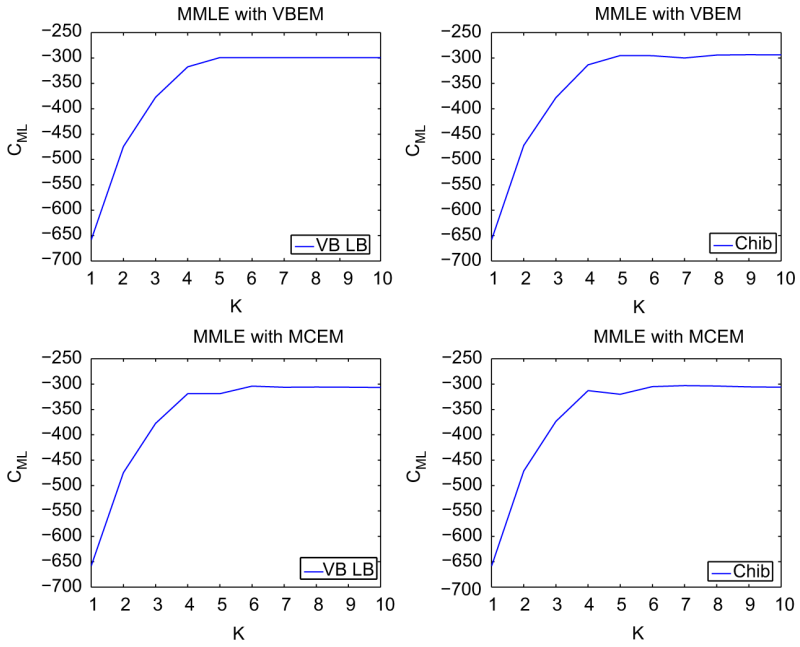


Fig. 1. Synthetical dataset. Marginal log-likelihood $C_{ML}(\hat{\boldsymbol{W}})$ for $K = 1, \ldots, 10$. Top row: $\hat{\boldsymbol{W}}$ is obtained with VBEM (4000 iterations), bottom-row: $\hat{\boldsymbol{W}}$ is obtained with MCEM (100 samples including burn in, 8000 iterations). Left column: $C_{ML}(\hat{\boldsymbol{W}})$ is approximated by the VB lower bound, right column: $C_{ML}(\hat{\boldsymbol{W}})$ is computed with Chib's method.

V. Experiments

We study the performances of MJLE and MMLE on real and synthetical data. We start with a synthetical data of small size to study the performances of the two MMLE methods, VBEM and MCEM. Next, we test MJLE and MMLE on the Swimmer dataset which is a benchmark dataset in dictionary learning problems. We also compare the two approaches on a face reconstruction experiment on CBCL dataset. As the last experiment, we compare MJLE, MMLE and LDA on the topic learning problem on text (lyrics) data. In the experiments, the prior hyperparameters are fixed to $\alpha_k = 1$ (exponential distribution) and $\beta_k = 1$, i.e., $p(h_{kn}) = \exp(-h_{kn})$, unless otherwise stated. In the algorithms, we initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ ($\langle \boldsymbol{H} \rangle$ in VBEM, $\boldsymbol{H}^{(0)}$ in MCEM) as $\mathtt{W} = \mathtt{abs}(\mathtt{randn}(\mathtt{F},\mathtt{K})) + \mathtt{ones}(\mathtt{F},\mathtt{K})$ and $\mathtt{H} = \mathtt{abs}(\mathtt{randn}(\mathtt{K},\mathtt{N})) + \mathtt{ones}(\mathtt{K},\mathtt{N})$, in Matlab notations. In MCEM one third of the samples are used as the burn in period and discarded in the estimations.

A. Synthetical Dataset

We fix a $\boldsymbol{W}^*$ matrix of size 10×5, of which columns are linearly independent and consist of zeros and tens. We generate data from model in Section II-A, with $N = 50$.

In Fig. 1, we investigate MMLE with the two possible EM algorithms (VB or MC) and the two possible means of estimating the marginal likelihood C_{ML} (variational lower bound or Chib's evaluation), as we increase the number of components, K. On the top row the dictionaries are estimated with VBEM, and with MCEM on the bottom row. On the left column the marginal likelihood is estimated with the variational lowerbound and on the right it is estimated with Chib's method. A first observation is that the four plots essentially coincides. This illustrates the equivalent performances of algorithms and likelihood evaluation techniques. A second observation is that the marginal likelihood ceases increasing when $K \geq 5$. As a matter of fact, visual inspection of the estimated dictionaries reveals that both VBEM and MCEM push the redundant columns of $\boldsymbol{W}$ to zero when K is greater than the ground truth value. In other words, MMLE performs an intrinsic automatic order selection during inference.

Fig. 2 displays the joint and marginal log-likelihood criteria, $C_{JL}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $C_{ML}(\hat{\boldsymbol{W}})$, when $\hat{\boldsymbol{W}}$ is learnt with MJLE. Right plot of Fig. 2 verifies that learning $\boldsymbol{W}$ from MJLE *does not* increase the marginal likelihood criterion as K increases, i.e., we verify experimentally that MJLE does not imply MMLE.

On this synthetical dataset, 4000 iterations of MJLE and MMLE with VBEM take 1.28 sec and 2.15 sec, respectively, on an average computer with a 2.66 GHz CPU and 4GB memory. However, 4000 iterations of MCEM which uses 100 samples per iteration including burn in take 673 sec, which means that running MCEM on large datasets can be prohibitive. In this section we showed that the dictionaries $\hat{\boldsymbol{W}}$ estimated with VBEM and MCEM behave similarly as K is increased and the marginal log likelihood values $C_{ML}(\hat{\boldsymbol{W}})$ of these dictionaries are very close. This means that approximate optimization with VBEM does not lead to a significant performance loss. Thus, we will not consider MCEM on the large datasets of the upcoming experiments.

B. Swimmer Dataset

To further investigate the automatic model order selection feature of MMLE, we consider the synthetical Swimmer dataset [25], for which a ground truth can be defined. The

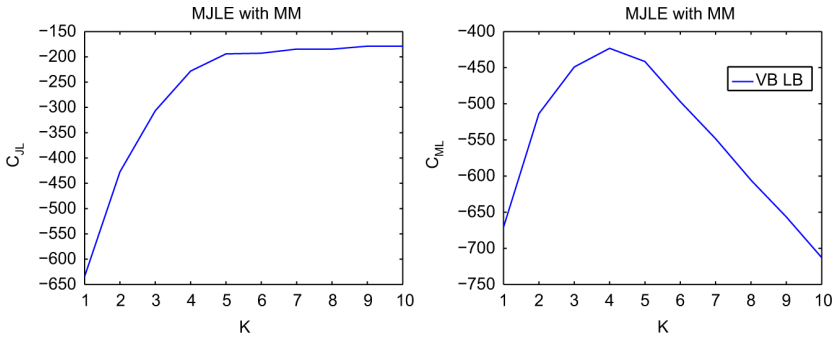


Fig. 2. Synthetical dataset. Joint and marginal log-likelihood criteria $C_{JL}(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ and $C_{ML}(\hat{\boldsymbol{W}})$ when $\hat{\boldsymbol{W}}$ is learnt with MJLE (4000 iterations), for $K = 1, \ldots, 10$.

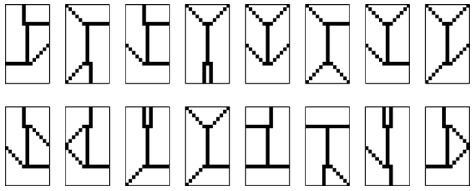


Fig. 3. Data samples consisting of images from the Swimmer dataset with Poisson noise.

dataset is composed of 256 images of size 13×22, representing a swimmer built of an invariant torso and 4 limbs. Each of the 4 limbs can be in one of 4 positions and the dataset is formed of all combinations. Hence, the ground truth dictionary corresponds to the collection of individual limb positions. As explained in [25] the torso is an unidentifiable component that can be paired with any of the limbs, or even split among the limbs. In our experiments, we mapped the binary values onto the range $[1, 100]$ with which we generated the data using the Poisson observation model, see some samples in Fig. 3.

The behavior of the marginal log likelihood, $C_{ML}(\hat{\boldsymbol{W}})$, against K on the noisy swimmer problem is given in Fig. 4. The marginal likelihood increases until $K = 16$, after that value redundant components are assigned to zero and the likelihood does not change. In Fig. 5, we compared the dictionaries learnt by MJLE and MMLE with $K = 20$ components. As can be seen from Fig. 5(a), MJLE produces spurious or duplicated components, i.e., overfits. In contrast, the ground truth is perfectly recovered with MMLE. We do not have a rigorous explanation for the self-ability of MMLE to prune columns of $\boldsymbol{W}$, but in Appendix we present a Laplace approximation of C_{ML} that gives an intuition of why this is happening.

C. Interpolation on CBCL Dataset

We now apply MJLE and MMLE on a missing data prediction task to further investigate overfitting. We used the CBCL face dataset [26] which is composed of 2429 face images. The raw images are 8 bits grayscale images with dimensions 19×19. Similar to [1], the grayscale intensities have been linearly scaled so that the pixel mean and standard deviation were equal to 64, and then clipped to the range $(0, 255]$, so as to homogenize illumination.

We define a Bernoulli mask variable m_{fn} with probability p. m_{fn} indicates whether v_{fn} is observed or not. In the presence of missing data, the observation model becomes

$$p(\boldsymbol{V}, \boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) = \prod_{fn} \left(\left(\mathbf{1}_{v_{fn} = \sum_k c_{k,fn}} \right) \times \prod_k p(c_{k,fn}|w_{fk}, h_{kn}) \right)^{m_{fn}}.$$

The algorithms of the previous sections can easily be derived using this observation model and the equations only slightly change. For example, in (17), one only needs to sum over the observed entries at the numerator and denominator of the update rule. This leads to more general algorithms for which the algorithms described in Sections III and IV for the completely observed data become special cases. We generated a mask $\boldsymbol{M} = \{m_{fn}\}$ with $p = 0.5$ and estimated $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ by MJLE and MMLE[4] with $K = 300$, $\alpha_k = 1$, $\beta_k = 1$. We used the same stopping criteria for both methods: the algorithms were exited when the relative increase in the likelihood is less than a threshold (10^{-6}) or the maximum number of iterations (4000) is reached. We repeated each experiment 5 times and reported the results of the run with the highest likelihood end value. We reconstructed the missing values using $\hat{\boldsymbol{V}} = \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}$ and computed the Peak Signal to Noise Ratio (PSNR) between original and reconstructed images which is defined as $20 \log_{10} \left(\frac{FP}{\|\boldsymbol{V} - \hat{\boldsymbol{V}}\|_2} \right)$, where P is the maximum possible pixel value (255 in this experiment). In Fig. 6, we present some examples of the faces in the dataset, their masked versions which were used as data in this experiment and the reconstructions with MMLE and MJLE, respectively. PSNR values for the reconstructions are displayed on top of them. The average PSNR values obtained are 27.2 dB with MMLE and 26.6 dB with MJLE. For 1640 faces (67% of the total number) the PSNR values of MMLE are higher. In general, the reconstructions obtained with MMLE are smoother and more pleasant to look at. This is because MJLE overfitted the data with $K = 300$ while in MMLE 230 out of 300 dictionary columns were assigned to very low values $(\|\boldsymbol{w}_k\| < 10^{-14})$.

We also investigated the effect of α_k on the performances of the methods. We repeated the same experiment with 11 values of α_k between 1 and 100. The PSNR values obtained with 10

[4] $\hat{\boldsymbol{H}}$ in MMLE was taken to be the mean of $q(\boldsymbol{H})$ at convergence.

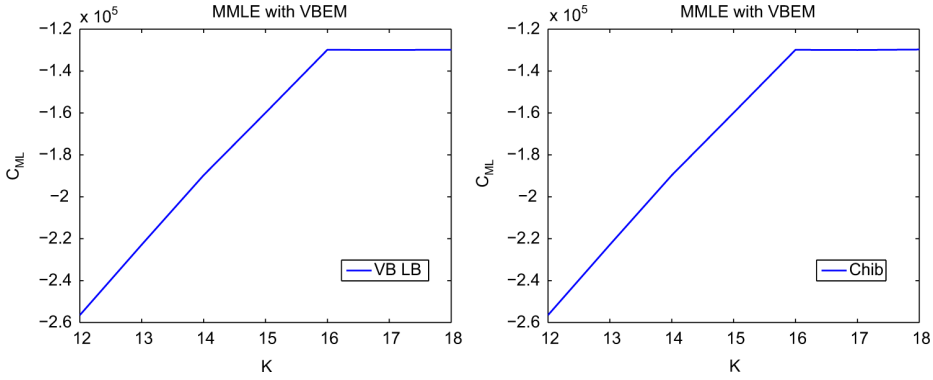


Fig. 4. Swimmer dataset. Marginal log-likelihood $C_{ML}(\hat{\boldsymbol{W}})$ for $K = 12, \ldots, 18$ estimated by VB lower bound (left), Chib's method (right). $\hat{\boldsymbol{W}}$ is obtained with VBEM (4000 iterations).

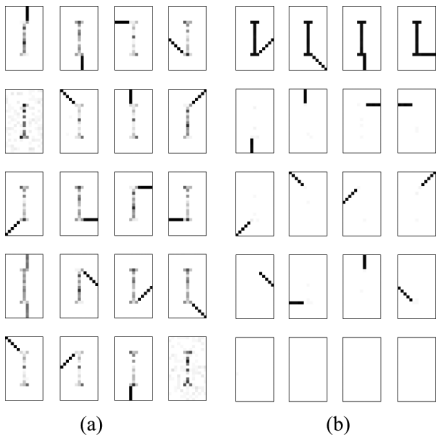


Fig. 5. Dictionaries learnt from the noisy Swimmer dataset with $K = 20$ (a) W_{MJLE} (b) W_{MMLE}.

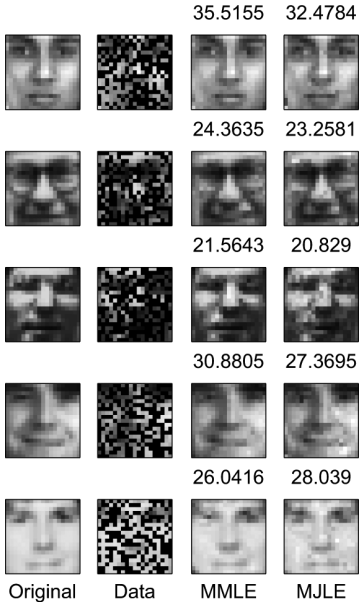


Fig. 6. Original faces from the CBCL dataset, masked data and reconstructions with MMLE and MJLE using 300 components. PSNR values (dB) are displayed on top of reconstructions.

repetitions of each method are presented in Fig. 7. As α_k increases, the means (and the variances) of the prior distributions

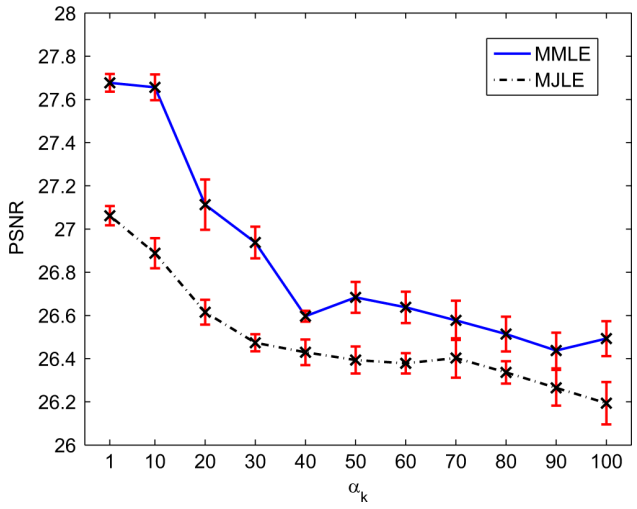


Fig. 7. Average PSNR values obtained with MMLE (solid) and MJLE (dotted) on the CBCL dataset with missing entries for various values of α_k in $[1, 100]$. Error bars represent standard deviations of 10 repetitions.

get bigger and more components (columns of $\boldsymbol{W}$) tend to converge to zero. The algorithm becomes more sensitive to initialization. Still, as can be seen from Fig. 7, the performance does not drop too drastically and MMLE gives consistently better results than MJLE. As a rule of thumb, $\alpha_k \in [1, 10]$ results in optimal degree of pruning in all of our experiments. Optimization of the hyperparameter α_k can be advantageous, but will make the algorithm slower than fixing it to a "good" value.

D. Musixmatch Lyrics Dataset

MusiXmatch (Million Song Dataset) [13] is a lyrics database with more than 230,000 songs. Each song constitutes a column of the data in a bag-of-words representation. The number of occurrences of the most common 5,000 words are used as the feature set. These 5,000 words are stemmed, i.e., related words are mapped onto their roots and cover 92% of all words that appear in the lyrics. The dataset also contains lyrics in other languages like Spanish, German, French, etc. We estimated dictionaries from a random subset of the dataset ($N = 10{,}000$) using MJLE, MMLE and LDA [10] with $K = 200$ components and $N_{iter} = 1000$ iterations. After estimating the dictionaries, in order to reconstruct components we estimated the expansion

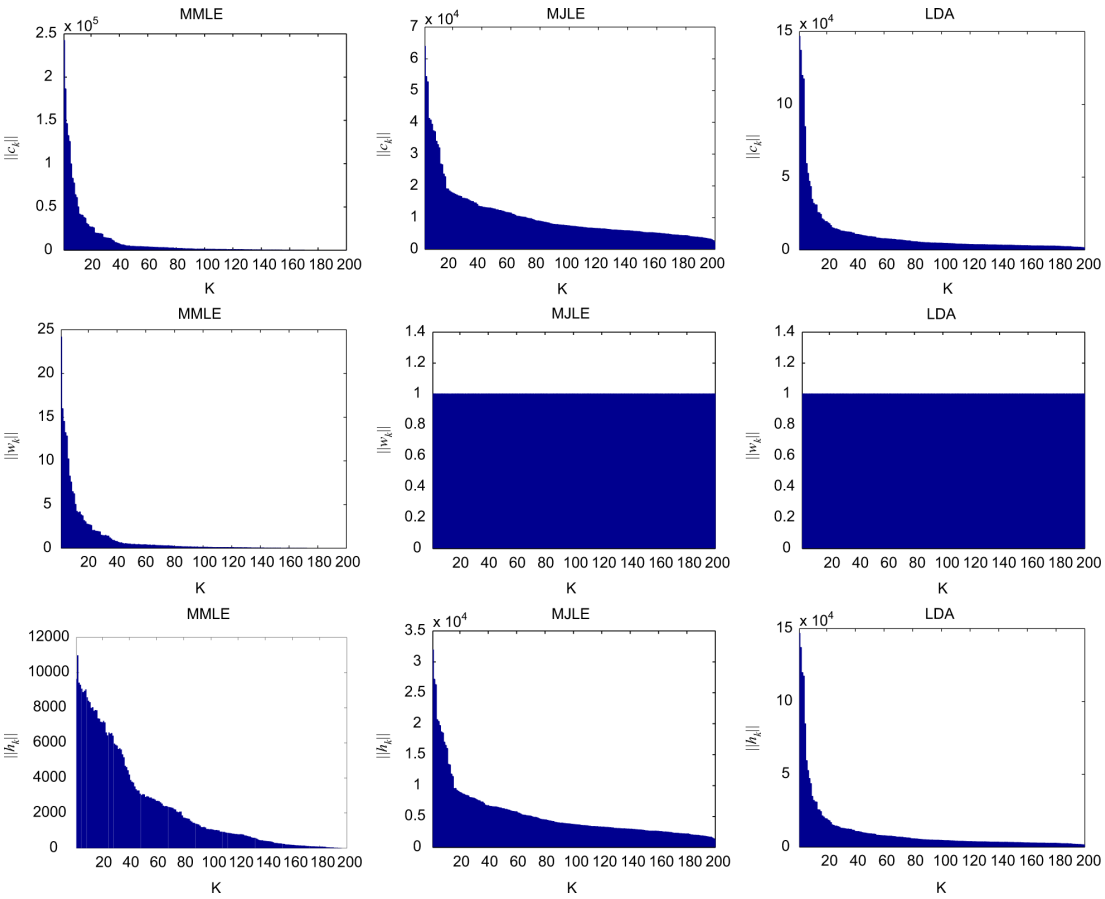


Fig. 8. MusiXmatch dataset analysis. Norms of components $\|\boldsymbol{C}_k\|$ (top), dictionary vectors $\|\boldsymbol{w}_k\|$ (middle) and activation vectors $\|\boldsymbol{h}_k\|$ (bottom), estimated by MMLE/VBEM (left) and MJLE (center) and LDA (right). $N_{iter} = 1000$, $K = 200$.

coefficients $\hat{\boldsymbol{H}}$ using the MAP estimator from corresponding models. The MAP estimator for $\boldsymbol{H}$ is a by product of the MJLE algorithm. For MMLE and LDA, we learnt them from GaP and LDA models with dictionaries set to their previously estimated values. The norms of dictionary vectors, $\|\hat{\boldsymbol{w}}_k\|$, expansion vectors, $\|\hat{\boldsymbol{h}}_k\|$, and components, $\|\hat{\boldsymbol{C}}_k\|$, are presented in Fig. 8.[5] The components are reconstructed using the posterior mean

$$\hat{c}_{k,fn} = \frac{\hat{w}_{fk}\hat{h}_{kn}}{\sum_j \hat{w}_{fj}\hat{h}_{jn}} v_{fn} \tag{23}$$

and components in all of these plots are sorted in descending order of $\|\hat{\boldsymbol{C}}_k\|$.

With MMLE/VBEM (left column in Fig. 8), we observe that 58 out of $K = 200$ dictionary columns have a norm less than 0.05, which is similarly reflected in the component norms. In MJLE and LDA, $\|\hat{\boldsymbol{w}}_k\|$ are not very informative because all dictionary columns sum to one by design and norms of the activation vectors, $\|\hat{\boldsymbol{h}}_k\|$, are also equal to those of the components, $\|\hat{\boldsymbol{C}}_k\|$. Comparing the components of MMLE, MJLE and LDA, we can see that MMLE explains the data making use of less components than MJLE and LDA. LDA also has this tendency but it is impossible to prune out components because every column of $\boldsymbol{W}$ is a discrete probability distribution and sum up to one.

For illustration, we now describe the results of decomposition with MMLE. In order to find the most representative songs of a topic, we propose to investigate the contribution of each topic in songs. We define the contribution of a topic as the number of words generated from that topic divided by the total number of words in that song

$$l_{kn} = \frac{\sum_f \hat{c}_{k,fn}}{\sum_k \sum_f \hat{c}_{k,fn}}.$$

The contributions estimated by MMLE are presented in Fig. 9. The components (rows) are again sorted in descending order of $\|\hat{\boldsymbol{C}}_k\|$. First 25–30 components mainly contain stop words and are found in most of the songs. The sparser rows among these (e.g., $k = 7, 13, 16\ldots$) belong to languages other than English. Manual inspection on the dictionary reveals that columns of the dictionary actually may correspond to themes of the songs, such as love, sex, religion, war, etc. In Table I, we present the most representative songs for a selection of components, i.e., songs for which l_{kn} are highest for a given k. We also present the most important words of the components and the most frequently used words in the songs. The importance, z_{fk}, is related

[5] $\hat{\boldsymbol{w}}_k = \{\hat{w}_{fk}\}_f$ and $\hat{\boldsymbol{h}}_k = \{\hat{h}_{kn}\}_n$ denote vectors of size F and N, respectively. $\hat{\boldsymbol{C}}_k = \{\hat{c}_{k,fn}\}_{fn}$ is a matrix of size $F \times N$.

TABLE I
MOST REPRESENTATIVE SONGS FOR FOUR OF THE COMPONENTS AND WORDS THAT APPEAR MOST FREQUENTLY.

($k = 2$) get nigga the ya shit like fuck em got hit bitch up off yall ass they that cmon money and	
UGK (Underground Kingz) - Murder	i the to nigga my a you got murder and it is am from we so with they yo cuz
Big Punisher - Nigga Shit	shit that nigga the i and my what to out am in on for love me with gettin you do
E-40 - Turf Drop [Clean]	gasolin the my i hey to a it on you some fuck spit of what one ride nigga sick gold
Cam'Ron - Sports Drugs & Entertainment	a the you i got yo stop shot is caus or street jump short wick either to on but in
Foxy Brown - Chyna Whyte	the nigga and you shit i not yall to a on with bitch no fuck uh it money white huh
($k = 8$) god of blood soul death die fear pain hell power within shall earth blind human bleed scream evil holi peac	
Demolition Hammer - Epidemic Of Violence	of pain death reign violenc and a kill rage vicious the to in down blue dead cold
Disgorge - Parallels Of Infinite Torture	of the tortur by their within upon flow throne infinit are no they see life eye befor
Tacere - Beyond Silence	silenc beyond a dark beauti i the you to and me it not in my is of your that do
Cannibal Corpse - Perverse Suffering	to my pain of i me for agoni in by and from way etern lust tortur crave the not be
Showbread - Sampsa Meets Kafka	to of no one die death loneli starv i the you and a me it not in my is your
($k = 26$) she her girl beauti woman & queen sex sexi cloth herself doll shes pink gypsi bodi midnight callin dress hair	
Headhunter - Sex & Drugs & Rock'N Roll	& sex drug rock roll n is good veri inde and not my are all need dead bodi brain i
Holy Barbarians - She	she of kind girl my is the a littl woman like world and gone destroy tiger me on an
X - Devil Doll	devil doll her she and a the in is of eye bone & shoe rag batter you to on no
Kittie - Paperdoll	her she you i now soul pain to is down want eat fit size and not in all dead bodi
Ottawan - D.I.S.C.O.	is she oh disco i o s d c super incred a crazi such desir sexi complic special candi
($k = 13$) je et les le pas dan pour des cest qui de tout mon moi au comm ne sur jai	
Veronique Sanson - Feminin	cest comm le car de bien se les mai a fait devant heur du et une quon quelqu etre
Nevrotic Explosion - Heritage	quon faut mieux pour nous qui nos ceux de la un plus tous honor parent ami oui
Kells - Sans teint	de la se le san des est loin peur reve pour sa sang corp lumier larm
Stille Volk - Corps Magicien	de les ell dan la se le du pass est sa par mond leur corp vivr lair voyag feu
Florent Pagny - Tue-Moi	si plus que un tu mon mes jour souvenir parc

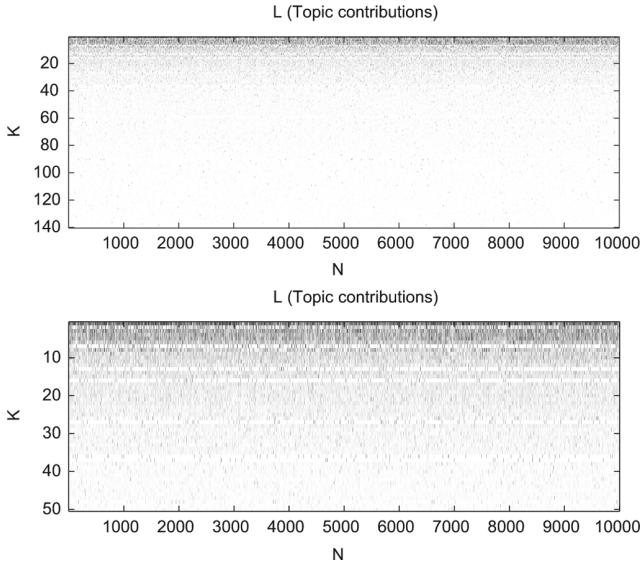


Fig. 9. Contribution matrix $\boldsymbol{L}$ estimated with MMLE on a the MusiXmatch dataset. Whole matrix (top), first 50 rows zoomed in (bottom). l_{kn} represents the contribution of kth topic in nth song. Higher values are represented by darker colors.

to how well that word is represented in that column and is defined by

$$z_{fk} = \frac{(\hat{w}_{fk} - \bar{w}_f)^2}{\bar{w}_f} \mathbf{1}_{\{\hat{w}_{fk} > \bar{w}_f\}} \tag{24}$$

where $\bar{w}_f$ denotes the average over components, i.e., $\bar{w}_f = \frac{1}{K}\sum_{k=1}^{K} \hat{w}_{fk}$.

Finally, we analyze lyrics of two songs by considering the topics of which contribution, l_{kn} for fixed n, are the highest. In Table II, we display ten highest contributing dictionary columns for the song "Do You Love Me?" by Nick Cave and the Bad Seeds. Each column is represented with the most important words according to (24). The same information for "California Love" by 2pac is presented in Table III. It is interesting to notice how different the significant dictionaries are in two songs which both contain "love" in the title.

VI. DISCUSSION AND CONCLUSION

In this paper we have challenged the standard NMF approach to nonnegative dictionary learning, which is based on maximum *joint* likelihood estimation (MJLE) and is ill-posed because the number of parameters to be learned increases with the data. Our approach here is maximum *marginal* likelihood estimation (MMLE), for which we proposed two EM algorithms, VBEM and MCEM. While both of these EM algorithms depend on approximating the functional in the E-step, VBEM maximizes a lower bound of the marginal likelihood, whereas MCEM is asymptotically exact. The same is true for the evaluation of the marginal likelihood, i.e., the approximation with Chib's method is asymptotically exact. Our experiments on synthetical data showed that the dictionaries estimated by these two EM algorithms show similar characteristics. In addition, the variational lower bound was tight with the results obtained with Chib's method.

Experiments on real and synthetical data have brought up a very attractive feature of MMLE, the self-ability of discarding "irrelevant" columns from the dictionary, i.e., performing automatic model order selection. This property is not by design as in other works of automatic relevance determination (e.g., [27], [28]), but stems from the objective function. The dictionaries estimated with MMLE lead to more accurate and interpretable components with no overfitting. In contrast with other model selection approaches in fully Bayesian settings (e.g., [9], [17]), which are based on the evaluation of the model evidence for every candidate value of K, our approach only requires to set K to a sufficiently large value and run the VBEM algorithm

TABLE II
TEN HIGHEST CONTRIBUTING DICTIONARY COLUMNS FOR "DO YOU LOVE ME?" BY NICK CAVE AND THE BAD SEEDS. EACH COLUMN IS REPRESENTED BY 15 MOST IMPORTANT WORDS. THE LAST ROW DISPLAYS THE CONTRIBUTION VALUE, l_{kn}, FOR THE CORRESPONDING DICTIONARY COLUMN. THESE 10 COLUMNS COVER 75% OF THE SONG.

the	you	love	not	i	me	she	god	was	so
in	your	buy	do	am	give	her	of	would	to
and	can	liar	wanna	myself	tell	girl	blood	could	for
of	if	tender	care	like	call	beauti	soul	were	now
world	know	dear	bad	know	mmm	woman	death	said	here
with	want	instrument	nobodi	need	show	&	die	had	again
they	make	mood	anyth	want	beg	queen	fear	thought	wait
from	when	treasur	want	feel	rescu	sex	pain	wish	long
as	see	emot	worri	caus	teas	sexi	hell	knew	too
by	yourself	untru	ai	and	squeez	cloth	power	came	home
at	need	surrend	treat	out	everytim	herself	within	made	and
out	with	deeper	but	sorri	knee	doll	shall	told	much
to	feel	sparkl	know	see	strife	shes	earth	took	alon
into	that	sweetest	money	in	contempl	pink	blind	saw	still
sky	how	diamond	hurt	swear	guarante	gypsi	human	then	how
0.15	0.10	0.09	0.08	0.08	0.08	0.06	0.04	0.04	0.03

TABLE III
TEN HIGHEST CONTRIBUTING DICTIONARY COLUMNS FOR "CALIFORNIA LOVE" BY 2 PAC. EACH COLUMN IS REPRESENTED BY 15 MOST IMPORTANT WORDS. THE LAST ROW DISPLAYS THE CONTRIBUTION VALUE, l_{kn}, FOR THE CORRESPONDING DICTIONARY COLUMN. THESE 10 COLUMNS COVER 87% OF THE SONG.

get	the	shake	it	you	we	come	yeah	around	i
nigga	in	motion	is	your	our	babi	five	goe	am
the	and	bump	take	can	us	magic	four	summer	myself
ya	of	groov	doe	if	togeth	til	woo	melt	like
shit	world	booti	make	know	both	lovin	summertim	wors	know
like	with	shakin	easi	want	higher	im	girlfriend	california	need
fuck	they	thigh	matter	make	ourselv	shi	wow	jone	want
em	from	oon	real	when	each	bodi	lala	scheme	feel
got	as	shiver	game	see	divid	sweat	grip	texa	caus
hit	by	panic	possibl	yourself	nation	cant	pine	dreamin	and
bitch	at	dick	play	need	unit	your	engin	screw	out
up	out	claw	chanc	with	other	birthday	feather	darker	sorri
off	to	opportun	give	feel	noel	wont	clap	careless	see
yall	into	collid	harder	that	standard	there	mornin	consol	in
ass	sky	ness	quit	how	rule	bella	gotta	giant	swear
0.49	0.09	0.08	0.07	0.03	0.03	0.02	0.02	0.02	0.02

once. This property is not shared by LDA although it is similar to MMLE in spirit. This is because in LDA the dictionary columns are constrained to sum to one.

The computational costs of MJLE and MMLE/VBEM are comparable, whereas MMLE/MCEM is very computationally demanding. Since two MMLE algorithms perform similarly, it is natural to use only VBEM on large datasets. In addition, VBEM can be made even faster by not updating the components which are already zero.

APPENDIX
LAPLACE APPROXIMATION OF $C_{ML}(\boldsymbol{W})$

In this section we give a Laplace approximation of the integral involved in $C_{ML}(\boldsymbol{W})$ whose expression provides an intuition for the pruning effect of MMLE. We can write that

$$C_{ML}(\boldsymbol{W}) = \sum_{n=1}^{N} \log \int_{\boldsymbol{h}_n} p(\boldsymbol{v}_n|\boldsymbol{W}\boldsymbol{h}_n)p(\boldsymbol{h}_n)d\boldsymbol{h}_n. \quad (25)$$

Let $\hat{\boldsymbol{H}}$ be the MAP estimation of $\boldsymbol{H}$ given $\boldsymbol{W}$, i.e.,

$$\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}} \log p(\boldsymbol{V}, \boldsymbol{H}|\boldsymbol{W}).$$

Then, a Laplace approximation of $C_{ML}(\boldsymbol{W})$ around its mode (which essentially consists in replacing the integrand in each term of the sum in (25) by a quadratic function with same mode and curvature at the mode) is given by

$$C_{ML}(\boldsymbol{W}) \approx C_{JL}(\boldsymbol{W}, \hat{\boldsymbol{H}}) - \frac{1}{2}\sum_{n} \log\det \boldsymbol{A}_n + \frac{KN}{2}\log 2\pi \quad (26)$$

where

$$\begin{aligned}\boldsymbol{A}_n &= -\nabla^2_{\boldsymbol{h}_n} \log p(\boldsymbol{v}_n, \boldsymbol{h}_n|\boldsymbol{W})\big|_{\boldsymbol{h}_n=\hat{\boldsymbol{h}}_n} \\ &= \boldsymbol{W}^T\boldsymbol{\Gamma}_{1,n}\boldsymbol{W} + \boldsymbol{\Gamma}_{2,n}\end{aligned}$$

where $\boldsymbol{\Gamma}_{1,n}$ and $\boldsymbol{\Gamma}_{2,n}$ are the diagonal matrices defined by

$$\begin{aligned}\boldsymbol{\Gamma}_{1,n} &= \mathrm{diag}[\boldsymbol{v}_n.(\boldsymbol{W}\hat{\boldsymbol{h}}_n^{.-2})] \\ \boldsymbol{\Gamma}_{2,n} &= \mathrm{diag}[(\boldsymbol{\alpha}-1).\hat{\boldsymbol{h}}_n^{.-2}]\end{aligned}$$

and where the '.' denotes MATLAB-like entry-wise operations. The penalty term $L(\boldsymbol{W}) = \sum_n \log\det \boldsymbol{A}_n$ in (26) will favor solutions such that $\det \boldsymbol{A}_n$ is small, ideally zero. A detailed analysis of $L(\boldsymbol{W})$, not presented here, reveals that it induces group-sparsity at the column level. This for example evident when $\alpha_k = 1$ and thus $\boldsymbol{\Gamma}_{2,n} = \boldsymbol{0}$. In this case, any zero column in $\boldsymbol{W}$ leads to $\det \boldsymbol{A}_n = 0$. While not giving a rigorous explanation, the Laplace approximation of $C_{ML}(\boldsymbol{W})$ hence suggests why MMLE induces self-regularization by pruning.

ACKNOWLEDGMENT

The authors would like to thank O. Cappé, A. Taylan Cemgil, and J. Le Roux for inspiring discussions related to this work.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[2] W. H. Richardson, "Bayesian-based iterative method of image restoration," *J. Opt. Soc. Amer.*, vol. 62, pp. 55–59, 1972.
[3] L. B. Lucy, "An iterative technique for the rectification of observed distributions," *Astron. J.*, vol. 79, pp. 745–754, 1974.
[4] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, pp. 337–365, 2000.
[5] J. F. Canny, "GaP: A factor model for discrete data," in *Proc. 27th ACM Int. Conf. Res. Develop. Inform. Retrieval (SIGIR)*, Sheffield, U.K., 2004, pp. 122–129.
[6] R. Nallapati, W. W. Cohen, S. Ditmore, J. Lafferty, and K. Ung, "Multiscale topic tomography," in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining (KDD'07)*, San Jose, CA, 2007.
[7] M. Titsias, "The infinite gamma-poisson feature model," presented at the Adv. Neural Inform. Process. Syst. (NIPS'07), Vancouver, BC, Canada, 2007.
[8] W. L. Buntine and A. Jakulin, "Discrete component analysis," *Lecture Notes in Comput. Sci.*, vol. 3940, pp. 1–33, 2006.
[9] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, p. 17, 2009, Article ID 785152, doi: 10.1155/2009/785152.
[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn.Res.*, vol. 3, pp. 993–1022, 2003.
[11] T. Hofmann, "Probabilistic latent semantic indexing," presented at the 22nd Int. Conf. Res. Develop. Inform. Retrieval (SIGIR), Berkeley, CA, 1999.
[12] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1313–1321, 1995.
[13] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," presented at the 12th Int. Soc. Music Inform. Retrieval Conf. (ISMIR'11), Miami, FL, 2011.
[14] O. Dikmen and C. Févotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning," in *Proc. Int. Conf. Acoust. Speech Signal Process.(ICASSP'11)*, Prague, Czech Republic, 2011.
[15] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th Eur. Signal Process. Conf. (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 1913–1917.
[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Comput.*, vol. 23, no. 9, Sep. 2011.
[17] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. 8th Int. Conf. Independent Component Anal. Signal Separation (ICA'09)*, Paraty, Brazil, Mar. 2009.
[18] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Budapest, Hungary, 2004, pp. 2529–2533.
[19] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011.
[20] A. P. Dempster, N. M. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 1, no. 39, pp. 1–38, 1977.
[21] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures," in *Bayesian Statistics 7*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Eds. London, U.K.: Oxford Univ. Press, 2003.
[22] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Amer. Statist. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.
[23] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Jun. 1984.
[24] E. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[25] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.
[26] B. Heisele, T. Poggio, and M. Pontil, "Face detection in still gray images," Center for Biological and Computational Learning, MIT, Cambridge, MA, A.I. Memo 1687, 2000.
[27] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. Workshop Signal Process. Adaptative Sparse Structured Representations (SPARS)*, St-Malo, France, 2009.
[28] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Low-rank matrix completion by variational sparse bayesian learning," in *Proc. Int. Conf. Acoust., Speech. Signal Process. (ICASSP'11)*, Prague, Czech Republic, 2011, pp. 2188–2111.
[29] 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, Mar. 2009.

Onur Dikmen (M'09) received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Bogaziçi University, Istanbul, Turkey.

He worked at Télécom ParisTech, France, as a CNRS Research Associate. He is currently with the Department of Information and Computer Science at Aalto University, Finland. His research interests include statistical signal processing, Bayesian statistics, and approximate inference. He works on Bayesian source modeling and nonnegative matrix factorization for source separation.

Cédric Févotte (M'09) received the State Engineering degree and the Ph.D. degree in control and computer science from École Centrale de Nantes, Nantes, France, in 2000 and 2003, respectively.

As a PhD. student he was with the Signal Processing Group at Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN) where he worked on time-frequency approaches to blind source separation. From 2003 to 2006, he was a Research Associate with the Signal Processing Laboratory at University of Cambridge (Engineering Deptartment) where he worked on Bayesian approaches to sparse component analysis with applications to audio source separation. He was then a Research Engineer with the start-up company Mist-Technologies (now Audionamix) in Paris, designing mono/stereo to 5.1 surround sound upmix solutions. In Mar. 2007, he joined Télécom ParisTech, first as a Research Associate and then as a CNRS Tenured Research Scientist in November, 2007. His research interests generally concern statistical signal processing and unsupervised machine learning and, in particular, applications to blind source separation and audio signal processing. He is the scientific leader of project TANGERINE (Theory and applications of nonnegative matrix factorization) funded by the French research funding agency ANR.

Dr. Févotte is a Member of the IEEE "Machine learning for signal processing" technical committee.

(Dikmen & Févotte, *NIPS*, 2011)

Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation

Onur Dikmen
CNRS LTCI; Télécom ParisTech
75014, Paris, France
dikmen@telecom-paristech.fr

Cédric Févotte
CNRS LTCI; Télécom ParisTech
75014, Paris, France
fevotte@telecom-paristech.fr

Abstract

In this paper we describe a maximum likelihood approach for dictionary learning in the multiplicative exponential noise model. This model is prevalent in audio signal processing where it underlies a generative composite model of the power spectrogram. Maximum joint likelihood estimation of the dictionary and expansion coefficients leads to a nonnegative matrix factorization problem where the Itakura-Saito divergence is used. The optimality of this approach is in question because the number of parameters (which include the expansion coefficients) grows with the number of observations. In this paper we describe a variational procedure for optimization of the marginal likelihood, i.e., the likelihood of the dictionary where the activation coefficients have been integrated out (given a specific prior). We compare the output of both maximum joint likelihood estimation (i.e., standard Itakura-Saito NMF) and maximum marginal likelihood estimation (MMLE) on real and synthetical datasets. The MMLE approach is shown to embed automatic model order selection, akin to automatic relevance determination.

1 Introduction

In this paper we address the task of nonnegative dictionary learning described by

$$V \approx WH, \tag{1}$$

where V, W, H are nonnegative matrices of dimensions $F \times N$, $F \times K$ and $K \times N$, respectively. V is the data matrix, where each column v_n is a data point, W is the dictionary matrix, with columns $\{w_k\}$ acting as "patterns" or "explanatory variables" reprensentative of the data, and H is the activation matrix, with columns $\{h_n\}$. For example, in this paper we will be interested in music data such that V is time-frequency spectrogram matrix and W is a collection of spectral signatures of latent elementary audio components. The most common approach to nonnegative dictionary learning is nonnegative matrix factorization (NMF) [1] which consists in retrieving the factorization (1) by solving

$$\min_{W,H} D(V|WH) \stackrel{\text{def}}{=} \sum\nolimits_{fn} d(v_{fn}|[WH]_{fn}) \quad \text{s.t.} \quad W, H \geq 0 \,, \tag{2}$$

where $d(x|y)$ is a measure of fit between nonnegative scalars, v_{fn} are the entries of V, and $A \geq 0$ expresses nonnegativity of the entries of matrix A. The cost function $D(V|WH)$ is often a likelihood function $-\log p(V|W,H)$ in disguise, e.g., the Euclidean distance underlies additive Gaussian noise, the Kullback-Leibler (KL) divergence underlies Poissonian noise, while the Itakura-Saito (IS) divergence underlies multiplicative exponential noise [2]. The latter noise model will be central to this work because it underlies a suitable generative model of the power spectrogram, as shown in [3] and later recalled.

A criticism about NMF is that little can be said about the asymptotical optimality of the learnt dictionary W. Indeed, because W is estimated jointly with H, the total number of parameters $FK + KN$ grows with the number of data points N. As such, this paper instead addresses optimization of the likelihood in the marginal model described by

$$p(V|W) = \int_H p(V|W,H)p(H)dH, \tag{3}$$

where H is treated as a random latent variable with prior $p(H)$. The evaluation and optimization of the marginal likelihood is not trivial in general, and this paper is precisely devoted to these tasks in the multiplicative exponential noise model.

The maximum marginal likelihood estimation approach we seek here is related to IS-NMF in such a way that Latent Dirichlet Allocation (LDA) [4] is related to Latent Semantic Indexing (pLSI) [5]. LDA and pLSI are two estimators in the same model, but LDA seeks estimation of the topic distributions in the marginal model, from which the topic weights describing each document have been integrated out. In contrast, pLSI (which is essentially equivalent to KL-NMF as shown in [6]) performs maximum *joint* likelihood estimation (MJLE) for the topics and weights. Blei *et al*. [4] show the better performance of LDA with respect to (w.r.t) pLSI. Welling *et al*. [7] also report similar results with a discussion, stating that deterministic latent variable models assign zero probability to input configurations that do not appear in the training set. A similar approach is Discrete Component Analysis (DCA) [8] which considers maximum marginal a posteriori estimation in the Gamma-Poisson (GaP) model [9], see also [10] for the maximum marginal likelihood estimation on the same model. In this paper, we will follow the same objective for the multiplicative exponential noise model.

We will describe a variational algorithm for the evaluation and optimization of (3); note that the algorithm exploits specificities of the model and is not a mere adaptation of LDA or DCA to an alternative setting. We will consider a nonnegative Generalized inverse-Gaussian (GIG) distribution as a prior for H, a flexible distribution which takes the Gamma and inverse-Gamma as special cases. As will be detailed later, this work relates to recent work by Hoffman *et al*. [11], which considers full Bayesian integration of W and H (both assumed random) in the exponential noise model, in a nonparametric setting allowing for model order selection. We will show that our more simple maximum likelihood approach inherently performs model selection as well by automatically pruning "irrelevant" dictionary elements. Applied to a short well structured piano sequence, our approach is shown to capture the correct number of components, corresponding to the expected note spectra, and outperforms the nonparametric Bayesian approach of [11].

The paper is organized as follows. Section 2 introduces the multiplicative exponential noise model with the prior distribution for the expansion coefficients $p(H)$. Sections 3 and 4 describe the MJLE and MMLE approaches, respectively. Section 5 reports results on synthetical and real audio data. Section 6 concludes.

2 Model

The generative model assumed in this paper is

$$v_{fn} = \hat{v}_{fn} \cdot \epsilon_{fn} \ , \tag{4}$$

where $\hat{v}_{fn} = \sum_k w_{fk}h_{kn}$ and ϵ_{fn} is a nonnegative multiplicative noise with exponential distribution $\epsilon_{fn} \sim \exp(-\epsilon_{fn})$. In other words, and under independence assumptions, the likelihood function is

$$p(V|W,H) = \prod\nolimits_{fn} (1/\hat{v}_{fn}) \exp(-v_{fn}/\hat{v}_{fn}) \, . \tag{5}$$

When V is a power spectrogram matrix such that $v_{fn} = |x_{fn}|^2$ and $\{x_{fn}\}$ are the complex-valued short-time Fourier transform (STFT) coefficients of some signal data, where f typically acts as a frequency index and n acts as a time-frame index, it was shown in [3] that an equivalent generative model of v_{fn} is

$$x_{fn} = \sum\nolimits_k c_{fkn}, \quad c_{fkn} \sim \mathcal{N}_c(0, w_{fk}h_{kn}) \ , \tag{6}$$

where $\mathcal{N}_c$ refers to the circular complex Gaussian distribution.[1] In other words, the exponential multiplicative noise model underlies a generative composite model of the STFT. The complex-valued matrix $\{c_{fkn}\}_{fn}$, referred to as k^{th}component, is characterized by a spectral signature w_k, amplitude-modulated in time by the frame-dependent coefficient h_{kn}, which accounts for nonstationarity. In analogy with LDA or DCA, if our data consisted of word counts, with f indexing words and n indexing documents, then the columns of W would describe topics and c_{fkn} would denote the number of occurrences of word f stemming from topic k in document n.

In our setting W is considered a free deterministic parameter to be estimated by maximum likelihood. In contrast, H is treated as a nonnegative random latent variable over which we will integrate. It is assigned a GIG prior, such that

$$h_{kn} \sim \mathcal{GIG}(\alpha_k, \beta_k, \gamma_k) \;, \tag{7}$$

with

$$\mathcal{GIG}(x|\alpha,\beta,\gamma) = \frac{(\beta/\gamma)^{\alpha/2}}{2\mathcal{K}_\alpha(2\sqrt{\beta\gamma})} x^{\alpha-1} \exp - \left(\beta x + \frac{\gamma}{x}\right) \;, \tag{8}$$

where $\mathcal{K}$ is a modified Bessel function of the second kind and x, β and γ are nonnegative scalars. The GIG distribution unifies the Gamma ($\alpha > 0$, $\gamma = 0$) and inverse-Gamma ($\alpha < 0$, $\beta = 0$) distributions. Its sufficient statistics are x, $1/x$ and $\log x$, and in particular we have

$$\langle x \rangle = \frac{\mathcal{K}_{\alpha+1}(2\sqrt{\beta\gamma})}{\mathcal{K}_\alpha(2\sqrt{\beta\gamma})}\sqrt{\frac{\gamma}{\beta}}, \qquad \left\langle \frac{1}{x} \right\rangle = \frac{\mathcal{K}_{\alpha-1}(2\sqrt{\beta\gamma})}{\mathcal{K}_\alpha(2\sqrt{\beta\gamma})}\sqrt{\frac{\beta}{\gamma}}, \tag{9}$$

where $\langle x \rangle$ denotes expectation. Although all derivations and the implementations are done for the general case, in practice we will only consider the special case of Gamma distribution for simplicity. In such case, β parameter merely acts as a scale parameter, which we fix so as to solve the scale ambiguity between the columns of W and the rows of H. We will also assume the shape parameters $\{\alpha_k\}$ fixed to arbitrary values (typically, $\alpha_k = 1$, which corresponds to the exponential distribution). Given the generative model specified by equations (4) and (7) we now describe two estimators for W.

3 Maximum joint likelihood estimation

3.1 Estimator

The joint (penalized) log-likelihood likelihood of W and H is defined by

$$C_{\mathrm{JL}}(W,H) \stackrel{\text{def}}{=} \log p(V|W,H) + \log p(H) \tag{10}$$

$$= -D_{\mathrm{IS}}(V|WH) - \sum\nolimits_{kn} (1-\alpha_k)\log h_{kn} + \beta_k h_{kn} + \gamma_k/h_{kn} + \mathrm{cst} \;, \tag{11}$$

where $D_{\mathrm{IS}}(V|WH)$ is defined as in equation (2) with $d_{\mathrm{IS}}(x|y) = x/y - \log(x/y) - 1$ (Itakura-Saito divergence) and "cst" denotes terms constant w.r.t W and H. The subscript JL stands for joint likelihood, and the estimation of W by maximization of $C_{\mathrm{JL}}(W,H)$ will be referred to as *maximum joint likelihood estimation* (MJLE).

3.2 MM algorithm for MJLE

We describe an iterative algorithm which sequentially updates W given H and H given W. Each of the two steps can be achieved in a *minorization-maximization* (MM) setting [12], where the original problem is replaced by the iterative optimization of an easier-to-optimize auxiliary function. We first describe the update of H, from which the update of W will be easily deduced. Given W, our task consists in maximizing $C(H) = -D_{\mathrm{IS}}(V|WH) - L(H)$, where $L(H) = \sum_{kn}(1-\alpha_k)\log h_{kn} + \beta_k h_{kn} + \gamma_k/h_{kn}$. Using Jensen's inequality to majorize the convex part of $D_{\mathrm{IS}}(V|WH)$ (terms in

[1] A complex random variable has distribution $\mathcal{N}_c(\mu,\lambda)$ if and only if its real and imaginary parts are independent and distributed as $\mathcal{N}(\Re(\mu),\lambda/2)$ and $\mathcal{N}(\Im(\mu),\lambda/2)$, respectively.

$v_{fn}/\hat{v}_{fn}$) and first order Taylor approximation to majorize its concave part (terms in $\log \hat{v}_{fn}$), as in [13], the functional

$$G(H, \tilde{H}) = -\left(\sum\nolimits_k p_{kn}/h_{kn} + q_{kn}h_{kn}\right) - L(H) + \text{cst}, \tag{12}$$

where $p_{kn} = \tilde{h}_{kn}^2 \sum_f w_{fk}v_{fn}/\tilde{v}_{fn}^2$, $q_{kn} = \sum_f w_{fk}/\tilde{v}_{fn}$, $\tilde{v}_{fn} = [W\tilde{H}]_{fn}$, can be shown to be a tight lower bound of $C(H)$, i.e., $G(H, \tilde{H}) \leq C(H)$ and $G(\tilde{H}, \tilde{H}) = C(\tilde{H})$. Its iterative maximization w.r.t H, where $\tilde{H} = H^{(i)}$ acts as the current iterate at iteration i, produces an ascent algorithm, such that $C(H^{(i+1)}) \geq C(H^{(i)})$. The update is easily shown to amount to solving an order 2 polynomial with a single positive root given by

$$h_{kn} = \frac{(\alpha_k - 1) + \sqrt{(\alpha_k - 1)^2 + 4(p_{kn} + \gamma_k)(q_{kn} + \beta_k)}}{2(q_{kn} + \beta_k)}. \tag{13}$$

The update preserves nonnegativity given positive initialization. By exchangeability of W and H when the data is transposed ($V^T = H^T W^T$), and dropping the penalty term ($\alpha_k = 1$, $\beta_k = 0$, $\gamma_k = 0$), the update of W is given by the multiplicative update

$$w_{fk} = \tilde{w}_{fk} \sqrt{\frac{\sum_n h_{kn} v_{fn}/\tilde{v}_{fn}^2}{\sum_n h_{kn}/\tilde{v}_{fn}}}, \tag{14}$$

which is known from [13].

4 Maximum marginal likelihood estimation

4.1 Estimator

We define the marginal log-likelihood objective function as

$$C_{\text{ML}}(W) \stackrel{\text{def}}{=} \log \int p(V|W, H)p(H)\, dH\,. \tag{15}$$

The subscript ML stands for marginal likelihood, and the estimation of W by maximization of $C_{\text{ML}}(W)$ will be referred to as *maximum marginal likelihood estimation* (MMLE). Note that in Bayesian estimation the term *marginal likelihood* is sometimes used as a synonym for the *model evidence*, which is the likelihood of data given the model, i.e., where all random parameters (including W) have been marginalized. This is not the case here where W is treated as a deterministic parameter and marginal likelihood only refers to the likelihood of W, where H has been integrated out. The integral in equation (15) is intractable given our model. In the next section we resort to a variational Bayes procedure for the evaluation and maximization of $C_{\text{ML}}(W)$.

4.2 Variational algorithm for MMLE

In the following we propose an iterative lower bound evaluation/maximization procedure for approximate maximization of $C_{\text{ML}}(W)$. We will construct a bound $B(W, \tilde{W})$ such that $\forall (W, \tilde{W}), C_{\text{ML}}(W) \geq B(W, \tilde{W})$, where $\tilde{W}$ acts as the current iterate and W acts as the free parameter over which the bound is maximized. The maximization is approximate in that the bound will only satisfy $B(\tilde{W}, \tilde{W}) \approx C_{\text{ML}}(\tilde{W})$, i.e., is loosely tight in the current update $\tilde{W}$, which fails to ensure ascent of the objective function like in the MM setting of Section 3.2.

We propose to construct the bound from a variational Bayes perspective [14]. The following inequality holds for any distribution function $q(H)$

$$C_{\text{ML}}(W) \geq \langle \log p(V|W, H) \rangle_q + \langle \log p(H) \rangle_q - \langle \log q(H) \rangle_q \stackrel{\text{def}}{=} B_q^{\text{vb}}(W)\,. \tag{16}$$

The inequality becomes an equality when $q(H) = p(H|V, W)$; when the latter is available in close form, the EM algorithm consists in using $\tilde{q}(H) = p(H|V, \tilde{W})$ and maximize $B_{\tilde{q}}^{\text{vb}}(W)$ w.r.t W, and iterate. The true posterior of H being intractable in our case, we take $q(H)$ to be a factorized,

parametric distribution $q_\theta(H)$, whose parameter θ is updated so as to tighten $B_q^{\mathrm{vb}}(\tilde{W})$ to $C(\tilde{W})$. Like in [11], we choose $q_\theta(H)$ to be in the same family as the prior, such that

$$q_\theta(H) = \prod\nolimits_{kn} \mathcal{GIG}(\bar{\alpha}_{kn}, \bar{\beta}_{kn}, \bar{\gamma}_{kn})\,. \tag{17}$$

The first term of $B_q^{\mathrm{vb}}(W)$ essentially involves the expectation of $-D_{\mathrm{IS}}(V|WH)$ w.r.t to the variational distribution $q_\theta(H)$. The product WH introduces some coupling of the coefficients of H (via the sum $\sum_k w_{fk}h_{kn}$) which makes the integration difficult. Following [11] and similar to Section 3.2, we propose to lower bound this term using Jensen's and Taylor's type inequalities to majorize the convex and concave parts of $-D_{\mathrm{IS}}(V|WH)$. The contributions of the elements of H become decoupled w.r.t to k, which allows for evaluation and maximization of the bound. This leads to

$$\langle \log p(V|H,W)\rangle_q \geq -\sum_{fn}\left(\sum_k \phi_{fkn}^2 \frac{v_{fn}}{w_{fk}}\left\langle \frac{1}{h_{kn}}\right\rangle_q\right) + \left(\log\psi_{fn} + \frac{1}{\psi_{fn}}\sum_k w_{fk}\langle h_{kn}\rangle_q - 1\right), \tag{18}$$

where $\{\psi_{fn}\}$ and $\{\phi_{fkn}\}$ are nonnegative free parameters such that $\sum_k \phi_{fkn} = 1$. We define $B_{\theta,\phi,\psi}(W)$ as $B_q^{\mathrm{vb}}(W)$ but where the expectation of the joint log-likelihood is replaced by its lower bound given right side of equation (18). From there, our algorithm is a two-step procedure consisting in 1) computing $\tilde{\theta}, \tilde{\phi}, \tilde{\psi}$ so as to tighten $B_{\theta,\phi,\psi}(\tilde{W})$ to $C_{\mathrm{ML}}(\tilde{W})$, and 2) maximizing $B_{\tilde{\theta},\tilde{\phi},\tilde{\psi}}(W)$ w.r.t W. The corresponding updates are given next. Note that evaluation of the bound only involves expectations of h_{kn} and $1/h_{kn}$ w.r.t to the GIG distribution, which is readily given by equation (9).

Step 1: Tightening the bound Given current dictionary update $\tilde{W}$, run the following fixed-point equations.

$$\phi_{fkn} = \frac{\tilde{w}_{fk}/\langle 1/h_{kn}\rangle_q}{\sum_j \tilde{w}_{fj}/\langle 1/h_{jn}\rangle_q}, \qquad \psi_{fn} = \sum_j \tilde{w}_{fj}\langle h_{jn}\rangle_q$$

$$\bar{\alpha}_{kn} = \alpha_k, \qquad \bar{\beta}_{kn} = \beta_k + \sum_f \frac{\tilde{w}_{fk}}{\psi_{fn}}, \qquad \bar{\gamma}_{kn} = \gamma_k + \sum_f \frac{v_{fn}\phi_{fkn}^2}{\tilde{w}_{fk}}\,.$$

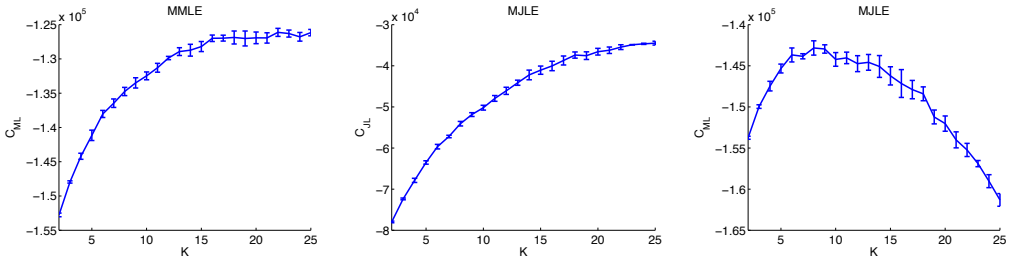
Step 2: Optimizing the bound Given the variational distribution $\tilde{q} = q_{\tilde{\theta}}$ from previous step, update W as

$$w_{fk} = \tilde{w}_{fk}\sqrt{\frac{\sum_n v_{fn}\left[\sum_j \tilde{w}_{fj}\langle 1/h_{jn}\rangle_{\tilde{q}}^{-1}\right]^{-2}\langle 1/h_{kn}\rangle_{\tilde{q}}^{-1}}{\sum_n\left[\sum_j \tilde{w}_{fj}\langle h_{jn}\rangle_{\tilde{q}}\right]^{-1}\langle h_{kn}\rangle_{\tilde{q}}}}\,. \tag{19}$$

The VB update has a similar form to the MM update of equation (14) but the contributions of H are replaced by expected values w.r.t the variational distribution.

4.3 Relation to other works

A variational algorithm using the activation matrix H and the latent components $C = \{c_{fkn}\}$ as hidden data can easily be devised, as sketched in [2]. Including C in the variational distribution also allows to decouple the contributions of the activation coefficients w.r.t to k but leads from our experience to a looser bound, a finding also reported in [11]. In a fully Bayesian setting, Hoffman *et al*. [11] assume Gamma priors for both W and H. The model is such that $\hat{v}_{fn} = \sum_k \lambda_k w_{fk}h_{kn}$, where λ_k acts as a component weight parameter. The number of components is potentially infinite but, in a nonparametric setting, the prior for λ_k favors a finite number of active components. Posterior inference of the parameters W, H, $\{\lambda_k\}$ is achieved in a variational setting similar to Section 4.2, by maximizing a lower bound on $p(V)$. In contrast to this method, our approach does not require to specify a prior for W, leads to simple updates for W that are directly comparable to IS-NMF and experiments will reveal that our approach embeds model order selection as well, by automatically pruning unnecessary columns of W, without resorting to the nonparametric framework.



(a) C_{ML} by MMLE (b) C_{JL} by MJLE (c) C_{ML} by MJLE

Figure 1: Marginal likelihood C_{ML} (a) and joint likelihood C_{JL} (b) versus number of components K. C_{ML} values corresponding to dictionaries estimated by C_{JL} maximization (c).

5 Experiments

In this section, we study the performances of MJLE and MMLE methods on both synthetical and real-world datasets.[2] The prior hyperparameters are fixed to $\alpha_k = 1$, $\gamma_k = 0$ (exponential distribution) and $\beta_k = 1$, i.e., $h_{kn} \sim \exp(-h_{kn})$. We used 5000 algorithm iterations and nonnegative random initializations in all cases. In order to minimize the odds of getting stuck in local optima, we adapted the deterministic annealing method proposed in [15] for MMLE. Deterministic annealing is applied by multiplying the entropy term $-\langle \log q(H) \rangle$ in the lower bound in (16) by $1/\eta^{(i)}$. The initial $\eta^{(0)}$ is chosen in $(0, 1)$ and increased through iterations. In our experiments, we set $\eta^{(0)} = 0.6$ and updated it with the rule $\eta^{(i+1)} = \min(1, 1.005\eta^{(i)})$.

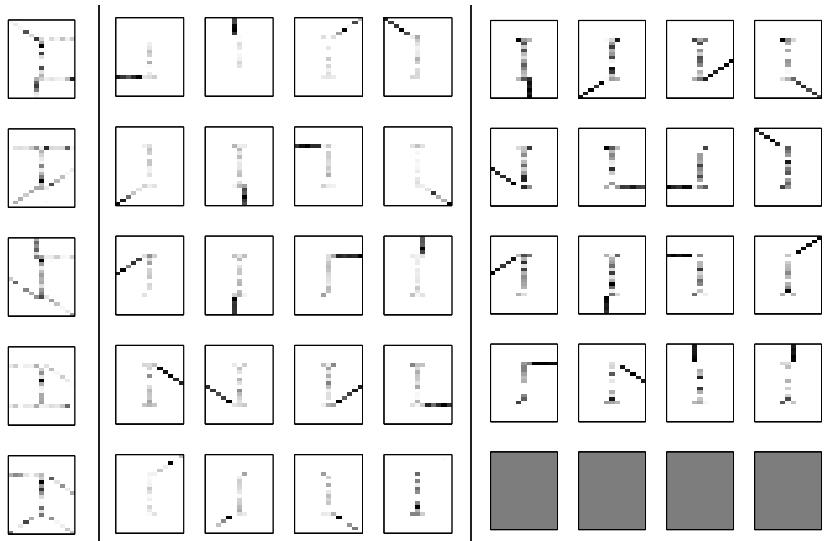
5.1 Swimmer dataset

First, we consider the synthetical `Swimmer` dataset [16], for which the ground truth of the dictionary is available. The dataset is composed of 256 images of size 32×32, representing a swimmer built of an invariant torso and 4 limbs. Each of the 4 limbs can be in one of 4 positions and the dataset is formed of all combinations. Hence, the ground truth dictionary corresponds to the collection of individual limb positions. As explained in [16] the torso is an unidentifiable component that can be paired with any of the limbs, or even split among the limbs. In our experiments, we mapped the values in the dataset onto the range $[1, 100]$ and multiplied with exponential noise, see some samples in Fig. 2 (a).

We ran the MM and VB algorithms (for MJLE and MMLE, respectively) for $K = 1 \ldots 20$ and the joint and marginal log-likelihood end values (after the 5000 iterations) are displayed in Fig. 1. The marginal log-likelihood is here approximated by its lower bound, as described in Section 4.2. In Fig. 1(a) and (b) the respective objective criteria (C_{ML} and C_{JL}) maximized by MMLE and MJLE are shown. The increase of C_{ML} stops after $K = 16$, whereas C_{JL} continues to increase as K gets larger. Fig. 1 (c) displays the corresponding marginal likelihood values, C_{ML}, of the dictionaries obtained by MJLE in Fig. 1 (b); this figure empirically shows that maximizing the joint likelihood does not necessarily imply maximization of the marginal likelihood. These figures display the mean and standard deviation values obtained from 7 experiments.

The likelihood values increase with the number of components, as expected from nested models. However, the marginal likelihood stagnates after $K = 16$. Manual inspection reveals that passed this value of K, the extra columns of W are pruned to zero, leaving the criterion unchanged. Hence, MMLE appears to embed automatic order selection, similar to automatic relevance determination [17, 18]. The dictionaries learnt from MJLE and MMLE with $K = 20$ components are shown in Fig. 2 (b) and (c). As can be seen from Fig. 2 (b), MJLE produces spurious or duplicated components. In contrast, the ground truth is well recovered with MMLE.

[2] MATLAB code is available at http://perso.telecom-paristech.fr/∼dikmen/nips11/



(a) Data (b) W_{MJLE} (c) W_{MMLE}

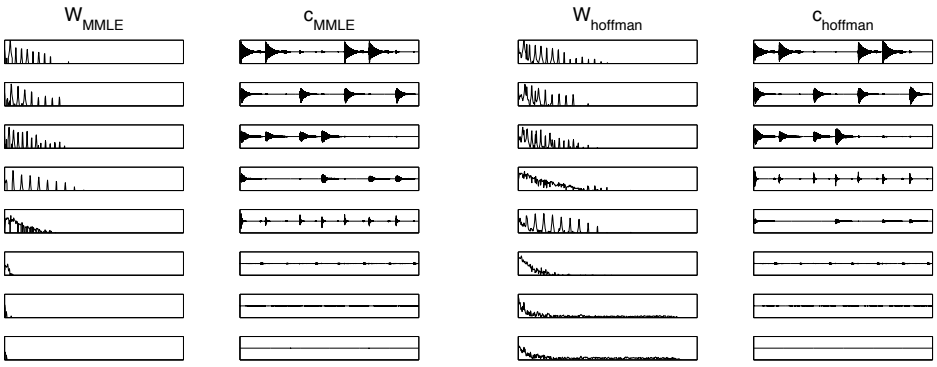
Figure 2: Data samples and dictionaries learnt on the swimmer dataset with $K = 20$.

5.2 A piano excerpt

In this section, we consider the piano data used in [3]. It is a toy audio sequence recorded in real conditions, consisting of four notes played all together in the first measure and in all possible pairs in the subsequent measures. A power spectrogram with analysis window of size 46 ms was computed, leading to $F = 513$ frequency bins and $N = 676$ time frames. We ran MMLE with $K = 20$ on the spectrogram. We reconstructed STFT component estimates from the factorization $\hat{W}\hat{H}$, where $\hat{W}$ is the MMLE dictionary estimate and $\hat{H} = \langle H \rangle_q$. We used the minimum mean square error (MMSE) estimate given by $\hat{c}_{fkn} = g_{fkn}.x_{fn}$, where g_{fkn} is the time-frequency Wiener mask defined by $\hat{w}_{fk}\hat{h}_{kn}/\sum_j \hat{w}_{fj}\hat{h}_{jn}$. The estimated dictionary and the reconstructed components in the time domain after inverse STFT are shown in Fig. 3 (a). Out of the 20 components, 12 were assigned to zero during inference. The remaining 8 are displayed. 3 of the nonzero dictionary columns have very small values, leading to inaudible reconstructions. The five significant dictionary vectors correspond to the frequency templates of the four notes and the transients. For comparison, we applied the nonparametric approach by Hoffman *et al*. [11] on the same data with the same hyperparameters for H. The estimated dictionary and the reconstructed components are presented in Fig. 3 (b). 10 out of 20 components had very small weight values. The most significant 8 of the remaining components are presented in the figure. These components do not exactly correspond to individual notes and transients as they did with MMLE. The fourth note is mainly represented in the fifth component, but partially appears in the first three components as well. In general, the performance of the nonparametric approach depends more on initialization, i.e., requires more repetitions than MMLE. For the above results, we used 200 repetitions for the nonparametric method and 20 for MMLE (without annealing, same stopping criterion) and chose the repetition with the highest likelihood.

5.3 Decomposition of a real song

In this last experiment, we decompose the first 40 seconds of *God Only Knows* by the Beach Boys. This song was produced in mono and we retrieved a downsampled version of it at 22kHz from the CD release. We computed a power spectrogram with 46 ms analysis window and ran our VB algorithm with $K = 50$. Fig. 4 displays the original data, and two examples of estimated time-frequency masks and reconstructed components. The figure also shows the variance of the reconstructed components and the evolution of the variational bound along iterations. In this example, 5 components out of the 50 are completely pruned in the factorization and 7 others are inaudible. Such decomposition can be used in various music editing settings, for example for mono to stereo remixing, see, e.g., [3].



(a) MMLE (b) Hoffman *et al*.

Figure 3: The estimated dictionary and the reconstructed components by MMLE and the nonparametric approach by Hoffman *et al*. with $K = 20$.

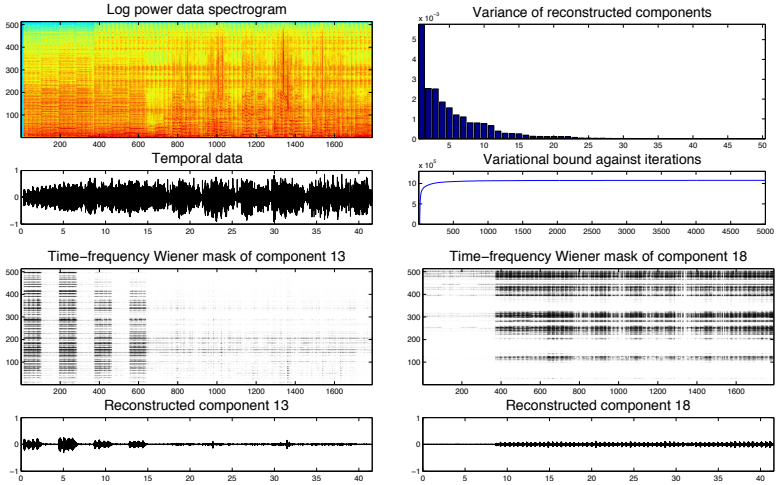


Figure 4: Decomposition results of a real song. The Wiener masks take values between 0 (white) and 1 (black). The first example of reconstructed component captures the first chord of the song, repeated 4 times in the intro. The other component captures the cymbal, which starts with the first verse of the song.

Acknowledgments

This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization).

6 Conclusions

In this paper we have challenged the standard NMF approach to nonnegative dictionary learning, based on maximum joint likelihood estimation, with a better-posed approach consisting in maximum marginal likelihood estimation. The proposed algorithm based on variational inference has comparable computational complexity to standard NMF/MJLE. Our experiments on synthetical and real data have brought up a very attractive feature of MMLE, namely its self-ability to discard irrelevant columns in the dictionary, without resorting to elaborate schemes such as Bayesian nonparametrics.

References

[1] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[2] C. Févotte and A. T. Cemgil. Nonnegative matrix factorisations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conference (EUSIPCO)*, pages 1913–1917, Glasgow, Scotland, Aug. 2009.

[3] 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, Mar. 2009.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

[5] Thomas Hofman. Probabilistic latent semantic indexing. In *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[6] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05)*, pages 601–602, New York, NY, USA, 2005. ACM.

[7] M. Welling, C. Chemudugunta, and N. Sutter. Deterministic latent variable models and their pitfalls. In *SIAM Conference on Data Mining (SDM)*, pages 196–207, 2008.

[8] W. L. Buntine and A. Jakulin. Discrete component analysis. In *Lecture Notes in Computer Science*, volume 3940, pages 1–33. Springer, 2006.

[9] John F. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th ACM international Conference on Research and Development of Information Retrieval (SIGIR)*, pages 122–129, 2004.

[10] O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*, Prague, Czech Republic, 2011.

[11] M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proc. 27th International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010.

[12] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30 – 37, 2004.

[13] Y. Cao, P. P. B. Eggermont, and S. Terebey. Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing*, 8(2):286–292, Feb. 1999.

[14] C. M. Bishop. *Pattern Recognition And Machine Learning*. Springer, 2008. ISBN-13: 978-0387310732.

[15] K. Katahira, K. Watanabe, and M. Okada. Deterministic annealing variant of variational Bayes method. In *International Workshop on Statistical-Mechanical Informatics 2007 (IW-SMI 2007)*, 2007.

[16] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[17] D. J. C. Mackay. Probable networks and plausible predictions – a review of practical Bayesian models for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[18] C. M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 382–388, 1999.